Voice-band Modem: A Device to Transmit Data over Telephone Networks

2. Advanced Ideas which made High Data Rates Possible

V Umapathi Reddy

V U Reddy is with the Electrical Communication Engineering Department, Indian Institute of Science. His research areas are adaptive signal processing, multirate filtering and wavelets, and multicarrier communication.

Over the last 40 years, there has been continuous evolution in the design of voice-band modems – starting at a data rate of 300 bits per second in the late 1950s, a rate of 33,600 bits per second has been achieved in 1995. Realising such high data rates over the voice band of 3400 Hz is a remarkable feat, made possible by combining sophisticated techniques from three disciplines, communication theory, signal processing and information theory. In this article, we briefly describe certain advanced ideas, which led to data rates very close to the channel capacity limit, established by Shannon.

Introduction

In Part 1[1], we gave a brief introduction to voice-band modems and presented basic principles of data transmission. We pointed out that V.34 modems send nearly 10 bps (bits per second) per Hz which is very close to the channel capacity limit, often called the Shannon bound after the founder of information theory. In this article, we introduce the advanced ideas which made possible such high data rates.

Channel Capacity

Given a channel (a communication medium) of bandwidth W Hz, one would like to know the maximum data rate (in bps) the channel can support. Shannon's capacity formula for an ideal bandlimited, additive white Gaussian noise channel is

[1] Part 1. Basic Principles of Data Transmission, *Resonance*, Vol.6, No.5, 2001.

$$C = W\log_2\left(1 + \frac{P_{\mathrm{av}}}{WN_0}\right) \text{ bps}, \qquad (1)$$

where W is the channel bandwidth, P_{av} is the average transmitted signal power and N_0 is the noise power spectral density (in Watts/Hz). When we say that the channel is ideal, its frequency transfer function (same as the frequency response), $H(f)$, is given by

$$H(f) = \begin{cases} 1 & |f| \le W \\ 0 & |f| > W. \end{cases} \qquad (2)$$

Additive white Gaussian noise (AWGN) channel means the noise added by the channel is white and Gaussian. A white Gaussian noise is the noise with uniform power distribution over all frequencies, and the samples of noise are Gaussian random variables. The ratio $\frac{P_{\mathrm{av}}}{WN_0}$ is referred to as the signal-to-noise ratio (SNR).

Though Shannon has established the capacity limit, he has not indicated the coding/modulation scheme which yields the bit rate equal to the capacity.

Consider the capacity of an ideal channel given by (1). Though Shannon has established the capacity limit, he has not indicated the coding/modulation scheme which yields the bit rate equal to the capacity. For a given SNR, a practical modulation (i.e., the modulation used in practice) yields bit rates less than the value given by the capacity formula. In other words, with a practical modulation, one needs more SNR (i.e., more power in the signal) than what is specified by the formula to achieve the rate equal to the capacity at an acceptable bit error rate (i.e., fraction of the transmitted bits that are received incorrectly). This additional SNR is the so-called SNR gap, which may also be viewed as the SNR penalty that we have to pay with a practical modulation scheme. *Figure* 1 provides comparison of the bit rates that are possible with several modulation schemes on an ideal bandlimited AWGN channel at 10^{-5} symbol error probability (i.e., probability that a transmitted symbol is received incorrectly), and the channel capacity limit. Here, the channel capacity limit is given in bps/Hz as a

