Trellis-Coded Modulation with Redundant Signal Sets Part II: State of the Art

Gottfried Ungerboeck

This article is intended to bring the reader up to the state of the art in trellis-coded modulation. The general principles that have proven useful in code design are explained. The important effects of carrier-phase offset and phase invariance are discussed. Finally, recent work in trellis-coded modulation with multi-dimensional signal sets is described

In this second part [1], a synopsis of the present state of the art in trellis-coded modulation (TCM) is given for the more interested reader. First, the general structure of TCM schemes and the principles of code construction are reviewed. Next, the effects of carrier-phase offset in carrier-modulated TCM systems are discussed. The topic is important, since TCM schemes turn out to be more sensitive to phase offset than uncoded modulation systems. Also, TCM schemes are generally not phase invariant to the same extent as their signal sets. Finally, recent advances in TCM schemes that use signal sets defined in more than two dimensions are described, and other work related to trellis-coded modulation is mentioned. The best codes currently known for one-, two-, four-, and eight-dimensional signal sets are given in an Appendix.

Design of Trellis-Coded Modulation Schemes

The trellis structure of the early hand-designed TCM schemes and the heuristic rules used to assign signals to trellis transitions suggested that TCM schemes should have an interpretation in terms of convolutional codes with a special signal mapping. This mapping should be based on grouping signals into subsets with large distance between the subset signals. Attempts to explain TCM schemes in this manner led to the general structure of TCM encoders/modulators depicted in Fig. 1. According to this figure, TCM signals are generated as follows: When m bits are to be transmitted per encoder/modulator operation, $\tilde{m} \leq m$ bits are expanded by a rate-$\tilde{m}/(\tilde{m}+1)$ binary convolutional encoder into $\tilde{m}+1$ coded bits. These bits are used to select one of $2^{\tilde{m}+1}$ subsets of a redundant 2^{m+1}-ary signal set. The remaining $m-\tilde{m}$ uncoded bits determine which of the $2^{m-\tilde{m}}$ signals in this subset is to be transmitted.

Set Partitioning

The concept of set partitioning is of central significance for TCM schemes. Figure 2 shows this concept for a 32-CROSS signal set [1], a signal set of lattice type "Z_2". Generally, the notation "Z_k" is used to denote an infinite "lattice" of points in k-dimensional space with integer coordinates. Lattice-type signal sets are finite subsets of lattice points, which are centered around the origin and have a minimum spacing of Δ_0.

Set partitioning divides a signal set successively into smaller subsets with maximally increasing smallest intra-set distances Δ_i, $i = 0,1, \ldots$. Each partition is two-way. The partitioning is repeated $\tilde{m}+1$ times until $\Delta_{\tilde{m}+1}$ is equal to or greater than the desired free distance of the TCM scheme to be designed. The finally obtained subsets, labeled D0, D1, ... D7 in the case of Fig. 2, will henceforth be referred to as the "subsets." The labeling of branches in the partition tree by the $\tilde{m}+1$ coded bits $z_n^{\tilde{m}}$, $\ldots, z_n^0$, in the order as shown in Fig. 2, results in a label $\underline{z}_n = [z_n^{\tilde{m}}, \ldots z_n^0]$ for each subset. The label reflects the position of the subset in the tree.

This labeling leads to an important property. If the labels of two subsets agree in the last q positions, but not in the bit z_n^q, then the signals of the two subsets are

0163-6804/87/0002-0012 $01.00 © 1987 IEEE

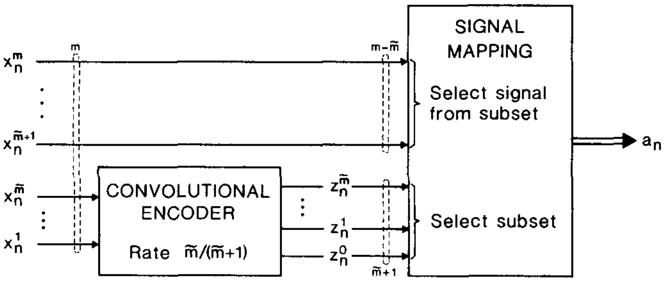


Fig. 1. General structure of encoder/modulator for trellis-coded modulation.

elements of the same subset at level q in the partition tree; thus they have at least distance Δ_q. This distance bound can be stated in a "set-partitioning lemma" and will be used in the next subsection.

The $m - \tilde{m}$ uncoded bits $x_n^m, \ldots, x_n^{\tilde{m}+1}$ are used to choose a signal from the selected subset. The specific labeling of subset signals by these bits is not particularly important at this point of the discussion. In the code trellis, the signals of the subsets become associated with $2^{m-\tilde{m}}$ parallel transitions.

The free Euclidean distance of a TCM code can now be expressed as

$$d_{free} = Min[\Delta_{\tilde{m}+1}, d_{free}(\tilde{m})],$$

where $\Delta_{\tilde{m}+1}$ is the minimum distance between parallel transitions and $d_{free}(\tilde{m})$ denotes the minimum distance between nonparallel paths in the TCM trellis diagram. In the special case of $\tilde{m} = m$, the subsets contain only one signal, and hence there are no parallel transitions.

Convolutional Codes for Trellis-Coded Modulation

At every time n, the rate-$\tilde{m}/(\tilde{m} + 1)$ convolutional encoder depicted in Fig. 1 receives $\tilde{m}$ input bits, and generates $\tilde{m} + 1$ coded bits which serve as the subset labels $\underline{z}_n = [z_n^{\tilde{m}}, \ldots z_n^0]$. The set of all possible sequences $\{\underline{z}_n\}$, which the encoder can generate, forms a convolutional code. A linear convolutional code of rate $\tilde{m}/(\tilde{m} + 1)$ is most compactly defined by a parity-check equation which puts a constraint on the code bits in a sliding time window of length $v + 1$:

$$\sum_{i=0}^{\tilde{m}}(h_v^i z_{n-v}^i \oplus h_{v-1}^i z_{n-v+1}^i \oplus \ldots h_0^i z_n^i) = 0.$$

In this equation, ⊕ denotes modulo-2 addition. The quantity ν is called the constraint length. The quantities h_ℓ^i, $\nu \geq \ell \geq 0$; $0 \leq i \leq \tilde{m}$, are the binary parity-check coefficients of the code. Valid code sequences satisfy this equation at all times n. The equation defines only the code sequences, not the input/output relation of an encoder. A later subsection deals with minimal encoder realizations with ν binary storage elements, which is equivalent to saying that the code has 2^ν trellis states.

From the parity-check equation, one can observe that code sequences $\{\underline{z}_n\}$ can have arbitrary values for each $\tilde{m}$-tuple $[z_n^{\tilde{m}}, \ldots z_n^1]$ with an appropriate choice of the sequence $\{z_n^0\}$ so that the parity-check equation is satisfied. This property can be expressed in a "rate-$\tilde{m}/(\tilde{m} + 1)$ code lemma."

