Transactions Letters

Constellation Shaping for Pragmatic Turbo-Coded Modulation With High Spectral Efficiency

Dan Raphaeli, *Senior Member, IEEE,* and Assaf Gurevitz

***Abstract*—We propose a new turbo-encoding scheme for high spectral efficiency with performance close to the Gaussian channel capacity. The scheme combines nonuniform signaling on a Gaussian channel with pragmatic turbo-coded modulation (TCM) for simple and flexible implementation. A variable-rate turbo code is followed by a Huffmann code mapping onto nonequiprobable points in a quadrature amplitude modulation constellation. The rate of the turbo code is matched to the Huffmann code by variable puncturing, such that both the input bit rate and the output symbol rate are constant. It is shown that the new scheme provides shaping gains of 0.6 and 0.9 dB, at rates 2 and 3 b/dimension, respectively, compared with the equiprobable pragmatic TCM, and reach about 1 dB from the continuous input Gaussian channel capacity.**

***Index Terms*—high spectral efficiency modulation, signal shaping, turbo codes.**

I. Introduction

THE idea behind constellation shaping is that signals with large norm are used less frequently than signals with small norm, thus improving the overall gain by adding shaping gain to their original coding gain [1]. The nonuniform signaling reduces the entropy of the transmitter output, and hence, the average bit rate. However, if points with small energy are chosen more often than points with large one, energy savings may compensate for this loss in bit rate. Theoretically, constellation points would be selected according to a continuous Gaussian distribution at every dimension, and thus achieve the maximum shaping gain. Practically, in finite constellations, a smaller gain can be achieved.

For Gaussian channels, turbo-coded modulation (TCM) techniques can be broadly classified into binary schemes and turbo trellis-coded modulation (TTCM) [2]. The first group can be further divided into "pragmatic" schemes with a single component binary turbo code, and multilevel binary turbo codes [3]. The pragmatic approach for a bandwidth-efficient turbo-coding scheme has been presented in [4]. This approach is simple and versatile, and is much less complex to design and to implement than TTCM. It uses only one turbo decoder, and by modifying its puncturing function and modulation signal constellation, it can obtain a large family of TCM schemes. It is possible to include shaping in the framework of multilevel codes [3]. The multilevel approach may be less attractive than the pragmatic approach due to its increased complexity, delay (from the multistage decoding), and sophisticated design rules. The shaping that was applied to multilevel codes is multidimensional trellis-shaping code, which is much more complex than the method proposed here. The performance obtained in the multilevel approach is similar (within 0.1 dB) to our results. TTCM was shown to obtain the highest performance in the coding of equiprobable quadrature amplitude modulation (QAM). Fragouli and Wesel [5] have shown that by careful code selection and interleaver design, it is possible to reach a signal-to-noise ratio (SNR) of 0.5–0.6 dB from the (constrained) capacity. Moreover, by combining nonuniform constellations [6] with the symbol-interleaved encoder, they achieved a shaping gain of approximately 0.2 dB [7].

Many authors suggested various forms of shaping within the framework of TCM, like the trellis shaping of Forney [8]. Another approach to achieve shaping was already proposed in the 1960s by Gallager [9]. He showed that optimal coding for the Gaussian channel can be made by grouping bits from a binary code, where the larger group is assigned to the less probable points in the constellation, and smaller to the more frequent. Obviously, such a method leads to a variable-rate input to the channel. A more detailed analysis of the method and close to optimal variable-rate code was suggested in [10], which showed that a close-to-optimal solution is the Huffmann prefix code. The variable-rate scheme of [10] is simple to encode, but requires *variable-rate input*, which results in system problems such as buffering and resynchronization.

