Concatenated LDPC-TCM Coding for Reliable Storage in Multi-Level Flash Memories

Quan Xu, Pu Gong and Thomas M. Chen
School of Engineering and Mathematical Sciences
City University London
Northampton Square, London EC1V 0HB, United Kingdom
Email: {Quan.Xu.1, Pu.Gong.1, Tom.Chen.1}@city.ac.uk

***Abstract*—In this paper, we present an efficient fault tolerant solution that concatenates trellis coded modulation (TCM) with an outer low-density parity-check (LDPC) code for multi-level per cell (MLC) flash memory. Traditional flash coding systems employ simple hard-decisions based codes, such as Bose-Chaudhuri-Hocquenghem (BCH) codes, that can correct a fixed, specified number of errors. Thanks to the Bahl, Cocke, Jelinek, and Raviv (BCJR) algorithm, the TCM decoder within the proposed design can provide soft decisions which make it possible to use the more powerful LDPC codes. Moreover, the error-correction performance is further improved since TCM can well decrease the raw error rate of MLC and hence relieve the burden of outer LDPC code. The effectiveness of concatenated LDPC-TCM systems has been successfully demonstrated through computer simulations.**

I. Introduction

Error correction codes (ECC) have become an important approach to enhance the data reliability of MLC flash memory. Due to the fact that flash systems only provide hard information to its decoders, simple algebraic codes such as BCH codes are generally employed in current design practice. As the storage density of MLC flash increases however, there is growing need for more advanced ECC techniques. LDPC codes are well known for their ability to approach the capacity limit in the additive white Gaussian noise (AWGN) channel, but LDPC codes are typically decoded with soft information. To realize the benefits of LDPC decoding in flash memory, researchers are trying to extract the soft information from the sensing outputs of flash systems [1, 2].

Trellis coded modulation is a powerful means for achieving coding gain in digital communication systems. Several studies have attempted to apply TCM to flash systems. Lou and Sundberg were among the first to use coded modulation in multilevel memories, but they did not consider outer ECC and sensing quantization [3]. Sun et al. successfully demonstrated that TCM can help to improve the performance of flash coding system, but they were focusing on short Hamming and convolutional codes [4]. Concatenation of TCM and BCH coding have been proposed, considering both the coded modulation and outer codes design [5]. However, the Viterbi algorithm was chosen to perform TCM decoding, which still results in hard decisions. In addition, 5-level MLC flash memory considered in the work is not readily available in the current market.

This paper investigates concatenated TCM and LDPC codes for flash memory that is modeled with pulse-amplitude modulation (PAM) plus Gaussian noise. We demonstrate that with the coded modulation, the storage reliability can be increased with the same signal-to-noise ratios (SNR). Furthermore, by performing the maximum *a posteriori* probability (MAP) decoding, we obtain soft information from TCM demodulator that can be utilized for LDPC decoding. Compared to flash memory with BCH coding, significant performance improvement of the concatenated system has been observed.

The rest of this paper is organized as follows. Section II concisely summarises the related background knowledge for understanding the topics in the forthcoming sections. In Section III, the proposed LDPC-TCM coding scheme is described and elaborated in detail. Theoretically achievable Asymmetric Coding Gain of the concatenated system is analysed in Section IV. Section V provides simulation results demonstrating the benefits of the proposed mechanism. The conclusion and further issues are drawn in Section VI.

II. Backgrounds

In this section, we briefly present the basics of NAND flash memory, Trellis Coded Modulation, and LDPC codes based on related works [2, 5–9]. The readers are referred to these literatures and references therein for detailed discussions on flash memory structures and related LDPC coding techniques.

A. NAND flash memory structure

Each NAND flash memory cell comprises a metaloxide semiconductor field effect transistor (MOSFET) with a floating gate [2]. The amount of charges stored during writing/programming in the floating gate is quantized to Ω levels to express $\log_2 \Omega$ bits of information.

The probability density function of the variation of threshold voltage in MLC flash memory cell is usually modeled by a Gaussian distribution. In this work, we assume an i.i.d. (independent and identically distributed) Gaussian threshold voltage for each level of the memory cell [1, 3]. Therefore an m-level flash cell is equivalent to an m-PAM communication system with additive white Gaussian noise. As an example, the threshold voltage distribution of 2 bits per cell flash memory are depicted in Fig. 1 which shows four distributions representing the memory levels with mean values of PV_i for $i \in \{0, 1, 2, 3\}$ and the same standard deviation of σ.