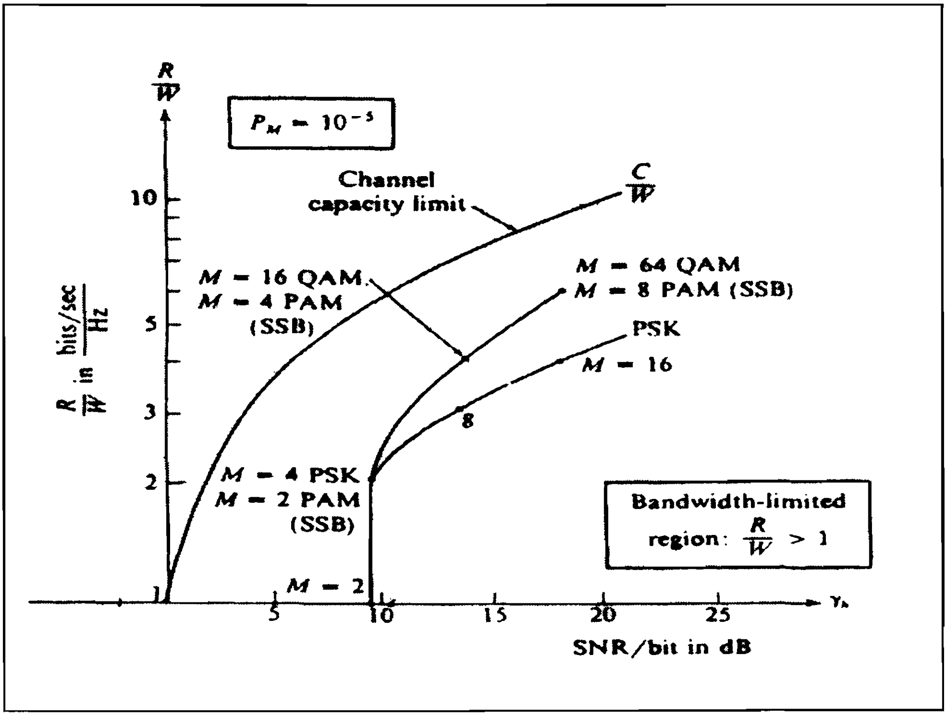


Figure 1. Comparison of several modulation methods (an ideal bandlimited AWGN channel with an ideal receiver is assumed for these plots)-R denotes the transmission rate, in bits/second, a modulation scheme yields at 10^{-5} symbol error probability, P_M. (From J G Proakis, Digital Communications, McGraw-Hill, 1995.)

function of SNR per bit. The SNR per bit can be readily obtained from (1). For example, at 30 dB SNR, C/W is approximately 10 bps/Hz. A 30 dB SNR spread over 10 bits gives 20 dB SNR per bit ($10\log_{10}(10^3/10)$), and this is what we obtain from *Figure* 1 for 10 bps/Hz.

For a given bandwidth, QAM (quadrature amplitude modulation), PSK (phase-shift-keying) and PAM(SSB) (single-sideband pulse amplitude modulation) yield the same number of bps/Hz. However, for more than 2 bps/Hz, the PSK needs more SNR per bit than the QAM (and PAM(SSB))[2]. For QAM, the SNR gap is nearly 9 dB at 10^{-5} symbol error probability, i.e., we need to increase the SNR per bit by about 9 dB to reach the capacity limit.

When we say that a modem is transmitting at a rate very close to the channel capacity, we mean that the underlying SNR at which the modem is yielding this rate is very close to the value specified by the capacity

[2] Though QAM and PAM(SSB) provide the same SNR efficiency, QAM is widely used in practice.

formula. In other words, the modem is operating at nearly zero SNR gap or zero SNR penalty.

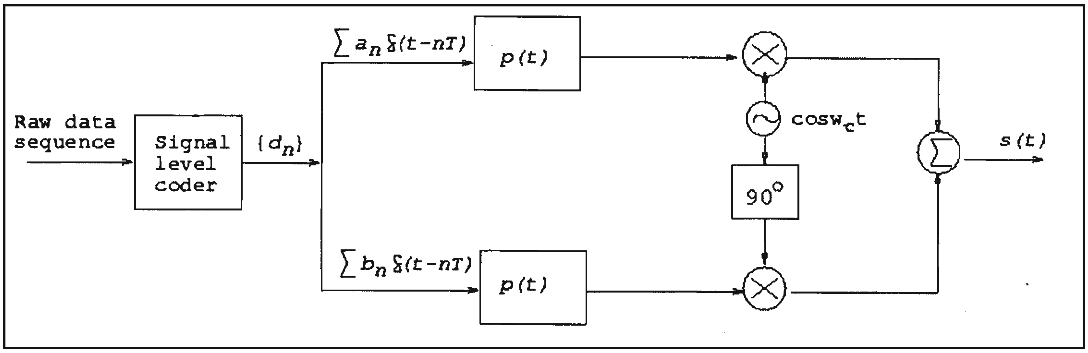
Can we reduce the SNR gap? The answer is 'yes', and this has been made possible by combining advanced ideas from communication theory, information theory and signal processing. We will briefly introduce some of these ideas which lead to significant reduction in the SNR gap. Since it is difficult to associate each idea with a particular discipline, we will not categorize them under the above mentioned areas. To keep our exposition simple, we assume ideal bandlimited AWGN channel with ideal receiver. But first, we describe the PSK and QAM.