Let now $\{\underline{z}_n\}$ and $\{\underline{z}'_n\} = \{\underline{z}_n \oplus \underline{e}_n\}$ be two code sequences, where $\{\underline{e}_n\}$ denotes the error sequence by which these sequences differ. Since the convolutional code is linear, $\{\underline{e}_n\}$ is also a code sequence. It follows from the "set-partitioning lemma" mentioned in the preceding subsection and the "rate-$\tilde{m}/(\tilde{m} + 1)$ code lemma" that the squared free distance between non-parallel paths in the TCM trellis is bounded by [2]

$$d_{free}^2(\tilde{m}) \geqq \underset{\{\underline{e}_n\} \neq \{\underline{0}\}}{Min} \sum_n \Delta_{q(\underline{e}_n)}^2.$$

Here $q(\underline{e}_n)$ is the number of trailing zeros in $\underline{e}_n$, that is, the number of trailing positions in which two subset labels $\underline{z}_n$ and $\underline{z}'_n = \underline{z}_n \oplus \underline{e}_n$ agree. For example, $q(\underline{e}_n) = 2$, if $\underline{e}_n = [e_n^m, \ldots, e_n^3, 1, 0, 0]$. The "set-partitioning lemma" states that the distance between signals in the subsets selected by $\underline{z}_n$ and $\underline{z}'_n$ is lower-bounded by $\Delta_{q(\underline{e}_n)}$. One must take $\Delta_{q(\underline{0})} = 0$, not $\Delta_{\tilde{m}+1}$. Minimization has to be carried out over all non-zero code (error) sequences $\{\underline{e}_n\}$ that deviate at, say, time 0 from the all-zero sequence $\{\underline{0}\}$ and remerge with it at a later time. The "rate-$\tilde{m}/(\tilde{m} + 1)$ code lemma" assures that for any given sequence $\{\underline{e}_n\}$ there exist two coded signal sequences whose signals have at any time n the smallest possible distance between the signals of subsets whose labels differ by $\underline{e}_n$. Usually, this smallest distance equals $\Delta_{q(\underline{e}_n)}$ for all $\underline{e}_n$. If this is the case, the above bound on $d_{free}(\tilde{m})$ becomes an equation. (Only when the signal subsets contain very few signals may the bound not be satisfied with equality. A similar always true equation can then be used to compute $d_{free}(\tilde{m})$ [2].)

This equation is of key importance in the search for optimum TCM codes. It states that free Euclidean distance can be determined in much the same way as free Hamming distance is found in linear binary codes, even though linearity does not hold for TCM signal sequences. It is only necessary to replace the Hamming weights of the e_n (number of 1's in $\underline{e}_n$) by the Euclidean weights $\Delta_{q(\underline{e}_n)}^2$. It is not necessary (as some authors seem to think) to compute distance between every pair of TCM signal sequences.

Search for Optimum TCM Codes

For the one- and two-dimensional signal sets depicted in Fig. 1 of Part I [1], the minimum intra-set distances are

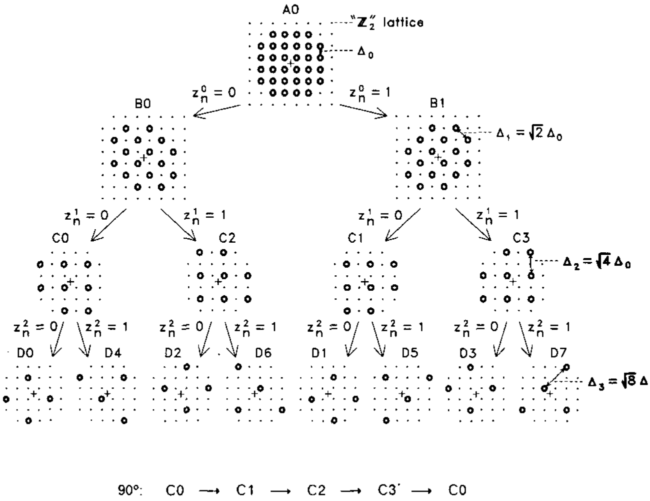


Fig. 2. Set partitioning of the 32-CROSS signal set (of lattice type "Z_2").

as follows. For 4-AM, 8-AM, ... (signal sets of type "Z_1"), $\Delta_{i+1} = 2\Delta_i$, i = 0.1, ... For 16-QASK, 32-CROSS, ... (signal sets of lattice type "Z_2"), $\Delta_{i+1} = \sqrt{2}\,\Delta_i$, i = 0,1, ... The non-lattice type signal sets 8-PSK and 16-PSK have special sequences of intra-set distances. The intra-set distances for higher-dimensional signal sets will be given when multi-dimensional TCM schemes are discussed later in this article.

For a given sequence of minimum intra-set distances $\Delta_0 \leq \Delta_1 \leq \ldots \Delta_{\tilde{m}}$, and a chosen value of ν, a convolutional code with the largest possible value of $d_{free}(\tilde{m})$ can be found by a code-search program described in [2]. The program performs the search for the $(\nu + 1) \cdot (\tilde{m} + 1)$ binary parity-check coefficients in a particular order and with a set of code-rejection rules such that explicit checks on the value of $d_{free}(\tilde{m})$ are very frequently avoided.

Tables of optimum codes for one-, two-, four-, and eight-dimensional TCM schemes are shown in the Appendix. Parity-check coefficients are specified in octal form, for example, $[h_6^0, \ldots, h_0^0] = [1,0,0,0,1,0,1]$ is written as $\underline{h}^0 = 105_8$. Equivalent codes in terms of free distance will be obtained if the parity-check coefficients of $\underline{h}^i$ are added modulo-2 to the coefficients of $\underline{h}^k$, for i > k [2]. If $\Delta_i = \Delta_k$, $\underline{h}^i$ and $\underline{h}^k$ may also be interchanged. When in the code tables the free distance of a code is marked by an asterisk (*), $d_{free}(\tilde{m})$ exceeds $\Delta_{\tilde{m}+1}$, and hence the free distance occurs only between the subset signals assigned to parallel transitions. These schemes have the smallest numbers of nearest neighbors. For example, the 256-state code for "Z_1"-type signals has this property. For large values of m, this code attains a full 6 dB coding gain with only two nearest neighbors.

Two Encoder Realizations

The parity-check equation specifies only the convolutional code. Encoders for the same code can differ in the input/output relation which they realize. Figure 3 illustrates two encoders for the 8-state linear code specified in Tables II and III ($\nu = 3$) in the Appendix. One is called a systematic encoder with feedback, the other a feedback-free encoder. Both encoders are minimal, that is, they are realized with ν binary storage elements. The transformation of one minimal encoder into the other follows from the structural properties of convolutional codes described in [3]. With a systematic encoder, the input bits appear unchanged at the output. Therefore, a systematic encoder cannot generate a catastrophic code, i.e., a code with no distance increase between two trellis paths that remain distinct for an unbounded length. This is also true, although far from being obvious, for an equivalent minimal feedback-free encoder [3].

The forward and backward connections in the systematic encoder are specified by the parity-check coefficients of the code. All codes presented in the Appendix have $h_\nu^0 = h_0^0 = 1$. This guarantees the realizability of an encoder in the form shown in Fig. 3a. The reader familiar with recursive digital filters will see that the parity-check equation is used (almost directly) to compute the bit z_n^0 from the other uncoded bits. Furthermore, all codes have $h_\nu^i = h_0^i = 0$, for i > 0. This

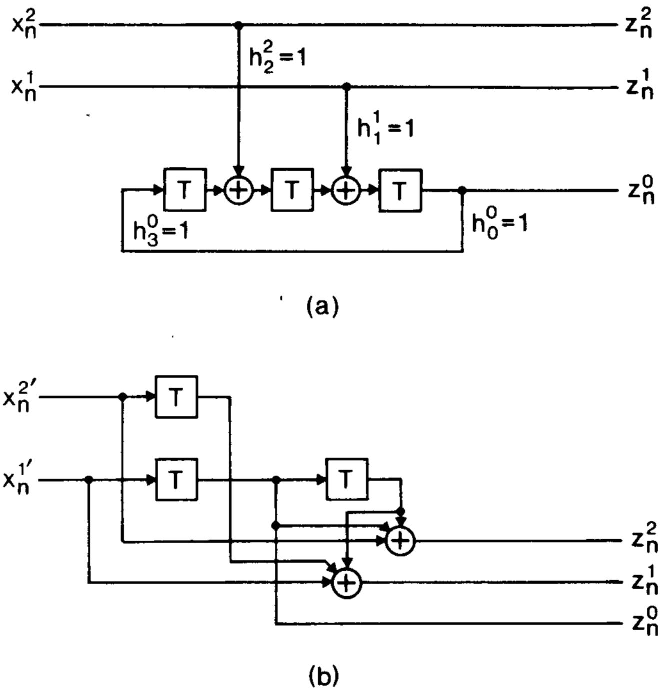


Fig. 3. Two encoders for a linear 8-state convolutional code with parity-check coefficients $\underline{h}^2 = [0,1,0,0]$, $\underline{h}^1 = [0,0,1,0]$, $\underline{h}^0 = [1,0,0,1]$ (cf. Tables II and III in the Appendix). (a) Minimal systematic encoder with feedback. (b) Minimal feedback-free encoder.

