INVITED ARTICLE

DIGITAL TRANSMISSION IN THE 21ST CENTURY: CONFLATING MODULATION AND CODING

EZIO BIGLIERI
POLITECNICO DI TORINO

The coding/modulation problem consists of finding practical ways of communicating discrete messages reliably on a real-world channel. Shannon's theorem about the existence of an optimum coding scheme achieving capacity gives no guidance as to how to find appropriate coding schemes or how complex they may be to implement. Thus, research activity in coding and modulation was aimed at selecting a transmission/reception scheme making the best possible use of the resources available for transmission, viz., bandwidth, power, and complexity, in order to approximate the ultimate performance. In the past decade methods of approaching capacity have been found for linear Gaussian channels (turbo codes, low-density parity-check codes). The wireless channel provides some additional challenges.

The coding/modulation problem consists of finding practical ways of communicating discrete messages reliably on a real-world channel. Shannon's theorem about the existence of an optimum coding scheme achieving capacity gives no guidance as to how to find appropriate coding schemes or how complex they may be to implement.

PROLEGOMENA

The coding/modulation (C/M) problem consists of finding practical ways of communicating discrete messages reliably on a real-world channel: this may involve space and satellite communications, data transmission over twisted-pair telephone wires, shielded cable-TV wire, data storage, digital audio/video transmission, mobile communication, terrestrial radio, deep-space radio, indoor radio, or file transfer. The channel is degraded, as it includes attenuation, thermal noise, intersymbol interference, multiple-user interference, multipath propagation, and power limitations.

The most general statement about the selection of a C/M scheme is that it should make the best possible use of the resources available for transmission, viz., bandwidth, power, and complexity, in order to achieve a required quality of service (QoS). Typically, the latter is expressed in terms of error probability, which in turn is a decreasing function of the signal-to-noise ratio (SNR). Sensible strategies for C/M design must be based on four factors:

- Error probability: this tells us how reliable the transmission is.
- Bandwidth efficiency: this measures the efficiency in bandwidth expenditure.
- Signal-to-noise ratio necessary to achieve the required QoS: this measures how efficiently the C/M scheme makes use of the available power.
- Complexity: this measures the cost of the equipment.

This article provides an overview of these strategies, and especially of the shift in their perspectives that has occurred in recent years. In particular, we shall see how Shannon started it all, how the paths of coding and modulation, after being separated for decades, have eventually merged, and how wireless channels pose new challenges.

THE SHANNON CHALLENGE

C/M design is based on a tradeoff: in fact, if we cannot transmit as many bits per second as we would like, it may be only because of complexity limitations, or because of choices made to meet the bandwidth or power constraints of the channel. To clarify this point, let us define two basic parameters. The first one is the spectral (or bandwidth) efficiency R_s/W, which tells us how many bits per second (R_s) can be transmitted in a given bandwidth W. Table 1 summarizes the spectral efficiencies achieved by some wireless systems and standards [1]. The second parameter is the *asymptotic power efficiency* γ of a C/M scheme. This is defined as follows. For high SNRs the error probability can be closely approximated by a decreasing function whose argument is γ times the ratio between the energy per transmitted information bit E_b and the noise power spectral density of the noise N_0. Thus, the parameter γ expresses how efficiently a modulation scheme makes use of the available signal energy to yield a given error probability. We may say that, at least for high SNRs, a C/M scheme is better than another if its asymptotic power efficiency is greater. (At low SNRs the situation is much more complicated than this, but the asymptotic power efficiency still plays some role.) Some pairs of values of R_s/W and that can be achieved by practical modulation schemes without coding are summarized in Table 2.

The fundamental tradeoff is that, for a given QoS requirement, increased spectral efficiency (demanded in a bandwidth-limited regime) can be reliably achieved only with a corresponding increase in the minimum required SNR. Conversely, the minimum required SNR can be reduced only by decreasing the spectral efficiency of the system. Roughly, we may say that we work in a bandwidth-limited regime if the chan-

System	Modulation	Spectral efficiency
GSM	GMSK	1.35 bit/s/Hz
IS-54/136	π/4-DQPSK	1.62 bit/s/Hz
IS-95	QPSK	0.96 bit/s/Hz
PDC	π/4-DQPSK	1.68 bit/s/Hz

TABLE 1. *Modulation and spectral efficiency of second-generation digital cellular standards. The spectral efficiency of IS-95 assumes 1024 chips/bit [1].*

nel constraints force us to work with a ratio R_s/W quite higher than 1, and in a power-limited regime if the opposite occurs.

ENTER SHANNON

The "Big Bang" of the C/M cosmogony occurred in 1948, when Claude Shannon [2] demonstrated that for any transmission rate less than or equal to a parameter called *channel capacity* there exists a coding scheme that achieves an arbitrarily small probability of error, and hence can make transmission over the channel perfectly reliable. With magnificent irony, Shannon exhibited a nonconstructive proof of his capacity theorem without any guidance as to how to find an actual coding scheme achieving the ultimate performance *with limited complexity*. The cornerstone of his proof was the fact that if we pick a long code *at random*, then its average probability of error will be satisfactorily low; moreover, there exists at least one code whose performance is at least as good as the average. Direct implementation of random coding, however, leads to a decoding complexity that prevents its actual use.

Since 1948 communication engineers have tried hard to develop practically implementable C/M schemes in an attempt to approach ideal performance, and hence capacity. In spite of some pessimism (for a long while the motto of coding theorists was "good codes are messy") the problem was eventually solved in the past decade, at least for an important special case, the linear Gaussian channel (additive white Gaussian noise, or AWGN, channel) [3–5].

A paradigm shift also occurred in the 1980s, prompted in part by a deeper understanding of the tenets of Shannon's theory: the conflation of modulation and coding, two disciplines that previously had evolved independently. Driven by the demand of higher values of R_s/W, increasingly sophisticated signals were advocated: QAM with large M values (especially for terrestrial radio links), lattice constellations [6], and signals designed for limited bandwidth occupancy [7, 8] were advocated. If high power efficiency was sought, more and more powerful error-correcting codes were made available. Eventually, the invention of trellis-coded modulation (TCM) [9, 10] made it clear that modulation and coding should be integrated in a single entity for better efficiency. In the following, we shall see how this was put into practice.