In this letter, we propose a practical method that uses this idea while overcoming the variable-rate input disadvantage. We combine the powerful but simple pragmatic turbo-code principle with this shaping method and supply a practical decoding scheme with good results. The matching is done by dynamic puncturing of a low-rate turbo code to match the symbol-dependent rate of the prefix code. For further improvement, we also apply a feedback to the log-likelihood ratio (LLR) calculation block, as suggested in [11]. We consider one-dimensional (1-D) pulse-amplitude modulation (PAM) constellations that are equivalent, on an additive white Gaussian

Paper approved by R. D. Wesel, the Editor for Coding and Communication Theory of the IEEE Communications Society. Manuscript received March 11, 2002; revised November 3, 2002 and September 5, 2003. This paper was presented in part at the IEEE Conference on Communications, Anchorage, AK, May 2003.

The authors are with the Department of Electrical Engineering—Systems, Faculty of Engineering, Tel-Aviv University, Tel-Aviv 66978, Israel (e-mail: danr@eng.tau.ac.il).

Digital Object Identifier 10.1109/TCOMM.2004.823564

noise (AWGN) channel, to square QAM constellations of twice the symbol rate.

Our method can be applied to any binary turbo or turbo-like code, including parallel concatenation, serial concatenation, and low-density parity-check (LDPC) codes.

II. Signal Shaping in a Finite Constellation

It is well known that the maximum theoretical shaping gain, or the maximum reduction in average transmit power for constellations of rates R bits/dimension (dim) $\rightarrow \infty$ is given by $G_{s,\max} = 10\log_{10}(\pi e/6) = 1.53$ dB.

In practice, when using signal sets with finite transmission rates, this gain can never be achieved. Therefore, in order to calculate the real gain that can be achieved when using small signal sets, we turn to the calculation of the *capacity gain* [3], which is the optimization of the mutual information of the AWGN channel with discrete inputs. Consider an AWGN channel having discrete inputs $\mathbf{c}$ taking on the values $\{c_j\}$ for $j = 0,\ldots,J-1$, with probabilities $\mathrm{P_r}(\mathbf{c}) = \{\mathrm{P_r}(c_0), \mathrm{P_r}(c_1), \ldots, \mathrm{P_r}(c_{J-1})\}$.

The capacity of the discrete input AWGN channel is given by the maximum of the mutual information

$$\begin{aligned} C &= \max_{\mathrm{P_r}(\mathbf{c})} I(\mathbf{c};Y) \\ &= \max_{\mathrm{P_r}(\mathbf{c})} \sum_{j=0}^{J-1} \mathrm{P_r}(c_j) \int_{-\infty}^{\infty} Q(Y\,|\,c_j) \\ &\quad \times \log \frac{Q(Y\,|\,c_i)}{\sum_{i=0}^{J-1} \mathrm{P_r}(c_i) Q(Y\,|\,c_i)}\, dY. \end{aligned} \tag{1}$$

The SNR is given by $(S/N) = (P/\sigma^2)$, where P is the average input signal power

$$P = \sum_{j=0}^{J-1} \mathrm{P_r}(c_j) \cdot c_j^2 \tag{2}$$

and σ^2 is the noise variance. Optimizing the mutual information with respect to the input probabilities $\mathrm{P_r}(\mathbf{c})$ will give the lowest required SNR when transmitting at various rates $R = C$ bits/dim. We consider the maximum SNR reduction, compared to the equiprobable transmission, as the desired capacity gain. In [12], the authors proposed a numerical method that optimizes the mutual information for a power-constrained finite constellation. We found out that the *Maxwell–Boltzmann* (MB) distribution suggested by [10] provides a very good approximation to the optimal solution, and is obtained by the distribution

$$\mathrm{P_r}(c_j) = K(\lambda) \cdot e^{-\lambda |c_j|^2}, \quad \lambda \geq 0 \tag{3}$$

where

$$K(\lambda) = \left(\sum_{c_j} e^{-\lambda |c_j|^2} \right)^{-1} \tag{4}$$

is the distribution normalization factor. This distribution maximizes the entropy of a finite constellation $H(\mathbf{c})$ under an energy constraint [3], [10], but does not necessarily optimize the mutual information. The parameter λ governs the tradeoff between the average power P of signal points and the entropy $H(\mathbf{c})$. For $\lambda = 0$, the uniform distribution arises, while increasing λ results in more concentrated distributions close to the origin. By selecting λ properly, the minimum average energy can be achieved for a given transmission rate, and consequently, the minimum SNR in the calculation of (1). The idea is best explained using an example.

