CONCEPTUAL BACKGROUND TO RADIO

John E. B. Ponsonby

emeritus, Jodrell Bank, UK
johnpon@supanet.com

1 Introduction

There are three types of electrical wires that one sees strung out over the countryside. They are:

1. Power lines. They have rather high voltages and currents that go up and down at 60 Hz. They go on and on and on with terrible monotony. But it is clear what they are conveying. It is power, and one can trace them from a power station to the final customer. We won't have much to say about them.

2. Telephone lines. These too can be traced from one place to another and they too carry power but not very much. What is interesting is that the voltages and currents are not monotonous. Instead they fluctuate. They fluctuate in a way that is predictable in its general character but which is quite unpredictable in detail. We shall try and understand what they are conveying.

3. Antennae. These carry high frequency voltages and currents that also may only be predictable in a general way but these wires don't go anywhere. They just stop in midair. We shall try and understand them as well.

Both telephones and antennae are in a sense the concern of, and are overseen by, the International Telecommunications Union: the ITU. It is through the coordinating work of the ITU that it is possible to make international telephones calls, have the Internet, and operate radio stations without mutual interference.

The ITU is one of the oldest international organizations. It was set up by the International Telegraph Conference held in Paris in 1865, convened by Emperor Napoleon III. Initially it was called the International Telegraphic Union at a time when there were already nearly a million kilometres of telegraph lines installed, but they couldn't cross international borders because of conflicting technical and operational standards. The ITU was set up before the invention of the telephone, before the invention of radio, and before the very word *telecommunication* was coined. Though Russia and Turkey and most of the European countries were represented, I regret to say that neither the USA nor the UK were in at the beginning.

The ITU became a *Specialized Agency* of the United Nations (UN) in 1947. But it has always held the UN somewhat at arm's length, firstly because the UN's predecessor organization, *The League of Nations* had collapsed, and also because there have been Member States of the ITU which have not been members of the UN. I have it by word of mouth in private conversation with the former Secretary-General of the ITU, Pekka Tarjanne, that the ITU does not receive instructions from, or report to, any higher body within the UN system of organizations. So it defines its own terms of reference.

We are going to hear a good deal this week about the ITU and its Recommendations and its Radio Regulations (RRs). Radio waves are no respecters of territorial sovereignty, but cross national borders, and can cause interference in other countries. It is the prime purpose of the ITU-R, the radio branch of the ITU, to manage the use of the Radio Spectrum in such a way that the various applications of radio can coexist and operate without causing mutual interference. We shall be learning how it addresses this mission, to what extent it is effective, and to what extent it is failing in this primary task.

It must never be forgotten that the ITU sees itself primarily as concerned with telecommunications. It is not primarily an international Spectrum Management organization, though in the absence of any other such body it has taken on that role. It views the electromagnetic spectrum as provided by nature for telecommunications and it rather grudgingly concedes that it has other uses, such as remote sensing. Radio Astronomy is a recognized *Radio Service,* one of about 40, but it is only recognized on the basis that it is a sort of "pretend" radio communication service. Most of our troubles stem from this pretence.

2 Information and its measure

The unpredictable fluctuating voltages on those telephone lines and on those antennae carry *information,* or at least they have the potential to do so. When English first acquired the word *inform* (via Old French *enfourmer)* it was used simply to mean to "give form or shape to". So one could inform a lump of clay. However it evolved from its primary notion of "shaping" and acquired the figurative meaning of "forming an idea of something" to "telling or instructing people about something". So *information* is strictly what shapes ideas in the mind, and *information technology,* about which we hear a lot, has the rather sinister meaning of being the technology for shaping ideas in people's minds.

I want to start by sketching out the rudiments of *Information Theory,* the theory of shaping ideas in people's minds. Claude Shannon, who was working for the great Bell Telephone Company, asked himself what is it that all their telephone lines were conveying. He knew it was information of course, the chatter on the wires certainly shaped ideas in minds, but he wanted to give a precise measure to it. He decided that a message conveys information to the extent that it is "News", that a message conveys information according to its surprise value.

A highly probable message tells us little and thus conveys little information: if the voice on the radio says "It will be sunny today with temperatures in the mid 70's", we are not astonished, we haven't learned much, indeed we might have guessed it. So very little information has been conveyed.

On the other hand if the voice says "The President has been shot", we sit up and take notice because it is an unlikely announcement. So more information has been conveyed. But alas it is the sort of thing that befalls presidents.

