# DATA TRANSMISSION A

For the purposes of protocol design, it can suffice to model a physical channel as a black box with just one interesting feature: it can distort the data that passes through it. We will understand a *channel* to be any medium capable of transferring signals from sender to receiver. The physical realization of the channel can be anything from a twisted pair of copper wires to a satellite link. The only thing we are interested in is the behavior of the channel in so far as it can modify the signals it transfers.

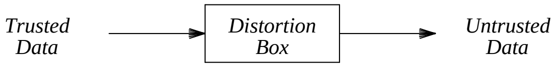

*Figure A.1 — Channel Behavior*

The distortion that is introduced by the channel is typically defined by an error distribution function with known characteristics. For a given medium, the average probability of bit errors can be looked up in a table (see Table A.1 at the end of this Appendix). With this information we can devise an error detection scheme that encodes the trusted data in such a way that its integrity can be checked after it has passed through the distortion box (Chapter 3). Such an error detection scheme intercepts most of the distortions, but is transparent to undistorted data. This transforms the *distortion box* into a *deletion box*.

Deletion errors can be dealt with straightforwardly in a flow control protocol that numbers messages (Chapter 4). Because the error control schemes are based on an estimate of the *average* bit error rate, there is always some probability that distorted data are not intercepted. The purpose of error control is to make the probability of these events acceptably small (Chapter 3).

To gain a better insight into the nature of transmission errors, however, we do take a peek into the distortion box in this appendix. We will see how the behavior of the physical channel is influenced by factors such as

- The data encoding method
- The channel quality (bandwidth, noise level)
- The physical dimensions of the channel

- ○ The signaling speed

With this background, it will be easier to make the right assessment about the protocol requirements for different types of channels. For instance, it would be utter folly to devise an elaborate protocol with forward error control (Chapter 3) on a 10 foot fiber-optic link. Similarly, it would be unwise to attempt to send data at 100 Mbps on a twisted pair cable, no matter what error control scheme is used.

### TYPES OF CHANNELS

In practice, three different types of data transmission channels are used. A *simplex* channel can only be used for data transfer in one direction. The sender typically has a ‘‘modulator’’ to translate binary data into analog signals, and the receiver has a ‘‘demodulator’’ for the reverse translation. A *duplex*, or *full-duplex*, channel can transfer information in both directions simultaneously. Each station has both a ‘‘modulator’’ and a ‘‘demodulator’’ combined into a single instrument called a ‘‘modem.’’ A *half-duplex* channel, finally, can transfer data in both directions, but not simultaneously. The stations have to be switched from sending to receiving or back. The switch usually takes about 200 msec.

### SERIAL AND PARALLEL

Depending on the available hardware, the raw data bits may be transmitted on a physical channel with several bits at a time in parallel, or one bit at a time in series. Parallel transmission is normally only used on short distances, e.g., from a machine to a peripheral. In parallel transmissions one extra line is used to carry a special clock or ‘‘strobe’’ signal that will indicate when precisely the signals on the other lines constitute a valid data word. Due to variations in propagation delays, and the range of possible distortions, it becomes increasingly difficult on longer lines to synchronize the strobe signal and the various bit-streams. For long distances serial transmission is therefore more common.

### ASYNCHRONOUS AND SYNCHRONOUS

On a serial line both the sender and the receiver have a separate clock that sets the transmission rate. The sender uses its clock to *drive* the line (i.e., to transmit the bits), and the receiver uses its clock to *scan* it. In asynchronous transmissions the two clocks need not be in perfect synchrony when no data are transmitted. Data is transmitted in chunks of, for instance, 7 or 8 bits, preceded by a special *start* symbol and followed by a *stop* symbol. The receiver uses the start symbol to synchronize its clock with the sender.

It is sufficient if the two clocks can stay in synchrony for only the 7 or 8 bits that make up a data word. The length of the data word is sometimes called the ‘‘synchronization gap,’’ the period of time that the two clocks must stay in synchrony. The stop symbol is usually either 1.5 or 2 bits long, to allow the receiver to process the data and catch up with the sender and to restore synchrony at the next start symbol. The period of time that passes between the stop symbol and the next start symbol, however, need not be an integral number of bit times.

