IEEE Information Theory Society Newsletter

December 1990 (USPS 360-250) Editor: Nader Mehravari

CODED MODULATION FOR BAND-LIMITED CHANNELS*

G. David Forney, Jr.
Codex/Motorola
Mansfield, MA 02048.

Abstract:
In 1948, Shannon developed fundamental limits on communications efficiency over band-limited Gaussian noise channels. Practical modulation techniques have until recently made only modest progress toward the Shannon limit. In the past few years, however, with the introduction of coded modulation methods, rapid advances have been made, to the point where practical systems are being developed that approach the theoretical Shannon limit. This paper is a general survey of theory and practice in the field from 1948 to date.

1. Introduction

In 1948 Shannon [1] introduced his famous formula for the channel capacity C in bits per second of an ideal band-limited Gaussian channel:

$$C = W \log_2 (1 + SNR) \text{ bps},$$

where W is the channel bandwidth in Hz, and SNR is the channel signal-to-noise ratio. For SNR large, Shannon's result implies that there exist coding schemes that can achieve arbitrarily low error probabilities at signal-to-noise ratios approximately 9 dB lower than than those required to achieve error rates of the order of 10^{-5} - 10^{-6} with conventional quadrature amplitude modulation (QAM); or, equivalently, that can achieve rates that are 3 bps/Hz larger than can be achieved with conventional QAM. Nonetheless, for about 30 years, little progress was made toward closing this gap, despite the commercial importance of improved rates and/or SNR margins for such band-limited applications as telephone-line data modems. Indeed, the common wisdom in the 1970's was that the highest practically possible data rates had already been reached.

Ungerboeck's invention of trellis-coded modulation (TCM) dramatically altered this view, particularly after the publication of his pioneering paper [2] in 1982, which showed that easily implemented TCM schemes could yield 3 to 4 dB of coding gain. TCM was rapidly adopted for implementation in high-speed telephone-line modems in the mid-1980's. The field remains an active research area, and further advances in theory and practice continue to occur.

This paper surveys the history of advances in modulation for band-limited channels, with particular focus on telephone-line modems, where these advances have generally first been implemented. The latest advances in coded modulation are discussed.

* *Editor's Note: This paper represents a summary of talks given by the author in the IEEE Region 10 Speakers Tour, October 17-26, 1989.*

continued on page 3

CODED MODULATION

continued from front cover

2. Fundamental limits

In 1928, Nyquist showed that under certain ideal symmetry conditions, a channel of nominal (Nyquist) bandwidth W could support *pulse amplitude modulation* (PAM) at a rate of 2W samples per second, with no intersymbol interference. Alternatively, such a channel can support *quadrature amplitude modulation* (QAM) of W pairs of samples per seconds, using two carriers in quadrature—e.g., $\cos 2\pi f_c t$ and $\sin 2\pi f_c t$, where f_c is the center frequency in the band. QAM is generally preferred in telephone-line modems.

To send data using QAM, the sample pairs$\{x_k, y_k\}$ are chosen from a discrete alphabet of two-dimensional points, called a *signal constellation*. In *strict-sense QAM*, the two coordinates x_k and y_k are chosen independently from a standard 2^b-point PAM constellation, so that the QAM constellation is a square $2^b \times 2^b$ constellation with points on a rectangular grid. Such a constellation supports a rate of R = 2b bits per two-dimensional symbol. For example, a 4x4 strict-sense QAM constellation supports 4 bits per symbol.

At high signal-to-noise ratios, the symbol error probability for strict-sense QAM is well approximated by

$$\Pr(E) \approx 4\, Q[3 \bullet SNR/2^R]^{1/2},$$

where Q[y] is the Gaussian probability of error function, $Q[y] = \int_y^\infty p(x)dx$, where p(x) is a Gaussian distribution with mean 0 and variance 1. Thus there is a universal performance curve for Pr(E) vs. the *high-SNR normalized signal-to-noise ratio*, defined as $SNR_{norm} = SNR/2^R$. This curve is shown in Figure 1, with SNR_{norm} expressed in dB.