However if the voice were to say: "The Martians landed this morning near Soccoro, New Mexico", we would be very astonished indeed. That is so unlikely that it really is NEWS, and a substantial amount of information has been conveyed.

We see that the amount of information is not related to the number of words or symbols, but must be some function of the probability of the message.

Suppose message A has probability p_a and conveys information I_a, and message B has probability p_b and conveys information I_b

Then, as it seems reasonable to suppose that information should be additive, so that receipt of both messages conveys information $I_a + I_b$, we look for a function $f(p)$ such that

$$f(p_a) + f(p_b) = f(p_a \times p_b) \qquad (1),$$

since the joint probability of two independent events is the product of their individual probabilities. We don't have to look far. The function with this property is the logarithm. So Shannon defined the information I of a message of probability p as:

$$I = -\log_2(p) \qquad (2).$$

The minus sign is there because $p < 1$ and the log of a number <1 is negative. The unit of information is the bit. Thus a message of probability 1 % conveys 6.644 bits. In this context one is not restricted to an integer number of bits.

At this point I should point out that Shannon was not the sole inventor of *Information Theory*. The same shape formed in the mind of V.A. Kotelnikov in Russia.

Suppose in some communication system there are N possible messages with probabilities p_n. [In the early days of the ITU when the cables only conveyed telegrams, there was a set of four books containing all the 1.9 million words officially recognized by the ITU and those were the only words one was allowed to send!] Then the nth message conveys $-\log_2(p_n)$ bits when it is sent. But it is sent with average frequency (in the statistical sense) p_n. So in the long run that particular message conveys information $-p_n \log_2(p_n)$, (our 1 % message conveys on average 0.06644 bits per message), and the mean information rate of the system is

$$H = -\sum_{n=1}^{n=N} p_n \log_2\left(p_n\right) \text{ bits per message} \qquad (3).$$

Notice that this is only a function of the message probability distribution. The larger N is, the smaller the p_n's become, and the bigger H (which is not upper case h but upper case Greek eta) becomes. H is in fact the Entropy of the message probability distribution. Two examples are shown in Fig. 1. The irregular probability distribution has entropy of 2.76 bits and the Gaussian 4.04 bits. In both cases the horizontal axis has no significance and the same result is obtained whatever the order in which the individual ordinates are plotted.