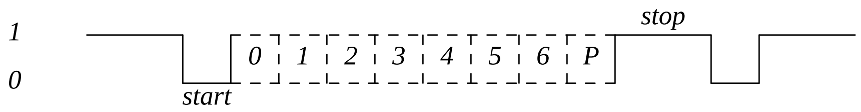

*Figure A.2 — Asynchronous Transmission*

Figure A.2 shows the asynchronous transmission of an 8 bit ASCII character: 7 bits of data followed by a parity bit, labeled *P* (see Chapter 3). The idle state of the line is indicated by a high voltage, a logical one. The one symbol is sometimes called a *mark* and the zero symbol a *space*.

The asynchronous transmission method is self-stabilizing, even when the receiver erroneously starts its clock at a data bit instead of the start symbol. The number of data bits scanned will come out wrong, producing a ‘‘framing error.’’ But since the assumed start of a data word can only move forward in time sooner or later the receiver will re-synchronize.

In synchronous transmission the sender and receiver’s clock must stay in synchrony at all times. When no data are transmitted, the two clocks can be kept synchronous with special ‘‘SYNC’’ characters.

Data can also be encoded in such a way that the signal always has a sufficient number of transitions to keep the receiver’s clock synchronized with the sender’s. With this method the bits are encoded in the *transitions* of a binary signal, rather than in absolute signal levels. The best known method of this type is the Manchester encoding. A one symbol is encoded in the Manchester code by a downward transition (one to zero) and a zero is encoded by an upward (zero to one) transition. This method uses two Baud (signal elements) to encode one bit of information. Figure A.3 illustrates this process.

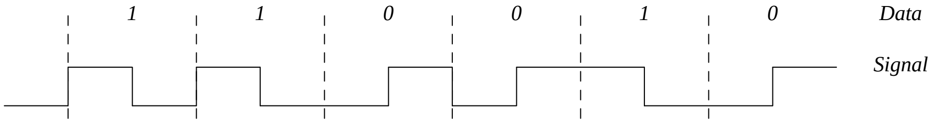

*Figure A.3 — Manchester Encoding*

The Manchester code is called a ‘‘self-clocking code.’’ The receiver’s clock can synchronize on the transition that is guaranteed to occur in the middle of each symbol. The Manchester code has another important property: it creates a ‘‘balanced’’ signal. The average value of the signal over time approaches zero, even if a continuous sequence of equal bits is transmitted. The distortion of a balanced signal on the physical data link is generally smaller than that of an unbalanced signal. The electrical properties of media such as a twisted pair or a coaxial cable are relatively unfavorable for DC (direct current) signals, but more favorable for AC (alternating current), or balanced signals.

Experiment has shown that the maximum signaling speed on a twisted pair cable can be increased by a factor of *ten* if an unbalanced code is replaced with a balanced one.

## SIGNALING SPEED

Signals are normally transmitted on channels as sequences of signal elements of some fixed duration[1]. Each signal element can have a finite value chosen from $V$ distinct signal levels. When $V = 2$, the signal is called a *binary* signal. The duration of each signal determines the signaling speed. This speed is expressed in the unit *Baud*[2] which is defined as the number of signal elements that can be transmitted per second. The signaling speed of a channel, however, is more appropriately measured by the rate at which ''information'' can be transferred. A *bit*[3] is the smallest unit of information. It has one of two possible values. If one signal level is used to encode one symbol, $V$ discrete signal levels trivially allow the encoding of $\log_2 V$ bits of information per signal element, so

$$1 \text{ Baud} = \log_2 V \text{ bits per second (bps)}$$

For binary signals the signaling speed in Baud therefore always equals the signaling speed in bps. Note, however, that in the Manchester code a sequence of two signal levels is used to encode a single symbol. For the Manchester codes, therefore, 2 Baud = 1 bps.

It is understandable that these units are easily confused. Note carefully what the difference is between a signaling speed of, for instance, 1200 Baud, 1200 bps, and 1200 char/sec.

## SIGNAL PROPAGATION

Information can be transferred over many different signal carriers, ranging from copper wires, coaxial cables, and optical fibers, to satellite links. Each channel has a characteristic behavior and requires a specific coding of the information into electrical or electromagnetic signals. Theoretically, the signal propagation time on each channel will set an upper limit to the maximum obtainable signaling speed. In practice, we will see that other factors, such as ''noise'' and bandwidth limitations, have a larger limiting effect. For electromagnetic waves, e.g., satellite links and optical fibers, the signal propagation time is roughly $3 \cdot 10^8$ meter/sec. For electrical signals in cables it is about a factor of ten less.

