Trellis-Coded Modulation with Redundant Signal Sets Part I: Introduction

Gottfried Ungerboeck

Simple four-state trellis-coded modulation (TCM) schemes improve the robustness of digital transmission against additive noise by 3 dB without reducing data rate or requiring more bandwidth than conventional uncoded modulation schemes. With more complex schemes, coding gains up to 6 dB can be achieved. This article describes how TCM works

0163-6804/87/0002-0005 $01.00 © 1987 IEEE

Trellis-Coded Modulation (TCM) has evolved over the past decade as a combined coding and modulation technique for digital transmission over band-limited channels. Its main attraction comes from the fact that it allows the achievement of significant coding gains over conventional uncoded multilevel modulation without compromising bandwidth efficiency. The first TCM schemes were proposed in 1976 [1]. Following a more detailed publication [2] in 1982, an explosion of research and actual implementations of TCM took place, to the point where today there is a good understanding of the theory and capabilities of TCM methods. In Part I of this two-part article, an introduction into TCM is given. The reasons for the development of TCM are reviewed, and examples of simple TCM schemes are discussed. Part II [15] provides further insight into code design and performance, and addresses recent advances in TCM.

TCM schemes employ redundant nonbinary modulation in combination with a finite-state encoder which governs the selection of modulation signals to generate coded signal sequences. In the receiver, the noisy signals are decoded by a soft-decision maximum-likelihood sequence decoder. Simple four-state TCM schemes can improve the robustness of digital transmission against additive noise by 3 dB, compared to conventional uncoded modulation. With more complex TCM schemes, the coding gain can reach 6 dB or more. These gains are obtained without bandwidth expansion or reduction of the effective information rate as required by traditional error-correction schemes. Shannon's information theory predicted the existence of coded modulation schemes with these characteristics more than three decades ago. The development of effective TCM techniques and today's signal-processing technology now allow these gains to be obtained in practice.

Signal waveforms representing information sequences are most impervious to noise-induced detection errors if they are very different from each other. Mathematically, this translates into the requirement that signal sequences should have large distance in Euclidean signal space. The essential new concept of TCM that led to the aforementioned gains was to use signal-set expansion to provide redundancy for coding, and to design coding and signal-mapping functions jointly so as to maximize directly the "free distance" (minimum Euclidean distance) between coded signal sequences. This allowed the construction of modulation codes whose free distance significantly exceeded the minimum distance between uncoded modulation signals, at the same information rate, bandwidth, and signal power. The term "trellis" is used because these schemes can be described by a state-transition (trellis) diagram similar to the trellis diagrams of binary convolutional codes. The difference is that in TCM schemes, the trellis branches are labeled with redundant nonbinary modulation signals rather than with binary code symbols.

The basic principles of TCM were published in 1982 [2]. Further descriptions followed in 1984 [3-6], and coincided with a rapid transition of TCM from the research stage to practical use. In 1984, a TCM scheme with a coding gain of 4 dB was adopted by the International Telegraph and Telephone Consultative Commit-

tee (CCITT) for use in new high-speed voiceband modems [5,7,8]. Prior to TCM, uncoded transmission at 9.6 kbit/s over voiceband channels was often considered as a practical limit for data modems. Since 1984, data modems have appeared on the market which employ TCM along with other improvements in equalization, synchronization, and so forth, to transmit data reliably over voiceband channels at rates of 14.4 kbit/s and higher. Similar advances are being achieved in transmission over other bandwidth-constrained channels. The common use of TCM techniques in such applications, as satellite [9-11], terrestrial microwave, and mobile communications, in order to increase throughput rate or to permit satisfactory operation at lower signal-to-noise ratios, can be safely predicted for the near future.

Classical Error-Correction Coding

In classical digital communication systems, the functions of modulation and error-correction coding are separated. Modulators and demodulators convert an analog waveform channel into a discrete channel, whereas encoders and decoders correct errors that occur on the discrete channel.

In conventional multilevel (amplitude and/or phase) modulation systems, during each modulation interval the modulator maps m binary symbols (bits) into one of $M = 2^m$ possible transmit signals, and the demodulator recovers the m bits by making an independent M-ary nearest-neighbor decision on each signal received. Figure 1 depicts constellations of real- or complex-valued modulation amplitudes, henceforth called signal sets, which are commonly employed for one- or two-dimensional M-ary linear modulation. Two-dimensional carrier modulation requires a bandwidth of 1/T Hz around the carrier frequency to transmit signals at a modulation rate of 1/T signals/sec (baud) without intersymbol interference. Hence, two-dimensional 2^m-ary modulation systems can achieve a spectral efficiency of about m bit/sec/Hz. (The same spectral efficiency is obtained with one-dimensional $2^{m/2}$-ary baseband modulation.)