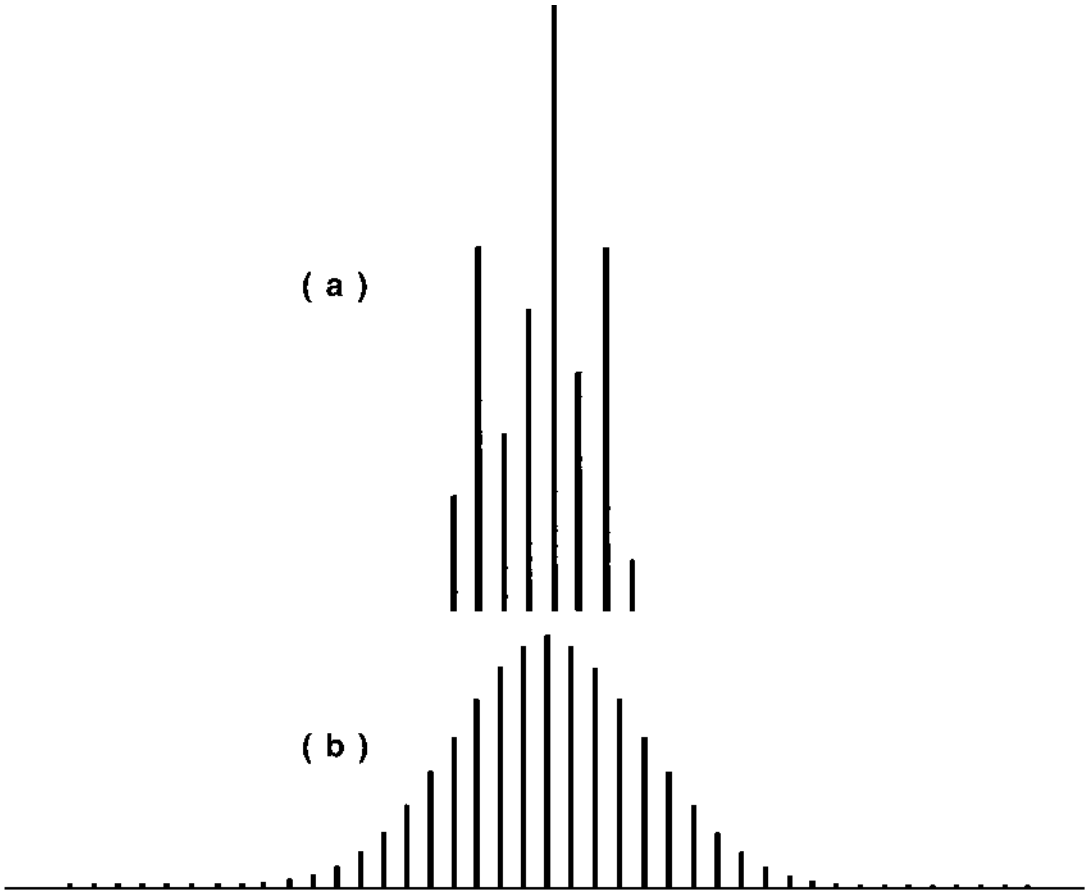

Fig. 1: Two message probability distributions; (a) has entropy 2.76 bits; (b) has entropy 4.04 bits.

I won't develop this further but I do want to impress on you that if the function of a communication system is to convey information, then it must emit unpredictable signals. The more unpredictable they are, the more information is **potentially** conveyed. That doesn't mean that all unpredictable signals contain a lot of information. They may be garbage or noise. But a wholly predictable signal, as on our 60 Hz power line, conveys no information.

An unpredictable message must start at some instant of time. It can have no precursor. That means that the signal symbols or fundamental elements in the coding and modulation system must start from absolutely zero at some instant of time. The waveforms must therefore be discontinuous at some instant.

3 Physical properties of signals

3.1 Degrees of freedom

All the electrical signals with which we have to deal can be viewed on an oscilloscope and seen as wiggly lines of various forms. They are continuous functions of time. Since they can take myriad forms one might at first think that to describe an arbitrary signal would require the specification of an infinite number of parameters. However it is not so. One may see this as follows. Imagine that you record on a length of magnetic tape your favourite piece of music. Then join the two ends of the tape to make an endless loop and then play it. What you will hear is an endless repetition of your piece. You may get fed up with it because after a time it no longer has surprise value. Be that as it may, the audio signal has become a periodic function with repeat time T.

The signal can be represented as a Fourier series of the form

$$v(t) = \sum_{n=1}^{n=N} A_n \cos\left(2\pi n t / T\right) + B_n \sin\left(2\pi n t / T\right) \qquad (4),$$

where the fundamental frequency is $1/T$, and the highest frequency N/T may be determined by your hearing. We call this upper limit the *bandwidth B* of the signal. It is measured in Hz

$$B = N/T \qquad (5).$$

Now we notice that for each harmonic there is an A coefficient and a B coefficient, so the total number of coefficients that have to be written down to completely describe the signal is *2N*. Thus far from needing an infinite number of parameters to describe a continuous signal we see that a signal of *bandwidth B* and of *duration T* can be completely described by *2BT* independent parameters.

A signal of bandwidth *B* and of duration *T* is said to possess *2BT degrees of freedom*. The degrees of freedom may be enumerated, as we have done, in frequency space, or equally well in time. Thus a signal of *bandwidth B* possesses *2B* degrees of freedom per unit time and is completely described if only its values at intervals of 1/(*2B*) are recorded. Given these regularly spaced values, the complete continuously varying original can be recovered. This result is due to Shannon and is called *Shannon's Sampling Theorem,* though the attribution is frequently dropped. *2B* is frequently described as the Nyquist sampling rate for a signal of bandwidth *B*.

The concept of the degrees of freedom of a signal is analogous to the concept of the mechanical degrees of freedom of, for instance, a molecule in a gas. In fact it is more than an analogy, they are the same. So, just as each degree of freedom of a molecule has, on average, energy kT/2, where k is Boltzmann's constant ($1.38\text{x}10^{-23}$ Joule per degree Kelvin) and T is now the absolute temperature (in degrees Kelvin), so the average energy of a thermally generated electrical signal is kT/2 per degree of freedom. Since for bandwidth *B* these come in the time domain at *2B* per second, the mean available noise power of a thermal signal of bandwidth *B* is kT*B* Joule/sec. This is the so-called Johnson noise.