Consider, in Figure A.4, the projection $p$ of an imaginary dot that moves around the circle. On the right it is shown how the projection on the $y$-axis changes with time when the dot moves with constant velocity: a perfect *sine* curve. One complete

[1] A notable exception is the Morse code. The familiar dot and dash signals are of unequal length.

[2] The word ''Baud'' honors the French telegraph operator Emil Baudot who invented a five bit code for telegraph transmissions in 1874.

[3] The term ''bit'' was coined by J.W. Tukey of AT&T Bell Laboratories as a shorthand for 'binary digit' Shannon [1948].

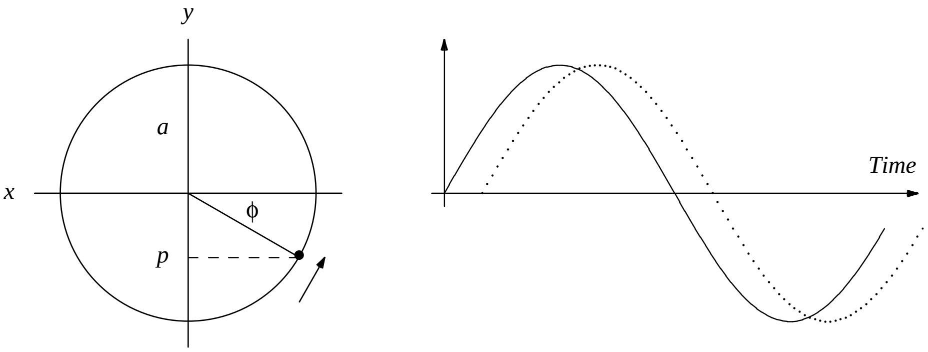

*Figure A.4 — Sine Curve*

traversal of circle produces one ‘‘period’’ or ‘‘cycle’’ of the sine. The maximum ‘‘amplitude’’ of the curve equals the radius of the circle *a*. If the curve is interpreted as an electrical signal, the velocity of the dot determines the signal ‘‘frequency.’’ The unit for measuring frequency is Hertz (Hz). One Hertz equals one cycle per second.

Figure A.4 also shows a dotted curve that would correspond to the projection of a second dot that would follow the first one at a fixed distance, given by the angle φ. The angle is called the ‘‘phase-shift’’ between the first and the second signal. Obviously, the maximum phase-shift will be one complete circle traversal, or 2π radians. Formally, a sine curve is described by

$$a\ sin(2\pi ft\ -\ \phi)$$

where *a* is the amplitude, *f* the frequency, *t* the time, and φ the phase shift.

The sine curve has two properties that make it attractive to theoreticians: it is continuous and it is periodic. The signal in Figure A.5, for instance, is neither, but it does seem to be a more likely representation of a binary bit stream.

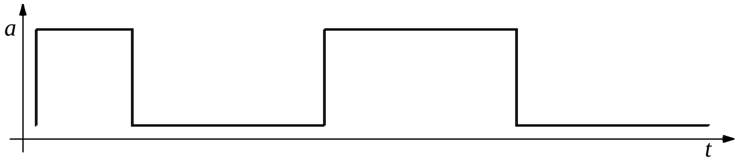

*Figure A.5 — Discrete, Non-periodic Signal*

## FOURIER SERIES

Fortunately, when we study the characteristics of transmission channels we do not have to consider all possible waveforms, like the complicated one in Figure A.5. We can achieve a very good approximation by considering only sine waves. Let us consider an arbitrary periodic signal like the one in Figure A.5. There are two problems with this signal: it is not periodic, and it is not continuous. The first problem is easy to fix, at least for modeling purposes. If we want to describe this fragment of the signal elegantly, we can model it as part of a longer, periodic, signal that is obtained by repeating the signal fragment infinitely often. The second problem is a non-problem: the ideal discontinuous square wave is just an abstraction. In practice, any change in signal levels takes a non-zero amount of time, and no discontinuity exists.