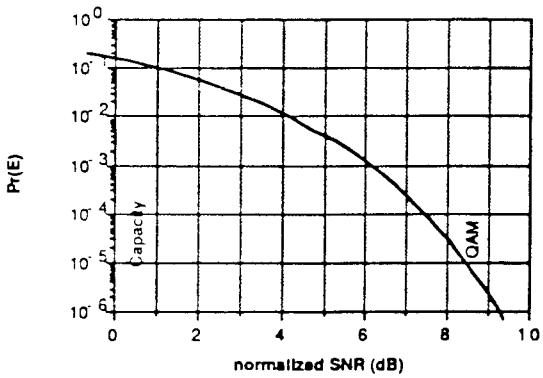

Figure 1: Pr(E) vs. SNR_{norm} for QAM

The capacity of an ideal discrete-time Gaussian channel is $\log_2(1+SNR)$ bits per two-dimensional symbol, or approximately $\log_2 SNR$ at high SNRs. This means that arbitrarily low error probabilities can be achieved for any rate $R < \log_2 SNR$; or, alternatively, whenever $SNR_{norm} > 1$ (0 dB). This ultimate limit, called the *Shannon limit*, is also shown in Figure 1. We see that there is an SNR gap of about 9 dB for error rates of the order of 10^{-5} - 10^{-6}; or, equivalently, a rate gap of about 3 bits per symbol.

3. Telephone-line modems

Voice-grade telephone channels are good examples of high-SNR band-limited channels. Happily, the model of a linear filter with additive Gaussian noise applies rather well to these channels. The usable bandwidth typically extends from about 300-500 Hz to about 3000-3400 Hz. On private lines in the U.S., the signal-to-noise ratio is guaranteed by tariff to be at least 28 dB, and typically can be 40 dB or more.

These parameters imply that the channel capacity can be from 20,000 to 30,000 bps. Nonetheless, through about 1980, it was generally accepted that 9600 bps was the highest practical data rate for telephone-line modems. This is another manifestation of the gap mentione above.

Although telephone channels can generally be modeled as linear filters, they are generally not ideal Nyquist filters, so they generate intersymbol interference (ISI). ISI may be reduced or eliminated by a compensating filter, called an equalizer. Moreover, different channels have different characteristics, which may vary slowly over time, so that ideally the equalizer should be automatic and adaptive. Much of the early work toward high-speed data transmission over telephone channels was directed to the problem of automatic adaptive equalization; by the late 1960's, it was essentially solved.

Telephone channels do have other disturbances, such as nonlinearities, phase jitter, echo (on two-wire dial lines), and the kinds of non-Gaussian noise one hears on poor lines. Once the ISI problem was solved by adaptive equalizers, phase jitter became the next limiting impairment on private lines, and was solved in the early 1970's by jitter-immune constellations or, alternatively, by high-performance jitter trackers. The problem of echo has been solved in the 1980's (not without difficulty) by precise echo cancellation. There has been little progress on combating other impairments, but fortunately there effects do not tend to be large, particularly on private lines.

The most advanced modulation and equalization techniques have often been developed and first implemented in telephone-line modems, because of the applicability of a linear Gaussian model, the commercial importance of the billion-dollar modem industry, the significance of higher data rates or improved SNR margin to the customer, and the relatively low symbol rates (≈ 2400 symbols per second) of modems, which permit use of sophisticated digital signal processing algorithms with thousands of operations per symbol.

Table I is an extension of the "Modem Milestones" table of [3], which attempts to identify the first commercially

continued on page 4

CODED MODULATION

continued from page 3

successful private-line modems that achieved various data rates. The table gives the modem type, the year of introduction, the data rate in bps, the symbol rate W (nominal Nyquist bandwidth in Hz), the rate R in bits per symbol, and the signal constellation/modulation method. An asterisk identifies those modems whose modulation schemes were subsequently adopted in international standards.

Year	Model	Speed	W	R	Mod.
1962	Bell 201B	2400	1200	2	4-PSK*
1967	Milgo 4400/48	4800	1600	3	8-PSK*
1971	Codex 9600C	9600	2400	4	16-QAM*
1980	Paradyne 14400	14400	2400	6	64-QAM
1981	Codex/ESE SP14.4	14400	2400	6	64-QAM
1984	Codex 2660	14400	2400	6	128-TCM*
1985	Codex 2680	19200	2743	7	160-TCM

Table I. Modem Milestones

The Bell 201B was the first widely-used synchronous modem. Using simple 4-phase modulation and 1200 symbols per second, it achieved 2400 bps on private lines. Because of the narrow nominal bandwidth of 1200 Hz, fixed compromise equalization was adequate. The international 2400 bps modem standard (CCITT Recommendation V.26, formally adopted in 1968) is modeled on the 201B.

