A NEW IMPLEMENTATION OF TRELLIS CODED QUANTIZATION BASED DATA HIDING

Xiaofeng Wang, Xiao-Ping Zhang

Ryerson University, Department of Electrical and Computer Engineering
350 Victoria Street, Toronto, Ontario, Canada, M5B 2K3
{xwang,xzhang}@ee.ryerson.ca

ABSTRACT

This paper discusses the construction and implementation problem of trellis coded quantization (TCQ) based data hiding. We explore the robustness and distortion of data hiding by analyzing its duality with distributed source coding. Based on our analysis a new implementation of the powerful trellis coded modulation (TCM) and TCQ data hiding scheme is presented. It simplifies the construction process with only one trellis and achieves a good tradeoff between robustness and distortion by embedding the information in the middle input of TCQ. Simulation is conducted for Gaussian, Laplacian and real image sources under additive Gaussian noise attack. The results demonstrate the effectiveness of the new implementation.

***Index Terms*—** Data hiding, trellis coded quantization, trellis coded modulation, Gaussian noise attack

1. INTRODUCTION

Data hiding is an information theory problem that is similar to channel coding with side information [1][2]. The distributed source coding is its dual problem [3]. Although many methods have been proposed for the construction of data hiding codes, most suffer from high complexity. For example, the lattice code proposed by P. Moulin [4] is a good way to approach the theory limit but in high dimensions its computational cost prevents it from practical use. There is still a large gap between the theory [1] and practical codes [4].

To build practical codes for data hiding a relatively new source coding method trellis coded quantization (TCQ) [5] has attracted much attention over the past several years [2][6][7]. Based on the duality between information embedding and channel coding with side information, Chou [2] proposes a combination of trellis coded modulation (TCM) and TCQ method. But the difficulty is how to find a strong TCM code that could partition into good TCQ code subsets. Simple methods like TCQ by path selection and initial state selection are proposed by the authors [6][7]. However to our best knowledge there is no comprehensive construction and implementation analysis of TCQ based data hiding. This paper analyzes TCQ based data hiding in distributed source coding with syndrome framework. The construction similarity and contrast between distributed source coding and data hiding are investigated for distortion and robustness analysis. Then a new implementation that embeds information in the middle input of the trellis and simplifies the hiding process by using just one trellis is introduced according to our analysis. The bit error rate (BER) simulation is conducted and compared to prove the effectiveness of the new implementation in various host distributions under Gaussian noise attack. The results show the new implementation achieves good robustness and distortion performance by hiding information in the middle of trellis inputs.

2. DATA HIDING PROBLEM

Data hiding is both source coding and channel coding problem shown in Fig. 1. From host side it is the source coding problem with the constrain that information M needs to be embedded inside the host S. The different data maps the host to the most similar one X inside different subset Λ_M. So the decoder could estimate the information $\hat{M}$ from the marked and possibly corrupted signal Y. From information side, it is channel coding problem with side (host) information available only at the encoder side. We only consider the common case blind detection that the information is not available at the decoder side.

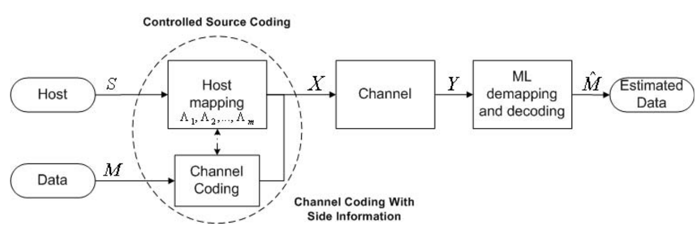

Fig. 1. Data Hiding System

Quantization is one of popular and effective subset separation methods in data hiding. In quantization based data hiding, the cover S of length n is represented by a quantized version X. There are several alternative quantized sequence subsets so the information sequence M could be embedded by choosing a subset. The encoding is the process of the quantizer selection. For simplicity, we consider the transmission of binary sequence $M \in \{0,1\}^n$ and all codes (quantized sequences) could be used is Λ. The process is as follows:

1. Identify all possible sequences $\Lambda_M \subset \Lambda$ that could represent message M.
2. Compute the closest sequence $X \in \Lambda_M$ to S.
3. The quantized sequence is corrupted by attacks. For example additive Gaussian noise V with mean 0 and variance δ_V^2, the channel output is $Y = X + V$.
4. Detect the hidden message $\hat{M}$ from Y by searching most possible subset Λ_M that Y may belong to.