Example 1: Consider the transmission of $R = 2.0$ b/dim using a 16-PAM signal constellation. In this constellation, the signal set $\mathbf{c}$ consists of the 1-D signals $\{-15,-13,-11,\ldots,-1,1,3,5,\ldots,13,15\}$. We first assume uniform distribution for each constellation point of $\mathrm{P_r}(c_j) = 1/16$, which gives an average input signal power $P = 85$. Applying (1) for a rate-$R = 2.0$ b/dim, we get $\sigma^2 = 4.74$, which gives an SNR $S/N = 17.9$. We now apply the MB distribution (3) and (4) to the constellation. If we apply a discrete Gaussian distribution (3) with an optimized value of $\lambda_{\mathrm{opt}} = 1/9$ to the constellation points $\mathbf{c}$ and use (1) again, we get the average power $P = 16.975$ and a minimum SNR $S/N = 15.0$. Therefore, the capacity gain that we achieve here is

$$10\log_{10}\left(\frac{17.9}{15.0}\right) = 0.768 \text{ dB}. \tag{5}$$

Similarly, using the same procedure for rate-$R = 3.0$ b/dim with an optimal discrete distribution of the 16-PAM constellation with $\lambda_{\mathrm{opt}} = 0.047$, we achieve a capacity gain

$$10\log_{10}\left(\frac{82.6}{64.5}\right) = 1.074 \text{ dB}. \tag{6}$$

The losses in (5) and (6), with respect to the gains achieved by the ideal continuous AWGN channel at similar rates, are 0.001 and 0.106 dB, respectively.

III. Combining Shaping and the Pragmatic Binary TCM

A. Construction of a Binary Distribution

We apply the theoretical considerations of the previous section to practical schemes. For practical reasons, the probabilities of the constellation points from the MB distribution (3) are rounded to the closest 2^{-k}. The results for a 16-ary PAM constellation is (the result holds for both $\mathrm{R} = 2.0$ b/symbol and $\mathrm{R} = 3.0$ b/symbol)

$$\begin{aligned} P_r(c = \pm 1, \pm 3) &= 2^{-3} \\ P_r(c = \pm 5, \pm 7, \pm 9) &= 2^{-4} \\ P_r(c = \pm 11) &= 2^{-5} \\ P_r(c = \pm 13, \pm 15) &= 2^{-6}. \end{aligned} \tag{7}$$

If we use this probability mapping again in (1) for transmission rates $R = 2.0$ and 3.0 b/dim, we achieve gains of 0.682 and 0.948 dB, which are quite close to the capacity gains of the MB distribution achieved in (5) and (6). Moreover, these probabilities can easily be implemented by using a table that maps equiprobable input words of six bits into nonequiprobable words of four bits having the probabilities specified in (7). Table I shows a way to do it. The columns $b_0, b_1, \ldots, b_5$ represent the input bits. The signal points in the first column show