ensures that at time n the uncoded bits have no influence on the bit z_n^0, nor on the input to the first binary storage element in the encoder. Hence, whenever in the code trellis two paths diverge from or merge into a common state, the bit z_n^0 must be the same for these transitions, whereas the other bits differ in at least one bit. Signals associated with diverging and merging transitions therefore have at least distance Δ_1 between them, which reflects the second heuristic rule for good TCM codes mentioned in Part I [1].

TCM schemes for two-dimensional carrier modulation (with 8-PSK signal sets and "Z_2"-type signal sets) have up to the present time attracted the most attention. Practical realizations of these systems indicated that the effects of transmission impairments other than additive Gaussian noise on their performance need to be studied, in particular those of carrier offset.

Effects of Carrier-Phase Offset

This section addresses the problems that arise when a carrier-modulated two-dimensional TCM signal is demodulated with a phase offset $\Delta\phi$. The soft-decision decoder then operates on a sequence of complex-valued signals $\{r_n\} = \{a_n \cdot \exp(j\Delta\phi) + w_n\}$, where the a_n are transmitted TCM signals and the w_n denote additive Gaussian noise. The phase offset $\Delta\phi$ could be caused, for instance, by disturbances of the carrier phase of the received signal which the phase-tracking scheme of the receiver cannot track instantly.

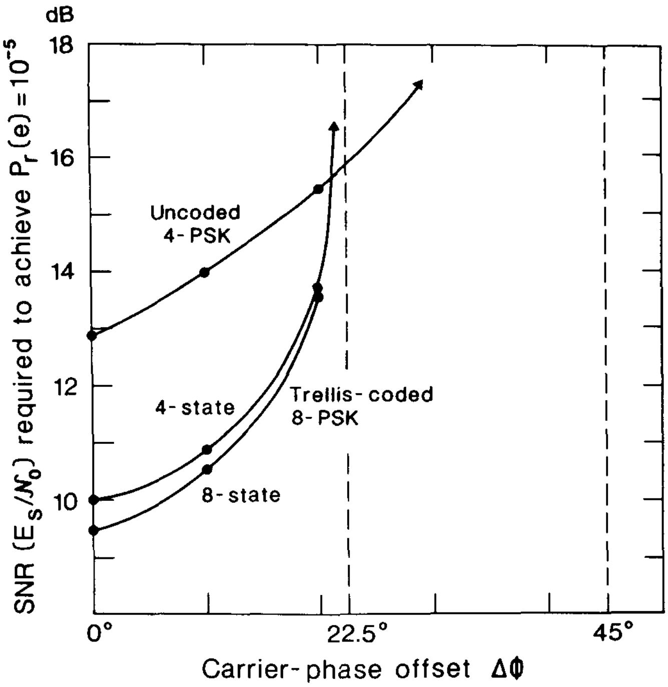


Fig. 4. Error performance of coded 8-PSK and uncoded 4-PSK in the presence of carrier-phase offset Δφ.

Performance Degradation

The error performance of 4-state and 8-state coded 8-PSK systems in the presence of phase offset (based on unpublished work) is illustrated in Fig. 4. The figure shows the signal-to-noise ratio needed to sustain an error-event probability of 10^{-5} as a function of Δφ. For the coded 8-PSK systems, the required signal-to-noise ratio increases with increasing values of Δφ until both systems fail at $\Delta\phi = 22.5°$, even in the absence of noise. In contrast, uncoded 4-PSK requires a higher signal-to-noise ratio at small phase offsets, but has an operating range up to $\Delta\phi = 45°$ in the absence of noise. These results are typical for TCM schemes.

The greater susceptibility of TCM schemes to phase offset can be explained as follows. In the trellis diagrams of TCM schemes, there exist long distinct paths with low growth of signal distance between them, that is, paths which have either the same signals or signals with smallest distance Δ_0 assigned to concurrent transitions. In the absence of phase offset, the non-zero squared distances Δ_0^2 and the squared larger distances of diverging or merging transitions add up to at least the squared free distance. However, if phase offset rotates the received signals such that received signals become located halfway between the signals of the original signal set, the difference in distance between received signals and the signals on distinct transitions that are Δ_0 apart may be reduced to zero. There may then be no difference in distance between a long segment of received signals and two distinct trellis paths, just as though the code were catastrophic. At this point, the decoder begins to fail.

Behavior of Carrier-Phase Tracking Loops

Nowadays, in most digital carrier-modulation systems, decision-directed loops are employed for carrier-phase tracking. In these loops, the phase offset is estimated from the received signal and the decoder decisions. The estimated phase offset controls the demodulating carrier phase. In a TCM receiver, if the phase offset exceeds a critical value, for example, 22.5° in the case of coded 8-PSK, the decoder decisions become essentially uncorrelated with the received signal and the mean value of the phase estimate drops to zero. Figure 5 illustrates, for 4-state coded 8-PSK [2], the mean estimate of Δφ ("S-curve") and its variance as a function of the actual value of the phase offset. A vanishing mean estimate, as occurs for Δφ between 22.5° and 157.5°, leaves the carrier-phase tracking loop in an undriven random-walk situation which can last for long periods. Eventually, the system resynchronizes when the randomly-fluctuating demodulating carrier phase approaches a value for which the received signal again resembles a valid TCM sequence. This behavior is in significant contrast to the short phase skips and rapid recovery observed in uncoded 4-PSK or 8-PSK systems. It suggests that in some cases TCM systems may require special methods to force rapid resynchronization.

Invariance of Two-Dimensional TCM Codes under Phase Rotation

TCM codes are not usually invariant to all phase rotations under which the signal set is phase invariant. Figure 5 indicates a phase symmetry of 4-state coded 8-PSK only at $\Delta\phi = 180°$, but not at other multiples of 45°. This symmetry can be verified by inspection of the code trellis presented in Fig. 2b of Part I [1]. Coded 8-PSK schemes which are invariant to phase shifts of all multiples of 45° have been found [4], but these schemes require more than four states to achieve a coding gain of 3 dB.

In general, it is desirable that TCM codes have as many phase symmetries as possible to ensure rapid carrier-phase resynchronization after temporary loss of synchronization. On the other hand, such phase invariances must be made transparent to the transmitted user

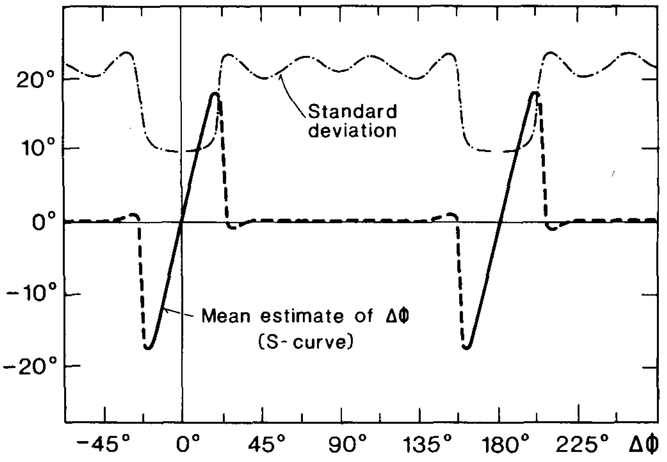


Fig. 5. Mean ("S-curve") and variance of the estimated phase offset $\Delta\hat{\phi}$ in a decision-directed carrier-phase tracking loop for 4-state coded 8-PSK versus the actual phase offset Δφ, at a signal-to-noise ratio of 13 dB (tentative decisions used with zero delay).