3. TRELLIS BASED SOLUTION FOR DATA HIDING

We use the set partition of trellis code to find optimal mapping for this information coding (space partition) problem. The source coding and channel coding are expected to be like sphere so as to take advantage of the gains of these codes. In this section we analyze the construction of the data hiding code based on its duality with distributed source coding as in [8]. In spit of their duality there are also fundamental differences. The inter coset distance is more important rather than the intra coset distance which is emphasized in the distributed source coding paper.

3.1. The scalar quantizer and memoryless coset construction

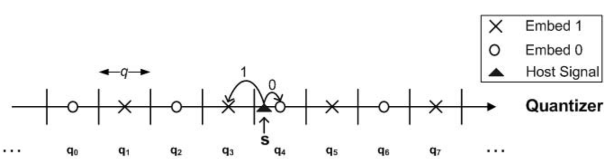

Fig. 2. The scalar quantizer and memoryless coset.

The simplest solution is that the source coding and set partition are both memoryless as the standard quantizaiton indexed modulation (QIM) approach shown in Fig. 2. For example 8 quantization levels are divided into two cosets $\{q_0, q_2, q_4, q_6\}$ and $\{q_1, q_3, q_5, q_7\}$. If "0" is embedded, the coset q_0, q_2, q_4 and q_6 is the quantizer of the host. For "1" the remaining coset q_1, q_3, q_5 and q_7 is used. The rate of the source $R_s = I(X, S)$ is 2 bit/sample. The rate of the embedded data $R = I(M, Y)$ is 1 bit/sample. The total information rate of the channel is $R_c = I(X, Y) = 3$. The goal of data hiding is to achieve the best robustness against attacks which is indicated by the distance between the cosets at certain distortion $D(X, S)$ introduced by source coding at certain embedding rate R. In single element case, the scalar quantizer with 2^{R_c} levels is partitioned into 2^R cosets. Each coset have 2^{R_s} code words. For the length n sequence, it is the set partition $\{1, 2, ..., 2^{n*R_c}\}/\{1, 2, ..., 2^{n*R_s}\}$ which controls the robustness and distortion. Usually square-error distortion $D(X, S) = \frac{1}{n}\|S - X\|^2$ is used.

3.2. The scalar quantizer and coset with memory

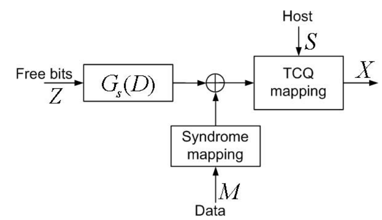

Fig. 3. The scalar quantizer and coset with memory.

The scalar QIM is not optimal because the host is a sequence not a single random variable. One approach is using syndrome of trellis code to do the set partition. As in Fig. 3, the message M is used to change the syndrome of the trellis code. For example a rate $2/3$ symmetric trellis code $G_s(D)$, the syndrome mapping of the data sequence M is $[0|0|S(D)]$. The TCQ output sequence X is,

$$X = Q(z(D)G_s(D) \oplus [0|0|S(D)]) \qquad (1)$$

Where Q is the quantization reconstruction function and $z(D)$ is input sequence. In this example $Q[\zeta] = q_\mu$. $\zeta \in \{0,1\}^3$ is the binary representation of μ. If we construct the coset this way we could not recognize the data embedded successfully because the difference of syndrome is not really protected by channel coding. The fundamental difference is that data hiding is space partition from channel code to source code while distributed source coding is the partition from source code to channel code. So the inter distance rather than the intra distance between cosets plays the more important role.

3.3. TCQ and memoryless coset construction

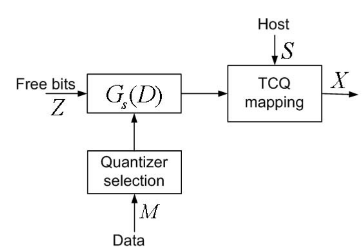

Fig. 4. TCQ and memoryless coset construction.

Fig. 4 is TCQ and the memoryless coset selection method. For example we use the cosets $\{q_0, q_1\}$ and $\{q_2, q_3\}$ to embed

"0" and "1" respectively as shown in Fig. 5(a). In the encoder side, if $2/3$ TCQ quantizer is employed, we will go through all possible representations of the sequence allowed by the quantizer selection to find the minimum distortion one. The information bit is protected by the distance of two cosets.

