Chapter 8

The information carrying capacity of a channel

8.1 Signals look like noise!

One of the most important practical questions which arises when we are designing and using an information transmission or processing system is, 'What is the *Capacity* of this system? — i.e. How much information can it transmit or process in a given time?' We formed a rough idea of how to answer this question in an earlier chapter. We can now go on to obtain a more well defined answer by deriving *Shannon's Equation.* This equation allows us to precisely determine the information carrying capacity of <u>any</u> signal channel.

Consider a signal which is being efficiently communicated (i.e. no redundancy) in the form of a time-dependent analog voltage, $V\{t\}$. The pattern of voltage variations during a specific time interval, T, allows a receiver to identify which one of a possible set of messages has actually been sent. At any two moments, t_1 and t_2, during a message the voltage will be $V\{t_1\}$ and $V\{t_2\}$.

Using the idea of intersymbol influence we can say that — since there is no redundancy — the values of $V\{t_1\}$ and $V\{t_2\}$ will appear to be independent of one another provided that they're far enough apart (i.e. $|t_1 - t_2| > \frac{1}{2B}$) to be worth sampling separately. In effect, we can't tell what one of the values is just from knowing the other. Of course, for any <u>specific</u> message, both $V\{t_1\}$ and $V\{t_2\}$ are determined in advance by the content of that particular message. But the receiver can't know which of all the possible messages has arrived until it <u>has</u> arrived. If the receiver <u>did</u> know in advance which voltage pattern was to be transmitted then the message itself wouldn't provide any new information! That is because the receiver wouldn't know any more after its arrival than before. This leads us to the remarkable conclusion that a signal which is efficiently communicating information will vary from moment to moment in an unpredictable, apparently random, manner. An efficient signal looks very much like random noise!

This, of course, is why random noise can produce errors in a received message. The statistical properties of an efficiently signalled message are similar to those of random noise. If the signal and noise were obviously different the receiver could easily separate the noise from the signal and avoid making any errors.

To detect and correct errors we therefore have to make the real signal less 'noise-like'. This is what we're doing when we use parity bits to add redundancy to a signal. The redundancy produces predictable relationships between different sections of the signal pattern. Although this reduces the system's information carrying efficiency it helps us distinguish signal details from random noise. Here, however, we're interested in discovering the maximum possible information carrying capacity of a system. So we have to avoid any redundancy and allow the signal to have the 'unpredictable' qualities which make it statistically similar to random noise.

The amount of noise present in a given system can be represented in terms of its mean noise power

$$N = V_N^2 / R \qquad \qquad \text{...(8.1)}$$

where R is the characteristic impedance of the channel or system and V_N is the rms noise voltage. In a similar manner we can represent a typical message in terms of its average signal power

$$S = V_S^2 / R \qquad \qquad \text{...(8.2)}$$

where V_S is the signal's rms voltage.

A real signal must have a finite power. Hence for a given set of possible messages there must be some maximum possible power level. This means that the rms signal voltage is limited to some range. It also means that the instantaneous signal voltage must be limited and can't be beyond some specific range, V_S'. A similar argument must also be true for noise. Since we are assuming that the signal system is efficient we can expect the signal and noise to have similar statistical properties. This implies that if we watched the signal or noise for a long while we'd find that their level fluctuations had the same peak/rms voltage ratio. We can therefore say that, during a typical message, the noise voltage fluctuations will be confined to some range

$$\pm V_N' = \pm \eta V_N \qquad \qquad \text{...(8.3)}$$

where the *form factor*, η, (ratio of peak to rms levels) can be defined from the signal's properties as

62

$$\eta \equiv \frac{V'_S}{V_S} \qquad \qquad \text{...(8.4)}$$

When transmitting signals in the presence of noise we should try to ensure that S is as large as possible so as to minimise the effects of the noise. We can therefore expect that an efficient information transmission system will ensure that, for every typical message, S is almost equal to some maximum value, P_{max}. This implies that in such a system, most messages will have a similar power level. Ideally, every message should have the same, maximum possible, power level. In fact we can turn this argument on its head and say that only messages with mean powers similar to this maximum are 'typical'. Those which have much lower powers are unusual — i.e. rare.

8.2 Shannon's equation

The signal and noise are *Uncorrelated* — that is, they are not related in any way which would let us predict one of them from the other. The total power obtained, P_T, when combining these uncorrelated, apparently randomly varying quantities is given by

$$P_T = S + N \qquad \qquad \text{...(8.5)}$$

i.e. the typical combined rms voltage, V_T, will be such that

$$V_T^2 = V_S^2 + V_N^2 \qquad \qquad \text{...(8.6)}$$

Since the signal and noise are statistically similar their combination will have the same form factor value as the signal or noise taken by itself. We can therefore expect that the combined signal and noise will generally be confined to a voltage range $\pm\eta V_T$.