SNR gap has been reduced by combining advanced ideas from communication theory, information theory and signal processing.

Two-Dimensional Modulation

Digital PAM is a one-dimensional modulation where the symbols are one-dimensional (see Part 1)[1], i.e., each symbol takes a real value. On the other hand, in two-dimensional modulation, the signal points are two-dimensional, i.e., each symbol is a vector with two components where each component assumes a real value. Alternatively, a two-dimensional symbol can be viewed as a complex number. Digital phase modulation, which is usually called phase-shift keying (PSK), and QAM are two-dimensional modulations. In PSK and QAM, the data bits are carried by in-phase and quadrature carriers. *Figure* 2 shows a basic two-dimensional modulator where the encoded symbol sequence is viewed as a sequence of complex numbers.

Figure 2. A basic two-dimensional modulator.



Let $d_n = a_n + jb_n$. Then, the modulated signal can be expressed as

$$s(t) = \sum_n a_n p(t-nT)\cos\omega_c t - \sum_n b_n p(t-nT)\sin\omega_c t \quad (3)$$

where $p(t)$ is the transmitter pulse shaping filter's impulse response (see Part 1[1]), T is the symbol interval, i.e., the time interval between the successive symbols and $\delta(t)$ is the Dirac delta function. $\{a_n\}$ and $\{b_n\}$ denote two real amplitude level sequences impressed on in-phase and quadrature carriers, respectively. For example, in 16-point QAM (see *Figure* 3), each of $\{a_n\}$ and $\{b_n\}$ takes $\sqrt{16}$ possible amplitude levels. In general, if M denotes the number of signal points with $M = M_1M_2$ where $M_1 = 2^{m_1}$ and $M_2 = 2^{m_2}$, then $\{a_n\}$ and $\{b_n\}$ take M_1 and M_2 possible amplitude levels, respectively. The QAM signal can thus be viewed as two PAM signals[3] modulated by in-phase and quadrature carriers, $\cos\omega_c t$ and $\sin\omega_c t$. Each QAM symbol d_n carries $\log_2 M$ data bits. For M=16, each block of 4 data bits is mapped to one of the signal points (shown as circles) in *Figure* 3.

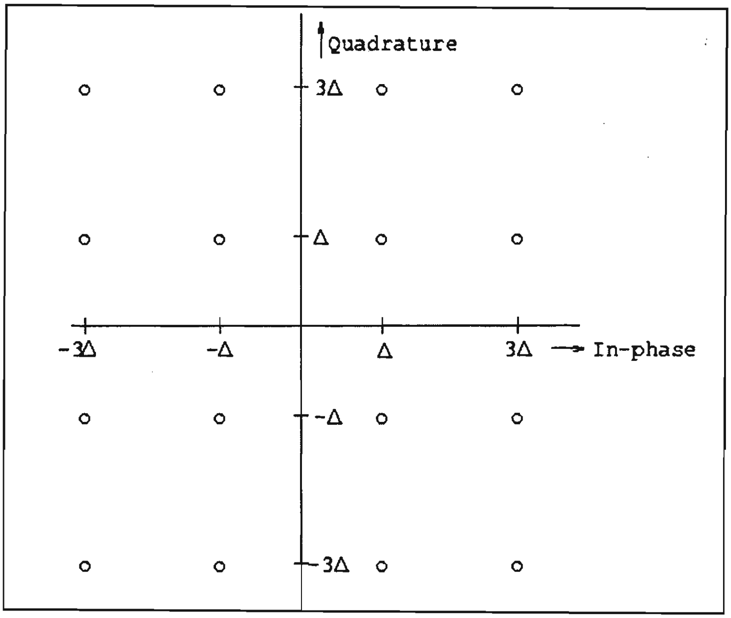


Figure 3. A rectangular 16-point QAM signal constellation.

[3] A more general form of QAM signals is obtained by treating them as combined PAM and PSK modulated signals, and the corresponding signal constellations are called cross-constellations.