The Milgo 4400/48 appears to have been the first commercially successful 4800 bps modem, using 8-phase modulation at 1600 symbols per second, and a manual adaptive equalizer. Although an 8-phase signal constellation is about 1.35 dB less SNR-efficient than an optimized 8-phase signal constellation, the choice of 8-phase was made for simplicity of implementation (analog implementation, in 1967). This modulation scheme is embodied in CCITT Recommendation V.27 (1972).

The Codex 9600C expanded the symbol rate and bandwidth to 2400 Hz by use of a digital adaptive equalizer, and a QAM 16-point constellation optimized for immunity to combined noise and phase jitter, at a sacrifice of SNR margin of about 1.3 dB relative to the square 4x4 constellation. This constellation was adopted in CCITT Recommendation V.29 (1976).

For the next decade, the industry was preoccupied with reducing size and cost, and it was generally accepted that 9600 bps was the highest achievable rate, even on private lines. This era of complacency was ended by Paradyne's announcement of a 14,400 bps modem in 1980. This modem was similar in all respects to earlier 9600 bps QAM modems with 2400 Hz symbol rates, except that it used a 64-point QAM constellation to support 6 bits per symbol. Due to advances in modem implementation and in the general quality of the telephone network, this modem worked reliably over a large percentage of private lines.

The constellation used by Paradyne was an 8x8 QAM constellation, with the 4 corner points moved to the axes. This modification improves SNR only by about 0.1 dB, but also improves the peak-to-average ratio and immunity to phase jitter and other signal-dependent impairments. From today's perspective, it can be seen as an effort to shape the constellation more like a circle than a square, and therefore can be regarded as the first embodiment of the concept of shaping (see below).

Codex and others then scrambled to develop comparable modems. The Codex/ESE SP14.4 was similar to the Paradyne modem in most respects, but used a 64-point constellation with points from a hexagonal rather than a rectangular grid. This gives a SNR improvement of about 0.6 dB, due to the greater packing efficiency of the hexagonal ('penny-packed') grid, which can be regarded as the first 'coding gain' achieved in a commercial modem.

Neither of these modems was standardized. By the time that the CCITT began to consider a standard for 9600 bps dial modems in 1983, Ungerboeck's paper [2] had appeared, and it was recognized that the 3 or more dB of coding gain that TCM could provide would be essential for reliable 9600 bps operation over the dial network. A variant of Ungerboeck's 8-state 2- dimensional code, due to Wei [4], was adopted in CCITT Recommendation V.32, with a coding gain of about 4 dB, and also subsequently in the V.33 standard for 14,400 bps private-line modems. At 6 bits per symbol, this code requires a 128-point constellation. The Codex 2660 was merely the first of many such modems.

In 1985, however, the Codex 2680 was able to achieve reliable 19,200 bps operation by expanding the bandwidth to 2743 Hz and using a multidimensional TCM scheme, also due to Wei [5], to support 7 bits per symbol with a signal constellation of only 160 points. This remains (in 1989) the highest achievable rate. The CCITT has not commenced any standardization activities for this rate. (No longer true, V. Fast)

4. Lattice codes

Before considering trellis codes, let us spend a moment on lattice codes, which have a longer history, and which illustrate many of the principles of coded modulation.

An N-dimensional *lattice* Λ is an array of discrete points in N-space that form an algebraic group under vector addition. For example, the set Z^N of all N-tuples of integers is a lattice.

A *lattice code* $C(\Lambda,R)$ may be defined as the set of all points in some lattice Λ (or a translate of Λ) that lie in some bounding region R. For example, a strict-sense QAM constellation is the set of all points in a translate of

continued on page 5

CODED MODULATION

continued from page 4

Z^2 that lie in a square region R of appropriate size.

At high signal-to-noise ratios, the SNR improvement of a lattice code over an uncoded strict-sense QAM constellation can be shown [6] to be separable into a *coding gain* $\gamma_c(\Lambda)$ that depends only on the packing density of the lattice Λ, and a *shaping gain* $\gamma_s(R)$ that depends only on the shape of the region R.