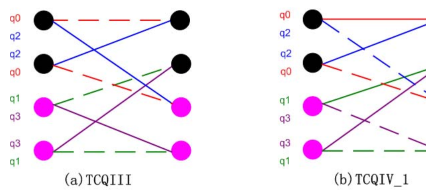

Fig. 5. Two 2/3 TCQ data hiding cases.

3.4. TCQ and coset construction with memory

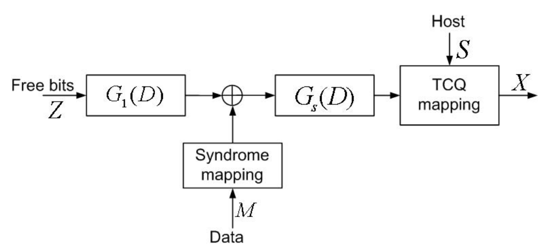

Fig. 6. TCQ and coset construction with memory.

The powerful TCM-TCQ method is shown in Fig. 6. For example a rate $2/3$ symmetric trellis code $G_1(D)$ with a rate $3/4$ $G_s(D)$, the mapping of the data sequence is $[0|0|S(D)]$. The TCQ output sequence X is,

$$X = Q(z(D)G_1(D)G_s(D) \oplus [0|0|S(D)]G_s(D)). \quad (2)$$

There are 2 free bits in this case. We will find these two free bits to generate the best sequence that could quantize the host sequence with the lowest distortion. In this case the information bit is protected by the trellis of $G_s(D)$.

4. A NEW TCQ DATA HIDING IMPLEMENTATION

Based on the analysis a new implementation of TCM-TCQ coset construction is discussed. We consider applying $G_1(D)$ and $G_s(D)$ together. For example as in equation (2), the information M controls one of the three inputs of $G_s(D)$. We could simplify the process by using only one trellis to do the job of both $G_1(D)$ and $G_s(D)$ and let the data determine one input of the trellis. One question is where to put the data inside the input bits. In order to take advantage of both source coding and channel coding the message bit should be in the middle of the inputs. If the embedded bit is most significant bit (MSB), it is bad source coding and will introduce more distortion which is not preferable in data hiding. In the other hand, if least significant bit (LSB) is used, it is bad channel code and will have less protection of the message. That means it is not robust in data hiding. When the embedding rate is fixed, the robustness and fidelity are two contradict goals of data hiding. To better tradeoff these two characters the coset should be carefully constructed. As shown in Fig. 7 the message bit is in the middle of two free bits. The four output bits are used to select cosets or quantizers. In the example if the free bit1 is 1, free bit2 is 0 , the message is 1 and current state is [0 0 0], the output will be [1 0 1 0] and the eleventh quantizer will be selected to quantize the host signal. The free bits, the message bit and the current state together select the quantizer. The embedding process could be summarized as follows.

- Step 1 - Select a powerful trellis code with the generator matrices $G(D)$.
- Step 2 - Choose the middle input of the trellis $G(D)$ to be bits of the information M.
- Step 3 - Allow the other free bits to adapt to find X which is the closest to the host S. The Viterbi algorithm could be used here to reduce the complexity.

This is trellis coded quantization process with one input used as the data carrier. At the detector side the information could be extracted from Y by decoding the middle input of the trellis $G(D)$. The whole TCM-TCQ data hiding construction and implementation is easily realized by using only one trellis which is favorable in practical scenario.

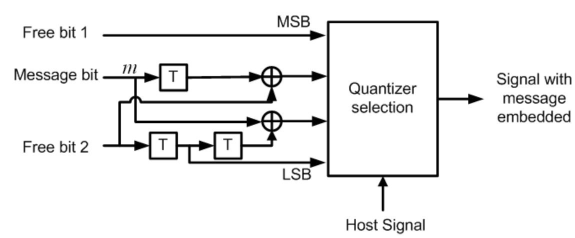

Fig. 7. Realization of TCQ embedding with rate 3/4 code.