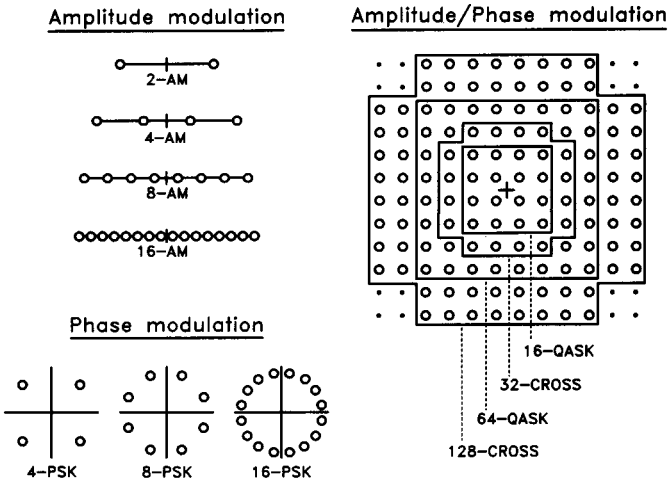

Fig. 1. Signal sets for one-dimensional amplitude modulation, and two-dimensional phase and amplitude/phase modulation.

Conventional encoders and decoders for error correction operate on binary, or more generally Q-ary, code symbols transmitted over a discrete channel. With a code of rate $k/n < 1$, $n - k$ redundant check symbols are appended to every k information symbols. Since the decoder receives only discrete code symbols, Hamming distance (the number of symbols in which two code sequences or blocks differ, regardless of how these symbols differ) is the appropriate measure of distance for decoding and hence for code design. A minimum Hamming distance d^H_{min}, also called "free Hamming distance" in the case of convolutional codes, guarantees that the decoder can correct at least $[(d^H_{min} - 1)/2]$ code-symbol errors. If low signal-to-noise ratios or non-stationary signal disturbance limit the performance of the modulation system, the ability to correct errors can justify the rate loss caused by sending redundant check symbols. Similarly, long delays in error-recovery procedures can be a good reason for trading transmission rate for forward error-correction capability.

Generally, there exist two possibilities to compensate for the rate loss: increasing the modulation rate if the channel permits bandwidth expansion, or enlarging the signal set of the modulation system if the channel is band-limited. The latter necessarily leads to the use of nonbinary modulation ($M > 2$). However, when modulation and error-correction coding are performed in the classical independent manner, disappointing results are obtained.

As an illustration, consider four-phase modulation (4-PSK) without coding, and eight-phase modulation (8-PSK) used with a binary error-correction code of rate 2/3. Both systems transmit two information bits per modulation interval (2 bit/sec/Hz). If the 4-PSK system operates at an error rate of 10^{-5}, at the same signal-to-noise ratio the "raw" error rate at the 8-PSK demodulator exceeds 10^{-2} because of the smaller spacing between the 8-PSK signals. Patterns of at least three bit errors must be corrected to reduce the error rate to that of the uncoded 4-PSK system. A rate-2/3 binary convolutional code with constraint length $\nu = 6$ has the required value of $d^H_{min} = 7$ [12]. For decoding, a fairly complex 64-state binary Viterbi decoder is needed. However, after all this effort, error performance only breaks even with that of uncoded 4-PSK.

Two problems contribute to this unsatisfactory situation.

Soft-Decision Decoding and Motivation for New Code Design

One problem in the coded 8-PSK system just described arises from the independent "hard" signal decisions made prior to decoding which cause an irreversible loss of information in the receiver. The remedy for this problem is soft-decision decoding, which means that the decoder operates directly on unquantized "soft" output samples of the channel. Let the samples be $r_n = a_n + w_n$ (real- or complex-valued, for one- or two-dimensional modulation, respectively), where the a_n are the discrete signals sent by the modulator, and the w_n represent samples of an additive white Gaussian noise process. The decision rule of the optimum sequence decoder is to

determine, among the set C of all coded signal sequences which a cascaded encoder and modulator can produce, the sequence $\{\hat{a}_n\}$ with minimum squared Euclidean distance (sum of squared errors) from $\{r_n\}$, that is, the sequence $\{\hat{a}_n\}$ which satisfies

$$|r_n - \hat{a}_n|^2 = \min_{\{\hat{a}_n\} \epsilon C} \sum |r_n - a_n|^2.$$

The Viterbi algorithm, originally proposed in 1967 [13] as an "asymptotically optimum" decoding technique for convolutional codes, can be used to determine the coded signal sequence $\{\hat{a}_n\}$ closest to the received unquantized signal sequence $\{r_n\}$ [12,14], provided that the generation of coded signal sequences $\{a_n\} \epsilon C$ follows the rules of a finite-state machine. However, the notion of "error-correction" is then no longer appropriate, since there are no hard-demodulator decisions to be corrected. The decoder determines the most likely coded signal sequence directly from the unquantized channel outputs.