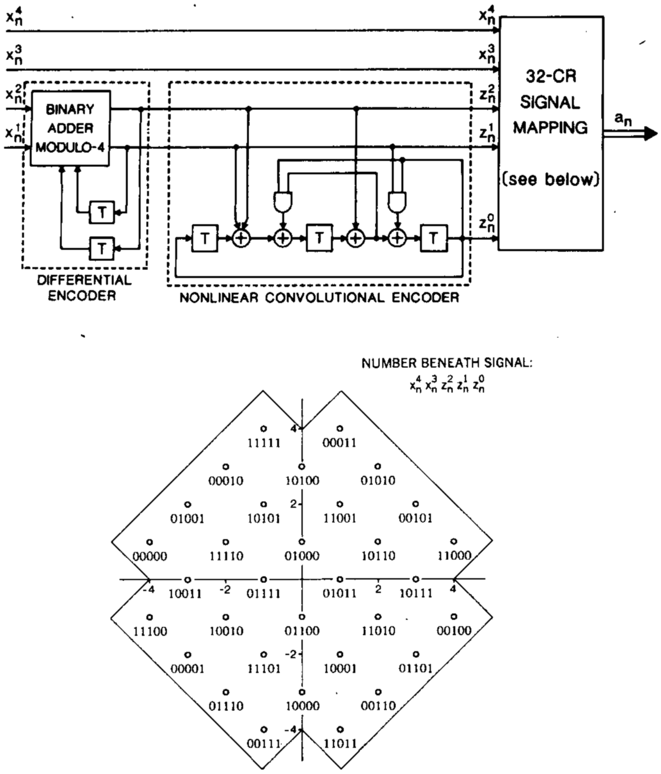


Fig. 6. Nonlinear 8-state encoder/modulator with 32-CROSS signal set and differential encoding, as in CCITT Recommendation V.32.

information by some form of differential encoding and decoding. If loss of phase synchronization is very unlikely, one may argue that TCM codes without phase invariances may have the advantage that the receiver can establish absolute phase from the received signal, so that no differential encoding/decoding is required.

The problems of phase invariance and differential encoding/decoding attracted considerable attention in work toward a TCM code for use in CCITT Recommendations for voice-band modems operating full-duplex at up to 9.6 kbit/s over two-wire telephone circuits, and at up to 14.4 kbit/s over four-wire circuits. There was considerable interest in a two-dimensional 8-state code that can achieve, with 90°-symmetric QASK and CROSS signal sets, a coding gain of about 4 dB over uncoded modulation. With the known linear code (cf. Table III in the Appendix, $\nu = 3$), it was only possible (by adding parity-check coefficients in a way which does not change free distance, as mentioned in the subsection on optimum-code search) to have either no phase symmetry or a symmetry at 180° [5], [4, Part I]. A breakthrough was finally accomplished by L.F. Wei, who introduced nonlinear elements into the convolutional encoder of the 8-state code. This made the code invariant to 90° rotations while maintaining its coding gain of 4 dB [6], [5, Part I]. Figure 6 shows the resulting encoder/modulator with its differential encoder, nonlinear convolutional encoder, and signal mapping for a 32-CROSS signal set ($m = 4$), as finally adopted in the CCITT V.32 Recommendation [7, Part I]. The labeling of subsets differs slightly from that indicated in Fig. 2, but the subsets are the same. The same code was also chosen for the CCITT V.33 Draft Recommendation [8, Part I], but with 64-QASK and 128-CROSS signal sets ($m = 5,6$). In the limit of large signal sets, the number of nearest neighbors in the 8-state linear and the CCITT nonlinear code is 16.

In a late contribution to the CCITT [7], illustrated in Fig. 7, an alternative 8-state nonlinear encoder with the differential-encoding function integrated into the encoder was proposed. The coding gain and the number of nearest neighbors are identical to those of the other 8-state schemes. The trellis diagram of the alternative nonlinear code was shown in Fig. 6 of Part I [1]. Differential decoding requires that the receiver compute $x_n^1 = z_n^0 \oplus z_{n+1}^0$. Subsets are labeled as indicated in Fig. 2. The selection of signals within the subsets by the uncoded bits x_n^4, x_n^3 is worth mentioning. If $x_n^4 = 0$, only signals of the inner 16-QASK set are transmitted ($m = 3$). With non-zero values of x_n^4, outer signals of the larger 32-CROSS set are also selected ($m = 4$). Extension of this concept to larger signal sets resulted in one general signal mapping for all data rates, e.g., for $3 \leq m \leq 7$ [7]. The mapping has the additional property that it can just as well be used for uncoded modulation with modulo-4 differential encoding of the bits z_n^1, z_n^0.

The nonlinear 8-state TCM codes appear to be special cases. Similar nonlinear phase-invariant codes with 16 and more states can be constructed. However, at least for 16 states, it does not seem possible to find a code with the same 4.8 dB coding gain as can be obtained with a linear code.

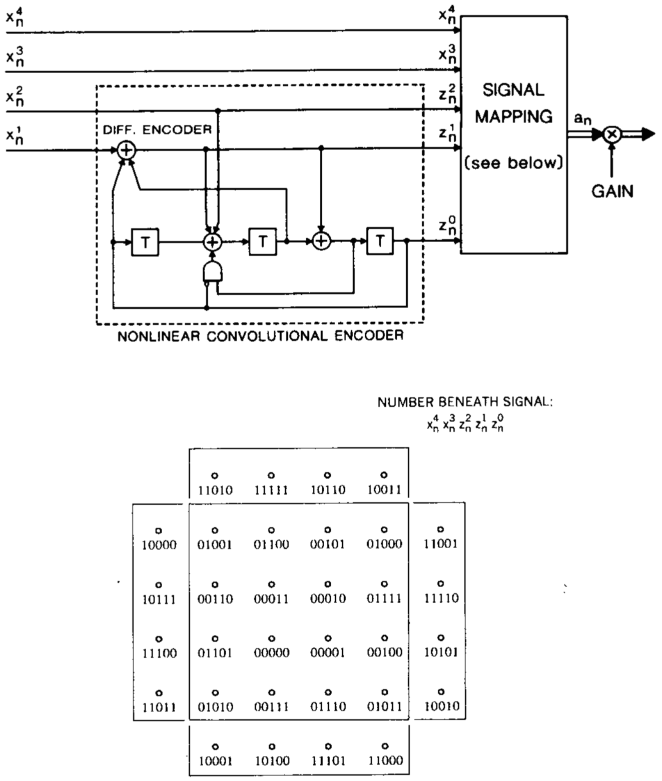


Fig. 7. Alternative nonlinear 8-state encoder/modulator with integrated differential encoding and general signal mapping for 16-QASK, 32-CROSS, etc., signal sets.