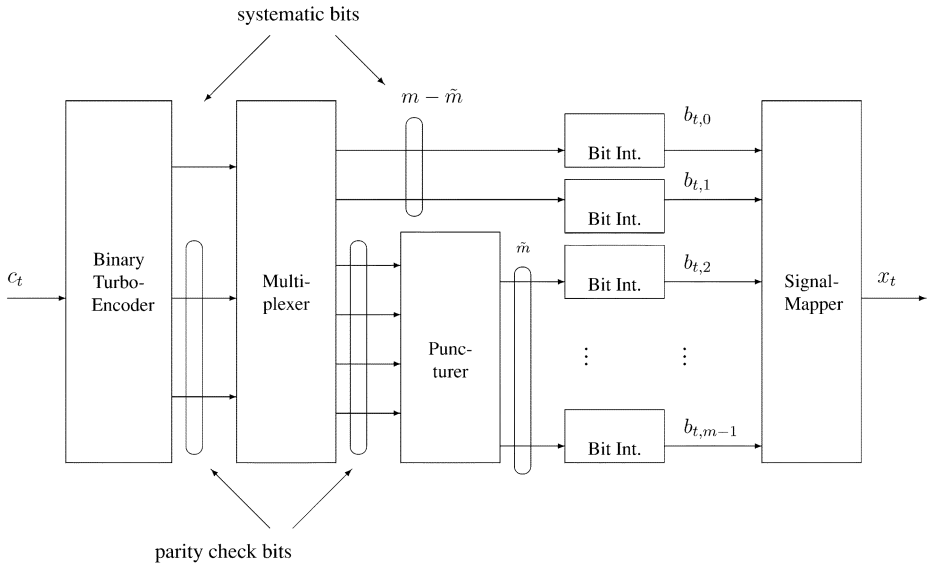

Fig. 1. Pragmatic binary TCM encoder.

TABLE I
SIGNAL MAPPER WITH BINARY PROBABILITIES OF TYPE $2^{-k}, k = 3, 4, 5,$ AND 6, FOR A 16-PAM CONSTELLATION

Signal Point	b_0	b_1	b_2	b_3	b_4	b_5
15	0	0	0	0	0	0
13	0	0	0	0	0	1
11	0	0	0	0	1	$\times 2$
9	0	0	0	1	$\times 2$	$\times 2$
7	0	0	1	1	$\times 2$	$\times 2$
5	0	0	1	0	$\times 2$	$\times 2$
3	0	1	1	$\times 2$	$\times 2$	$\times 2$
1	0	1	0	$\times 2$	$\times 2$	$\times 2$
-1	1	1	0	$\times 2$	$\times 2$	$\times 2$
-3	1	1	1	$\times 2$	$\times 2$	$\times 2$
-5	1	0	1	0	$\times 2$	$\times 2$
-7	1	0	1	1	$\times 2$	$\times 2$
-9	1	0	0	1	$\times 2$	$\times 2$
-11	1	0	0	0	1	$\times 2$
-13	1	0	0	0	0	1
-15	1	0	0	0	0	0

how to map each input word into one of the 16 signals. The notation $[\times 2]$ in the table means that a certain input bit in that location can take on the values zero or one. In this way, four words are mapped, for example, to the signal point 9.0, and eight are mapped to 1.0, and so on. Clearly, the probability of each input word is $1/64$, whereas probabilities of the output words become $8/64, 4/64, 2/64$, and $1/64$. As requested, the output probabilities are equal to the binary probabilities in (7).

It should be noticed that Table I actually performs prefix code mapping onto the constellation. This can be seen by disregarding the $[\times 2]$ entries of each input word. What is left is a variable-length Huffman prefix code.

B. Applying to TCM

We now apply the nonequiprobable distribution derived above to TCM. In pragmatic binary TCM [4], a single binary turbo code of rate $1/3$ is used as the component code. Its encoder outputs are suitably multiplexed and punctured to obtain $\tilde{m}$ parity bits and $m - \tilde{m}$ information bits, as shown in Fig. 1. The encoded bits are mapped onto an M-ary phase-shift keying (PSK) or M-QAM signal set. For simplicity, we used an M-PAM signal set, which is equivalent to an M^2-QAM having a spectral efficiency of $2(m - \tilde{m})$ b/s/Hz.

The signal mapper associates each word of m encoded bits to one of the M-PAM channel symbols available in the modulator. In an equiprobable signaling scheme, we map m encoded bits onto one of the $M = 2^m$ symbols using Gray code. Otherwise, in a nonequiprobable scheme, we apply a table (e.g., Table I) that maps m-bits equiprobable input words onto nonequiprobable M-PAM symbols.

It is clear that the mapping through the table onto the nonequiprobable constellation is a form of an additional puncturing because of the $[\times 2]$ ambiguity of some of the bits. On the other hand, as long as we keep the systematic bits unpunctured, the gain achieved by the nonequiprobable signaling scheme more than compensates for this loss in code strength. Thus, our scheme actually implements a variable-rate turbo code, because the number of parity bits given to each information bit varies as a function of the input combination.