The most probable errors made by the optimum soft-decision decoder occur between signals or signal sequences $\{a_n\}$ and $\{b_n\}$, one transmitted and the other decoded, that are closest together in terms of squared Euclidean distance. The minimum squared such distance is called the squared "free distance:"

$$d^2_{free} = \min_{\{a_n\} \neq \{b_n\}} \sum |a_n - b_n| \; ; \quad \{a_n\}, \{b_n\} \epsilon C.$$

When optimum sequence decisions are made directly in terms of Euclidean distance, a second problem becomes apparent. Mapping of code symbols of a code optimized for Hamming distance into nonbinary modulation signals does not guarantee that a good Euclidean distance structure is obtained. In fact, generally one cannot even find a monotonic relationship between Hamming and Euclidean distances, no matter how code symbols are mapped.

For a long time, this has been the main reason for the lack of good codes for multilevel modulation. Squared Euclidean and Hamming distances are equivalent only in the case of binary modulation or four-phase modulation, which merely corresponds to two orthogonal binary modulations of a carrier. In contrast to coded multilevel systems, binary modulation systems with codes optimized for Hamming distance and soft-decision decoding have been well established since the late 1960s for power-efficient transmission at spectral efficiencies of less than 2 bit/sec/Hz.

The motivation of this author for developing TCM initially came from work on multilevel systems that employ the Viterbi algorithm to improve signal detection in the presence of intersymbol interference. This work provided him with ample evidence of the importance of Euclidean distance between signal sequences. Since improvements over the established technique of adaptive equalization to eliminate intersymbol interference and then making independent signal decisions in most cases did not turn out to be very significant, he turned his attention to using coding to improve performance. In this connection, it was clear to him that codes should be designed for maximum free Euclidean distance rather than Hamming distance, and that the redundancy necessary for coding would have to come from expanding the signal set to avoid bandwidth expansion.

To understand the potential improvements to be expected by this approach, he computed the channel capacity of channels with additive Gaussian noise for the case of discrete multilevel modulation at the channel input and unquantized signal observation at the channel output. The results of these calculations [2] allowed making two observations: firstly, that in principle coding gains of about 7-8 dB over conventional uncoded multilevel modulation should be achievable, and secondly, that most of the achievable coding gain could be obtained by expanding the signal sets used for uncoded modulation only by the factor of two. The author then concentrated his efforts on finding trellis-based signaling schemes that use signal sets of size 2^{m+1} for transmission of m bits per modulation interval. This direction turned out to be succesful and today's TCM schemes still follow this approach.

The next two sections illustrate with two examples how TCM schemes work. Whenever distances are discussed, Euclidean distances are meant.

Four-State Trellis Code for 8-PSK Modulation

The coded 8-PSK scheme described in this section was the first TCM scheme found by the author in 1975 with a significant coding gain over uncoded modulation. It was designed in a heuristic manner, like other simple TCM systems shortly thereafter. Figure 2 depicts signal sets and state-transition (trellis) diagrams for a) uncoded 4-PSK modulation and b) coded 8-PSK modulation with four trellis states. A trivial one-state trellis diagram is shown in Fig. 2a only to illustrate uncoded 4-PSK from the viewpoint of TCM. Every connected path through a trellis in Fig. 2 represents an allowed signal sequence. In

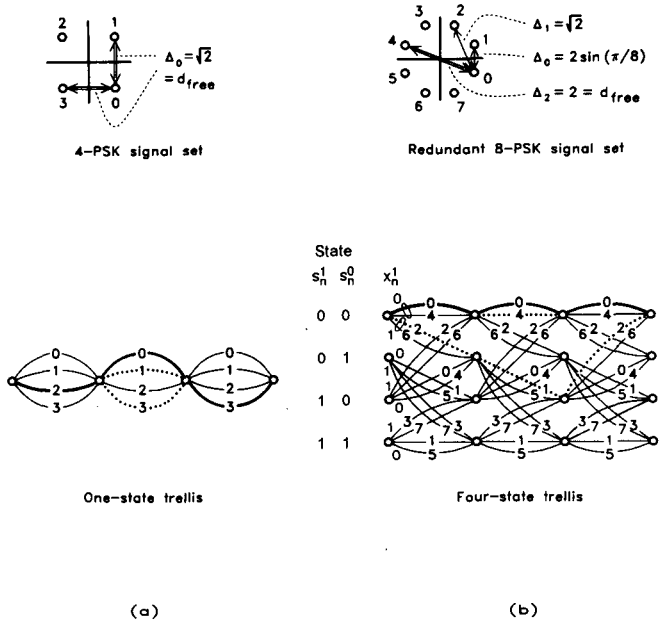

Fig. 2. (a) Uncoded four-phase modulation (4-PSK), (b) Four-state trellis-coded eight-phase modulation (8-PSK).

both systems, starting from any state, four transitions can occur, as required to encode two information bits per modulation interval (2 bit/sec/Hz). For the following discussion, the specific encoding of information bits into signals is not important.