In PSK, the modulated signal can be expressed as

$$s_{\mathrm{psk}}(t) = \sum_n \Delta \cos\phi_n p(k - nT) \cos\omega_c t$$

$$- \sum_n \Delta \sin\phi_n p(k - nT) \sin\omega_c t, \qquad (4)$$

where $\phi_n = \frac{2\pi(n-1)}{M}$, $n = 1, 2, \quad , M$ with M denoting the number of PSK signals, and $(\Delta)^2$ denoting the energy in each signal point. *Figure* 4 shows PSK signal constellations for $M = 2$ and M=8. The PSK with $M = 2$ is called binary PSK (BPSK). Thus, in PSK, $\{a_n\}$ and $\{b_n\}$ (see (3)) take real values from the sets $\{\Delta\cos\frac{2\pi(n-1)}{M}\}$ and $\{\Delta\sin\frac{2\pi(n-1)}{M}\}$, respectively, with the condition that $\sqrt{a_n^2 + b_n^2} = \Delta$. As in QAM, M has to be an integer power of 2 and for $M = 2^k$, each block of k data bits is mapped to one of the M PSK signal points.

For a given channel noise, probability of correct decision of a received symbol depends on the Euclidean distance between two closest signal points, which is referred to as the minimum distance of the signal constellation. Let this be denoted by d_{min}. For the signal constellation of *Figure* 3, d_{min}=2Δ, and average signal energy, assuming that all the signal points are transmitted with equal probability, is

$$E_{\mathrm{av}} = \frac{1}{16}[4(2\Delta^2 + 10\Delta^2 + 10\Delta^2 + 18\Delta^2)] = 10\Delta^2. \quad (5)$$

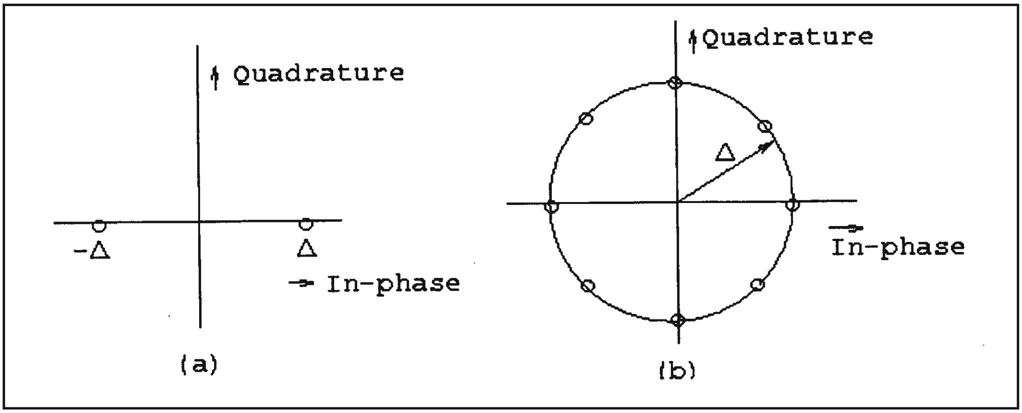


Figure 4. M-point PSK signal constellation.

Obviously, higher the $d_{\min}$, larger is the separation between the signal points, and for the same level of channel noise, the probability of correct decision of a received symbol goes up. But, note that the transmitted signal energy also goes up when $d_{\min}$ is increased.

Our interest here is to find ways to bring down the transmitted signal energy from the value given by the plots of *Figure* 1 for the underlying modulation, keeping the symbol error probability at the pre-specified level (which is 10^{-5} for the plots of *Figure* 1).

Techniques for Reducing the SNR Gap

We will now describe briefly some ideas for reducing the SNR gap. Elaboration of these ideas is beyond the scope of this article. Our intention here is to convey that these advanced ideas yield substantial gain in SNR or reduction in the SNR gap.

Error-Control Coding: An error-control coding (also known as channel coding) adds a controlled amount of redundancy to the raw data bits, i.e., information bits, before transmission, and the decoder at the receiver exploits this redundancy to correct errors introduced during the transmission. For example, a (n, k) block code with $n > k$ (this is also characterized as rate k/n code) maps each block of k information bits to a codeword of n bits [4] where $(n - k)$ bits are the redundant bits. Depending on the (n, k) code, one may correct up to one-, two-, ..., l-bit error patterns in each received codeword. What do we gain by such error-control coding?