Though the values given to the various degrees of freedom are independent, the values follow a definite distribution law. Johnson noise viewed in the time domain follows a *Gaussian amplitude probability distribution*. If such noise is sampled at the Nyquist rate and a histogram of the values plotted, it will be found to tend to a Gaussian. Likewise artificial band-limited Gaussian noise may be generated by choosing numbers at random from a source following a Gaussian distribution and using them to construct a continuous signal.

3.2 Reconstruction

A short length of artificial band-limited Gaussian noise is shown in Fig. 2. The sample values were drawn from a source of random numbers following the Gaussian amplitude probability distribution also shown. The smooth curve that passes through all the sample values was constructed by convolution with the *Queen of Functions,* the sinc function:

$$\text{sinc}\,(x) \;=\; \frac{\sin(\pi\, x)}{\pi\, x} \qquad (6).$$

This has value 1 at t = 0 and value 0 for all integer x ≠ 0. We set $x = 2B$t. Then since the interval between the samples is 1/Nyquist rate = 1/(2B), it means that the sinc

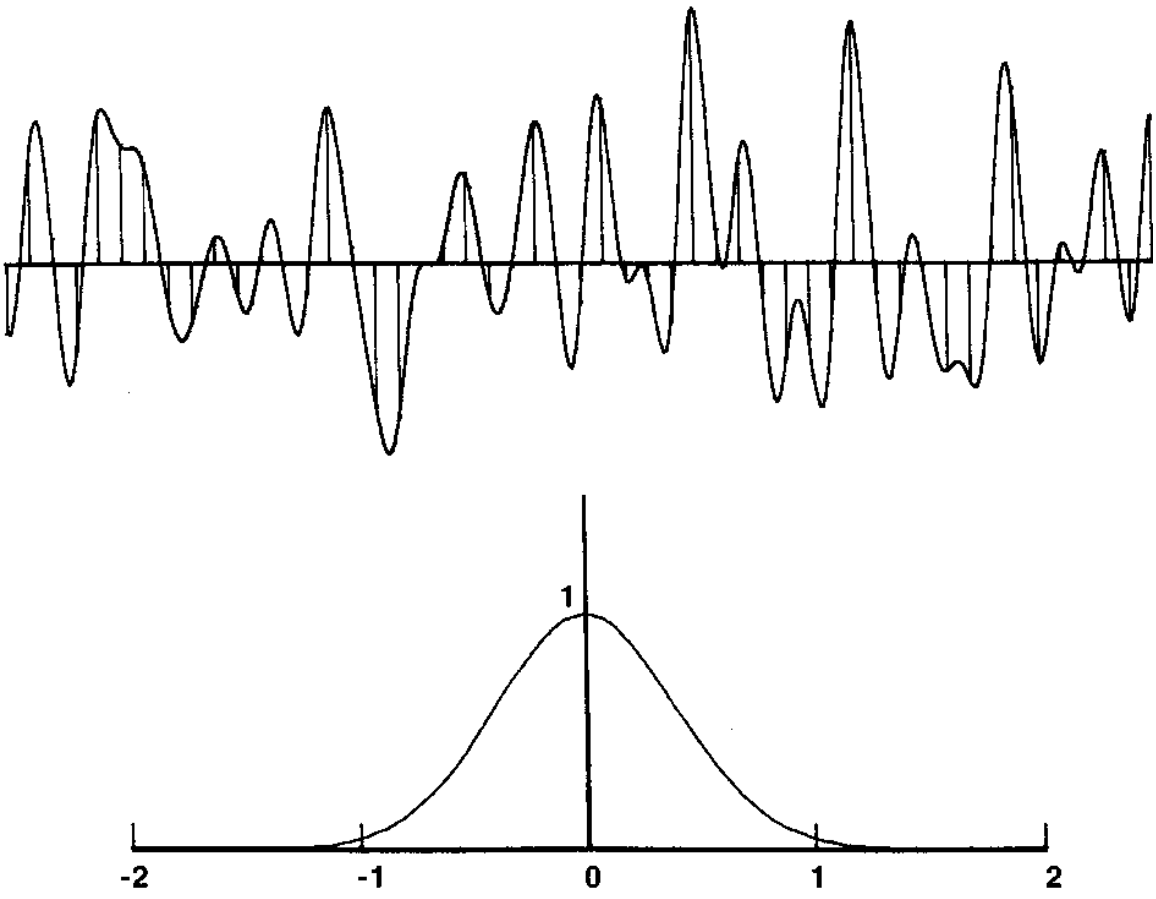

Fig. 2: Sample of band-limited Gaussian noise constructed from the samples shown. These were selected at random from the Gaussian amplitude probability distribution shown below.