The four "parallel" transitions in the one-state trellis diagram of Fig. 2a for uncoded 4-PSK do not restrict the sequences of 4-PSK signals that can be transmitted, that is, there is no sequence coding. Hence, the optimum decoder can make independent nearest-signal decisions for each noisy 4-PSK signal received. The smallest distance between the 4-PSK signals is $\sqrt{2}$, denoted as Δ_0. We call it the "free distance" of uncoded 4-PSK modulation to use common terminology with sequence-coded systems. Each 4-PSK signal has two nearest-neighbor signals at this distance.

In the four-state trellis of Fig. 2b for the coded 8-PSK scheme, the transitions occur in pairs of two parallel transitions. (A four-state code with four distinct transitions from each state to all successor states was also considered; however, the trellis as shown with parallel transitions permitted the achievement of a larger free distance.) Fig. 2b shows the numbering of the 8-PSK signals and relevant distances between these signals: $\Delta_0 = 2\sin(\pi/8)$, $\Delta_1 = \sqrt{2}$, and $\Delta_2 = 2$. The 8-PSK signals are assigned to the transitions in the four-state trellis in accordance with the following rules:

a) Parallel transitions are associated with signals with maximum distance Δ_2(8-PSK) = 2 between them, the signals in the subsets (0,4), (1,5), (2,6), or (3,7).
b) Four transitions originating from or merging in one state are labeled with signals with at least distance Δ_1(8-PSK) = $\sqrt{2}$ between them, that is, the signals in the subsets (0,4,2,6) or (1,5,3,7).
c) All 8-PSK signals are used in the trellis diagram with equal frequency.

Any two signal paths in the trellis of Fig. 2(b) that diverge in one state and remerge in another after more than one transition have at least squared distance $\Delta_1^2 + \Delta_0^2 + \Delta_1^2 = \Delta_2^2 + \Delta_0^2$ between them. For example, the paths with signals 0-0-0 and 2-1-2 have this distance. The distance between such paths is greater than the distance between the signals assigned to parallel transitions, Δ_2(8-PSK) = 2, which thus is found as the free distance in the four-state 8-PSK code: $d_{free} = 2$. Expressed in decibels, this amounts to an improvement of 3 dB over the minimum distance $\sqrt{2}$ between the signals of uncoded 4-PSK modulation. For any state transition along any coded 8-PSK sequence transmitted, there exists only one nearest-neighbor signal at free distance, which is the 180° rotated version of the transmitted signal. Hence, the code is invariant to a signal rotation by 180°, but to no other rotations (cf., Part II). Figure 3 illustrates one possible realization of an encoder-modulator for the four-state coded 8-PSK scheme.

Soft-decision decoding is accomplished in two steps: In the first step, called "subset decoding", within each subset of signals assigned to parallel transitions, the signal closest to the received channel output is determined. These signals are stored together with their squared distances from the channel output. In the second step, the Viterbi algorithm is used to find the signal path through the code trellis with the minimum sum of squared distances from the sequence of noisy channel outputs received. Only the signals already chosen by subset decoding are considered.

Tutorial descriptions of the Viterbi algorithm can be found in several textbooks, for example, [12]. The essential points are summarized here as follows: assume that the optimum signal paths from the infinite past to all trellis states at time n are known; the algorithm extends these paths iteratively from the states at time n to the states at time n + 1 by choosing one best path to each new state as a "survivor" and "forgetting" all other paths that cannot be extended as the best paths to the new states; looking backwards in time, the "surviving" paths tend to merge into the same "history path" at some time n − d; with a sufficient decoding delay D (so that the randomly changing value of d is highly likely to be smaller than D), the information associated with a transition on the common history path at time n − D can be selected for output.

Let the received signals be disturbed by uncorrelated Gaussian noise samples with variance σ^2 in each signal dimension. The probability that at any given time the decoder makes a wrong decision among the signals associated with parallel transitions, or starts to make a sequence of wrong decisions along some path diverging for more than one transition from the correct path, is called the error-event probability. At high signal-to-noise ratios, this probability is generally well approximated by

$$Pr(e) \simeq N_{free} \cdot Q[d_{free}/(2\sigma)],$$

where Q(.) represents the Gaussian error integral

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty exp(-y^2/2)dy,$$

and N_{free} denotes the (average) number of nearest-neighbor signal sequences with distance d_{free} that diverge at any state from a transmitted signal sequence, and remerge with it after one or more transitions. The above approximate formula expresses the fact that at high

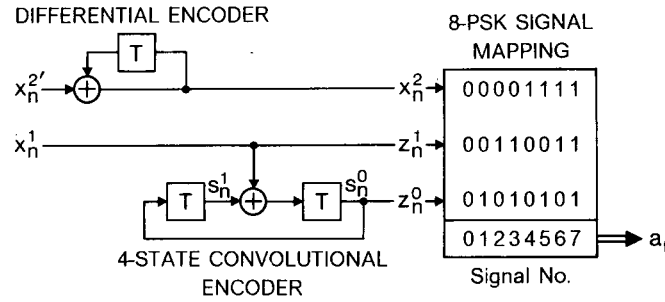

Fig. 3. Illustrates an encoder for the four-state 8-PSK code.

signal-to-noise ratios the probability of error events associated with a distance larger then d_{free} becomes negligible.