Suppose, we want to transmit information bits at R bits/second, at a bit error rate of 10^{-5}. We may choose a symbol interval of T seconds such that $2/T = R$. Then, an uncoded system with 4-QAM (which corresponds to 2 bits/symbol) requires nearly 10 dB SNR per bit. If we now use an error-control coding to encode each block of 2 information bits into a codeword of n bits such that

[4] In general, a block code maps k information symbols to a codeword of n symbols with $n > k$, where the symbol in this context means a group of bits. However, to avoid confusion between this terminology and the symbols we introduced in the context of modulation, we use bits in our presentation here.

the code corrects up to 3-bit error patterns, then the coded bits can be transmitted at a bit error rate of 10^{-2} and after decoding at the receiver, the bit error rate (of information bits) comes back to 10^{-5}. Suppose, we transmit the encoded bits using BPSK. The SNR per bit required for bit error rate of 10^{-2} is about 4 dB. This means, we have reduced the SNR requirement by about 6 dB. But, note that this reduction is at the cost of increase in transmission bandwidth or decrease in information bit rate, as explained next. If we want to keep the symbol rate at $1/T$ we need n symbol intervals to transmit the n bits of codeword which effectively gives a rate of $2/(nT)$ information bits/second – the information bit rate comes down to $(1/n)^{\text{th}}$ of the rate without coding. If we want to ensure the same information bit rate as in the uncoded case, we have to increase the symbol rate by n times implying that the bandwidth has to be increased by n times.

If we want to ensure the same information bit rate as in the uncoded case, we have to increase the symbol rate by *n* times implying that the bandwidth has to be increased by *n* times.

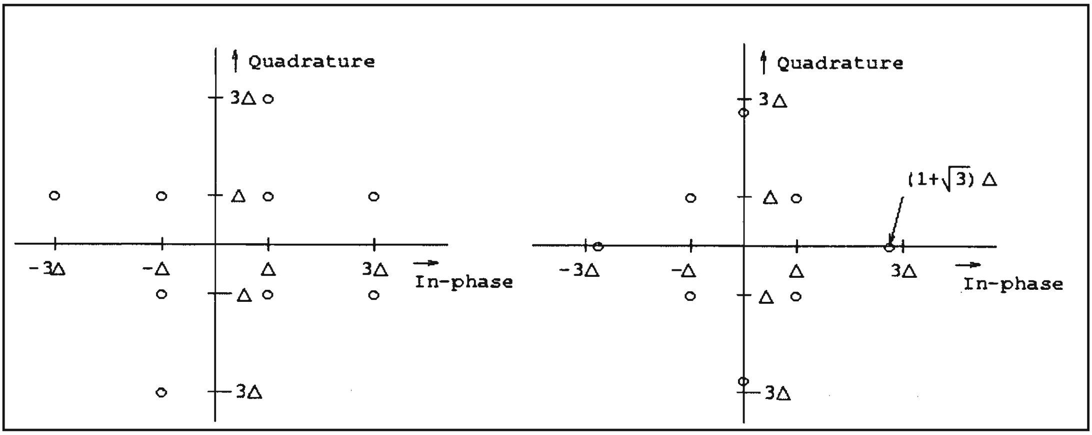
In the above example, we considered a simple $(n, 2)$ block code to illustrate how the error correcting capability can be traded for a reduction in the transmitted SNR. The reduction in data rate (rate of the information bits) or increase in bandwidth, pointed out above, can be brought down substantially if we have a (n, k) code for which k/n is close to unity while at the same time the code can correct large number of errors. Cyclic block codes possess these properties and the popular Reed–Solomon (R–S) codes belong to this class (a (n, k) code is cyclic if a cyclic shift of a codeword is also a codeword). For example, R–S(255, 239) code can correct up to 8-bit error patterns (in fact, it is 8-symbol error patterns where a symbol represents a group of bits[4]) while the rate loss, given by $1 - k/n$, is about 6%. With this code, for the same bandwidth, we obtain a 3 dB gain in SNR, i.e., a reduction of 3 dB in SNR per bit, compared to that without coding, at 10^{-5} bit error rate. Of course, this is at the cost of 6% loss in information bit rate.

Coded modulation refers to combining of coding and modulation. The most popular coding used in coded modulation is the convolutional coding, and the corresponding scheme is known as the trellis-coded modulation

R–S codes are popular for their ability to correct a burst of errors. A convolutional code, on the other hand, is popular for correcting random errors. Since this code can be described by a trellis diagram, it is also called a trellis code. A rate 2/3 binary convolutional code with constraint length 6 (this gives a 64-state trellis) can correct up to 3-bit error patterns.