The nominal coding gain of a lattice is a measure of its efficiency for sphere packing in N dimensions. Finding the best N-dimensional sphere packing is an old mathematical problem. Table II shows the best known lattices for sphere packing in for various values of $N \leq 24$, from Conway and Sloane [7], with their coding gains in dB. Note that the best sphere packings in 16 and 24 dimensions were not known at the time of Shannon's paper [1].

Λ	N	Name	Date	$\gamma_c(\Lambda)$
Z	1	Integer lattice		0.00
A_2	2	Hexagonal lattice		0.62
D_4	4	Schläfli	≈1850	1.51
E_8	8	Gosset	≈1900	3.01
Λ_{16}	16	Barnes-Wall	1959	4.52
Λ_{24}	24	Leech	1967	6.02

Table II. Coding Gains of Lattices

As $N \rightarrow \infty$, the nominal coding gain increases without limit, whereas, from the capacity result, only about 9 dB of gain is actually possible. The effective coding gain for these dense lattices is reduced from the nominal gain by the large number of nearest neighbors to each lattice point. For example, in the Leech lattice, which has actually been used in a 19,200 bps coded modem [8], the number of nearest neighbors is 196,240, which reduces the effective coding gain by about 2 dB. These high-dimensional lattices also tend to have large decoding complexity and constellation expansion (see below).

As for shaping gain, the optimum shape in N dimensions is an N-sphere. The shaping gain of an N-sphere is about 0.20 dB for N = 2, about 1.10 dB for N = 24, and as $N \rightarrow \infty$, the shaping gain approaches $\pi e/6$ (1.53 dB) [3]. Therefore, of the total 9 dB gap, only about 1.5 dB can possibly be achieved by shaping, and the remainder must be achieved by coding. However, shaping gain is quite independent of coding gain, and shaping gains of greater than 1 dB can be achieved quite simply with a technique called trellis shaping [9].

DeBuda [10] has shown that there exist lattice codes that can achieve capacity, which implies that effective coding gains of up to about 7.5 dB at error rates of 10^{-5} - 10^{-6} are possible.

5. Trellis codes

Trellis codes are to lattice codes as convolutional codes are to block codes, and have many of the same advantages in practice. They tend to achieve better effective coding gains for the same implementation complexity, due to their generally much lower number of nearest neighbors (not to better nominal coding gains). They are naturally suited to sending continuous data sequences, which is the form of input data into modems, whereas lattice codes are more naturally suited to blocks of data. Trellis codes have been almost universally adopted in the modem industry.

As an example, Ungerboeck's 2-dimensional 4-state trellis code works as follows. To send R bits per symbol, a rectangular-grid QAM signal constellation with 2^{R+1} points is used; the *constellation expansion ratio* is thus a factor of 2. The constellation is divided into 4 subsets, each with 2^{R-1} points, in a regular way, such that the distance between points within a subset is twice the distance between points in the original constellation, and the distance between certain subsets is at least $2^{1/2}$ times that in the original constellation.

A rate-1/2, 4-state convolutional encoder is used to encode 1 input bit per symbol into 2 coded bits, which are used to select one of the 4 subsets of the signal constellation. An additional R-1 uncoded bits per symbol then select the actual point to be sent from the selected subset. Thus the constellation expansion of a factor of 2 is due to the code redundancy of 1 bit per symbol.

The set of all possible sequences of coded bits may be specified by a 4-state trellis diagram for the convolutional code, with each branch labeled by the corresponding two coded bits. If each pair of coded bits is replaced by the corresponding subset, then the trellis diagram specifies the set of all possible subset sequences. The Viterbi algorithm (VA) is an efficient way of searching such a trellis, and is commonly used for decoding trellis codes. The VA complexity is proportional to the number of states of the convolutional encoder.

Ungerboeck's 4-state code is arranged so that the minimum distance between signal point sequences in different subset sequences—i.e., corresponding to different trellis paths—is greater than the minimum within-subset distance. This property depends only on subset distance properties, and not on any other feature of the signal constellation. Therefore the minimum distance between different possible signal point sequences is simply that of one within-subset difference on the same path ('parallel transition'), which is a factor of 2 greater than the distance between signal points in the original constellation. This gives a gain in SNR margin of 6 dB. However, the constellation expansion of a factor of 2 costs 3 dB, so the net coding gain is 3 dB for this very simple code.