For uncoded 4-PSK, we have $d_{free} = \sqrt{2}$ and $N_{free} = 2$, and for four-state coded 8-PSK we found $d_{free} = 2$ and $N_{free} = 1$. Since in both systems free distance is found between parallel transitions, single signal-decision errors are the dominating error events. In the special case of these simple systems, the numbers of nearest neighbors do not depend on which particular signal sequence is transmitted.

Figure 4 shows the error-event probability of the two systems as a function of signal-to-noise ratio. For uncoded 4-PSK, the error-event probability is extremely well approximated by the last two equations above. For four-state coded 8-PSK, these equations provide a lower bound that is asymptotically achieved at high signal-to-noise ratios. Simulation results are included in Fig. 4 for the coded 8-PSK system to illustrate the effect of error events with distance larger than free distance, whose probability of occurrence is not negligible at low signal-to-noise ratios.

Figure 5 illustrates a noisy four-state coded 8-PSK signal as observed at complex baseband before sampling in the receiver of an experimental 64 kbit/s satellite modem [9]. At a signal-to-noise ratio of $E_s/N_0 = 12.6$ dB (E_s: signal energy, N_0: one-sided spectral noise density), the signal is decoded essentially error-free. At the same signal-to-noise ratio, the error rate with uncoded 4-PSK modulation would be around 10^{-5}.

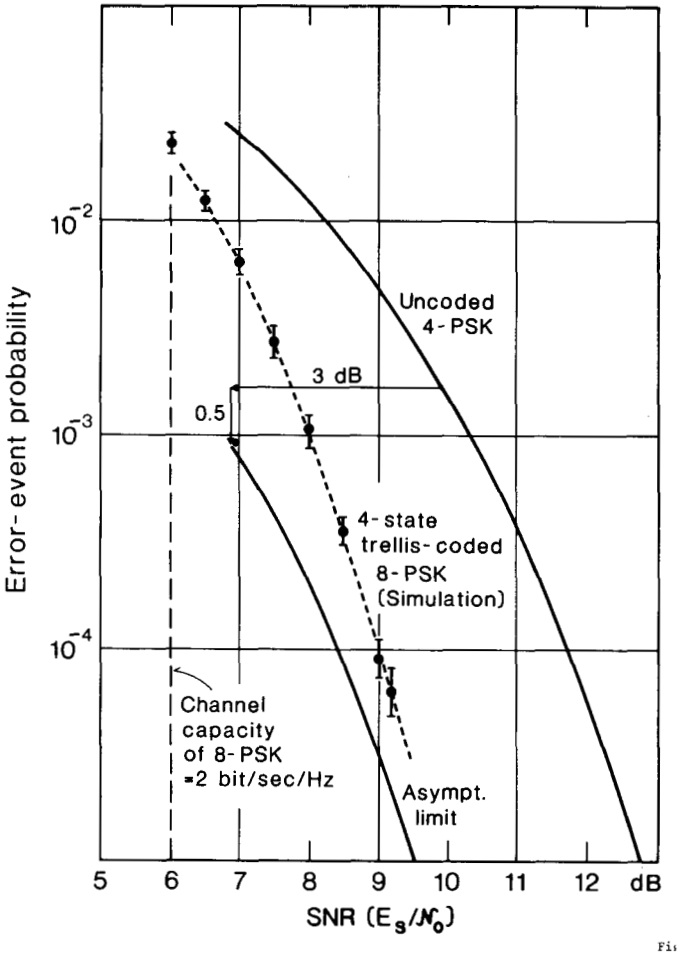

Fig. 4. Error-event probability versus signal-to-noise ratio for uncoded 4-PSK and four-state coded 8-PSK.

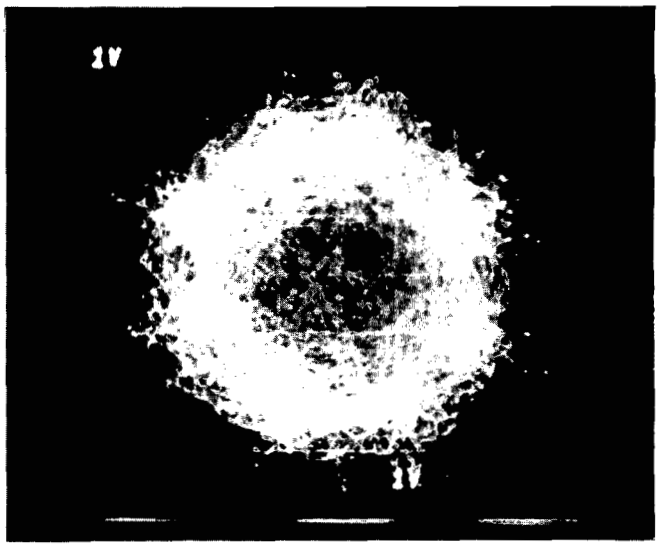

Fig. 5. Noisy four-state coded 8-PSK signal at complex baseband with a signal-to-noise ratio of $E_sN_0 = 12.6$ dB.