function whose peak coincides with one sample has value zero at the positions of every other sample. This convolution amounts to erecting a sinc function at every sample and adding them together, so

$$v(t) = \sum_{all\ m} D_m \operatorname{sinc}(2Bt - m) \qquad (7),$$

where D_m is the m^{th} sample.

This same construction allows a continuous band-limited signal to be reconstructed exactly from its equispaced sample values, provided the sampling rate exceeds or is at least equal to the Nyquist rate.

3.3 Thermal Gaussian noise

Thermal Gaussian noise is all-pervading. The mean power kT*B* is the *available power* at the terminals or port of any lossy, that is to say dissipative, electrical circuit. Obviously if the terminals aren't connected to a load the power doesn't flow. So kT*B* is the maximum power that would flow if the terminals were connected to a matched load. The temperature T is the temperature of the lossy element. If the lossy element is an identifiable resistor, then T is its physical temperature as measured by a thermometer in contact with it.

Of particular interest to us are the terminals of an antenna. This is a lossy structure in so far as the power fed in by a transmitter doesn't come back. An antenna has a *radiation resistance* and it too manifests thermal Gaussian noise. If one imagines an antenna as enclosed in a huge box with the walls at temperature T then, in equilibrium, the box will be filled with Black-Body radiation characteristic of that temperature. The antenna couples to the field and makes power kT*B* available at its terminals. With a narrow-beam antenna, ideally, the antenna temperature is the temperature of the surface at which the beam is directed.

4 Shannon's Channel Capacity Theorem

We have seen how the apparently infinitely parametered variation of a continuous band-limited signal has in fact only a finite number of degrees of freedom. This is well known. Less well known is that a continuous band-limited signal has a finite potential for conveying information. Shannon was able to show that in the presence of additive Gaussian noise of mean power *N* (Watts), a communication channel of bandwidth *B* (Hz) can convey information at the rate *R without any error* according to

$$R = B \log_2 (1 + P / N) \text{ bits /sec} \qquad (8).$$

Here *P* is the signal power. This is *Shannon's Channel Capacity Theorem*. I make no attempt at a potted derivation of this very profound result. It is in several regards analogous to the Second Law of Thermodynamics. It defines the limit of the possible. The proof is a non-constructive "existence proof", so it is not known how to construct a system that achieves this limiting rate of transmission. What is known is that the signal will have the appearance and characteristics of Gaussian noise.

Just as one can use the Second Law of Thermodynamics to define a Carnot efficiency against which the efficiency of a real steam engine can be compared, so the limit defined by the Channel Capacity Theorem allows one to see how efficient a real communication system is in comparison with the theoretical limit. This ability should be a part of every Spectrum Manager's mental tool kit.

Denoting the Signal to Noise ratio $P / N = x$ we can write:

$$R = B \log_2 (1 + x) = B \ln (1 + x) / \ln (2) \qquad (9),$$

the latter form being more convenient. We shall study two cases of particular interest.

4.1 The Bandwidth limited case

This is the case where one has an allocated band and one must stick to it. B is fixed, so the noise power $N = kTB$ defined in §3.1 is fixed and proportional to bandwidth.

Since $\ln (1 + x) \approx x$ for small x, the information rate at low signal-to-noise ratio is proportional to signal power. So

$$R = B x / \ln (2) \qquad (10),$$

and what one wants, channel capacity, is proportional to what one has to pay for, which is signal power. That seems natural and good.