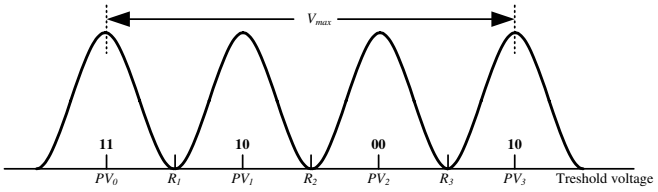

Fig. 1. The approximate Flash memory cell threshold voltage distribution model

B. Low Density Parity Check Codes

Low-density parity check code has been developed by Gallager [6] in the early 1960's. An LDPC code is defined as the null space of a parity check matrix **H** with the following structural properties: (1) each row consists of ρ "ones"; (2) each column consists of η "ones"; (3) the number of "ones" in common between any two columns, denoted ζ, is no greater than 1; (4) both ρ and η are small compared to the length of the code and the number of rows in **H**. Since ρ and η are small, **H** has a small density of "ones" and hence is a sparse matrix.

The class of LDPC codes have held the attention of coding theorists in the past decade not only because of their near-capacity performance on data transmission and storage channels, but also because their decoders can be implemented with manageable complexity. In addition to introducing LDPC codes, Gallager also provided a decoding algorithm that is typically near optimal. Since that, other researchers have independently discovered several related algorithms, albeit sometimes for different applications. The class of decoding algorithms are collectively termed "message passing" algorithms since their operation can be explained by the passing of messages in graph-based model of LDPC codes. Generally, message passing decoders use soft reliabilities about the received bits; conversely, a quantization of the received information or hard decisions can degrade the performance of an LDPC code.

C. Trellis Coded Modulation

In digital communication, Trellis Coded Modulation is an attractive solution for improving band limited communication systems. This technique evolves from the end of 1970's, when Ungerboeck [7] addressed the issue of bandwidth expansion by combining coding and modulation. According to him, "redundancy" is now provided by using an expanded signal set and the coding is done directly on the signal sequences. TCM is highly efficient, i.e. high coding gain and little descending data rate by coding and error correction, and more efficient especially for multilevel modulation.

The output signal sequences of TCM systems are usually decoded by the Viterbi Algorithm (VA) [8], which computes the most-likely input sequence (i.e., performs hard-output detection). For concatenated LDPC-TCM system, however, the soft decisions are required. Among the approaches that can generate soft information, the BCJR algorithm achieves optimum soft-output performance, while being well-suited for hardware implementation. Thus, we consider the BCJR algorithm in our design.

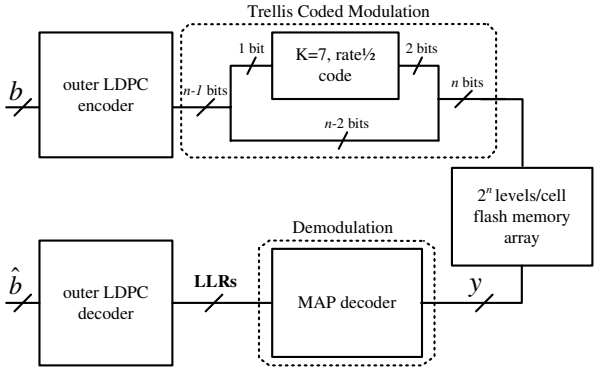

Fig. 2. Block diagram of TCM LDPC coding system

III. Concatenated LDPC-TCM Coding System

This sections provides detailed descriptions of the proposed concatenated LDPC-TCM coding system illustrated in Fig. 2. The analysis and verification will be presented in Section IV and V. Firstly, some redundancy is added to the information bit stream $\boldsymbol{b}$ by the LDPC encoder, and then the bit stream is passed to the TCM module after serial/parallel conversion. After being processed at the TCM module, the coded data bits are then programmed to the flash memory array. The threshold voltage within each MLC cell are sensed and quantized during the reading process, and the quantized value $\boldsymbol{y}$ is utilized in TCM demodulation and LDPC decoding.

For ease of practical implementation, we adopt the industrial standard pragmatic approach to TCM [8] (this approach will be referred to as the pragmatic TCM for convenience in the rest of the paper). In the design here, $n-1$ encoded bits are to be stored in one memory cell, among which $n-2$ bits are stored directly while 1 bit is fed to a constraint-length 7 (64 states), rate $1/2$ convolutional code. The output n bits are stored in 2^n levels; as a result, the number of storage levels per cell must increase from 2^{n-1} to 2^n.