Ungerboeck's two-dimensional codes have nominal

continued on page 6

CODED MODULATION

continued from page 5

coding gains of 3 dB for 4 states, 4 dB for 8 states, and up to 6 dB at 256 states. In all cases the code redundancy is 1 bit per symbol, so the constellation expansion ratio is 2. For 8 or more states, an 8-way partition of a QAM constellation is used. Ungerboeck also gives a comparable set of 1-dimensional codes based on a 4-way partition of a PAM signal constellation, which also achieve from 3 to 6 dB as the number of states goes from 4 to 256, but these codes have a code redundancy of 2 bits per two-dimensional symbol, and thus a constellation expansion ratio of 4.

Wei's multidimensional codes achieve lower constellation expansion ratios by only using one bit of redundancy for every N dimensions, where N > 2. For example, if N = 4, then the constellation expansion is only about $2^{1/2}$, and if N = 8, it is only about $2^{1/4}$. Wei gives a number of families of such codes in 4, 8 and 16 dimensions, which also achieve nominal coding gains up to about 6 dB. These codes generally have a low number of nearest neighbors, and are very well suited for implementation.

Figure 2 (from [11]) shows the effective coding gains versus a normalized measure of decoding complexity for Ungerboeck's one-dimensional and two-dimensional codes, and for Wei's codes. Wei's 16-state 4-dimensional code is more than 0.5 dB better than the 8-state 2-dimensional Ungerboeck-type code used in the V.32 and V.33 modems, but generally the curves for the various classes of codes are quite close. The major advantage of Wei's codes is their reduced constellation expansion.

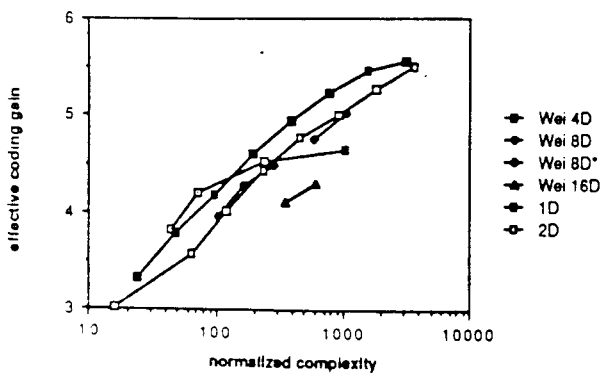

Figure 2: Performance vs. complexity for Ungerboeck 1D and 2D codes, and for Wei 4D, 8D, and 16D codes.

Calderbank and Sloane [12] have also developed broad classes of trellis codes based on partitions of lattices into cosets of sublattices. The constellation expansion ratio of the Calderbank-Sloane codes tends to be higher than that of the Wei codes, and the performance vs. complexity curve is about the same. However, the introduction of the lattice/coset viewpoint in this work was probably the most important advance in the theory of trellis codes after Ungerboeck [2].

6. Conclusion

The performance attainable with practical trellis codes is now approaching the Shannon limit. Ungerboeck's 256-state codes obtain effective coding gains of about 5.5 dB, at the cost of high but not unthinkable complexity. With trellis shaping, another 1 dB can be obtained. This is most of the 7.5 dB of coding gain and 1.5 dB of shaping gain that is possible in principle. DeBuda's result shows that with more complicated codes of the same type, it should be possible to get as close to the Shannon limit as desired.

Does this mean that now there is nothing more to be done in coded modulation? Not in our opinion. Ungerboeck [2] speculated that there was little to be gained by going to multidimensional codes, but Wei [5] showed that there was a gain, not in performance but in constellation expansion (and rotational invariance). The concept of shaping gain, and practical techniques for obtaining most of the possible shaping gain, have been developed since Wei.

On band-limited channels, such as the telephone channel, it is important to be able to combine coding with equalization. Very recently, with a technique called trellis precoding [13], it has been shown that coding, shaping, and equalization can all be combined in such a way that one can get as close to capacity on ISI channels as on ideal channels. (Another way of doing this is by multi-channel techniques, such as are used in Telebit's modems.) This development, in combination with line probing techniques that determine the optimum transmission rate and frequency band, will make possible the achievement of data rates up to 24 Kbps on good enough channels.