Fourier discovered that every continuous periodic signal can be described by a sum of simple sine waves, each with a frequency that is an integer multiple of a ‘‘base frequency’’ *f*.

$$\sum_{n=1}^{\infty} a_n \sin(2\pi nft - \phi_n)$$

In this formula, $a_n$ is a coefficient that determines the amplitude of the *n*-th frequency component and $\phi_n$ is the corresponding phase shift. For aperiodic signals the discrete series of frequency components changes into a continuum of frequencies, but in principle the same type of analysis can be performed.

Figure A.6 gives an example of the approximation of a discrete square wave by the sum of two sine components.

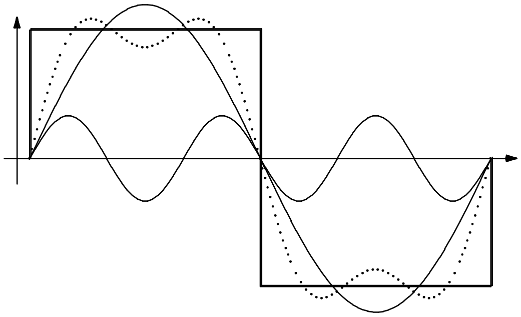

*Figure A.6 — Fourier Series*

The more sine components we add, the better the approximation. The composite signal is again constructed from one base frequency and a range of ‘‘harmonics,’’ of which we have used only the first one. The complete composite is defined by

$$\sum_{n=0}^{\infty} \frac{1}{2n+1} \sin((2n+1)2\pi ft)$$

## BANDWIDTH

If we set out signal frequency along the x-axis and amplitude along the y-axis, we can describe this signal in the ‘‘frequency domain’’ as shown in Figure A.7.

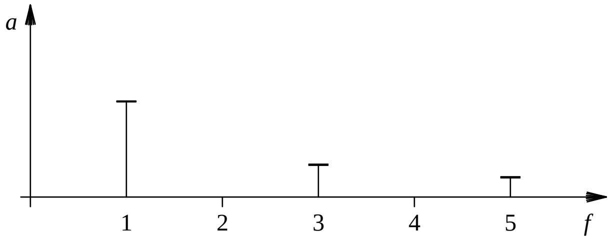

*Figure A.7 — Frequency Domain*

If we increase the signaling speed, the base frequency and all its harmonics will also increase. Unfortunately, a real transmission channel can only transfer a limited range of signal frequencies. A voice-grade telephone line, for instance, can only transfer signals between 300 Hz and 3400 Hz. If we increase the signaling speed, the higher frequency components may fall outside the signaling band and disappear from the signal transmitted. If we decrease the signaling speed the same may happen with the lower frequency components, having an even more detrimental effect on the signal quality.

The ‘‘bandwidth’’ of the channel determines its quality. Bandwidth is defined as the difference between the highest and the lowest frequency that the channel can reliably transfer. The larger the bandwidth, the more information the channel can carry.

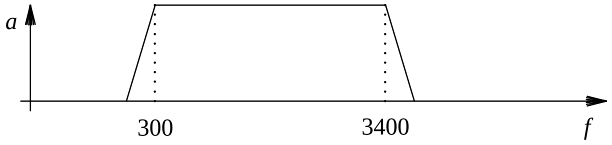

*Figure A.8 — Cutoff Frequencies*

In general, if we transmit a composite signal over a bandwidth-limited channel some frequency components will be attenuated more than others, and some will be lost completely. The result will be a distorted signal. If we try to transmit the binary signal from Figure A.5 directly as an electrical signal, the distorted signal that will arrive at the receiver may well look like the dotted line in Figure A.6.

Figure A.8 shows the bandwidth of a standard switched telephone line. No signal with a frequency less than 300 Hz will get through it, and no signal with a frequency higher than 3400 Hz. The bandwidth is 3.1 kHz. To transfer an arbitrary binary signal across a telephone channel, it must be translated into frequencies that do pass the channel effortlessly. This process, called *modulation*, is discussed below. For now, it should be noticed that every physical transmission medium has a finite bandwidth, and consequently distorts the signals transmitted on it. An ordinary wire pair has a bandwidth of roughly 250 kHz (see Table A.1). The cutoff frequency is roughly at 200 kHz, with the attenuation of signals of a higher frequency rising exponentially. For coaxial cables the high cutoff frequency is about an order of magnitude higher.