Consider now dividing this range into 2^b bands of equal size. (i.e. each of these bands will cover $\Delta V = 2\eta V_T / 2^b$.) To provide a different label for each band we require 2^b symbols or numbers. We can then always indicate which ΔV band the voltage level occupies at any moment in terms of a unique b-bit binary number. In effect, this process is another way of describing what happens when we take digital samples with a b-bit analog to digital convertor working over a total range $2V_T$.

There is no real point in choosing a value for b which is so large that ΔV is smaller than $2\eta V_N$. This is because the noise will simply tend to randomise the actual voltage by this amount, making any extra bits meaningless. As a result the maximum number of bits of information we can obtain regarding the level at any moment will given by

$$2^b = \frac{V_T}{V_N} \qquad \qquad \dots(8.7)$$

i.e.

$$2^b = \sqrt{\frac{V_T^2}{V_N^2}} = \sqrt{\left(\frac{V_N^2}{V_N^2}\right) + \left(\frac{V_S^2}{V_N^2}\right)} = \sqrt{1 + (S/N)} \qquad \dots(8.8)$$

which can be rearranged to produce

$$b = \mathrm{Log}_2\left\{\left(1 + \frac{S}{N}\right)^{1/2}\right\} \qquad \qquad \dots(8.9)$$

If we make *M, b*-bit measurements of the level in a time, *T*, then the total number of bits of information collected will be

$$H = Mb = M.\mathrm{Log}_2\left\{\left(1 + \frac{S}{N}\right)^{1/2}\right\} \qquad \qquad \dots(8.10)$$

This means the information transmission *rate, I,* bits per unit time, will be

$$I = \left(\frac{M}{T}\right)\mathrm{Log}_2\left\{\left(1 + \frac{S}{N}\right)^{1/2}\right\} \qquad \qquad \dots(8.11)$$

From the Sampling Theorem we can say that, for a channel of bandwidth, *B*, the highest practical sampling rate, *M / T*, at which we can make independent measurements or samples of a signal will be

$$\frac{M}{T} = 2B \qquad \qquad \dots(8.12)$$

Combining expressions 8.11 and 8.12 we can therefore conclude that the maximum information transmission rate, *C*, will be

$$C = 2B\,\mathrm{Log}_2\left\{\left(1 + \frac{S}{N}\right)^{1/2}\right\} = B\,\mathrm{Log}_2\left\{1 + \frac{S}{N}\right\} \qquad \dots(8.13)$$

This expression represents the maximum possible rate of information transmission through a given channel or system. It provides a mathematical proof of what we deduced in the first few chapters. The maximum rate at which we can transmit information is set by the bandwidth, the signal level, and the noise level. *C* is therefore called the channel's information carrying *Capacity*. Expression 8.13 is called *Shannon's Equation* after the first person to derive it.

8.3 Choosing an efficient transmission system

In many situations we are given a physical channel for information transmission (a set of wires and amplifiers, radio beams, or whatever) and have to decide how we can use it most efficiently. This means we have to

assess how well various information transmission systems would make use of the available channel. To see how this is done we can compare transmitting information in two possible forms — as an analog voltage and a serial binary data stream — and decide which would make the best use of a given channel.

When doing this it should be remembered that there are a large variety of ways in which information can be represented. This comparison only tells us which out of the two we've considered is better. If we really did want to find the 'best possible' we might have to compare quite a few other methods. For the sake of comparison we will assume that the signal power at our disposal is the same regardless of whether we choose a digital or an analog form for the signal. It should be noted, however, that this isn't always the case and that any variations in available signal power with signal form will naturally affect the relative merits of the choices.

Noise may be caused by various physical processes, some of which are under our control to some extent. Here, for simplicity, we will assume that the only significant noise in the channel is due to unavoidable thermal noise. Under these conditions the noise power will be

$$N = kTB \qquad \qquad ...(8.14)$$

where T is the physical temperature of the system, and k is Boltzmann's constant.