Coded Modulation: In describing the SNR gain with error-control coding, we considered coding and modulation as separate operations. A more meaningful approach is to combine these two operations. Coded modulation refers to combining of coding and modulation. The most popular coding used in coded modulation is the convolutional coding, and the corresponding scheme is known as the trellis-coded modulation (TCM) (a coded modulation employing a block code, like R–S code, is called block-coded modulation). Decoding of the trellis-coded modulation is performed with Viterbi algorithm. Of course, the computational complexity of a coded modulation and decoding is significantly more than that of uncoded case, and this is the penalty that we have to pay if we want to reduce the SNR gap without increasing the bandwidth or reducing the information bit rate. A gain of about 3 dB at 10^{-6} bit error rate can be obtained by employing TCM schemes of moderate complexity.

Signal Constellation Shaping: Recall that the $d_{\min}$ and average signal energy are related. For a given $d_{\min}$, the average signal energy $(\Delta)^2$ is fixed in the case of PSK. But, this is not so in QAM. This suggests that there is a scope to design a QAM-signal constellation which for a specified $d_{\min}$, has the least average signal energy. *Figure* 5 gives two 8-point QAM signal constellations (circles represent the signal points), where the average signal energy of the constellation on the left is $6(\Delta)^2$ and of the one on the right is $4.73(\Delta)^2$, while the $d_{\min}=2\Delta$ in both the cases. Thus, the constellation on the right gives a gain of approximately 1 dB in SNR. In



other words, with this constellation, the SNR gap can be reduced by about 1 dB.

Figure 5. Two 8-point QAM signal constellations (d_{min} is same in both the cases, but the average signal energies are different).

Thus, by incorporating R–S coding, TCM and constellation shaping, we can obtain about 7 dB gain in SNR, i.e., bring down the SNR gap from 9 dB to about 2 dB. Of course, the overall gain may not be equal to the sum of individual gains when all the schemes are implemented in a communication system.

Signal Processing Advances: While discussing the coded modulation, we pointed out that the computational complexity of the coded modulation and decoding will be much higher than that of the uncoded scheme. This is where the high-speed digital signal processing (DSP) has played a key role – it made possible to implement computationally complex modulators and decoders in real time.

Recall that the channel capacity given by (1) is for an ideal bandlimited AWGN channel. In practice, however, the channels are non-ideal and the channel noise is non-white, asnd one would like to know the maximum bit rates these channels can support. Let $\Phi_{nn}(f)$ and $S(f)$ denote the channel noise and transmitted signal power as a function of frequency (in Watts/Hz), respectively. Then, the capacity of the non-ideal channel, with frequency transfer function $H(f)$ and bandwidth W Hz, is

> The computational complexity of the coded modulation and decoding will be much higher than that of the uncoded scheme.

Summary of the Basic Ideas which made High Data Rates Possible

1. Apply error-control coding, such as Reed–Solomon codes, to the raw data which is to be transmitted.
2. Use coded modulation, such as TCM, to transmit the encoded data.
3. Use two-dimensional modulation, such as QAM (QAM is widely used in modems), for the TCM, with appropriately designed signal constellation.
4. Use equalization and Viterbi decoding.

(due to Shannon)

$$C' = \int_W \log_2 \left[1 + \frac{S(f)|H(f)|^2}{\Phi_{nn}(f)}\right] df \;\; \text{bps.} \qquad (6)$$

Note that (6) reduces to (1) when $H(f)$ is ideal over the bandwidth of W Hz and $\Phi_{nn}(f)= N_0$, with $P_{\text{av}} = S(f)W$

Practical channels introduce intersymbol interference (ISI) and an equalizer needs to be incorporated in the receiver to combat ISI (see Part 1[1] for discussion on ISI and the role of equalizer). The SNR gap in this case can be much more depending on the frequency response of the channel and the effectiveness of the equalization. The recent signal processing advances that made possible high data rates are in the area of equalization. The popular schemes are fractionally-spaced and decision-feedback equalization.

Address for Correspondence
V Umapathi Reddy
Electrical Communication
Engineering
Indian Institute of Science,
Bangalore 560012, India.
Email: vur@ece.iisc.ernet.in

Suggested Reading

[1] J G Proakis, *Digital Communications* (small third edition), McGraw-Hill, 1995.

[2] R D Gitlin, J F Hayes and S B Weinstein, *Digital Communication Principles*, Plenum Press, 1992.