The distortion will increase with the signaling speed, simply because the higher data rates cause higher signal frequencies.

A sequence of binary signals will deteriorate from a nice clean square wave to a smooth waveform in which the individual bits may be hard to recognize. The dotted line in Figure A.9 shows the ‘‘decision level’’ below which a signal is classified as a zero. The accuracy of the receiver is severely tested by the signal distortion. A small amount of noise can immediately cause classification errors in the receiver. Note also that the presence of the two one signals surrounding the isolated zero signal in Figure A.9, contribute to the distortion of the zero. This ‘‘inter-symbol interference’’ becomes worse as the signaling speed goes up, and the ‘‘symbols’’ are more closely spaced.

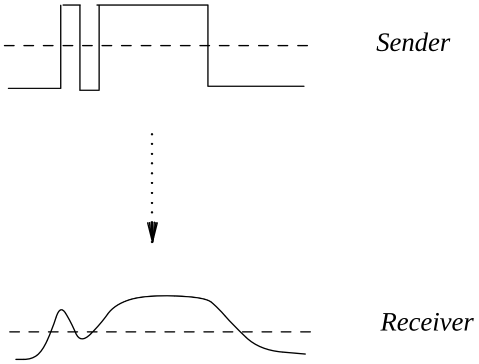

*Figure A.9 — Distorted Signal*

### MODULATION

Modulation is used to adapt the signals to the characteristics of a channel. On a phone line, for instance, we can transmit a binary one as a frequency (a sine) of 1270 Hz, and a zero as 1070 Hz. To make a full-duplex channel, we can choose 2225 Hz and 2025 Hz for the transmission of respectively a one and a zero on the return channel[4]. All these frequencies are within the range that is transmitted with little or no signal attenuation on a phone line (Figure A.8), in order to avoid some of the effects of harmonic distortions on signal quality.

This modulation method is known as *frequency shift keying*, or also simply as *frequency modulation*. As we noted earlier, not too many channels can transmit DC signals conveniently. A balanced, or AC, signal can survive the damage done by the channel much better. If we take a standard sine wave as a basic carrier signal to transmit the data, there are three different ways in which we can change (modulate)

[4] These are in fact the frequencies used on a 300 Baud Bell 108 modem.

that carrier to encode the information. We can use the data signal to vary the carrier's

- ○ Amplitude
- ○ Frequency
- ○ Phase angle

Amplitude modulation for a binary signal would be achieved if we chose two representative amplitudes, e.g., 5 Volts and 10 Volts, to encode binary data. The frequency transmitted is constant, and can be chosen in the middle of the band of frequencies that is accepted by the channel. Any noise on the channel, however, is added to the signal as transmitted and can cause bit errors. Signal attenuation, especially time dependent variations in attenuations can cause extra errors.

Frequency modulation is more robust against noise and direct attenuation of the signal. But now, frequency dependent propagation delays and subtle frequency interference patterns caused by echo and cross talk can cause problems (see below). By using multiple frequencies, however, it is easy to increase the signaling speed in bits per second, for a given baud rate.

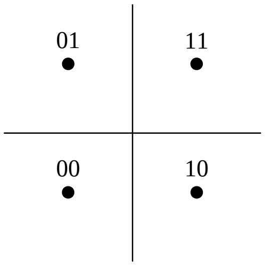

*Figure A.10 — Quadrature Amplitude Modulation*

The third method, using phase shift keying, or phase modulation, is the most complicated one of the three. Every signal element is now encoded by a phase shift from the previous signal element. In these *quadrature amplitude modulation* techniques (Figure A.10) a combination of amplitude and phase modulation is used. A simple version of this uses four different phase shifts: at 90° increments: 45°, 135°, 225°, and 315°. Since this is a one out of four choice, every new symbol now encodes two bits of information, and the data rate in bits per second will be twice the data rate measured in Baud.

### DISTORTION

A signal transmitted on a bandwidth-limited channel incurs a frequency-dependent attenuation. This type of signal distortion is a linear distortion. It can be measured and can, to an extent, be compensated for with special filters that flatten the response curve in the frequency domain.