Modulation	R_s/W	γ
PAM	$2\log_2 M$	$\frac{3\log_2 M}{M^2-1}$
PSK	$\log_2 M$	$\sin^2\frac{\pi}{M}\cdot\log_2 M$
QAM	$\log_2 M$	$\frac{3}{2}\frac{\log_2 M}{M-1}$
FSK	$2\frac{\log_2 M}{M}$	$\frac{1}{2}\log_2 M$

TABLE 2. *Maximum bandwidth- and power-efficiency of some modulation schemes: PAM, PSK, QAM, and orthogonal FSK. Here M is the number of available signals.*

POWER-LIMITED REGIME

In the power-limited (i.e., wide-bandwidth, low-SNR) regime we should use error-control codes, which increase the power efficiency by adding extra bits to the transmitted symbol sequence. The communication limits for an AWGN channel, when we are willing to tolerate a certain non-zero BER while using a rate-R_c code (R_c is measured in information bits

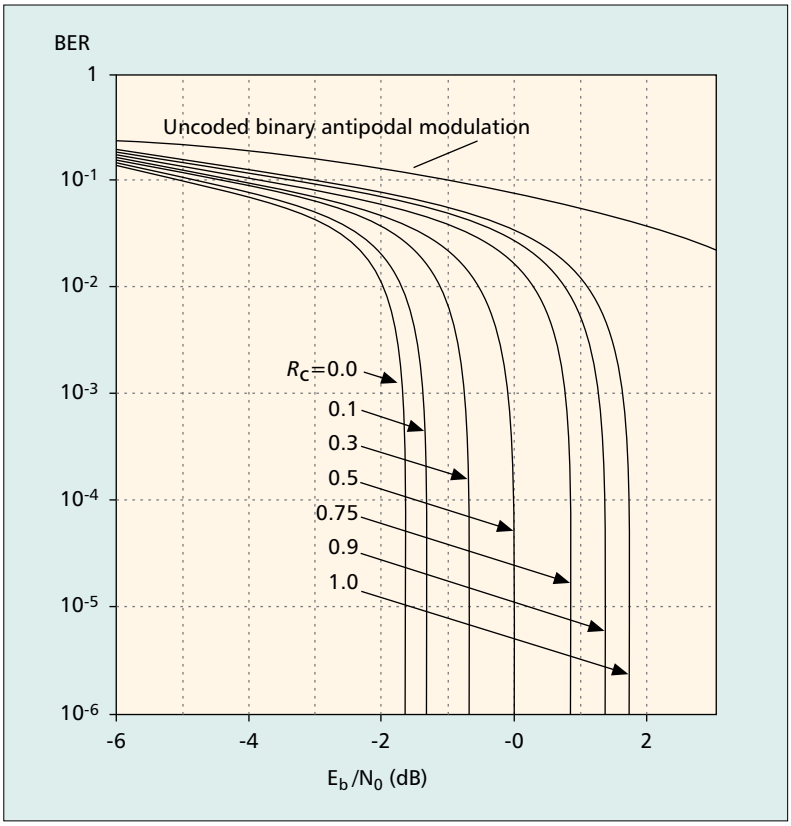

FIGURE 1. *Admissible region for the pair BER, E_b/N_0. For a given code rate R_c, only the region above the curve labeled R_c is admissible. R_c = 0 refers to the limiting value as the code rate decreases.*

The invention of "turbo codes," (parallel concatenated convolutional codes) took the coding community by storm. In spite of their reduced complexity, they approach Shannon's performance limit to an unprecedented extent.

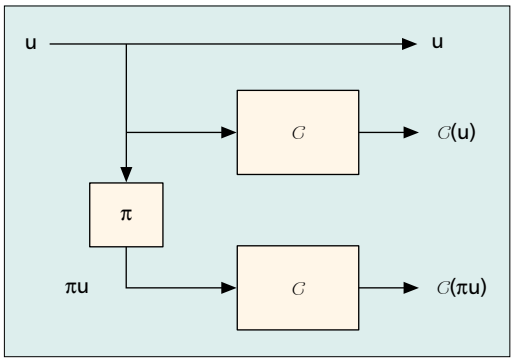

FIGURE 2. *"Turbo" encoder.*

per transmitted symbol), are shown in Fig. 1 [6]. Observe how even with a vanishingly small code rate no reliable transmission of information is possible unless E_b/N_0 does exceed –1.6 dB. These limits are derived with no restriction on the signaling scheme used; for a given choice of the modulation (for example, a binary constellation), the admissible region will be further reduced.

Hard vs. Soft Decoding — The first quantum leap in error-control coding practice occurred when system engineers realized in full that the separation between the functions of the demodulator and the decoder, far from being a neutral choice, was actually considerably harmful. The theory of "error-correcting codes" was originally motivated by the need to improve the performance of a modem by introducing a device that could compensate for the errors introduced by the demodulator. In this framework, the demodulator judges first what the modulator input was, and then passes its decision on to the decoder; this in turn uses the known code structure to judge which code word was sent. This procedure is called "hard" (or "algebraic") decoding; it cannot be an optimum strategy, because for every hard decision the demodulator discards some information that could be used, and it is known that we should "never discard information prematurely that may be useful in making a decision until after all decisions related to that information have been completed" [11].

With an integrated view of coding and modulation, the demodulator does not make mistakes that the decoder is expected to correct. Rather, it supplies only tentative estimates (often referred to as "soft decisions") for the various symbols, without discarding any information that the decoder may need for optimum operation. An "optimum" decoder either minimizes the word error rate (under the standard assumption that all code words were generated with equal probability), or the bit error rate (BER). If BER is minimized, the decoder is called MAP ("maximum a posteriori"). Soft decoding often provides a considerable improvement in performance; a figure often quoted for the SNR advantage of soft versus algebraic decoders is 2 dB [6]. This synergy between the demodulator part and the decoder part of the receiver makes it more appropriate to talk about *error-control* rather than *error-correcting* codes.

Maximization of the Hamming distance between code word pairs (the number of components in which they differ) is a sensible criterion for selecting a code to be used for hard decoding; for soft decoding, the geometric ("Euclidean") distance between code words, interpreted as signal points, should be used instead. When soft decoding is used, it is often said that coding occurs "in the signal space."

Graphical Models for Codes — The invention of "turbo codes," (parallel concatenated convolutional codes) [12] took the coding community by storm in the mid-1990s. In spite of their reduced complexity, they approach Shannon's performance limit to an unprecedented extent. Turbo codes are among the very best codes known: they combine a random-like behavior (which is attractive in light of Shannon's coding theorem) with a relatively simple structure, obtained by concatenating low-complexity compound codes. They can be decoded by MAP-decoding their component codes using partially available information from all others. The encoder structure of a simple turbo code is shown in Fig. 2. A block **u** of data to be transmitted enters a systematic encoder which produces three sequences: one is **u** itself, the second is the parity-check sequence generated from the convolutional encoder C, and the third is generated by applying an interleaver (corresponding to a permutation π) to the input stream, then applying the permuted stream to a second convolutional encoder (which may be equal to the first one). The result is a code with rate 1/3 (three bits are output for every input bit).