5. EXPERIMENTAL RESULTS

Four TCQ based data hiding methods are applied to three types of hosts as shown in Fig. 8, Fig. 9 and Fig. 10 in order to prove the effectiveness of the new implementation. Two synthesis sources, one with Gaussian distribution and the other with Laplacian distribution, and several representative images including Lena, bridge, cameraman and hill are used as host in our experiment. The data to be embedded is

0s and 1s with same chances. The embedding rates are set to be the same one bit per host element. After embedding, the distortion is measured and then the Gaussian noise is added accordingly for DNR vs BER performance. The DNR is defined as the real distortion to noise ration. The distortion is the quantization introduced distortion. The BERs are computed by averaging the BER of 1000 random experiments. The length for Viterbi coding and decoding is 256. One TCQIII method used is Fig. 5(a). Two TCQVI methods represent the rate 2/3 and 3/4 code as in Fig. 5(b) and Fig. 7 respectively. Generally more states mean more robustness against gaussian noise. 8 states TCQVI_2 in the Gaussian and real image case is about 0.5db better than 4 states TCQVI_1 and TCQIII at BER 10^{-3}. TCQVI is not always better than TCQIII. That is because of only two inputs in rate 2/3 code the message bit could only select the LSB for embedding. But if there are three inputs as in Fig. 7, the middle bit could be selected to balance the robustness and distortion that ensures better robustness. The new implementation of TCQVI_2 are better than all other cases which proves our analysis.

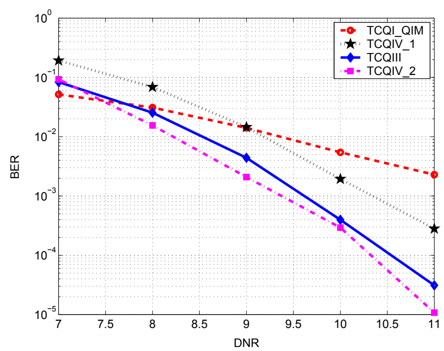

Fig. 8. BER with Laplacian host (mean 0 and variance 1).

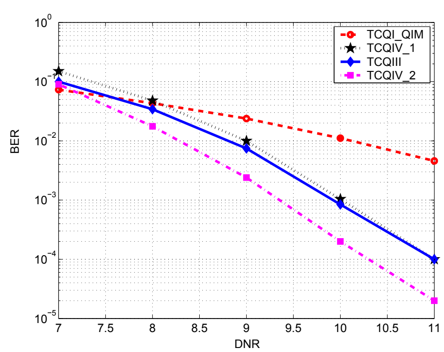

Fig. 9. BER with Gaussian host (mean 0 and variance 1).

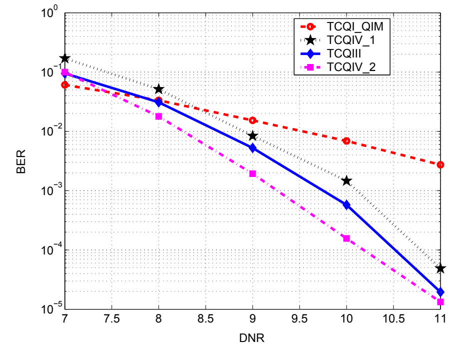

Fig. 10. BER performance with hill image as host.

6. CONCLUSION

In this paper we identify the construction duality and differences between the data hiding problem and the distributed source coding. Based on robustness and distortion analysis a new construction and implementation procedure of the famous TCM-TCQ data hiding is given. A variety of new simulation results for the TCQ based data hiding prove the effectiveness of the implementation. Future work includes finding more powerful trellis codes to deal with various kinds of attacks and approach theory capacity limit.

7. REFERENCES

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

[2] J. Chou, S. S. Pradhan, and K. Ramchandran, "On the duality between data hiding and distributed source coding," in *Proc. 33rd Asilomar conference on Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 1999.

[3] Richard J. Barron, B. Chen, and G.W. Wornell, "The duality between information embedding and source coding with side information and some applications," *IEEE Trans. Inform. Theory*, vol. 49, pp. 1159 – 1180, May 2003.

[4] P. Moulin and R. Koetter, "Data-hiding codes," *Proc. of IEEE*, vol. 93, pp. 2083–2127, Dec. 2005.

[5] M.W. Marcellin and T.R.Fischer, "Trellis coded quantization of memoryless and gauss-markov sources," *IEEE Trans. Commun.*, vol. 38, pp. 82 – 93, Jan. 1990.

[6] E. Esen and A. A. Alatan, "Data hiding using trellis-coded quantization," in *ICIP*, SINGAPORE, Oct. 2004.

[7] Xiaofeng Wang and Xiao-Ping Zhang, "Generalized trellis coded quantization for data hiding," in *ICASSP*, Apr. 2007, vol. 2, pp. 269–272.

[8] S. Sandeep Pradham and Kannan Ramchandran, "Distributed souce coding using syndroms(discus): Design and construction," *IEEE Trans. Inform. Theory*, vol. 49, pp. 626 – 643, 2003.