In TCM schemes with more trellis states and other signal sets, d_{free} is not necessarily found between parallel transitions, and N_{free} will generally be an average number larger than one, as will be shown by the second example.

Eight-State Trellis Code for Amplitude/Phase Modulation

The eight-state trellis code discussed in this section was designed for two-dimensional signal sets whose signals are located on a quadratic grid, also known as a lattice of type "Z_2". The code can be used with all of the signal sets depicted in Fig. 1 for amplitude/phase modulation. To transmit m information bits per modulation interval, a signal set with 2^{m+1} signals is needed. Hence, for m = 3 the 16-QASK signal set is used, for m = 4 the 32-CROSS signal set, and soforth. For any m, a coding gain of approximately 4 dB is achieved over uncoded modulation.

Figure 6 illustrates a "set partitioning" of the 16-QASK and 32-CROSS signal sets into eight subsets. The partitioning of larger signal sets is done in the same way. The signal set chosen is denoted by A0, and its subsets by D0, D1, . . . D7. If the smallest distance among the signals in A0 is Δ_0, then among the signals in the union of the subsets D0,D4,D2,D6 or D1,D5,D3,D7 the minimum distance is $\sqrt{2}\ \Delta_0$, in the union of the subsets D0,D4; D2,D6; D1,D5; or D3,D7 it is $\sqrt{4}\ \Delta_0$, and within the individual subsets it is $\sqrt{8}\ \Delta_0$. (A conceptually similar partitioning of the 8-PSK signal set into smaller signal sets with increasing intra-set distances was implied in the example of coded 8-PSK. The fundamental importance

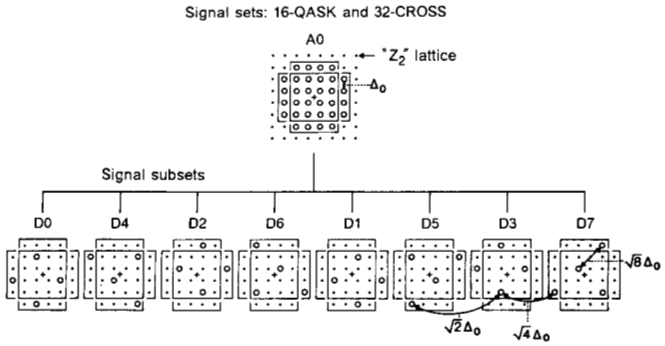

Fig. 6. Set partitioning of the 16-QASK and 32-CROSS signal sets.

of this partitioning for TCM codes will be explained in Part II.)

In the eight-state trellis depicted in Fig. 7, four transitions diverge from and merge into each state. To each transition, one of the subsets D0, . . . D7 is assigned. If A0 contains 2^{m+1} signals, each of its subsets will comprise 2^{m-2} signals. This means that the transitions shown in Fig. 7 in fact represent 2^{m-2} parallel transitions in the same sense as there were two parallel transitions in the coded 8-PSK scheme. Hence, 2^m signals can be sent from each state, as required to encode m bits per modulation interval.

The assignment of signal subsets to transitions satisfies the same three rules as discussed for coded 8-PSK, appropriately adapted to the present situation. The four transitions from or to the same state are always assigned either the subsets D0,D4,D2,D6 or D1,D5,D3,D7. This guarantees a squared signal distance of at least $2\Delta_0^2$ when sequences diverge and when they remerge. If paths remerge after two transitions, the squared signal distance is at least $4\Delta_0^2$ between the diverging transitions, and hence the total squared distance between such paths will be at least $6\Delta_0^2$. If paths remerge after three or more transitions, at least one intermediate transition contributes an additional squared signal distance Δ_0^2, so the squared distance between sequences is at least $\sqrt{5}\ \Delta_0$.

Hence, the free distance of this code is $\sqrt{5}\ \Delta_0$. This is smaller than the minimum signal distance within in the subsets D0, . . . D7, which is $\sqrt{8}\ \Delta_0$. For one particular code sequence D0-D0-D3-D6, Fig. 6 illustrates four error paths at distance $\sqrt{5}\ \Delta_0$ from that code sequence; all starting at the same state and remerging after three or four transitions. It can be shown that for any code

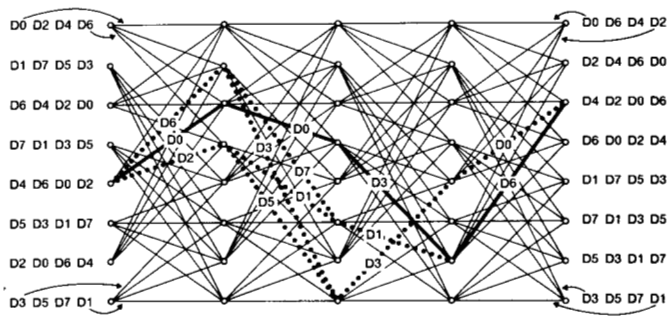

Fig. 7. Eight-state trellis code for amplitude/phase modulation with "Z_2"-type signal sets; $d_{free} = \sqrt{5}\ \Delta_0$.

sequence and from any state along this sequence, there are four such paths, two of length three and two of length four. The most likely error events will correspond to these error paths, and will result in bursts of decision errors of length three or four.