While codes intended for hard decoding are best described by looking at their algebraic structure, codes to be soft-decoded can be usefully described by graphical models, on which appropriate decoding algorithms can operate. Trellises used to describe convolutional codes are still the most popular graphical models. The celebrated Viterbi decoding algorithm is usually described as a way to find the shortest path through one such trellis, where the length of each trellis branch is measured by the distance

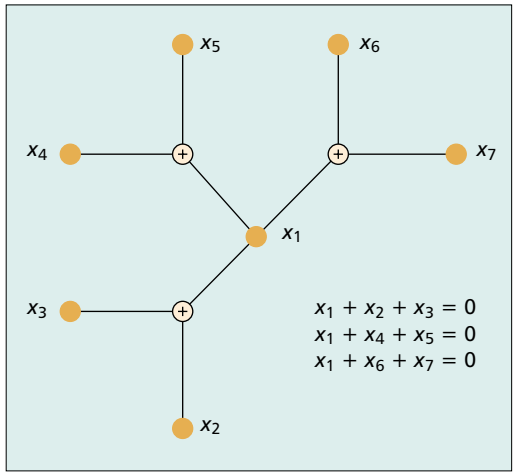

FIGURE 3. *Graph of a linear binary code with words of length 7. The parity-check equations of the code that the graph represents are also shown (sums are modulo-2).*

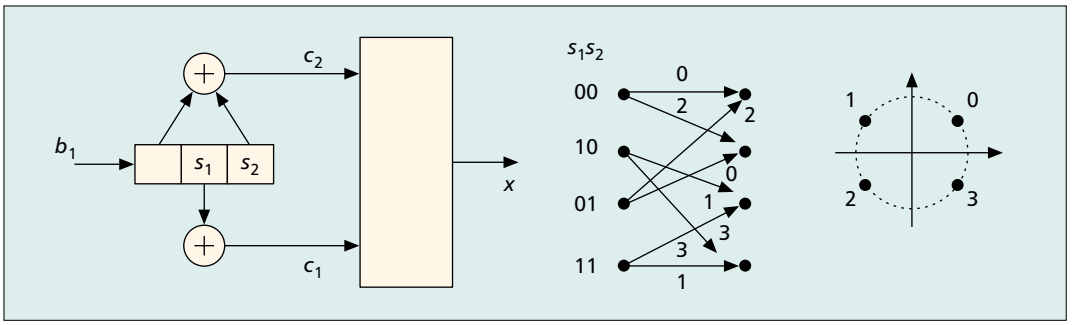

FIGURE 4. *A TCM encoder/modulator based on 4-PSK signals. The convolutional encoder transforms the binary symbol b_1 into the binary pair (c_1, c_2), while the modulator transforms this pair into a signal x chosen from 4-PSK. The trellis section describes how signals are generated, and can be used to decode/demodulate them using the Viterbi algorithm. The pair (s_1, s_2) denotes the state of the encoder; there are four states here.*

between a candidate symbol and the signal received. More recently, the exceedingly general tool of "factor graphs" was introduced to describe families of codes and a way to decode them [13–15]. When a code can be represented by a cycle-free factor graph, then the structure of the code graph lends itself directly to the specification of a finite algorithm (the "sum-product" and the "min-sum" algorithms) for soft decoding. If cycles are present, then iterative algorithms can be implemented. See Fig. 3 for an example of a linear binary length-7 code described by its graph. Its three parity-check equations (the constraints that a binary 7-tuple must satisfy in order to be a word of this code) are represented by a graph in which filled circles denote the symbols, and sum nodes the parity-check constraints. In the code of Fig. 3 we see for example that (0111111) is a valid code word, while (0011111) is not. In this framework, turbo codes can be represented by graphs that are made up of simple, easily decodable codes connected by a long, pseudo-random interleaver. Since they have long cycles, decoding of turbo codes is accomplished iteratively by a sum-product algorithm. (No one yet fully understands just why turbo decoding works as well as it does. This is an active area of research.)

Low-density parity-check (LDPC) codes [16] are also codes defined in terms of a graph. LDPC codes, like turbo codes, are very effectively decoded using the sum-product algorithm. Gallager invented low-density parity-check codes in his 1963 thesis nearly 40 years ago, an invention long before his time ("a bit of 21st-century coding that happened to fall in the 20th century" [13]). He also invented what is now known as the sum-product iterative decoding algorithm. Following in the footsteps of Gallager, nearly 20 years ago Tanner started plowing what is now known as the field of codes on graphs, one that is quickly achieving a tropical richness. About five years ago, the study of turbo codes and the independent rediscovery of the power and efficiency of LDPC codes rekindled an interest in this field as a conceptual umbrella under which to bring turbo codes as well as LDPC codes. Nowadays, turbo and LDPC decoding algorithms are described as instances of iterative decoding on graphs. More generally, coding on graphs has led to the belief that "almost any simple code interconnected by a large pseudo-random interleaver and decoded with sum-product algorithm will yield near-Shannon performance." [13]

Performance — Figure 1 illustrates how far uncoded transmission is from the theoretical limit, and hence how much can be gained with coding. With uncoded transmission at a BER of 10^{-5} we are about 9.4 dB away from Shannon's limit. With a powerful convolutional code we may obtain an improvement close to 5.7 dB over uncoded transmission. Binary convolutional codes with sequential decoding were shown in the 1960s to be an implementable solution for operating about 3 dB away from the Shannon limit, and in recent years this 3-dB barrier was eventually broken. Up until the last few years, a code obtained by concatenating a Reed-Solomon code with a convolutional code was considered to be the state of the art; at a BER of 10^{-5}, this system was roughly 2.3 dB from Shannon's limit. Turbo codes with properly designed interleavers can now achieve an error performance extremely close to the limit. The first "turbo code," introduced in 1993, was roughly 0.5 dB from the limit at a BER of 10^{-5}. At the time of this writing, the record holder among the codes whose performance has been simulated is a rate-1/2 LDPC code that achieves within 0.04 dB of the Shannon limit at a BER of 10^{-6} using a block length of 10^7 symbols [17].