Multi-Dimensional Trellis Codes

Recently, there have been a number of investigations into trellis coding with signal sets defined in more than two dimensions [3, Part I], [8-11]. In practical systems, multi-dimensional signals can be transmitted as sequences of constituent one- or two-dimensional (1-D or 2-D) signals. In this section, 2K-D TCM schemes are considered which transmit m bits per constituent 2-D signal, and hence mK bits per 2K-D signal. The principle of using a redundant signal set of twice the size needed for uncoded modulation is maintained. Thus, 2K-D TCM schemes use 2^{Km+1}-ary sets of 2K-D signals. Compared to 2-D TCM schemes, this results in less signal redundancy in the constituent 2-D signal sets.

For 2-D TCM schemes with "Z_2"-type signal sets, the minimum signal spacing Δ_0 must be reduced by approximately the factor $\sqrt{2}$ ($-$ 3 dB) to have the same average signal power as for uncoded modulation. This loss in signal spacing needs to be more than compensated for by coding to obtain an overall improvement in free distance. The lower signal redundancy of multi-dimensional TCM schemes with "Z_{2K}"-type signal sets results only in a reduction of the minimum signal spacing by the 2K-th root of 2 (-1.5 dB for $K = 2$; and -0.75 dB for $K = 4$), so coding has to contribute less than in the case of 2-D TCM to obtain the same gain in free distance. The larger signal spacing should also make multi-dimensional TCM systems less sensitive to phase offset. Finally, it has been found that multi-dimensional TCM schemes with 90° phase invariance can be obtained with linear codes.

Four-Dimensional Trellis-Coded Modulation

The 4-D TCM schemes ($K = 2$) described in this subsection employ compact sets of 2^{2m+1} signals chosen from a lattice of type "Z_4" with minimum signal spacing Δ_0. Figure 8 illustrates the set partitioning of a signal set A_4^0 of type "Z_4". The general idea is to derive the set partitioning of a higher-dimensional signal set from the set partitioning of constituent lower-dimensional signal sets. In the present case, A_4^0 and its subsets are characterized by two constituent "Z_2"-type signal sets A0 and their subsets, such as introduced in Fig. 2. This leads to a partition tree with signal sets of types "Z_4" → "D_4" → "Z_4" → "D_4" → "Z_4", etc., with minimum intra-set distances Δ_0, $\Delta_1 = \Delta_2 = \sqrt{2}\,\Delta_0$, $\Delta_3 = \Delta_4 = \sqrt{4}\,\Delta_0$, etc. The next paragraph describes the details of the partitioning process (and may be skipped by readers without specific interest in this process).

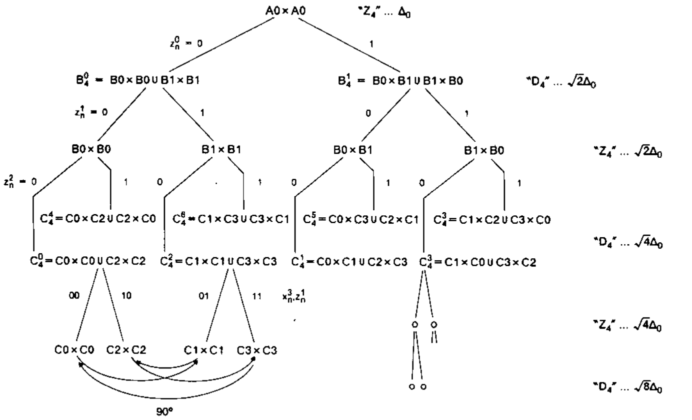


Fig. 8. Set partitioning of four-dimensional signal sets of lattice type "Z_4", also showing the effect of a 90° rotation.

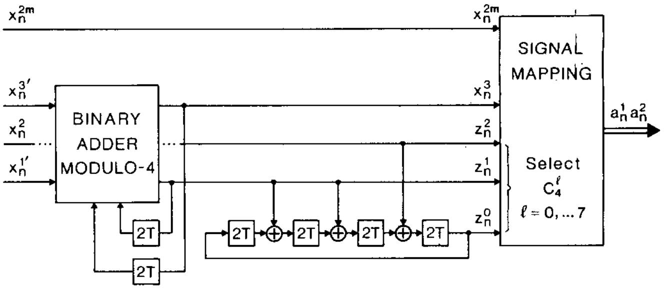
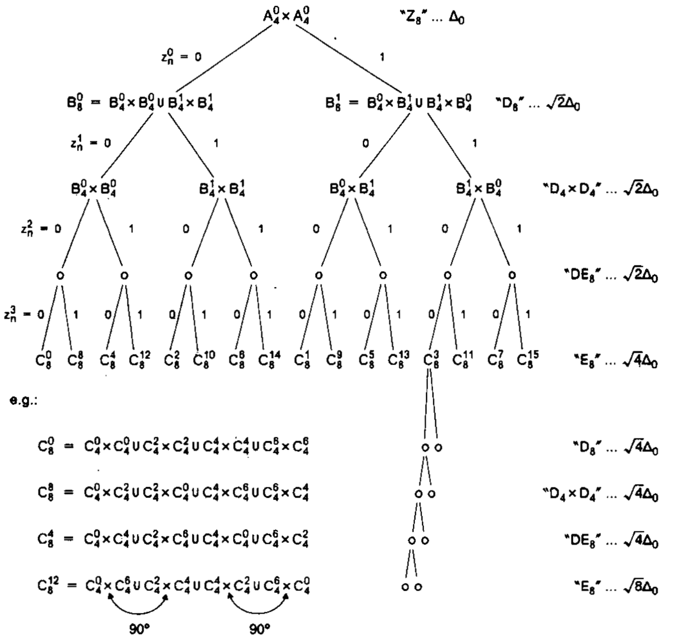


Fig. 9. Sixteen-state encoder/demodulator for four-dimensional "Z_4"-type trellis-coded modulation with differential encoding.

Set partitioning begins by writing $A_4^0 = A0 \times A0$ ($\times$ denotes set-product operation: the product set consists of all concatenations of elements of the first set with the elements of the second set). Substitution of $A0 = B0 \cup B1$ ($\cup$ denotes set union) yields $A_4^0 = (B0 \cup B1) \times (B0 \cup B1) = (B0 \times B0) \cup (B0 \times B1) \cup (B1 \times B0) \cup (B1 \times B1)$. The first partition divides A_4^0 into the two subsets $B_4^0 = (B0 \times B0) \cup (B1 \times B1)$ and $B_4^1 = (B0 \times B1) \cup (B1 \times B0)$. These subsets are of type "D_4", where "D_4" denotes the densest lattice known in 4-D space [12]. The minimum intra-set distance in B_4^0 and B_4^1 is $\sqrt{2}\,\Delta_0$, which is the minimum distance between constituent 2-D signals in B0 or B1, and also between one 4-D signal in B0×B0 and another in B1×B1. On the next binary partition, e.g., when B_4^0 is partitioned into subsets B0×B0 and B1×B1, no distance increase is obtained. These subsets are of type "Z_4", like A_4^0, from which they differ only in their orientation, position with respect to the origin, and scaling. Hence, their partitioning is conceptually similar to that of A_4^0. The minimum intra-set distance increases to $\sqrt{4}\,\Delta_0$ when, e.g., B0×B0 is split into subsets $C_4^0 = (C0 \times C0) \cup (C2 \times C2)$ and $C_4^4 = (C0 \times C2) \cup (C2 \times C0)$, which are now again of type "D_4".

Optimum convolutional codes are found by using the obtained sequence of minimum intra-set distances in the code-search program mentioned earlier. The codes and their asymptotic coding gains over uncoded modulation with "Z_2"-type signals are given in Table IV in the Appendix. The gains are valid for large signal sets which fill the same volume in signal space as the signal sets used for uncoded modulation. Thus, the comparison is made for the same average signal power and the same peak power of 2-D signals.