The question of how to obtain even higher coding gains is still open. Multi-stage codes look promising from the point of view of performance vs. complexity, but their promise has not yet been proved. Sequential decoding is another technique capable of high performance. The theory of trellis codes is still in a very formative stage, with attention now turning to geometric theories. It has always been the case in the past that advances in the theory have led to practical improvements.

Finally, these codes are just beginning to be considered for application to other band-limited channels, such as radio channels, satellite channels, or the high-rate digital subscriber loop (HDSL) channel. Each one of these applications will present its own problems, including that of high-speed VLSI implementation, and may require the development of new types of codes in response. In summary, there will be room for work in trellis codes for some time to come.

continued on page 7

ISIT'91 Update

Anthony Ephremides

The next IEEE International Symposium on Information Theory will take place, of course, in Budapest, Hungary next June from the 23rd to the 28th as most of us already know. By now, it is of course too late to send in a paper, since the deadline for submissions is already past. The Program Committee is working diligently and hard to sort out the final program based on the large number of submissions it has received. However, it is not too late to contemplate attending the Symposium if you haven't done so already. Not only is it the first time that a major IEEE Society Symposium is taking place in what used to be Eastern Europe, but, in addition, the exciting changes taking place in that part of the world promise to make the visit to Budapest next year a most interesting affair from many points of view.

An interesting related development is that the annual Communication Theory Workshop that is sponsored by the IEEE Communication Society has been planned to take place during the week immediately following our Symposium on the Greek island of Rhodes. Those who wish to attend both meetings can schedule their trip conveniently so as to include both places.

The Organizing Committee is in the process of seeking funds to support a small number of attendants with travel expenses. When the available amount is finalized, there will be an announcement and solicitation for application from interested parties. As usual, the expected limited funds will be used to assist authors who are either students or young researchers who do not have other means of supporting their travel.

Budapest, as many of you know, is a beautiful city with rich historical heritage and many points of interest. It can be reached by air or rail or car easily from most cities in Europe. When the program is finalized, there will be detailed additional information that will facilitate your travel plans. There is a new, very up-to-date Conference Center in Budapest that will host our Symposium. It is conveniently located in a nice part of the town near many hotels and restaurants and adjacent to a beautiful park. Start planning now for an unforgettable experience next June.

CODED MODULATION

continued from page 6

References

1. C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379-423 and 623-656, 1948.
2. G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 55-67, 1982.
3. G. D. Forney, Jr., R. G. Gallager, G. R. Lang, F. M. Longstaff and S. U. Qureshi, "Efficient modulation for band-limited channels," *IEEE J. Select. Areas Commun.*, vol. SAC-2, pp. 632-647, 1984.
4. L.-F. Wei, "Rotationally invariant convolutional channel coding with expanded signal space. Part II: Nonlinear codes," *IEEE J. Select. Areas Commun.*, vol. SAC-2, pp. 672-686, 1984.
5. L.-F. Wei, "Trellis-coded modulation with multidimensional constellations," *IEEE Trans. Inform. Theory*, vol. IT-33, pp.483-501, 1987.
6. 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. SAC-7, pp. 877-892, 1989.
7. J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*. New York: Springer-Verlag, 1988.
8. G. R. Lang and F. M. Longstaff, "A Leech lattice modem," *IEEE J. Select. Areas Commun.*, vol. SAC-7, pp. 968-973, 1989.
9. G. D. Forney, Jr., "Trellis shaping," IEEE Workshop on Information Theory, Ithaca, N.Y., June 1989.
10. R. deBuda, "Some optimal codes have structure," *IEEE J. Select. Areas Commun.*, vol. SAC-7, pp. 893-899, 1989.
11. G. D. Forney, Jr., "Coset codes—Part I: Introduction and geometrical classification," *IEEE Trans. Inform. Theory*, vol. IT-34, pp. 1123-1151, 1988.
12. A. R. Calderbank and N. J. A. Sloane, "New trellis codes based on lattices and cosets," *IEEE Trans. Inform. Theory*, vol. IT-33, pp 177-195, 1987.
13. M. V. Eyuboglu and G. D. Forney, Jr., "Trellis precoding," 1990 IEEE International Symposium on Information Theory, San Diego, Calif., January 1990.