Turbo codes perform extremely well for BERs above 10^{-4}–10^{-5} (the "waterfall" region); however, they have a significantly weakened performance at lower BERs, due to the fact that their component codes have a relatively poor minimum Euclidean distance, which manifests its effects at these BERs. The fact that these codes do not have large minimum distances causes the BER curve to decrease its slope at BERs below 10^{-5}, a phenomenon known as *error floor*. It has been argued that the presence of this error floor makes turbo codes not suitable for applications requiring extremely low BERs. Their poor minimum distance, and their lack of error-detection capability (due to the fact that in turbo decoding only information bits are decoded) make these codes perform badly in terms of block error probability. In turn, poor block error perfor-

Nowadays, turbo and LDPC decoding algorithms are described as instances of iterative decoding on graphs. More generally, coding on graphs has led to the belief that "almost any simple code interconnected by a large pseudo-random interleaver and decoded with sum-product algorithm will yield near-Shannon performance."

In the bandwidth-limited regime the early solution was uncoded multilevel modulation, as nonbinary signal alphabets must be used to approach capacity; in the mid 1970s the invention of trellis-coded modulation showed a different way.

Number of states	Coding gain (8-PSK)	Coding gain (16-QAM)
4	3.0	4.4
8	3.6	5.3
16	4.1	6.1
32	4.6	6.1
64	4.8	6.8
128	5.0	7.4
256	5.4	7.4

TABLE 3. *Asymptotic coding gains of TCM (in dB).*

mance also makes these codes not suitable for certain communication applications [18]. Another relevant factor that may provide guidance in the choice of a C/M scheme is the decoding delay that one should allow. In fact, turbo codes and LDPC codes suffer from a considerable decoding delay, and hence their application might be useful for data transmission more than for real-time speech.

Bandwidth-Limited Regime

The use of error-control codes requires the modulator to operate at a higher data rate and, hence, requires a larger bandwidth. In a bandwidth-limited environment, increased efficiency in both power and frequency utilization can be obtained by choosing an integrated C/M solution, where higher-order modulation schemes (e.g., 8-PSK instead of 4-PSK) are combined with high-rate coding schemes. In the bandwidth-limited regime the early solution was uncoded multilevel modulation, as nonbinary signal alphabets must be used to approach capacity; in the mid 1970s the invention of trellis-coded modulation showed a different way. The *trellis-coded modulation* (TCM) solution [9, 10] combines the choice of a modulation scheme with that of a convolutional code (Fig. 4), while the receiver, instead of performing demodulation and decoding in two separate steps, combines the two operations into one.

The modulator has memory here; in standard (memoryless) modulation, for each symbol emitted by the source the modulator chooses a signal that depends only on that symbol. With TCM, the signal chosen depends also on a limited number of past symbols. We say that the past symbols drive TCM into a certain *state* (defined as the contents of the two rightmost positions of the shift register in Fig. 4), and that the signal generated depends on the source symbol and on the state. The transition between states is described by a trellis, whose branches are labeled by the signals generated as the state sequence evolves. In "standard" error-control coding the redundancy necessary to power savings is obtained by increasing the number of transmitted symbols, which entails a bandwidth expansion. Here the redundancy occurs as a factor-of-2 expansion of the signal constellation size, *with no bandwidth expansion*. For example, the scheme of Fig. 4 carries 1 bit per signal (in fact, one signal is chosen for each value of b_1), but uses, instead of a binary constellation, a 4-signal constellation (which could carry two bits per signal if used with no coding).

The number of states of the modulator is directly proportional to the number of computations needed to decode TCM, and hence is the complexity factor here. Yet, increasing the number of states leads to better performance. Table 3 summarizes some of energy savings ("coding gains") in dB that can be obtained by doubling the constellation size and using TCM. These are considered for coded 8-PSK (relative to uncoded 4-PSK) and for coded 16-QAM (relative to uncoded 8-PSK). These gains can actually be achieved only for high SNRs, and decrease as the latter decrease.

Choosing the Modulation — Shannon's theorem assumes that, for a given spectral efficiency, one is free to choose the modulation scheme that results in the best possible performance. How-ever, in real communication systems there are many practical considerations that come into play in picking the modulation. For example, radio communication systems that use nonlinear power amplifiers in their transmitter require constant-envelope signaling such as *M*-PSK or continuous-phase modulation (which has a compact spectrum with narrow sidelobes [7, 8]) in order to avoid signal distortion. If operating in a radio environment where several users share the spectrum, in addition to having a high spectral efficiency, a modulation scheme should be robust to co-channel interference, and hence have a compact power density spectrum, with a narrow main lobe and fast roll-off of side-lobes. Also, for data transmission over voice-grade telephone channels, the linear distortion caused by bandwidth limitations ("intersymbol interference") forces the use of large signal constellations that employ a combination of amplitude and phase modulation to achieve high data rates.

A technique called *multilevel coding* can achieve the capacity by generating signal constellations by combining several relatively simple *linear binary codes*. The observation that for the AWGN channel the capacity is achieved by a Gaussian signaling distribution has led to a concept called *shaping* [5]. A signal constellation has good shape if its distributions over the coordinate axes are nearly Gaussian.

The Wireless Channel

C/M choices are strongly affected by the channel model. We have first examined the Gaussian channel, because this has shaped the discipline of C/M. Other important channel models arise in digital wireless transmission. The consideration of wireless channels where nonlinearities, Doppler shifts, fading, shadowing, and interference from other users make the simple AWGN channel model far from realistic, and forces one to revisit the Gaussian-channel paradigms described in the previous section. Over wireless channels, due to fading and interference, the sig-

nal-to-disturbance ratio becomes a random variable, which brings into play a number of new issues [19].

Among the most popular wireless channel models we recall the flat independent fading channel (where the signal attenuation is constant over one symbol interval, and changes independently from symbol to symbol), the block-fading channel (where the signal attenuation is constant over an N-symbol block, and changes independently from block to block), and a channel operating in an interference-limited mode. This last model takes into consideration the fact that in a multi-user environment a central concern is overcoming interference, which may limit the transmission reliability more than noise.

The Flat Fading Channel

This simplest fading channel model assumes that the duration of a modulated symbol is much greater than the delay spread caused by multipath propagation. If this occurs, then all frequency components in the transmitted signal are affected by the same random attenuation and phase shift, and the channel is frequency-flat. If in addition the channel varies very slowly with respect to the symbol duration, then the fading level remains approximately constant during the transmission of one symbol (if this does not occur the fading process is called *fast*.)