The coding gains asymptotically achieved at high signal-to-noise ratios are calculated in decibels by

$$G_{c\,u} = 10\ log_{10}\ [(d^2_{free,c}/d^2_{free,u})/E_{s,c}/E_{s,u})],$$

where $d^2_{free,c}$ and $d^2_{free,u}$ are the squared free distances, and $E_{s,c}$ and $E_{s,u}$ denote the average signal energies of the coded and uncoded schemes, respectively. When the signal sets have the same minimum signal spacing Δ_0, $d^2_{free,c}/d^2_{free,u} = 5$, and $E_{s,c}/E_{s,u} \simeq 2$ for all relevant values of m. Hence, the coding gain is $10\ log_{10}(5/2) \simeq 4$ dB.

The number of nearest neighbors depends on the sequence of signals transmitted, that is N_{free} represents an average number. This is easy to see for uncoded modulation, where signals in the center of a signal set have more nearest neighbors than the outer ones. For uncoded 16-QASK, N_{free} equals 3. For eight-state coded 16-QASK, N_{free} is around 3.75. In the limit of large "Z_2"-type signal sets, these values increase toward 4 and 16 for uncoded and eight-state coded systems, respectively.

Trellis Codes of Higher Complexity

Heuristic code design and checking of code properties by hand, as was done during the early phases of the development of TCM schemes, becomes infeasible for codes with many trellis states. Optimum codes must then be found by computer search, using knowledge of the general structure of TCM codes and an efficient method to determine free distance. The search technique should also include rules to reject codes with improper or equivalent distance properties without having to evaluate free distance.

In Part II, the principles of TCM code design are outlined, and tables of optimum TCM codes given for one-, two-, and higher-dimensional signal sets. TCM encoder/modulators are shown to exhibit the following general structure: (a) of the m bits to be transmitted per encoder/modulator operation, $\tilde{m} \leq m$ bits are expanded into $\tilde{m} + 1$ coded bits by a binary rate-$\tilde{m}/(\tilde{m}+1)$ convolutional encoder; (b) the $\tilde{m} + 1$ coded bits select one of $2^{\tilde{m}+1}$ subsets of a redundant 2^{m+1}-ary signal set; (c) the remaining $m-\tilde{m}$ bits determine one of $2^{m-\tilde{m}}$ signals within the selected subset.

New Ground Covered by Trellis-Coded Modulation

TCM schemes achieve significant coding gains at values of spectral efficiency for which efficient coded-modulation schemes were not previously known, that is, above and including 2 bit/sec/Hz. Figure 8 shows the free distances obtained by binary convolutional coding with 4-PSK modulation for spectral efficiencies smaller than 2 bit/sec/Hz, and by TCM schemes with two-dimensional signal sets for spectral efficiencies equal to or larger than 2 bit/sec/Hz. The free distances of uncoded modulation at the respective spectral effi-

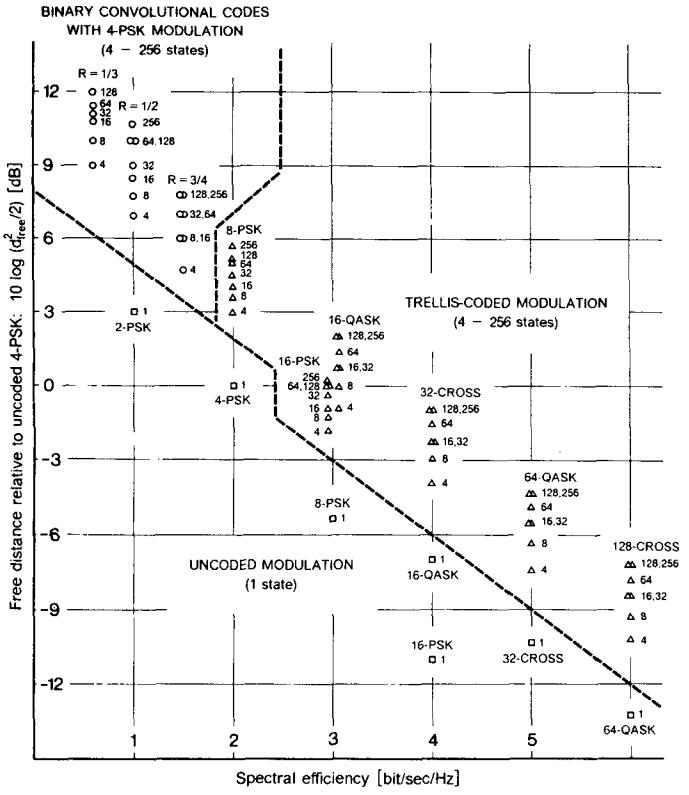

Fig. 8. Free distance of binary convolutional codes with 4-PSK modulation, and TCM with a variety of two-dimensional modulation schemes, for spectral efficiencies from 2/3 to 6 bit/sec/Hz.

ciencies are also depicted. The average signal energy of all signal sets is normalized to unity. Free distances are expressed in decibels relative to the value $d^2_{free} = 2$ of uncoded 4-PSK modulation. The binary convolutional codes of rates 1/3, 1/2, and 3/4 with optimum Hamming distances are taken from textbooks, such as, [12]. The TCM codes and their properties are found in the code tables presented in Part II (largely reproduced from [2]).