The receiver, shown in Fig. 2, calculates the LLR calculation block for each encoded bit based on the received noisy symbol. The stream of the bit-likelihood values is then bit deinterleaved, demultiplexed, and depunctured before passing to the binary turbo decoder which is based on the maximum *a posteriori* (MAP) algorithm, e.g., [13]. As suggested in [11], the LLR calculation block is used in the turbo decoder iterations. The extrinsic coded data shown in Fig. 2 are the new values of the output turbo decoder MAP LLRs, for both systematic and parity bits, in each turbo iteration. This data will be used by the LLR calculation block as *a priori* input to the next iteration. To illustrate this, consider the calculation of the LLR for each encoded bit b_t

$$\Lambda(b_t) = \log \frac{\sum_{x:b_t=1} \exp\left(-\frac{1}{2\sigma^2}(r-x)^2\right) \cdot P_r(x \mid b_t = 1)}{\sum_{x:b_t=0} \exp\left(-\frac{1}{2\sigma^2}(r-x)^2\right) \cdot P_r(x \mid b_t = 0)} \tag{8}$$

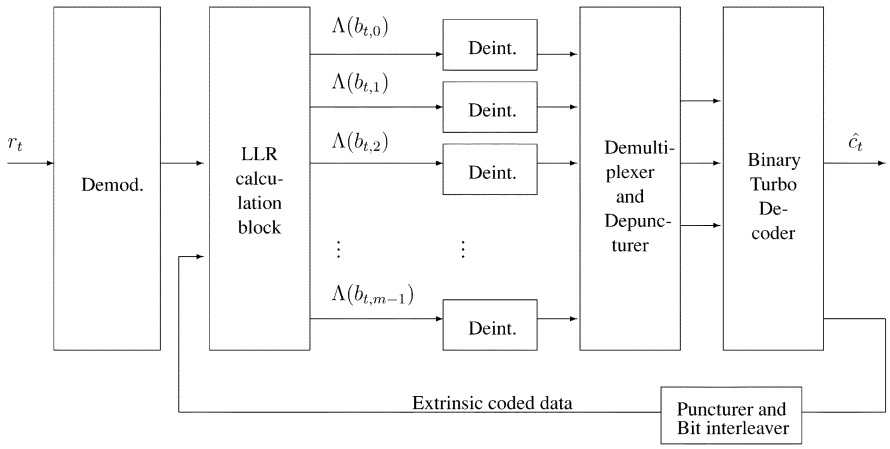

Fig. 2. Pragmatic binary TCM decoder.

where x is a symbol at the output of the mapping table, having inputs $b_t = 0$ or 1, r is the received symbol, and σ^2 is the noise variance. Each new turbo iteration will update the variable $P_r(x \mid b_t)$ and thus improve the stream of bit LLRs entering the decoder for the next iteration.

IV. Simulation Results

We used the standard rate-1/3 turbo-encoder, e.g., [14], made up of two elementary encoders with memory size four and the same generator polynomial 37-23 (octal number 23 represents feedback connections, and 37 represents feed-forward connections). The generator polynomial was taken from [15], which states the optimal encoders for various turbo-code rates and memory sizes. Turbo decoding was performed in 18 iterations on blocks of 32 768, 16 384, and 8192 information bits using pseudorandom interleaving. We applied two schemes for spectral efficiencies of 2 and 3 b/dim. The first one used a rate-1/3 turbo encoder and nonequiprobable signaling. This scheme applied Table I for a mapping of six-bit input words including two information bits onto 16-PAM symbols having a binary distribution, as in (7). It was compared to a rate-2/3 turbo code using equiprobable signaling with three-bit 8-PAM symbols, as in the standard pragmatic binary TCM technique [4]. The second scheme used a rate-1/2 turbo encoder and a similar nonequiprobable signaling technique, where this time, three information bits were mapped onto six-bit input words. It was compared with a rate-3/4 equiprobable 16-PAM pragmatic binary TCM. The bit-error rate (BER) versus E_b/N_0 for the two schemes are shown in Figs. 3 and 4. The capacity limits for the 16-PAM constellation and the continuous-input AWGN channel at both transmission rates are also reported.