Let X_l denote the trellis state at time l; the coded 2 bits can be decided by the state transition from X_l to X_{l+1}. For the mapping of pragmatic TCM, 2 coded bits choose a cell voltage level within a subconstellation according to the Gray code, and $n-2$ uncoded bits choose the subconstellation lexicographically, thus there are in total 2^{n-2} parallel transitions for given two coded bits, as shown in Fig. 3. Consider the pragmatic TCM on 8-level MLC for example, "111", "110", "100", "101", "011", "010", "000" and "001" are mapped to the voltage levels $PV0$, $PV1$, $PV2$, $PV3$, $PV4$, $PV5$, $PV6$ and $PV7$, respectively.

Due to the fact that LDPC decoder only accepts soft information, a demodulation module that can provide soft decisions is required, although the Viterbi algorithm is generally used to perform TCM decoding for the maximum likelihood sequence estimation (MLSE) which results in hard decisions. In this paper, the BCJR algorithm [9] is employed in the MAP

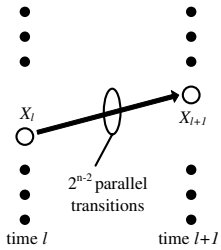

Fig. 3. State transitions in pragmatic TCM

decoder, which generates soft information of each bit, in terms of log-likelihood ratios (LLRs), to the following LDPC decoder.

Suppose the demodulator receives $\boldsymbol{y} = (y_0, y_1, \ldots, y_{L-1})$ from each page of the memory block, where y_l is the quantized voltage sensed from one memory cell, and let c_l denote the expected n-bit symbol along with the transition from time l to time $l+1$. For each bit $c_{l,k}, k = 1, 2, ...n$, the LLR is defined as

$$\text{LLR}(c_{l,k}) = \ln\left(\frac{\Pr(c_{l,k}=1|\boldsymbol{y})}{\Pr(c_{l,k}=0|\boldsymbol{y})}\right) \tag{1}$$

As a demodulated noisy value y_l is received from the flash channel with AWGN noise with variance σ^2, the likelihood function becomes

$$p(y_l|X_l, X_{l+1}) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{|y_l - c_l|^2}{2\sigma^2}\right) \tag{2}$$

The BCJR algorithm finds the *a posteriori* probabilities using the forward and backward recursions, respectively, as:

$$\alpha(X_{l+1}) = \sum_{X_l} \alpha(X_l)\gamma(X_l \to X_{l+1}) \tag{3}$$

$$\beta(X_{l+1}) = \sum_{X_{l+1}} \beta(X_{l+1})\gamma(X_l \to X_{l+1}) \tag{4}$$

where the values for $\alpha(X_0)$ and $\beta(X_L)$ should be determined according to the initial conditions, and $\gamma(X_l \to X_{l+1})$ is the branch metric that is given by

$$\gamma(X_l \to X_{l+1}) = \begin{cases} \Pr(X_{l+1}|X_l)p(y_l|X_l, X_{l+1}), & \text{valid transition;} \\ 0, & \text{invalid transition.} \end{cases} \tag{5}$$

The first term in the above equation corresponds to the *a prior* probability of the transition, $(X_l \to X_{l+1})$, which is known at the TCM encoder. The *a posteriori* probability of c_l is given by

$$\Pr(c_l|\boldsymbol{y}) = \sum_{\Lambda_l(c_l)} \alpha(X_{l+1})\beta(X_l)\gamma(X_l \to X_{l+1}) \tag{6}$$

where $\Lambda_l(c_l)$ denotes the set of the state transitions, $(X_i \to X_{i+1})$, for given c_l. With the calculated probabilities of all the symbols in the trellis, we can further obtain the *a posteriori* probability of each bit and the corresponding LLRs.

IV. Theoretical Analysis of Coded Modulation in Flash Memory

It is well known that the error performance of TCM systems in terms of SNR is measured by the free Euclidean Distances, where SNR is defined as the ratio between the average signal power, E_s, and the average noise power, $2\sigma^2$. Assuming the mean value of the threshold voltage at erase level (PV_0) is 0, the programming voltage at the highest level becomes V_{max} (see Fig. 1), and the peak power E_p equals V_{max}^2. In this work, both the TCM and non-TCM systems use the same peak power for fair comparisons.