Thermal noise has a 'white' spectrum — i.e. the noise power spectral density is the same at all frequencies. Many of the other physical processes which generate noise also exhibit white spectra. As a consequence we can often describe the overall noise level of a real system in terms of a *Noise Temperature, T*, which is linked to the observed total noise by expression 8.14. The concept of a noise temperature is a convenient one and is used in many practical situations. Its important to remember, however, that a noisy system may have a noise temperature of, say, one million Kelvins, yet have a physical temperature of no more than 20 °C! The noise temperature <u>isn't</u> the same thing as the 'real' temperature. A very noisy amplifier doesn't have to glow in the dark or emit X-rays!

Most real signals begin in an analog form so we can start by considering an analog signal which we wish to transmit. The highest frequency component in this signal is at a frequency, W Hz. The Sampling Theorem tells us that we would therefore have to take at least $2W$ samples per second to convert all the signal information into another form. If we

choose to transmit the signal in analog form we can place a low-pass filter in front of the receiver which rejects any frequencies above *W*. This filter will not stop any of the wanted signal from being received, but rejects any noise power at frequencies above *W*. Under these conditions the effective channel bandwidth will be equal to *W* and the received noise power, *N*, will be equal to kTW. Using Shannon's equation we can say that the effective capacity of this analog channel will be

$$C_{analog} \;=\; W \, \mathrm{Log}_2\left\{1 \,+\, \frac{S}{kTW}\right\} \qquad\qquad …(8.15)$$

In order to communicate the same information as a serial string of digital values we have to be able to transmit two samples of *m* bits each during the time required for one cycle at the frequency, *W* — i.e. we have to transmit $2mW$ bits per second. The frequencies present in a digitised version of a signal will depend upon the details of the pattern of '1's and '0's. The highest frequency will, however, be required when we alternate '1's and '0's. When this happens each pair of '1's and '0's will look like the high and low halves of a signal whose frequency is *mW* (<u>not</u> $2mW$). Hence the digital signal will require a channel bandwidth of *mW* to carry information at the same rate as the analog version.

Various misconceptions have arisen around the question of the bandwidth required to send a serial digital signal. The most common of these amongst students (and a few of their teachers!) are:-

 i) *'Since you are sending 2mW bits per second, the required digital bandwidth is 2mW.'*

 ii) *'Since digital signals are like squarewaves, you have to provide enough bandwidth to keep the 'edges square' so you can tell they're bits, not sinewaves.'*

<u>Neither of the above statements are true</u>. The required signal bandwidth is determined by how quickly we have to be able to switch level from '1' to '0' and vice versa. The digital receiver doesn't have to see 'square' signals, all it has to do is decide which of the two possible levels is being presented during the time allotted for any specific bit.

In order to allow all the digital signal into the receiver whilst rejecting 'out of band' noise we must now employ a noise-rejecting filter in front of the receiver which only rejects frequencies above *mW*. The effective capacity of this digital channel will then be

$$C_{digital} \;=\; mW \, \mathrm{Log}_2\left\{1 \,+\, \frac{S}{kTmW}\right\} \qquad\qquad …(8.16)$$

This shows the capacity of the channel at our disposal <u>if</u> we can set the bandwidth to the value required to send the data in digital serial form.

Note that this is <u>not</u> the actual rate at which we wish to send data! The digital data rate is

$$I = 2mW \qquad \qquad ...(8.17)$$

It will only be possible to transmit the data in digital form <u>if</u> we can satisfy two conditions:

  i) The channel must actually be able to transmit frequencies up to mW.

  ii) The capacity of the channel must be greater or equal to I.

The digital form of signal will only communicate information at a higher rate than the analog form if

$$I > C_{analog} \qquad \qquad ...(8.18)$$

so there is no point in digitising the signal for transmission unless this inequality is true. The number of bits per sample, m, must therefore be such that

$$m > \left(\frac{1}{2}\right) \text{Log}_2 \left\{ 1 + \frac{S}{kTW} \right\} \qquad \qquad ...(8.19)$$

Otherwise the precision of the digital samples will be worse than the uncertainty introduced into an analog version of the signal by the channel noise. As a result, if the digital system is to be better than the analog one, the number of bits per sample must satisfy 8.19. (Note that this also means the initial signal has to have a S/N ratio good enough to make it worthwhile taking m bits per sample!)