All coded systems achieve significant distance gains with as few as 4, 8, and 16 code states. Roughly speaking, it is possible to gain 3 dB with 4 states, 4 dB with 8 states, nearly 5 dB with 16 states, and up to 6 dB with 128 or more states. The gains obtained with two-state codes usually are very modest. With higher numbers of states, the incremental gains become smaller. Doubling the number of states does not always yield a code with larger free distance. Generally, limited distance growth and increasing numbers of nearest neighbors, and neighbors with next-larger distances, are the two mechanisms that prevent real coding gains from exceeding the ultimate limit set by channel capacity. This limit can be characterized by the signal-to-noise ratio at which the channel capacity of a modulation system with a 2^{m+1}-ary signal set equals m bit/sec/Hz [2] (see also Fig. 4).

Conclusion

Trellis-coded modulation was invented as a method to improve the noise immunity of digital transmission systems without bandwidth expansion or reduction of data rate. TCM extended the principles of convolutional coding to nonbinary modulation with signal sets of arbitrary size. It allows the achievement of coding gains of 3-6 dB at spectral efficiencies equal to or larger than 2 bit/sec/Hz. These are the values at which one wants to operate on many band-limited channels. Thus, a gap in the theory and practice of channel coding has been closed.

References

[1] G. Ungerboeck and I. Csajka, "On improving data-link performance by increasing the channel alphabet and introducing sequence coding," 1976 Int. Symp. Inform. Theory, Ronneby, Sweden, June 1976.

[2] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Trans. Information Theory*, vol. IT-28, pp. 55-67, Jan. 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 Trans. Selected Areas in Comm.*, vol. SAC-2, pp. 632-647, Sept. 1984.

[4] L. F. Wei, "Rotationally invariant convolutional channel coding with expanded signal space—Part I: 180 degrees," *IEEE Trans. Selected Areas in Comm.*, vol. SAC-2, pp. 659-672, Sept. 1984.

[5] L. F. Wei, "Rotationally invariant convolutional channel coding with expanded signal space—Part II: nonlinear codes," *IEEE Trans. Selected Areas in Comm.*, vol. SAC-2, pp. 672-686, Sept. 1984.

[6] A. R. Calderbank and J. E. Mazo, "A new description of trellis codes," *IEEE Trans. Information Theory*, vol. IT-30, pp. 784-791, Nov. 1984.

[7] CCITT Study Group XVII, "Recommendation V.32 for a family of 2-wire, duplex modems operating on the general switched telephone network and on leased telephone-type circuits," Document AP VIII-43-E, May 1984.

[8] CCITT Study Group XVII, "Draft recommendation V.33 for 14400 bits per second modem standardized for use on point-to-point 4-wire leased telephone-type circuits," Circular No. 12, COM XVII/YS, Geneva, May 17, 1985.

[9] G. Ungerboeck, J. Hagenauer, and T. Abdel Nabi, "Coded 8-PSK experimental modem for the INTELSAT SCPC system," Proc. 7th Int. Conf. on Digital Satellite Communications (ICDS-7), pp. 299-304, Munich, May 12-16, 1986.

[10] R. J. F. Fang, "A coded 8-PSK system for 140-Mbit/s information rate transmission over 80-MHz nonlinear transponders," Proc. 7th Int. Conf. on Digital Satellite Communications (ICDS-7), pp. 305-313, Munich, May 12-16, 1986.

[11] T. Fujino, Y. Moritani, M. Miyake, K. Murakami, Y. Sakato, and H. Shiino, "A 120 Mbit/s 8PSK modem with soft-Viterbi decoding," Proc. 7th Int. Conf. on Digital Satellite Communications (ICDS-7), pp. 315-321, Munich, May 12-16, 1986.

[12] G. C. Clark and J. B. Cain, *Error-Correction Coding for Digital Communications*, Plenum Press, New York and London, 1981.

[13] A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Information Theory*, vol. IT-13, pp. 260-269, April 1967.

[14] G. D. Forney, Jr., "The Viterbi algorithm," *Proc. of the IEEE*, vol. 61, pp. 268-278, March 1973.

[15] G. Ungerboeck, "Trellis-coded modulation with redundant signal sets, Part II: State of the art," *IEEE Communications Magazine*, vol. 25, no. 2, Feb. 1987.