It may be helpful to discuss the 16-state code of Table IV, which achieves an asymptotic coding gain of 4.52 dB, in more detail. The code uses the eight 4-D subsets C_4^0, ... C_4^7 shown in Fig. 8, and has 64 distinct transitions in its trellis diagram. The only nearest-neighbor signals are those associated with parallel transitions, and their number at any transition is 24 (the number of nearest neighbors in a "D_4" lattice). Figure 9 depicts one possible realization of an encoder/modulator with differential



$A_4^0 \times A_4^0$
"Z_8" ... Δ_0
$z_n^0 = 0$
1
$B_8^0 = B_4^0 \times B_4^0 \cup B_4^1 \times B_4^1$
$B_8^1 = B_4^0 \times B_4^1 \cup B_4^1 \times B_4^0$
"D_8" ... $\sqrt{2}\Delta_0$
$z_n^1 = 0$
1
0
1
$B_4^0 \times B_4^0$
$B_4^1 \times B_4^1$
$B_4^0 \times B_4^1$
$B_4^1 \times B_4^0$
"$D_4 \times D_4$" ... $\sqrt{2}\Delta_0$
$z_n^2 = 0$
1
0
1
0
1
0
1
"DE_8" ... $\sqrt{2}\Delta_0$
$z_n^3 = 0$
1
C_8^0 C_8^8 C_8^4 C_8^{12} C_8^2 C_8^{10} C_8^6 C_8^{14} C_8^1 C_8^9 C_8^5 C_8^{13} C_8^3 C_8^{11} C_8^7 C_8^{15}
"E_8" ... $\sqrt{4}\Delta_0$
e.g.:
$C_8^0 = C_4^0 \times C_4^0 \cup C_4^2 \times C_4^2 \cup C_4^4 \times C_4^4 \cup C_4^6 \times C_4^6$
"D_8" ... $\sqrt{4}\Delta_0$
$C_8^8 = C_4^0 \times C_4^2 \cup C_4^2 \times C_4^0 \cup C_4^4 \times C_4^6 \cup C_4^6 \times C_4^4$
"$D_4 \times D_4$" ... $\sqrt{4}\Delta_0$
$C_8^4 = C_4^0 \times C_4^4 \cup C_4^2 \times C_4^6 \cup C_4^4 \times C_4^0 \cup C_4^6 \times C_4^2$
"DE_8" ... $\sqrt{4}\Delta_0$
$C_8^{12} = C_4^0 \times C_4^6 \cup C_4^2 \times C_4^4 \cup C_4^4 \times C_4^2 \cup C_4^6 \times C_4^0$
"E_8" ... $\sqrt{8}\Delta_0$
90°
90°

Fig. 10. Set partitioning of eight-dimensional signal sets of lattice type "Z_8", also showing the effect of a 90° rotation.

encoding. The code from Table IV was first made invariant to inversion of the bit z_n^1 by interchanging the parity-check coefficient vectors $\underline{h}^1$ and $\underline{h}^2$. Invariance to 90° rotations and the required differential encoding follow from the 90° symmetries indicated in Fig. 8, which in turn are based on the 90° symmetries in the constituent 2-D signal subsets. The subsets C_4^0, ... C_4^7, each composed of two subsets CiXCk, must be chosen individually for each value of m. The subset C0XC0 contains 2^{2m-3} signals, and may be constructed first. The other subsets CiXCk are then obtained by 90° rotations of the two constituent subsets C0 in C0XC0. For the specific case of m = 4.5, C0XC0 contains 8X8 signals, and hence the 8-ary subset C0 of Fig. 2 can be used. This construction of the 4-D subsets also suggests an efficient subset-decoding method that begins with signal decisions within the constituent 2-D subsets C0, ... C3. In general, the design of signal sets can be more complicated. References [3, Part I] and [11] discuss mapping techniques for cases where signal-set sizes are not powers of 2.

Eight-Dimensional Trellis-Coded Modulation

The technique of set partitioning of a higher-dimensional signal set based on the known partitioning of lower-dimensional sets is now applied to 8-D signal sets (K=4) of type "Z_8" = "$Z_4 \times Z_4$". Figure 10 illustrates the details. The sequence of minimum intra-set distances Δ_0, $\Delta_1 = \Delta\ 2 = \Delta_3 = \sqrt{2}\,\Delta_0$, $\Delta_4 = \Delta_5 = \Delta_6 = \Delta_7 = \sqrt{4}\,\Delta_0$, etc., is obtained, corresponding to a chain of lattice types "Z_8" → "D_8" → "$D_4 \times D_4$" → "DE_8" → "E_8" → "D_8", etc., where "E_8" denotes the famous Gosset lattice, the densest lattice known in 8-D space [12]. (The nomenclature "DE_8" was introduced in [9]; [11] uses "D^+_8".)

Codes obtained by the code-search program are given in Table V in the Appendix. The codes use 2^{4m+1} 8-D signals partitioned into 16 subsets C_8^0, ... C_8^{15} of type "E_8". In the limit of large signal sets, the codes achieve an asymptotic coding gain of 5.27 dB over uncoded "Z_2"-type modulation. If code complexity is increased to 64 states, the only nearest neighbors are those associated with parallel transitions, and their number is 240, which is the number of nearest neighbors in an "E_8" lattice. The "E_8"-type subsets can be further partitioned into two subsets with 90° symmetries as indicated in Fig. 10. This property can be verified by observing the 90° symmetries among the constituent 4-D signals as shown in Fig. 8. Hence, 8-D codes are inherently 90° phase invariant, because their subsets have this property. Differential encoding/decoding can be performed entirely within the subsets, decoupled from the convolutional encoding function.

Other 8-D TCM schemes are obtained by choosing the 2^{4m+1} signals from another lattice type than "Z_8" in the chain of types encountered in Fig. 10, and performing the code search for the sequence of minimum intra-set distances that originates from this type. Codes with signals from "DE_8" or "E_8" are of some interest [9]-[11], although it does not seem that these codes exhibit significant advantages over the "Z_8"-type codes, if code complexities, asymptotic coding gains, and numbers of nearest neighbors are compared. This is also true for 4-D codes with "D_4" signals, as compared to codes with "Z_4" signals.

Discussion

The number of distinct transitions in the trellis diagrams of TCM codes is $2^{\nu+\tilde{m}}$. This so-called "trellis complexity" represents a measure of code (decoding) complexity. A fair comparison of TCM schemes with different signal dimensionalities requires normalization of trellis complexities and numbers of nearest neighbors to the same number of signal dimensions. In the following, normalization to two dimensions is assumed. Hence, normalized trellis complexity specifies the number of distinct trellis transitions to be dealt with by the decoder per 2-D signal or two 1-D signals received. Similarly, a normalized number of nearest neighbors indicates the number of error events with free distance that could start (on average) during the same time interval.

In Fig. 11, asymptotic coding gains of TCM schemes with large 1-D (K = 0.5) to 8-D (K = 4) signal sets are plotted versus normalized trellis complexity, $2^{\nu+\tilde{m}}/K$. Normalized numbers of nearest neighbors, N_{free}/K, are given in parentheses. At a normalized trellis complexity of 8, the "Z_2"-type 4-state code is without competition. The "Z_4"-type 16-state code, whose encoder/modulator was illustrated in Fig. 9, shows a 0.5 dB advantage over a "Z_2"-type 8-state code, e.g., the nonlinear CCITT code, and also a slightly reduced number of nearest neighbors, at the same normalized complexity of 32. Next in the order of increasing complexities, the "Z_2"-type 16-state code may be of interest, but it cannot be made invariant to 90° rotations. At a normalized complexity of 128, i.e., four times the complexity of the CCITT code, the "Z_8"- and "E_8"-type 64-state codes are found as attractive 90° phase-invariant codes. Finally, at a 32 times higher complexity than the CCITT code, the "Z_1"-type 256-state code stands out for its asymptotic coding gain and low number of nearest neighbors.