Unfortunately, we can't just choose a value for m which is as large as we would always wish. This is because the data rate, I, cannot exceed the digital channel capacity, $C_{digital}$. From 8.16 and 8.17 this is equivalent to requiring that

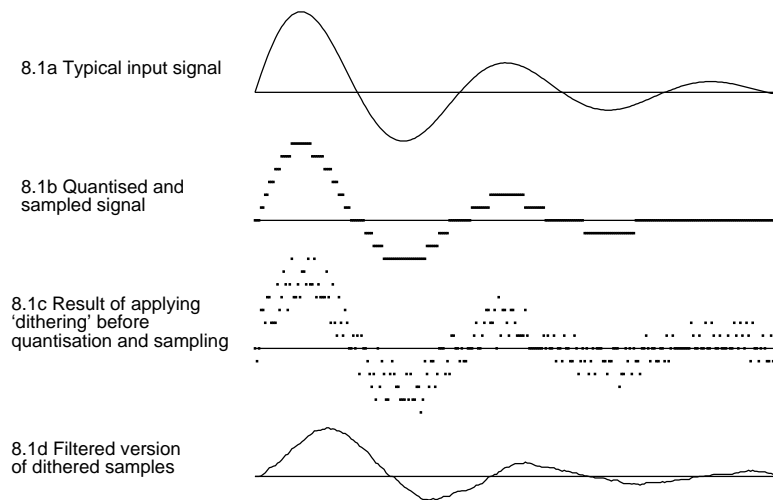$$2mW \leqslant mW \; \text{Log}_2 \left\{ 1 + \frac{S}{kTmW} \right\} \qquad \qquad ...(8.20)$$

i.e.

$$m \leqslant \frac{S}{3kTW} \qquad \qquad ...(8.21)$$

We can therefore conclude that a digitised form of signal will convey more information than an analog form over the available channel <u>if</u> we can choose a value for m which simultaneously satisfies conditions 8.19 and 8.21, and the available channel can carry a bandwidth, mW. If we can't satisfy these requirements the digital signalling system will be poorer than the analog one.

8.4 Noise, quantisation, and dither

An unavoidable feature of digital systems is that there must always be a finite number of bits per sample. This affects the way details of a signal will be transmitted.

8.1a Typical input signal

8.1b Quantised and sampled signal

8.1c Result of applying 'dithering' before quantisation and sampling

8.1d Filtered version of dithered samples

Figure 8.1   The use of 'dithering' to overcome quantisation distortion.

Figure 8.1a represents a typical example of an input analog signal. In this case the signal was obtained from the function $\mathrm{Sin}\{ax\}\,\mathrm{Exp}\{-bx\}$ — i.e. an exponentially decaying sinewave. Figure 8.1b shows the effect of converting this into a stream of 4-bit digital samples and communicating these samples to a receiver which restores the signal into an analog form. Clearly, figures 8.1a and 8.1b are <u>not</u> identical! The received signal (figure 8.1b) has obviously been *Distorted* during transmission and is no longer a precise representation of the input. This distortion arises because the communication system only has $2^4 = 16$ available code symbols or levels to represent the variations of the input signal. The output of the system is said to be *Quantised*. It can only produce one of the sixteen available possible levels at any instant. The difference between adjacent levels is called the *Quantisation Interval*. Any smooth changes in the input become converted into a 'staircase' output whose steps are one quantisation interval high.

This form of distortion is particularly awkward when we are interested in

the small details of a signal. Consider, for example, the low-amplitude fluctuations of the 'tail' of the signal shown in figure 8.1a. These variations are totally absent from the received signal shown in figure 8.1b. This is because the digitising system uses the same symbol for all of the levels of this small tail. As a result we can expect that any details of the signal which involve level changes smaller than a quantisation interval may be entirely lost during transmission.

At first sight these quantisation effects seem unavoidable. We can reduce the severity of the quantisation distortion by increasing the number of bits per sample. In our 4-bit example the quantisation interval is $1/2^4$ th of the total range (6·25%). If were to replace this with a Compact Disc standard system using 16-bit samples the quantisation interval would be reduced to $1/2^{16}$th (0.0015%). This reduces the staircase effect, but doesn't banish it altogether. As a result, small signal details will, it seems, always be lost. Fortunately, there <u>is</u> a way of dealing with this problem. We can add some random noise to the signal <u>before</u> it is sampled. Noise which has been deliberately added in this way to a signal before sampling is called *Dither.*

Figure 8.1c shows the kind of received signal we will obtain if some noise is added to the initial signal before sampling. This noise has the effect of superimposing a random variation onto the staircase distortion. Figure 8.1d shows the effect of passing the output shown in figure 8.1c through a filter which smooths away the higher frequencies. This essentially produces a 'moving average' of the received signal plus noise. This filtering action can be carried out by passing the output from the receiver's digital-to-analog convertor through a low-pass analog filter (e.g. a simple *RC* time constant). Alternatively, filtering can be carried out by performing some equivalent calculations upon the received digital values before reconversion into an analog output. This 'numerical' approach was adopted for the example shown in figure 8.1.