Let M represent the number of levels in a MLC flash memory, the minimum squared Euclidean distances (MSED) which equals the programming voltage difference between two adjacent levels and the mean value of each level can be expressed as

$$d_{min}^2 = \frac{V_{max}^2}{(M-1)^2} \tag{7}$$

and

$$PV_i = \frac{V_{max}}{M-1} i, i = 0, 1, 2..., M-1. \tag{8}$$

The average power E_s is calculated as

$$E_s = \frac{1}{M}\sum_{i=0}^{M-1} (PV_i)^2. \tag{9}$$

Substituting (7) and (8) into the above equation, we obtain

$$E_s = \frac{V_{max}^2(2M-1)}{6(M-1)} = \frac{(M-1)(2M-1)}{6} d_{min}^2 \tag{10}$$

With the flash channel model presented in Section II-A, the SNR is given by

$$\text{SNR}(dB) = 10\log_{10}\frac{E_s}{2\sigma^2} \tag{11}$$

As for the TCM system, the free squared Euclidean distance d_{free}^2 [8] is given by

$$d_{free}^2 = 2(2d_{min})^2 + (d_f - 4)(d_{min})^2 = d_{min}^2(d_f + 4) \tag{12}$$

where d_f is a factor that depends on the convolutional codes used, and is equal to 10 for the constraint-length 7 (64 states) convolutional code of pragmatic TCM [8]. Substituting (7) into the above equation, d_{free}^2 can be rewritten as

$$d_{free}^2 = \frac{14V_{max}^2}{(M-1)^2} \tag{13}$$

The asymptotic coding gains (ACG) [7] (achieved at high SNR) of the proposed system is computed as

$$\text{ACG} = 10\log_{10}\left(\frac{d_{free}^2 E_{s/non-tcm}}{d_{min}^2 E_{s/tcm}}\right) \tag{14}$$

Substituting (7), (10) and (13) into the above equation and noting that the values of M for non-TCM and TCM system are 2^{n-1} and 2^n respectively, we obtain

$$\text{ACG} = 10\log_{10}\left(\frac{14(2^{n-1}-1)}{2^{n+1}-1}\right) \tag{15}$$

Increasing the number of flash memory cell levels reduces the Euclidean distances between the multiple levels; nevertheless, the trellis coded modulation could offer a coding gain that overcomes such disadvantage and further improve performance over the non-TCM system. In the next section, we will observe this error performance improvement versus the derived SNR.

V. Performance Evaluation

For purpose of comparison, we briefly present the asymptotic coding gain achieved for M-PAM modulation [8], which is generally used in digital communication with a similar model as Fig. 1.

$$\begin{aligned}\text{ACG} &= 10\log_{10}\left(4\times\frac{M^2-4}{M^2-1}\times\frac{7}{8}\right)\\ &= 10\log_{10}\left(\frac{7(2^{2n-1}-2)}{2^{2n}-1}\right)\end{aligned} \tag{16}$$

where n becomes the number of bits per transmitted symbol corresponding to the bits per cell in MLC flash memory. Fig. 4 shows that the value of ACG increases with n and the M-PAM modulation achieves higher gain than flash memory at certain value of n. The reason is because the average energies have been normalized to unity in the transmitter of digital communication systems, while the peak energies rather than average energies are normalized in flash memories. In the current market, 4-level and 8-level MLC are the prevalent flash storage media, so these two types were chosen to be used in the following simulations. As shown in Fig. 4, for TCM system on 8-level MLC, the asymptotic coding gain can be achieved as much as 4.5 dB in SNR.

To evaluate performance improvements due to the LDPC-TCM concatenated coding provided in Fig.2, we consider two flash memory systems which have the same page length of 4K bytes but using 4-level and 8-level MLC respectively. BCH coding and standard Gray mapping are applied in the first system while LDPC coding and the pragmatic TCM mapping are used in the second one. We store the same size of information bits that are randomly generated into these two memory systems and compare the bit error rate (BER) after reading and decoding.