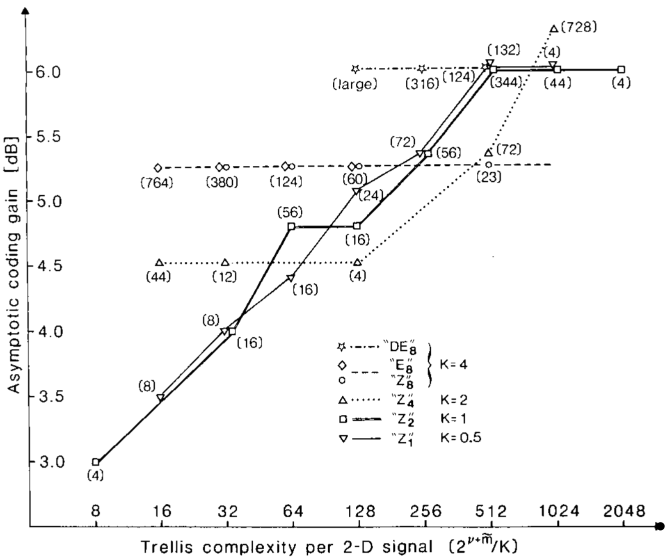


Fig. 11. Asymptotic coding gains versus trellis complexity per 2-D signal ($2^{\nu+\tilde{m}}/K$) for large 2K-D signal sets of type "Z_{2K}" for K = 0.5, 1, 2, 4; "E_8"; and "DE_8". Numbers of nearest neighbors per 2-D signal (N_{free}/K) are given in parentheses.

The asymptotic coding gain of the "DE_8" codes exceeds that of the "Z_8" and "E_8" codes by 0.75 dB, but the "DE_8" codes also have many more nearest neighbors. Hence, one may question their usefulness. Similarly, the "Z_1"-type 128-state code with the highest asymptotic coding gain of 6.28 dB shown in Fig. 11 may not be of practical interest, because of its large number of nearest neighbors.

Figure 11 gives important information about the ranking of TCM codes. However, the picture also remains somewhat incomplete. Real coding gains at given error probabilities, considering nearest and next-nearest neighbors and the boundary effect of finite signal sets, are not included. In first approximation, one may use the rule that for error rates around 10^{-5} the real coding gain is reduced by 0.2 dB for every increase in the number of nearest neighbors by the factor of 2. There is also very little published information about the carrier-phase sensitivity (a possible advantage of the multi-dimensional TCM schemes) of the TCM schemes under discussion. The complexity of subset decoding and decoder-memory requirements are further important aspects that need to be considered.

In general, one can make the following observations. At low complexity, higher-dimensional TCM schemes exhibit larger asymptotic coding gains than the lower-dimensional schemes, however, these coding gains are compromised by large numbers of nearest neighbors. In the mid-range,. 4-D and 8-D TCM schemes achieve slightly larger real coding gains than the 1-D and 2-D schemes. Finally, at high trellis complexities lower-dimensional TCM schemes will eventually prevail in performance. This can be explained by the fact that these schemes have more signal redundancy available for coding than higher-dimensional TCM schemes. Overall, the differences in real coding gains are not very large, that is, they are smaller than 1 dB for the range of complexities considered.

Other Recent Work

Trellis codes have also been designed for 1-D and 2-D signal sets with nonequally-spaced ("asymmetric") signals [6, Part I], [13]. Some modest coding gains compared to schemes with equally-spaced signals are achieved when the codes have few states and small signal sets. These gains disappear for larger signal sets and higher code complexity. There are open questions about the number of nearest neighbors and sensitivity to carrier phase offset when signals are nonequally spaced.

While TCM schemes have been designed for linear modulation channels, similar developments took place in the field of continuous phase modulation (CPM) for channels requiring constant envelope signals. A summary on CPM schemes is given in [14].

Conclusion

It is probably fair to state that in recent years the theory of trellis-coded modulation has matured to the point where the achievement of further major gains seem less likely. However, there are still open questions concerning real coding gains, performance under channel impairments other than Gaussian noise, and actual implementation complexities.

The 8-state CCITT scheme was established only two years ago (1984). In the meanwhile, many manufacturers of voice-band modems and other transmission equipment have adopted the new combined coding and modulation technique. At least one manufacturer has already realized the sophisticated "Z_8"-type 64-state TCM scheme in a commercial product. In the struggle toward higher coding gains, application of more complexity is met with diminishing returns. For channels with Gaussian noise, the so-called "cut-off rate" R_0, which is smaller than channel capacity by the equivalent of about 3 dB, has been suggested as a more realistic limit [15]. TCM schemes have reached this barrier.

Acknowledgments

The author expresses his sincere appreciation for the many helpful comments and suggestions he obtained from colleagues and reviewers while writing this two-part article. Dr. G.D. Forney deserves special thanks for his many excellent technical comments and many detailed suggestions about the presentation of the material. By sending early versions of [9] and [11] to the author, Dr. L.F. Wei and Dr. D.G. Forney contributed significantly to the discussion of multi-dimensional TCM, as presented this part of the paper. Dr. V.M. Eyuboglu helped generously by verifying the correctness of the codes presented and providing still missing numbers of nearest neighbors.

Appendix: Code Tables

Tables I-III are largely reproduced from [2]. Tables IV and V have not been published previously; however, similar codes with up to 64 states were found by L.F. Wei [9]. In the tables, an asterisk (*) indicates that free distance occurs only among parallel transitions, i.e., $d_{free}(\tilde{m}) > \Delta_{\tilde{m}+1}$.

TABLE I
CODES FOR AMPLITUDE MODULATION WITH "Z_1" SIGNALS, $\{\Delta_i, 0 \leq i \leq 2\} = \Delta_0, 2\Delta_0, 4\Delta_0$.

No. of states		Parity check coefficients			Asympt. coding gain [dB]			
2^ν	$\tilde{m}$	$\underline{h}^1$	$\underline{h}^0$	d^2_{free}/Δ^2_0	$G_{4AM/2AM}$ $(m = 1)$	$G_{8AM/4AM}$ $(m = 2)$	$G_{c/u}$ $(m \rightarrow \infty)$	N_{free} $(m \rightarrow \infty)$
4	1	2	5	9.0	2.55	3.31	3.52	4
8	1	04	13	10.0	3.01	3.77	3.97	4
16	1	04	23	11.0	3.42	4.18	4.39	8
32	1	10	45	13.0	4.15	4.91	5.11	12
64	1	024	103	14.0	4.47	5.23	5.44	36
128	1	126	235	16.0	5.05	5.81	6.02	66
256	1	362	515	16.0*	—	5.81	6.02	2
256	1	362	515	17.0	5.30	—	—	

TABLE II
CODES FOR PHASE MODULATION
8-PSK: $\{\Delta_i, 0 \leq i \leq 2\} = 2 \sin(\pi/8), \sqrt{2}, 2$;
16-PSK: $\{\Delta_i, 0 \leq i \leq 3\} = 2 \sin(\pi/16), 2 \sin(\pi/8), \sqrt{2}, 2$.