The error-floor effect we have encountered occurred at BERs below 10^{-5} for the information block lengths we used. We can notice that for rate-2.0 b/dim and $P_b(e) = 10^{-5}$, the nonequiprobable scheme produces gains of 0.59 dB ($N = 32\,768$), 0.477 dB ($N = 16\,384$), and 0.57 dB ($N = 8192$) compared with the equiprobable one. The only power-constrained, continuous-input channel-capacity limit of the AWGN channel is $E_b/N_0 = 5.74$ dB. The performance of our decoder at $P_b(e) = 10^{-5}$ for the longest block length is about 1.1 dB from this limit. At 3.0 b/dim, the nonequiprobable

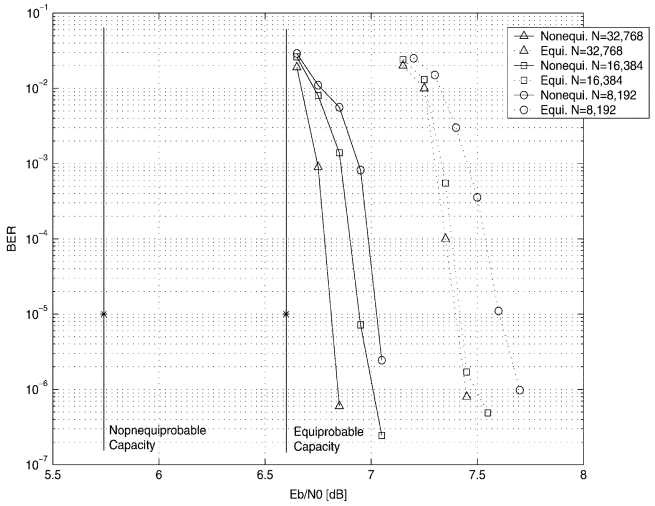

Fig. 3. Performance comparison between two schemes of nonequiprobable and equiprobable signaling at rate-2 b/dim using pragmatic binary TCM with 18 iterations and block lengths $N = 32\,768$, $N = 16\,384$, and $N = 8192$ b. Channel capacity limit is 5.74 dB, equiprobable 16-PAM capacity limit is 6.6 dB.

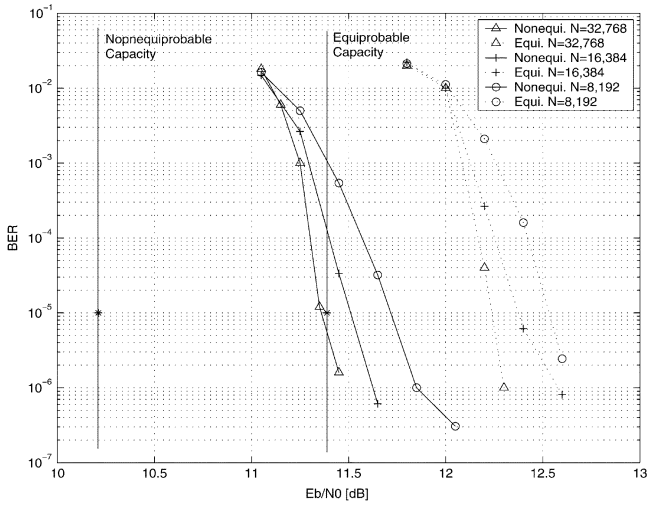

Fig. 4. Performance comparison between two schemes of nonequiprobable and equiprobable signaling at rate-3 b/dim using pragmatic binary TCM with 18 iterations and block lengths $N = 32\,768$, $N = 16\,384$, and $N = 8192$ b. Channel capacity limit is 10.2 dB, equiprobable 16-PAM capacity limit is 11.38 dB.

scheme produces gains of 0.87 dB ($N = 32\,768$), 0.86 dB ($N = 16\,384$), and 0.82 dB ($N = 8192$) compared with the equiprobable one for the same BER, the channel capacity limit is $E_b/N_0 = 10.2$ dB, and we achieve $P_b(e) = 10^{-5}$ for the longest block length within 1.1 dB of this limit.