For the error correction codes, we first design a rate-0.927 (17664, 16384) structured LDPC code [10] whose parity-check matrix is specified by a triangular plus dual-diagonal form to lower the error floor and encoding complexity. Additionally, the bipartite graph of the LDPC code used has been constructed to be free of cycle-4 with the bit-filling algorithm. Min-sum decoding algorithm is used to carry out LDPC code decoding. For the purpose of comparisons, we also consider a

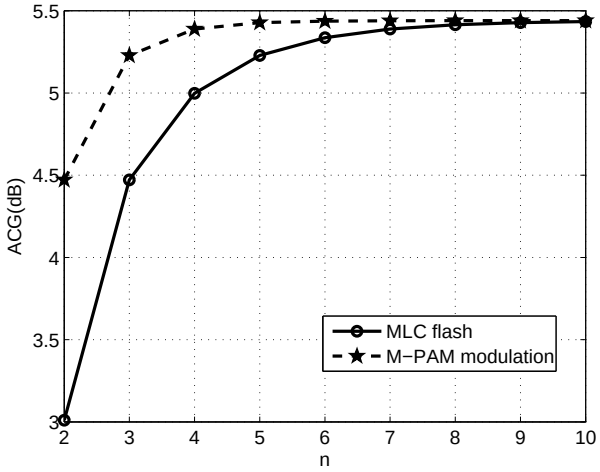

Fig. 4. Asymptotic Coding Gain of TCM for MLC flash memories and M-PAM modulation

$[n, k, t]$ = [16383, 15200, 85] binary BCH code with the same rate-0.927.

Fig. 5 shows the raw BER performance of the flash systems without ECC. Comparing to the flash system using 4-level MLC, the coding gain (CG) of TCM flash system using 8-level MLC is about 3.4 dB at the bit error probability of 10^{-6}. For the flash system at very low bit error probability (say less than 10^{-9}), the coding gain will reach the asymptotic coding gain of 4.5 dB that was derived earlier. In practical flash memories, the stored information (associated with cell threshold voltages) is usually sensed and quantized during the reading, so we assume two types of uniform sensing quantization schemes in our simulations: 16 levels and 32 levels, labelled as TCM-16Q and TCM-32Q, respectively. Performance loss has been observed after applying quantization to the TCM flash system. However, it still demonstrates a substantial performance improvement compared to the raw BER of 4-level flash system, as shown in Fig. 5.

Fig. 6 shows the BER performance of the flash systems with outer ECC. BCH coding is adopted in the flash system using 4-level MLC because for which only hard-decisions are available. As shown, the proposed LDPC-TCM coding provides a remarkable performance contribution (about 0.75 dB improvement over BCH coding at the bit error probability of 10^{-6}) to the error correction of flash systems. Similarly, some performance loss is introduced due to the quantization.

The outer ECC of the proposed design can be replaced with BCH coding if the TCM decoder simply outputs the hard reliabilities. For performance evaluation, we also compared our design with such BCH-TCM coding system. Fig. 7 illustrates the simulation results of these two schemes under the same memory parameters. The curves show that LDPC code outperforms the BCH code on the condition of same TCM parameters and quantization. Additionally, compared to the results shown in Fig. 6, it has been observed that BCH coding also benefits from the use of TCM if the quantization is not

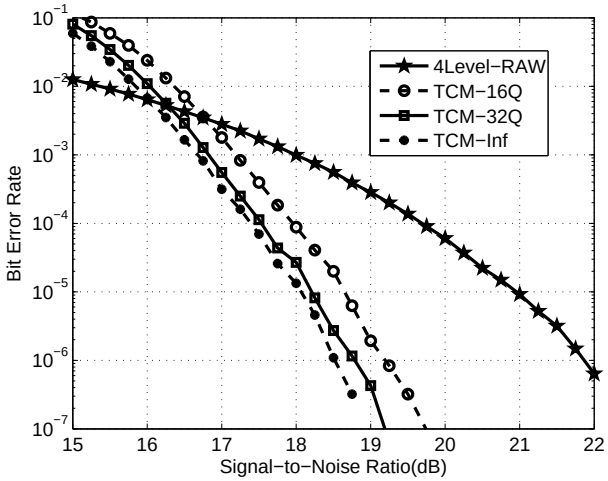

Fig. 5. Simulation results for flash systems without outer ECC

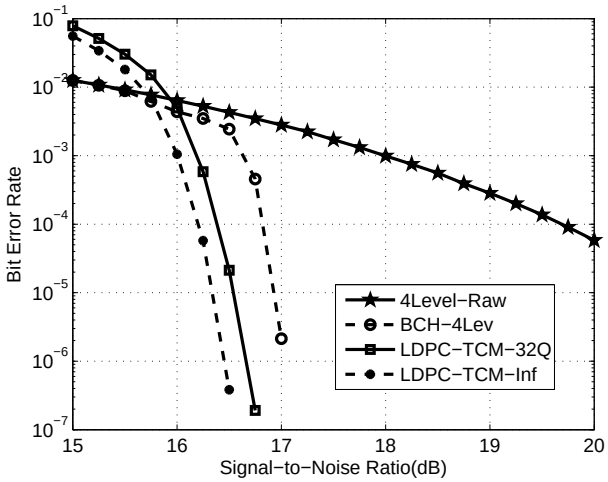

Fig. 6. Performance comparisons of concatenated LDPC-TCM and BCH coding systems((17664, 16384)LDPC code and (16383, 15200)BCH code)

too coarse.