No. of states		Parity-check coefficients				Asympt. coding gain [dB]		
2^ν	$\tilde{m}$	$\underline{h}^2$	$\underline{h}^1$	$\underline{h}^0$	d^2_{free}/Δ^2_0	$G_{8PSK/4PSK}$ $(m=2)$	$G_{16PSK/8PSK}$ $(m=3)$	N_{free} $(m \rightarrow \infty)$
4	1	—	2	5	4.000*	3.01	—	1
8	2	04	02	11	4.586	3.60	—	2
16	2	16	04	23	5.172	4.13	—	≈2.3
32	2	34	16	45	5.758	4.59	—	4
64	2	066	030	103	6.343	5.01	—	≈5.3
128	2	122	054	277	6.586	5.17	—	≈0.5
256	2	130	072	435	7.515	5.75	—	≈1.5
4	1	—	2	5	1.324	—	3.54	4
8	1	—	04	13	1.476	—	4.01	4
16	1	—	04	23	1.628	—	4.44	8
32	1	—	10	45	1.910	—	5.13	8
64	1	—	024	103	2,000*	—	5.33	2
128	1	—	024	203	2,000*	—	5.33	2
256	2	374	176	427	2.085	—	5.51	≈8.0

TABLE III
CODES FOR TWO-DIMENSIONAL MODULATION WITH "Z_2" SIGNALS, $\{\Delta_i, 0 \leq i \leq 3\} = \Delta_0, \sqrt{2}\,\Delta_0, \sqrt{4}\,\Delta_0, \sqrt{8}\,\Delta_0$.

No. of states		Parity-check coefficients				Asympt. coding gain [dB]				
2^ν	$\tilde{m}$	$\underline{h}^2$	$\underline{h}^1$	$\underline{h}^0$	d^2_{free}/Δ^2_0	$G_{16QA/8PSK}$ $(m=3)$	$G_{32CR/16QA}$ $(m=4)$	$G_{64QA/32CR}$ $(m=5)$	$G_{C/U}$ $(m \rightarrow \infty)$	N_{free} $(m \rightarrow \infty)$
4	1	—	2	5	4.0*	4.36	3.01	2.80	3.01	4
8	2	04	02	11	5.0	5.33	3.98	3.77	3.98	16
16	2	16	04	23	6.0	6.12	4.77	4.56	4.77	56
32	2	10	06	41	6.0	6.12	4.77	4.56	4.77	16
64	2	064	016	101	7.0	6.79	5.44	5.23	5.44	56
128	2	042	014	203	8.0	7.37	6.02	5.81	6.02	344
256	2	304	056	401	8.0	7.37	6.02	5.81	6.02	44
512	2	0510	0346	1001	8.0*	7.37	6.02	5.81	6.02	4

TABLE IV
CODES FOR FOUR-DIMENSIONAL MODULATION WITH "Z_4" SIGNALS,
$\{\Delta_i, 0 \le i \le 5\} = \Delta_0, \sqrt{2}\,\Delta_0, \sqrt{2}\,\Delta_0, \sqrt{4}\,\Delta_0, \sqrt{4}\,\Delta_0, \sqrt{8}\,\Delta_0$.

No. of		Parity-check coefficients						Asympt. coding	
states								gain [dB]	N_{free}
2^ν	$\tilde{m}$	$\underline{h}^4$	$\underline{h}^3$	$\underline{h}^2$	$\underline{h}^1$	$\underline{h}^0$	d^2_{free}/Δ^2_0	$(m \to \infty)$	$(m \to \infty)$
8	2	—	—	04	02	11	4.0	4.52	88
16	2	—	—	14	02	21	4.0*	4.52	24
32	3	—	30	14	02	41	4.0*	4.52	8
64	4	050	030	014	002	101	5.0	5.48	144
128	4	120	050	022	006	203	6.0	6.28	

TABLE V
CODES FOR EIGHT-DIMENSIONAL MODULATION WITH "Z_8" SIGNALS,
$\{\Delta_i, 0 \le i \le 5\} = \Delta_0, \sqrt{2}\,\Delta_0, \sqrt{2}\,\Delta_0, \sqrt{2}\,\Delta_0, \sqrt{4}\,\Delta_0, \sqrt{4}\,\Delta_0$.

No. of		Parity-check coefficients						Asympt. coding	
states								gain [dB]	N_{free}
2^ν	$\tilde{m}$	$\underline{h}^4$	$\underline{h}^3$	$\underline{h}^2$	$\underline{h}^1$	$\underline{h}^0$	d^2_{free}/Δ^2_0	$(m \to \infty)$	$(m \to \infty)$
16	3	—	10	04	02	21	4.0	5.27	
32	3	—	10	04	02	41	4.0	5.27	496
64	3	—	044	014	002	101	4.0*	5.27	240
128	4	120	044	014	002	201	4.0*	5.27	112

V.M. Eyuboglu and G.D. Forney [16] discovered typographical errors in the earlier published "Z_1"- and "Z_2"-type 256-state codes [2], which have now been corrected in Tables I and III.

Some of the 8-PSK codes of Table II were improved, compared to those published in [2], by using the exact expression for $d_{free}(\tilde{m})$ in the code search. The 16-PSK codes of Table II are new.

The exact numbers of nearest neighbors, N_{free}, given in the tables were taken from various sources, in particular [11] and [17]. The approximate values of N_{free}, given for some codes in Table II, are average values recently determined by the author.

References

[1] G. Ungerboeck, "Trellis-coded modulation with redundant signal sets—Part I: Introduction," *IEEE Communications Magazine*, vol. 25, no. 2, Feb. 1987.

[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., "Convolutional codes I: Algebraic structure," *IEEE Trans. Information Theory*, vol. IT-16, pp. 720-738, Nov. 1970.

[4] M. Oerder, "Rotationally invariant trellis codes for mPSK modulation," *1985 Internat. Commun. Conf. Record*, pp. 552-556, Chicago, June 23-26, 1985.

[5] IBM Europe, "Trellis-coded modulation schemes for use in data modems transmitting 3-7 bits per modulation interval," CCITT SG XVII Contribution COM XVII, No. D114, April 1983.

[6] AT&T Information Systems, "A trellis coded modulation scheme that includes differential encoding for 9600 bit/sec, full-duplex, two-wire modems," CCITT SG XVII Contribution COM XVII, No. D159, August 1983.

[7] IBM Europe, "Trellis-coded modulation schemes with 8-state systematic encoder and 90° symmetry for use in data modems transmitting 3-7 bits per modulation interval," CCITT SG XVII Contribution COM XVII, No. D180, October 1983.

[8] A. R. Calderbank and N. J. A. Sloane, "Four-dimensional modulation with an eight-state trellis code," *AT&T Tech. Jour.*, vol. 64, pp. 1005-1017, May-June 1985.

[9] L. F. Wei, "Trellis-coded modulation with multidimensional constellations," submitted to *IEEE Trans. Information Theory*, Aug. 1985.

[10] A. R. Calderbank and N. J. A. Sloane, "An eight-dimensional trellis code," *Proc. of the IEEE*, vol. 74, pp. 757-759, May 1986.

[11] G. D. Forney, Jr., *Coset Codes I: Geometry and Classification*, Aug. 25, 1986.

[12] N. J. A. Sloane, "The packing of spheres," *Scientific American*, vol. 250, pp. 116-125, Jan. 1984.

[13] M. K. Simon and D. Divsalar, "Combined trellis coding with asymmetric MPSK modulation," *JPL Publication* 85-24, May 1, 1985.

[14] C. E. Sundberg, "Continuous phase modulation," *IEEE Communications Magazine*, vol. 24, no. 4, pp. 25-38, April 1986.

[15] J. L. Massey, "Coding and modulation in digital communications," *Proc. 1974 Int. Zurich Seminar on Digital Communications*, Zurich, Switzerland, pp. E2(1)-(4), March 1974.

[16] V. M. Eyuboglu and G. D. Forney, Jr., private communications, Sept. 1984 and Sept. 1986.