Comparing figures 8.1d and 8.1b we can see that the combination of input dithering and output filtering can remove the quantisation staircase. We may therefore conclude that *Dithering* provides a way to overcome this form of distortion. It can also (as shown) allow the system to communicate signal details such as the small 'tail' of the waveform which are <u>smaller</u> than the quantisation. In reality any input signal will already contain some random noise, however small. In principle therefore we don't need to add any extra noise if, instead, we can employ an analog-to-digital convertor (ADC) which produces enough bits per sample to ensure that the quantisation interval is less than the pre-existing noise level. All that

matters is that the signal presented to the ADC varies randomly by an amount greater than the quantisation interval. In principle, the amount of information communicated is not significantly altered by using dithering. However, the form of information loss changes from a 'hard' staircase distortion loss to a 'gentle' superimposed random noise which is often more acceptable – for example, in audio systems, where the human ear is less annoyed by random noise than periodic distortions. The ability of dithered systems to respond to tiny signals well below the quantisation level is also useful in many circumstances. Hence dither is widely used when signals are digitised.

From a practical point of view using random noise in this way is quite useful. Most of the time engineers and scientists want to reduce the noise level in order to make more accurate measurements. Noise is usually regarded as an enemy by information engineers. However when digitising analog signals we <u>want</u> a given amount of noise to avoid quantisation effects. The noise allows us to detect small signal details by averaging over a number of samples. Without the noise these details would be lost since small changes in the input signal level would leave the output unchanged.

In fact, the use of dither noise in this way is a special case of a more general rule. Consider as an example a situation where you are using a 3-digit *Digital VoltMeter* (DVM) to measure a d.c. voltage. In the absence of any noise you get a steady reading, something like 1·29 V, say. No matter how long you stare at the DVM, the value remains the same. In this situation, if you want a more accurate measurement you may have to get a more expensive DVM which shows more digits! However, if there is a large enough amount of random noise superimposed on the d.c. you'll see the DVM reading vary from time to time. If you now regularly note the DVM reading you'll get some sequence like, 1·29, 1·28, 1·29, 1·27, 1·26, 1·29, etc... Having collected enough measurements you can now add up all the readings and take their average. This can provide a <u>more</u> accurate result than the steady 1·29 V you'd get from a steady level in the absence of any noise.

We'll be looking at the use of *Signal Averaging* in more detail in a later chapter. Here we need only note that, for averaging to work, we must have a random level fluctuation which is at least a little larger than the quantisation interval. In the case of the 3-digit DVM the quantisation level is the smallest voltage change which alters the reading — i.e. 0·01 Volts in this example. In the case of the 4-bit analog to digital/digital to analog system considered earlier it is $1/2^4$ of the total range. Although the

70

details of the two examples differ, the basic usefulness of dither and averaging remains the same.

Summary

You should now know that an efficient (i.e. no redundancy or repetition) signal provides information because its form is unpredictable in advance. This means that its statistical properties are the same as random noise. You should also now know how to use *Shannon's Equation* to determine the information carrying capacity of a channel and decide whether a digital or analog system makes the best use of a given channel. You should now know how quantisation distortion arises. It should also be clear that a properly dithered digital information system can provide an output signal which looks just like an analog 'signal plus noise' output without any signs of quantisation.

Questions

1) Explain what we mean by the *Capacity* of an information carrying channel. A channel carries a signal whose maximum possible peak-to-peak voltage is V_S = 1 V and has a peak-to-peak noise voltage, V_N = 0·001 V. The bandwidth of the channel is B = 10 kHz. Derive *Shannon's Equation* and use it to calculate the value of the channel's capacity. **[199,314 bits/ second.]**

2) Explain what we mean by the *Noise Temperature* of a system. A channel has a bandwidth of 100 kHz and is used to carry a serial digital signal. The signal is produced by an 8-bit analog to digital convertor fed by an analog input. How many samples per second can the system carry? The signal power level is 1 μW. What is the highest noise temperature value which would still let the system carry the digital signal successfully? **[25,000 samples/second. 2·4×10^{11} K.]**

3) Using the same channel as above, what is the highest noise temperature which would be acceptable if the channel were used to carry the information in its original <u>analog</u> form? **[8·8×10^7 K.]**

4) Explain what we mean by the term *Dither* and say how it can be used to overcome *Quantisation Distortion* effects.

71