The assumption of a frequency-flat fading allows it to be modeled as a process affecting the transmitted signal in a multiplicative form. The additional assumption of slow fading reduces this process to a random variable during each symbol interval. Let $x(t)$ denotes the complex envelope of the modulated signal transmitted during an interval of length T (this means that the signal is actually the real part of $x(t)\exp(j2\pi f_c t)$, f_c the carrier frequency). Then a common model for the complex envelope of the signal received at the output of a channel affected by slow, flat fading and additive white Gaussian noise can be expressed in the form

$$r(t) = Re^{j\Theta}x(t) + n(t), \qquad (1)$$

where $n(t)$ is a complex Gaussian noise, and $Re^{j\Theta}$ is a complex Gaussian random variable, with R a real random variable having a Rice or Rayleigh pdf. The value taken on by $Re^{j\Theta}$ is called the channel state information (CSI).

If we can further assume that fading is so slow that we can estimate the phase Θ with sufficient accuracy, and hence compensate for it, then coherent detection is feasible, and model (1) can be further simplified to

$$r(t) = Rx(t) + n(t). \qquad (2)$$

It should be immediately apparent that with this simple model of fading channel the only difference with respect to an unfaded AWGN channel, described by the input-output relationship

$$r(t) = x(t) + n(t), \qquad (3)$$

resides in the fact that R, instead of being a constant attenuation, is now a random variable, whose value affects the amplitude, and hence the power, of the received signal. A key role here is played by the channel state information, i.e., the fade level, which may be known at the transmitter, at the receiver, or both. Being privy to CSI allows the receiver to adapt its detection strategy, and the transmitter to adapt its transmission policy, for example by increasing its power in the presence of a deep fade (more on this later).

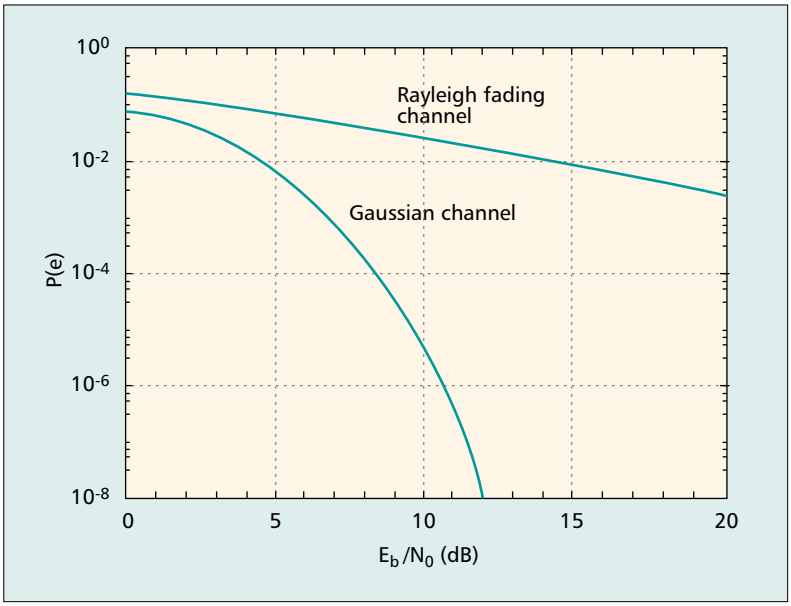

Figure 5. *Error probabilities of binary transmission over the Gaussian channel and over a Rayleigh fading channel.*

Figure 5 compares the error probability over the Gaussian channel with that over the Rayleigh fading channel with no CSI available at either transmitter or receiver (binary uncoded coherent PSK is assumed). This simple example shows how considerable the loss in energy efficiency due to fading can be. Moreover, in the power-limited environment typical of wireless channels, the simple device of increasing the transmitted energy to compensate for the effect of fading is not directly applicable. A solution is consequently the use of coding, which can compensate for a substantial portion of this loss. Even diversity, a well known technique to counteract fading, can be categorized as a special case of coding. In fact, diversity consists of transmitting the same information over a number of independent channels, and hence can be viewed as a special kind of "repetition" coding, one whose Hamming distance turns out to be equal to the number of diversity branches.

Coding for the Slow, Flat Rayleigh Fading Channel — Analysis of coding for the slow, flat Rayleigh fading channel proves that Hamming distance (also called "code diversity" in this context) plays the central role here. Assume transmission of a coded sequence $\mathbf{x} = (x_1, x_2, \ldots x_n)$ where the components of $\mathbf{x}$ are signals selected from a constellation. We also assume that, thanks to perfect (i.e., infinite-depth) interleaving, the fading random variables affecting the various signals x_k are independent. Finally, it is assumed that perfect CSI is available at the receiver and the detection is coherent, i.e., that the phase shift due to fading is estimated and removed.

The basic idea behind adaptivity consists of allocating transmitted power and code rate to take advantage of favorable channel conditions by transmitting at high speeds, while at the same time counteracting bad conditions by reducing the throughput.

We can calculate the probability that the receiver prefers the candidate code word $\hat{\mathbf{x}}$ to the transmitted code word $\mathbf{x}$ (this is called the "pairwise error probability" and is the basic building block of any error probability evaluation). This turns out to be roughly inversely proportional to the signal-to-noise ratio raised to a power which is the Hamming distance between $\mathbf{x}$ and $\hat{\mathbf{x}}$, the code diversity [20, 21].

Robustness — From the previous discussion, it is accepted that C/M schemes optimum for this channel should maximize the Hamming distance between code words. Now, if the channel model is uncertain, or not stationary enough to design a C/M scheme closely matched to it, then the best proposition may be that of a "robust" solution, that is, a solution that provides suboptimum (but close to optimum) performance on a wide variety of channel models. The use of antenna diversity with maximal-ratio combining [6] provides good performance on a wide variety of fading environments.

The standard approach to receive-antenna diversity is based on the fact that, since each antenna generates its own channel, the probability that the signal will be simultaneously faded on all channels can be made small, and hence the detector performance improves. Another perspective [22] is based upon the observation that, under fairly general conditions, a channel affected by fading can be turned into an additive white Gaussian noise (AWGN) channel by increasing the number of antenna-diversity branches and using maximum-ratio combining (which requires knowledge of CSI at the receiver). Consequently, it can be expected (and verified by analyses and simulations) that a coded modulation scheme designed to be optimal for the AWGN channel will perform asymptotically well also on a fading channel with diversity, at the only cost of an increased receiver complexity.

Another robust solution is offered by bit-interleaved coded modulation (BICM), which consists of introducing a bit interleaver between encoder and modulator [23].

Power Allocation — A strategy that can be used in conjunction with coding and diversity is based on the simple observation that the increase of signal power reflected in Fig. 5 is based on an average fading level, and consequently power increase may be allocated more efficiently on a symbol-by-symbol basis, provided that CSI is available at the transmitter.