VI. Conclusions

An error-correction scheme of concatenated LDPC-TCM coding for MLC flash memory is proposed. Compared to the flash coding system that provides hard-decisions and employs BCH codes only, results show remarkable BER performance improvements from the system equipped with (17664, 16384) LDPC codes and industrial pragmatic TCM. In this paper, we have also derived mathematical formulations to quantitatively analyse the asymmetric coding gain achieved in flash channels. BCJR algorithm is performed and associated with TCM, which has been demonstrated to be an alternative of converting the hard-decisions into soft-decisions. To further improve our work, better quantization schemes for memory sensing should be considered while designing the error correction system.

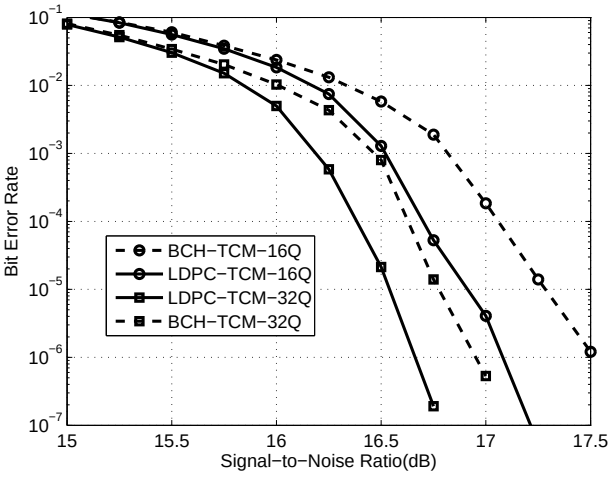

Fig. 7. Performance comparisons of concatenated LDPC-TCM and BCH-TCM coding systems

References

[1] J. Wang, T. Courtade, H. Shankar, and R. Wesel, "Soft information for LDPC decoding in flash: mutual-information optimized quantization," in *IEEE Globecom*, Dec. 2011, pp. 1–6.

[2] G. Dong, N. Xie, and T. Zhang, "On the Use of Soft-Decision Error-Correction Codes in NAND Flash Memory," *IEEE Trans. on Circuits and Systems I: Regular Papers*, vol. 58, no. 2, pp. 429–439, 2011.

[3] H.-L. Lou and C.-E. Sundberg, "Coded modulation to increase storage capacity of multilevel memories," in *IEEE Globecom*, 1998, pp. 3379–3384.

[4] F. Sun, S. Devarajan, K. Rose, and T. Zhang, "Multilevel flash memory on-chip error correction based on trellis coded modulation," in *2006 IEEE Int. Symp. on Circuits and Systems (ISCAS 2006)*, May 2006.

[5] S. Li and T. Zhang, "Improving multi-level NAND flash memory storage reliability using concatenated BCH-TCM coding," *IEEE Trans. on VLSI Systems*, vol. 18, no. 10, pp. 1412–1420, Oct. 2010.

[6] R. Gallager, "Low-density parity check codes," *IRE Trans. Information Theory*, pp. 21–28, Jan. 1962.

[7] G. Ungerboeck, "Trellis-coded modulation with redundant signal sets Part I: Introduction," *IEEE Com. Mag.*, vol. 25, no. 2, pp. 5–11, 1987.

[8] A. Viterbi, J. Wolf, E. Zehavi, and R. Padovani, "A pragmatic approach to trellis-coded modulation," *IEEE Com. Mag.*, vol. 27, no. 7, pp. 11–19, July 1989.

[9] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. on Info. Theory*, vol. 20, no. 2, pp. 284–287, Mar. 1974.

[10] Z. He, P. Fortier, and S. Roy, "A class of irregular LDPC codes with low error floor and low encoding complexity," *IEEE Com. Letters*, vol. 10, no. 5, pp. 372–374, 2006.