It may be interesting to compare our results with the recent results of [7], where the authors used an 8-PAM TTCM scheme and shaping using nonuniform constellations for rate-2 b/dim. Using an input block size of 16 384 bits, they achieved $P_b(e) = 10^{-5}$ at SNR of $E_b/N_0 = 7.05$ dB, which is about 0.15 dB higher than our results. By increasing the block length to $N = 32\,768$ and the number of decoder iterations, our results are better within 0.25 dB.

V. Conclusion

In this letter, we presented a new scheme for improving the performance of pragmatic binary TCM by using nonequiprobable signaling. We described a nonequiprobable signaling technique that makes it possible to approach the maximum capacity gain of a finite-constellation AWGN channel. Our nonuniform signaling scheme is very easy to implement, and adds negligible load on the turbo decoder. We showed for an example of 6 b/QAM symbol, a gain of 0.9 dB relative to the equiprobable scheme performance, and transmission within 1.1 dB of the Shannon limit. Note that the difference between the capacities of the equiprobable and nonequiprobable schemes is 1.07 dB.

References

[1] G. D. Forney, Jr. and L. F. Wei, "Multidimensional constellations—Part I: Introduction, figures of merit, and generalized cross constellations," *IEEE J. Select. Areas Commun.*, vol. 7, pp. 877–892, Aug. 1989.

[2] P. Robertson and T. Woerz, "Bandwidth-efficient turbo trellis-coded modulation using punctured component codes," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 206–218, Feb. 1998.

[3] U. Wachsmann, R. F. H. Fischer, and J. B. Huber, "Multilevel codes: Theoretical concepts and practical design rules," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1361–1391, July 1999.

[4] S. Le Goff, A. Glavieux, and C. Berrou, "Turbo codes and high-efficiency modulation," in *Proc. IEEE Int. Conf. Communications*, New Orleans, LA, May 1994, pp. 645–649.

[5] C. Fragouli and R. D. Wesel, "Turbo encoder design for symbol-interleaved parallel concatenated trellis-coded modulation," *IEEE Trans. Commun.*, vol. 49, pp. 425–430, Mar. 2001.

[6] D. Sommer and G. Fettweis, "Signal shaping by nonuniform QAM for AWGN channels and applications using turbo coding," in *Proc. ITG Conf. Source and Channel Coding*, Jan. 2000, pp. 81–86.

[7] C. Fragouli, R. D. Wesel, D. Sommer, and P. Fettweis, "Turbo codes with nonuniform constellations," in *Proc. IEEE Int. Conf. Communications*, vol. 1, 2001, pp. 70–73.

[8] G. D. Forney, Jr., "Trellis shaping," *IEEE Trans. Inform. Theory*, vol. 38, pp. 281–300, Mar. 1992.

[9] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[10] F. R. Kschischang and S. Pasupathy, "Optimal nonuniform signaling for Gaussian channels," *IEEE Trans. Inform. Theory*, vol. 39, pp. 913–929, May 1993.

[11] S. Benedetto and G. Montorsi, "Generalized concatenated codes with interleavers," in *Proc. IEEE Int. Symp. Turbo Codes*, Brest, France, Sept. 3–5, 1997, pp. 32–39.

[12] N. Varnica, X. Ma, and A. Kavcic, "Capacity of power-constrained memoryless AWGN channels with fixed input constellations," in *Proc. IEEE Global Telecommunications Conf.*, vol. 2, Nov. 17–21, 2002, pp. 1339–1343.

[13] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol-error rate," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 284–287, Mar. 1974.

[14] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near-optimum error-correcting coding and decoding: Turbo codes," *IEEE Trans. Commun.*, vol. 44, pp. 1261–1271, Oct. 1996.

[15] S. Benedetto, R. Garello, and G. Montorsi, "A search for good convolutional codes to be used in the construction of turbo codes," *IEEE Trans. Commun.*, vol. 46, pp. 1101–1105, Sept. 1998.