Consider the simplest such strategy. The flat, independent fading channel with coherent detection yields the received signal (2). Assume that the CSI R is known at the transmitter front-end, that is, the transmitter knows the value of R during the transmission (this assumption obviously requires that R is changing very slowly). Under these conditions, let the average transmitted power be S_0 and the transmitted signal

$$x(t) = \sqrt{S(R)}s(t), \tag{4}$$

where $s(t)$ has unit power. One possible optimization criterion (constant error probability over each symbol) requires that

$$S(R) = S_0 \frac{R^{-2}}{\mathrm{E}[R^{-2}]}. \tag{5}$$

where E denotes expectation. In this way the channel is transformed into an equivalent additive white Gaussian noise channel. This technique ("channel inversion") is simple to implement, since the encoder and decoder are designed as for the AWGN channel, independently of the fading statistics; for instance, it is common in spread spectrum systems with near-far interference imbalances. (These occur when a mobile device close to a base station is received at a higher power than the desired source, which is located further from the station.) However, it may suffer from a large capacity penalty. For example, with Rayleigh fading, $\mathrm{E}[R^{-2}]$ diverges. To avoid divergence of the average power (or an inordinately large value thereof) one may invert the channel only if the involved power expenditure is not too large; otherwise, compensate only for a part of the channel attenuation. Optimal power allocation is also described as "water-filling" in time, as it resembles the "water-filling" in frequency used to calculate the capacity of a Gaussian channel with colored noise [24].

Adaptive C/M Techniques

Since wireless channels exhibit a time-varying response, adaptive transmission strategies look attractive to prevent insufficient utilization of the channel capacity. The basic idea behind adaptivity consists of allocating transmitted power and code rate to take advantage of favorable channel conditions by transmitting at high speeds, while at the same time counteracting bad conditions by reducing the throughput. For an assigned QoS, the goal is to increase the average spectral efficiency by taking advantage of the transmitter having knowledge of the CSI. The amount of performance improvement provided by such knowledge can be evaluated in principle by computing the Shannon capacity of a given channel with and without it. However, it should be kept in mind that capacity results refer to a situation in which complexity and delay are not constrained. Thus, for example, for a Rayleigh fading channel the capacity with channel state information (CSI) at the transmitter and the receiver is only marginally larger than for a situation in which only the receiver has CSI. This implies that if very powerful and complex codes are used, then CSI at the transmitter can buy little. However, in a delay- and complexity-constrained environment a considerable gain can be achieved [25].

Adaptive techniques (see [26] and references therein) are based on two steps:

- Measurement of the parameters of the transmission channel.
- Selection of one or more transmission parameters based on the optimization of a preassigned cost function.

A basic assumption here is that the channel does not vary too rapidly, otherwise the parameters selected might be badly matched to the channel. Thus, adaptive techniques can only be beneficial

in a situation where the Doppler spread is not too wide. This makes adaptive techniques especially attractive in an indoor environment, where propagation delays are small and the relative speed between transmitter and receiver is typically low. In these conditions, adaptive techniques can work on a frame-by-frame basis. We list below a number of these techniques.

Adapting power level. Through power control, the transmission level is varied according to the channel fluctuations. This strategy increases the transmitter peak-power requirements, and in a multi-user environment, the level of cochannel interference, which may reduce channel capacity if coordination among users is not allowed.

Adapting constellation size. Among adaptive transmission techniques, adaptive modulation plays a central role, because it increases the data transmission efficiency without increasing the multi-access interference power. In its essence, adaptive modulation consists of transmitting at the highest possible rate compatible with an assigned QoS, as specified by higher-layer requirements such as packet error rate, packet delay, etc. This is obtained by using a hierarchy of different constellations of increasing size. Adaptive constellation size may be implemented so as to maintain a constant transmit power while providing a target QoS. The number of signals in the modulator constellation can be varied in such a way that the short-term BER is approximately constant while the short-term bit rate varies, or vice versa. In a single-user environment adaptive modulation can provide a 5-10 dB gain over a fixed-rate system having only power control.

Adapting code rate. The coding scheme can be changed so as to respond to the channel state by selecting the optimum code rate. Punctured convolutional codes are especially useful to this purpose, because they enable adaptive encoding and decoding without modifying the basic structure of the encoder and the decoder.

Adapting power level and constellation size. Both modulation scheme and transmit level can be adapted in a single-user environment or for a multi-user channel. This combination leads to a significant throughput increase as compared to no power control.

Adapting constellation size and symbol rate. Both constellation size and symbol transmission rate can be adapted. The system selects the optimum modulation parameters so as to maximize the bit rate while satisfying the required BER. Here a lower symbol rate is achieved by consecutively transmitting identical modulation symbols at the maximum symbol rate (this is equivalent to repetition coding).

Adapting power and transmission rate. Both the transmission rate and the power can be selected so as to maximize the spectral efficiency while satisfying average power and BER constraints.

Adapting modulation size and coding scheme. This adaptation strategy refers to trellis-coded modulation. By holding the number of information bits entering the encoder fixed, and adapting the number of bits left uncoded according to the estimate of the CSI, the trellis structure is changed. This adaptation scheme is not robust to estimation errors, especially on Rayleigh fading channels. If this is the case, BICM [23] may be a more attractive proposition, because it increases the time diversity. With this scheme, the code is left fixed, while the signal constellation is adapted to the channel conditions (hence, it could be categorized under the rubric "adapting the constellation size"). Analyses suggest that the code structure of BICM is more suitable for adaptive systems that must support highly mobile users.

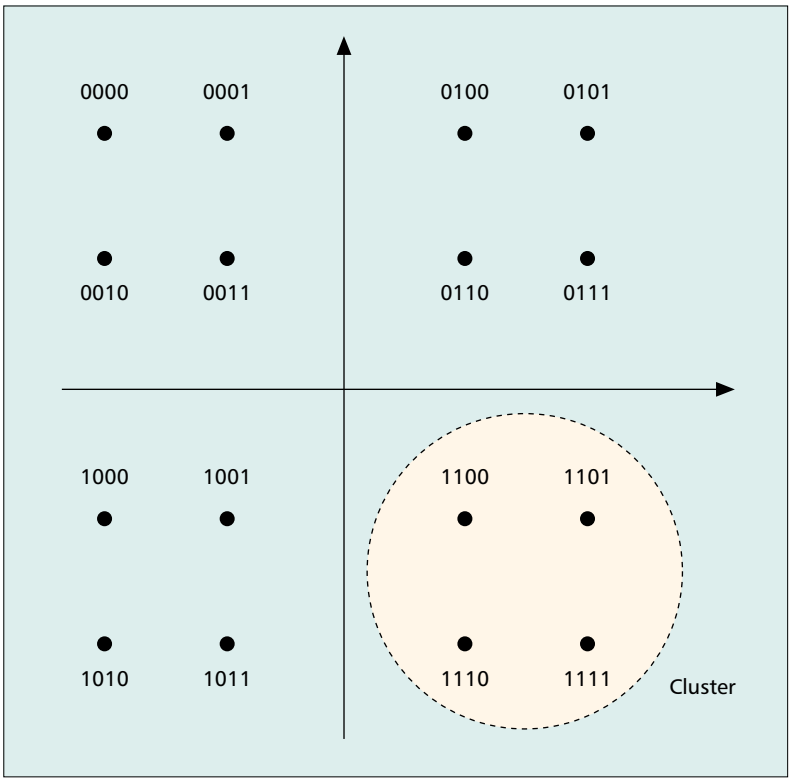

FIGURE 6. *A 16-point signal constellation for multiresolution modulation in which the first two bits identify the cluster and the last two identify the signal within it.*