The signal propagation time can be different for each frequency component in a composite signal. This causes an unintended phase shift between harmonics: the higher frequencies usually travel faster than the others. For a given channel, this phase distortion can also be corrected with special filters.

Transmission channels can also add new waveforms of varying frequencies to a signal. These non-linear distortions can be completely unrelated to the original signal and are much harder to counter.

Signal echoes are an example of non-linear distortions. Wherever there is a sudden change of impedance in the channel, e.g., at the terminals, the signal may bounce back onto the line and travel in the opposite direction, distorting the original signal. A similar type of non-linear distortion is caused by cross-talk. The distortion comes from other channels that are physically close enough to cause shadow signals by electromagnetic induction. In modulated signals the same type of problem can occur as inter-modulation noise.

Still more drastic causes of error are electric spikes and sparks: short, powerful, and unpredictable electric discharges. They can be caused by switches, engines, or simply by spontaneous discharges in the atmosphere. They are hard to avoid, other than by thorough insulation.

### NYQUIST´s SAMPLING THEOREM

The relation between signaling speed and bandwidth was first studied by H. Nyquist in 1924. He showed that if samples are taken from an arbitrary signal that is transmitted across a channel with a bandwidth $B$, the original signal can be completely reconstructed if at least $2B$ samples per second are taken. This *sampling theorem* can be used to determine the maximum signaling speed. $2B$ samples can maximally define $2B$ different signal elements. The maximum signaling speed on a channel with a bandwidth of $B$ Hz is then

$$2B \ \log_2 V \quad \text{bps}$$

According to this estimate, the signaling speed can be increased arbitrarily by increasing the number of signal values $V$. Below we will see that there is yet another factor that limits the signaling speed: noise.

### NOISE

Noise is a fundamental and unavoidable cause of signal distortion. Thermal noise is caused by thermal fluctuations of electrons in conductors. It has no preference for any particular frequency: it is equally present in all. It is therefore sometimes referred to as *white noise*. An important measure for the quality of a signal is the signal-to-noise ratio.

The strength, or power, of a signal is expressed in watts (energy per second). Signal ratios are most conveniently defined in *decibel*. If $P_1$ and $P_2$ give the power of two signals in watts, then

$$10 \log_{10} \frac{P_1}{P_2} \quad \text{dB}$$

is their ratio in decibels. Decibels are used, for instance, to express the signal attenuation on a channel. If $R_1$ is the signal attenuation on one channel in dB, and $R_2$ is the attenuation on another channel, the combined loss if both channels are used in series will simply be $R_1 + R_2$.

### SHANNON-HARTLEY LIMIT

In 1948, Claude E. Shannon studied the precise effect of the signal-to-noise ratio on data transmission. He showed, for instance, that the maximum signaling speed in on a channel with bandwidth *B* and signal-to-noise ratio *S/N*, with *S* and *N* in watts, is

$$C = B \log_2 (1 + \frac{S}{N}) \text{ bps}$$

This result is known as the Shannon-Hartley limit. *C* is called the channel capacity. For a telephone line we have $B = 3100$ Hz and a signal-to-noise ratio of 30 dB (1000:1), giving a maximum signaling speed of 30 Kbit/sec. Above this limit is in general not possible to distinguish the signal transmitted from the background noise: the information content of the signal is too low.

Figure A.11 shows the values that can be calculated from the Shannon-Hartley limit for a telephone line, for signal-to-noise ratios from 1 to 30 dB. The dotted line shows the asymptote $B \log_2 (S/N)$. Signaling speeds above the drawn line cannot be realized, not even with the most clever encoding of data one can imagine.