Adapting code rate, symbol rate, constellation size. Code rate, symbol rate, and constellation size can be adapted simultaneously. Code rate adaptation is obtained by puncturing a convolutional code, while the constellation size is selected by setting SNR thresholds. If the target BER cannot be achieved under any combination of parameters, the system transmits no data.

ADDITIONAL ISSUES

UNEQUAL ERROR PROTECTION

In some analog source coding applications, such as speech or video compression, the sensitivity of the source decoder to errors in the coded symbols is typically not uniform: the quality of the reconstructed analog signal is rather insensitive to errors affecting certain classes of bits, while it degrades sharply when errors affect other classes. This happens, for example, when analog source coding is based on some form of hierarchical coding, where a relatively small number of bits carries the "fundamental information" and a larger number of bits carries the "details," as in the case of MPEG standards.

By introducing correlation among signals transmitted by different antennas as well as transmitted at different times, a coding gain is obtained without any sacrifice in bandwidth and with a relatively simple receiver structure. Space-time codes are already finding their way to modern wireless system standards.

Assuming that the source encoder produces frames of binary-coded symbols, each frame can be partitioned into classes of symbols of different "importance" (i.e., of different sensitivity). Then it is apparent that the best coding strategy aims at achieving lower BER levels for the important classes while admitting higher BER levels for the unimportant ones. This feature is referred to as "unequal error protection." [27–30]

A conceptually similar solution to the problem of avoiding degradations of the channel that have a catastrophic effect on the transmission quality is "multiresolution modulation." This generates a hierarchical protection scheme by using a signal constellation consisting of subconstellations ("clusters") of points spaced at different distances. The minimum distance between two clusters is higher than the minimum distance within a cluster. The most important bits are assigned to the selection of a cluster, and the less important bits to signal points within that cluster. Fig. 6 shows an example of a 16-signal constellation for multiresolution modulation.

Using Multiple Antennas

As mentioned before, multiple receive antennas can be used as an alternative to coding, or in conjunction with it, to provide diversity. Recent work (see, e.g., [31, 32]) has explored the ultimate performance limits in a fading environment of systems in which multiple antennas are used at both transmitter and receiver side.

It has been shown that, in a system with t transmit and r receive antennas and a slow fading channel modeled by an $r \times t$ matrix with random i.i.d. complex Gaussian entries (the "independent Rayleigh fading" assumption), the average channel capacity with perfect CSI at the receiver is about $m \triangleq \min(t, r)$ times larger than that of a single-antenna system for the same transmitted power and bandwidth. The capacity increases by about m bit/s/Hz for every 3 dB increase in signal-to-noise ratio (SNR), and can reach levels that may not be achieved in any other way with current technology. A further performance improvement can be achieved under the assumption that CSI is available at the transmitter as well. Obtaining transmitter CSI from multiple transmitting antennas is particularly challenging, because the transmitter should achieve instantaneous information about the fading channel. On the other hand, if transmit CSI is missing, the C/M scheme employed should guarantee good performance with the majority of possible channel realizations.

Codes specifically designed for a multiple-transmit-antenna system use degrees of freedom in both space and time, and are usually called *space-time codes* [33, 34]. The symbols in a code word are distributed across transmit antennas and time. By introducing correlation among signals transmitted by different antennas as well as transmitted at different times, a coding gain is obtained without any sacrifice in bandwidth and with a relatively simple receiver structure. Space-time codes are already finding their way into modern wireless system standards.

Conclusions

We have examined some of the tradeoffs involved at the physical layer in digital transmission when the availability of a limited-energy source or a limited bandwidth is a constraint. Specifically, we have described the choice of coding/modulation schemes. Since the optimization of the latter is critically dependent on the channel model, we have examined the Gaussian channel (which, although a poor model for many of the wireless channels, has shaped the discipline of coding and modulation) and the wireless channel in some of its realizations. We have seen a paradigm in action here: the interaction among theoretical grounds (as offered by Shannon's information theory), the ever-increasing need for reliable and fast transmission, and the development of low-cost circuitry allowing the implementation of complex algorithms has made it possible to achieve ultimate limits in communication over the Gaussian channel. As prophetically pointed out in [4], as new technical challenges are waiting in the wings, the design of modulation and coding schemes for new mobile communication services and high-density data storage systems promises to keep the C/M research field active well in this century.

References

[1] T. S. Rappaport, *Wireless Communications, Principles and Practice*, Upper Saddle River, NJ: Prentice Hall, 1996.

[2] C. E. Shannon, "A Mathematical Theory of Communication," *Bell Sys. Tech. J.*, vol. 27, July-Oct. 1948, pp. 379–423, 623–56.

[3] A. R. Calderbank, "The Art of Signaling: Fifty Years of Coding Theory," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, Oct. 1998, pp. 2561–95.

[4] D. J. Costello *et al.*, "Applications of Error-Control Coding," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, Oct. 1998, pp. 2531–60.

[5] G. D. Forney, Jr., and G. Ungerboeck, "Modulation and Coding for Linear Gaussian Channels," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, Oct. 1998, pp. 2384–2415.

[6] S. Benedetto and E. Biglieri, *Principles of Digital Transmission with Wireless Applications*, New York: Kluwer/Plenum, 1999.

[7] J. B. Anderson, T. Aulin, and C.-E. W. Sundberg, *Digital Phase Modulation*, New York: Plenum, 1986.

[8] J. B. Anderson and C.-E. W. Sundberg, "Advances in Constant Envelope Coded Modulation," *IEEE Commun. Mag.*, vol. 29, no. 12, Dec. 1991, pp. 36–45.

[9] E. Biglieri *et al.*, *Introduction to Trellis-Coded Modulation with Applications*, New York: MacMillan, 1991.

[10] G. Ungerboeck, "Trellis-coded Modulation with Redundant Signal Sets - Part I: Introduction," *IEEE Commun. Mag.*, vol. 25, no. 2, Feb. 1987, pp. 5–11; "Trellis-coded, Modulation with Redundant Signal Sets – Part II: State of the Art," *Ibidem*, pp. 12–21.

[11] A. J. Viterbi, "Wireless Digital Communication: A View Based on Three Lessons Learned," *IEEE Commun. Mag.*, vol. 29, no. 9, Sept. 1991, pp. 33–36.

[12] C. Berrou and A. Glavieux, "Near Optimum Error Correcting Coding and Decoding: Turbo-Codes," *IEEE Trans. Commun.*, vol. 34, no. 10, Oct. 1996, pp. 1261–71.

[13] G. D. Forney, Jr., "Codes On Graphs: News and Views," *Proc. 2nd Int'l. Symp. Turbo Codes and Related Topics*, Brest, France, Sept. 4–7, 2000, pp. 9–16.

[14] B. J. Frey, *Graphical Models for Machine Learning and Digital Communication*, Cambridge, MA: MIT Press, 1998.

[15] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor Graphs and the Sum-Product Algorithm," *IEEE Trans. Inform. Theory*, vol. 47, no. 2, Feb. 2001, pp. 498–519.

[16] R. G. Gallager, *Low Density Parity Check Codes*, Cambridge, MA: MIT Press, 1963.
[17] S.-Y. Chung *et al.*, "On the Design of Low-Density Parity-Check Codes Within 0.0045 dB of the Shannon Limit," *IEEE Commun. Letters*, vol. 5, no. 2, Feb. 2001, pp. 58–60.
[18] Y. Kou, S. Lin, and M. P. C. Fossorier, "Low-Density Parity-Check Codes Based On Finite Geometries: A Rediscovery and New Results," *IEEE Trans. Inform. Theory*, vol. 47, no. 7, Nov. 2001, pp. 2711–36.
[19] E. Biglieri, J. Proakis, and S. Shamai (Shitz), "Fading Channels: Information-Theoretic and Communication Aspects," *IEEE Trans. Inform. Theory*, vol. 44, no. 6, Oct. 1998, pp. 2619–92.
[20] S. H. Jamali and T. Le-Ngoc, *Coded-Modulation Techniques for Fading Channels*, New York: Kluwer Academic Publishers, 1994.
[21] N. Seshadri and C.-E. W. Sundberg, "Coded Modulations for Fading Channels – An Overview," *European Trans. Telecomm.*, vol. ET-4, no. 3, May–June 1993, pp. 309–24.
[22] J. Ventura-Traveset *et al.*, "Impact of Diversity Reception on Fading Channels with Coded Modulation-Part I: Coherent Detection," *IEEE Trans. Commun.*, vol. 45, no. 5, May 1997, pp. 563–72.
[23] G. Caire, G. Taricco, and E. Biglieri, "Bit-interleaved Coded Modulation," *IEEE Trans. Inform. Theory*, vol. 44, no. 3, May 1998, pp. 927–46.
[24] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, New York: Wiley, 1991.
[25] G. Caire, G. Taricco, and E. Biglieri, "Optimal Power Control for the Fading Channel," *IEEE Trans. Inform. Theory*, vol. 45, no. 5, July 1999, pp. 1468–89.
[26] F. Alesiani *et al.*, "Performance of Adaptive Modulation Techniques in the UMTS System," *Proc. IEEE Globecom 2001*, San Antonio, TX, Nov. 26–28, 2001, pp. 1272–76.
[27] F. Burkert *et al.*, "'Turbo' Decoding with Unequal Error Protection Applied to GSM Speech Coding," *Proc. IEEE Globecom 1996*, London, UK, 18–22 Nov. 1996, pp. 2044–48.
[28] G. Caire and E. Biglieri, "Parallel Concatenated Codes with Unequal Error Protection," *IEEE Trans. Commun.*, vol. 46, no. 5, May 1998, pp. 565–67.
[29] G. Caire and G. Lechner, "Turbo Codes with Unequal Error Protection," *IEE Electronics Letters*, vol. 32, no. 7, 28 Mar. 1996,pp. 629–31.
[30] K. Fazel, "Matched Combined Channel Coding and Modulation to the Hierarchical TV Source Coding Scheme," R. De Gaudenzi and M. Luise (Editors), *Audio and Video Digital Radio Broadcasting Systems and Techniques*, pp. 265, Amsterdam, The Netherlands: Elsevier Science BV, 1994.
[31] G. J. Foschini, "Layered Space-Time Architecture for Wireless Communication in a Fading Environment When Using Multi-element Antennas," *Bell Labs Tech. J.*, vol. 1, no. 2, Autumn 1996, pp. 41–59.
[32] E. Telatar, "Capacity of Multi-antenna Gaussian Channels," *European Trans. Telecomm.*, vol. 10, no. 6, Nov./Dec. 1999, pp. 585–95.
[33] A. F. Naguib, N. Seshadri, and A. R. Calderbank, "Increasing Data Rate over Wireless Channels," *IEEE Signal Processing Mag.*, vol. 17, no. 3, May 2000, pp. 49–61.
[34] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time Codes for High Data Rate Wireless Communication: Performance Criterion and Code Construction," *IEEE Trans. Inform. Theory*, vol. 44, no. 2, Mar. 1998, pp. 744–65.

BIOGRAPHY

EZIO BIGLIERI (biglieri@polito.it) received his training in electrical engineering from Politecnico di Torino (Italy), where he received his Dr. Engr. degree in 1967. From 1968 to 1975 he was with Politecnico di Torino, first as a research engineer, then as an associate professor. In 1975 he became a professor of electrical engineering at the University of Napoli (Italy). In 1977 he returned to Politecnico di Torino as a professor in the department of electrical engineering. From 1987 to 1990 he was a professor of electrical engineering at the University of California, Los Angeles. Since 1990 he has again been a professor with Politecnico di Torino. He was elected three times to the Board of Governors of the IEEE Information Theory Society, and in 1999 he was the President of the Society. He is a distinguished lecturer for the IEEE Information Theory Society and the IEEE Communications Society. His honors include the IEEE Donald G. Fink Prize Paper Award (2000), the IEEE Third-Millennium Medal for outstanding contributions to the Information Theory area of technology (2000), and the IEEE Communications Society Edwin Howard Armstrong Achievement Award (2001). He is a Fellow of the IEEE.