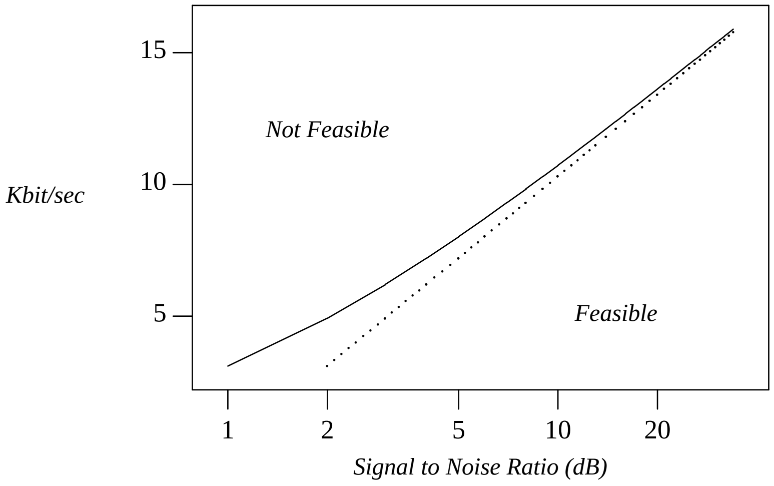

*Figure A.11 — Shannon-Hartley Limit*

For *binary* signals the Nyquist rate of *2B* bps (about 6 Kbit/sec) can be achieved, theoretically, for a signal-to-noise ratio of only 2.5 dB.

To even approach the Shannon-Hartley limit we must make optimal use of statistical information about the data to be transmitted. The transfer of English text, for instance, can be optimized by taking the frequency of occurrence of certain letters and letter combinations into account, assigning the shortest code to the most frequent ones.

The Nyquist rate for a bandwidth-limited channel was $2B$, or, for $B = 3100$ Hz, 6200 Baud. This means that to realize a signaling speed of 30 Kbit/sec we must also use 32 different signal levels: it cannot be realized with a binary signal.

In practice, the signaling speeds that are used are much lower than both the Nyquist and the Shannon-Hartley limit. One reason is that all other causes of distortion (echo, cross-talk, non-linear distortions, and so on) are not taken into account in these results. Furthermore, it is not always worthwhile or possible to include very elaborate coding schemes that can truly optimize the transmission rates. In practice the maximum signaling speed on voice-grade phone lines is not higher than 1200 to 2400 Baud.

The simplest way to obtain a higher signaling speed on a bandwidth-limited channel is, of course, to increase the bandwidth. This is precisely what the phone company does with the new *voice over data* (Co-Lan) services on specially equipped telephone lines, offering both normal phone service and simultaneous duplex data transfers at signaling speeds up to 19.2 Kbit/sec.

### OVERVIEW

A signal that is transmitted on a physical channel can be affected by two main types of distortion:

- ○ The transformation of the original signal
- ○ The addition of information unrelated to the original signal

Examples of the first type of distortion are frequency dependent attenuation, and the loss of high and low frequency signal components due to bandwidth limitations. Examples of the second type of distortion are noise, echoes, crosstalk, and interference patterns caused by non-linear signal distortions.

The effect of the first type of distortion can be reduced by using proper data encoding, modulation, and signal filtering techniques.

Typical data and error rates for three common types of physical media are given in Table A.1.

**Table A.1**

| | Twisted Pair | Coaxial Cable | Optical Fiber |
|---|---|---|---|
| Data Rate in Mbps | 10 | 100 | 1000 |
| Bit Error Rate | $10^{-5}$ | $10^{-6}$ | $10^{-9}$ |
| Bandwidth | 250 kHz | 350 MHz | 1 GHz |

Note, however, that many other factors besides bandwidth affect the data and error rates: the particular method of data encoding used, the length of the data line and hence its susceptibility to noise, echoes, cross-talk, non-linear distortions, etc. For a twisted pair cable, for instance, the quoted rate of 10 Mbps holds for a line length up to about 30 ft, for ‘‘balanced transmissions’’ (for example, with a Manchester encoding). At 300 ft, the data rate drops to 1 Mbps; at 3000 ft it drops to 100 Kbit/sec. Transmission at 1 Mbps on a 3000 ft twisted pair cable, therefore, requires signal regenerators (repeaters).

## BIBLIOGRAPHIC NOTES

A detailed study of line characteristics and data transmission theory is given in Bennet and Davey [1965]. An excellent tutorial on modems, data lines and protocol standards is McNamara [1982]; a well recommended practical reference book. An application oriented treatment of data transmission techniques is presented in Tugal and Tugal [1982]. Other solid treatments of data transmission theory and techniques can be found in Bertsekas and Gallager [1987], [Stallings ’85], and, of course, Tanenbaum [1981, 1988]. A pleasant introduction to some of the details of data transmission can also be found in Byte [1989].