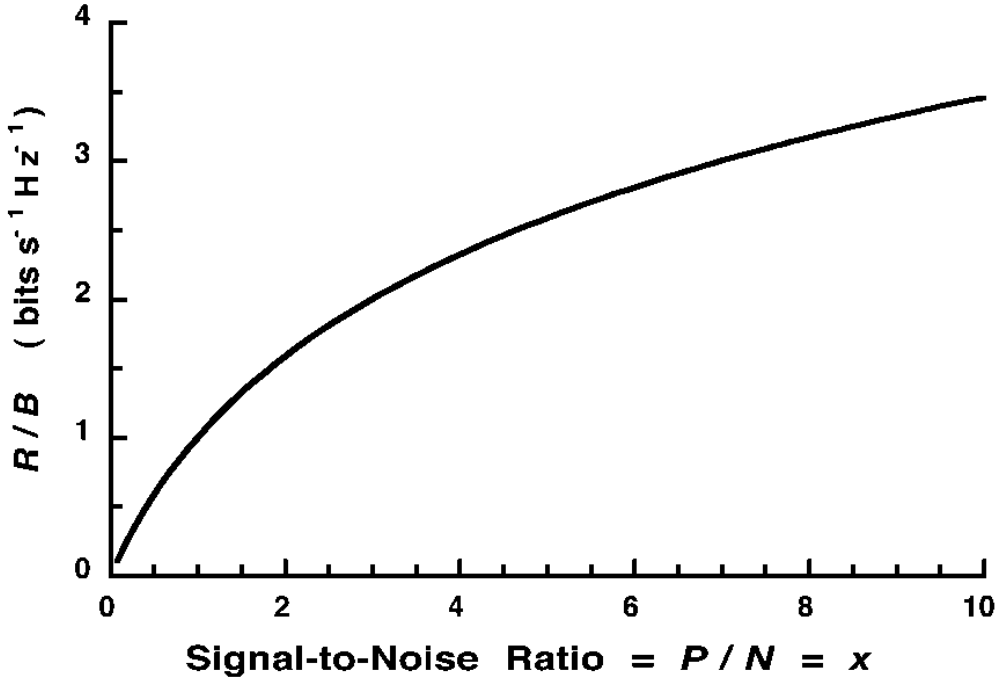

Fig. 3: Channel Capacity R/B (bits s^{-1} Hz^{-1}) versus Signal-to-Noise ratio P / N in bandwidth limited case.

However at high signal-to-noise ratio, when $\ln (1 + x) \approx \ln (x)$

$$R = B \ln(x) / \ln(2) \qquad (11),$$

and one gets ever smaller increments of capacity for equal increases in cost. This is an instance of what economists call the *Law of Diminishing Returns*. It is an unhappy regime to be in.

The general relationship is shown in Fig. 3. At unity signal-to-noise ratio the limiting capacity is 1 bit/sec/Hz. To increase the capacity to 2 bits/sec/Hz requires a three-fold increase of transmitter power. To increase the capacity to 3 bits/sec/Hz requires a 7-fold increase of power. In principle one may send an arbitrarily large number of bits/sec through any finite bandwidth, but one has to pay dearly to do so.

4.2 The Power limited case

We have seen that noise power is generally proportional to the bandwidth. Thus one may write

$$N = \nu B \text{ or } B = P / (\nu x) \qquad (12),$$

where now both P and v are constants, and v is the noise power per unit bandwidth. The limiting information rate is

$$R = \left(\frac{P}{\nu}\right) \frac{1}{x} \frac{\ln(1+x)}{\ln(2)} \text{ bits/sec} \qquad (13).$$

Counter-intuitively this increases as the signal-to-noise ratio decreases and achieves the extreme value of

$$R = \left(\frac{P}{\nu}\right) \frac{1}{\ln(2)} \text{ bits/sec} \qquad (14),$$

when the signal-to-noise ratio is vanishing and the bandwidth tends to infinity! This relationship is shown in Fig. 4. We see that actually the result isn't so alarming as it sounds. The limiting rate is very nearly achieved if the bandwidth is increased far enough to make $P/N \approx 0.1$.

There are many real practical situations in which signal power is limited. One thinks of spacecraft and indeed of mobile phones that have rather small batteries. The Channel Capacity Theorem says that if the power is well used one will find oneself using a modulation scheme which spreads the power rather thinly over a relatively wide band and one will operate at very low signal-to-noise ratio. There are indeed systems that have these characteristics: wideband FM broadcasting and Code Division Multiple Access (CDMA) systems in mobile phones. Perhaps they are on the right lines.

4.3 Concluding remarks

1. It is always correct to aim for the lowest possible noise power by making v as small as possible.

2. Channel capacity always increases with increase of signal power.

3. However if v and P are fixed, the channel capacity is maximized by increasing the bandwidth until the signal-to-noise ratio is much less than unity.

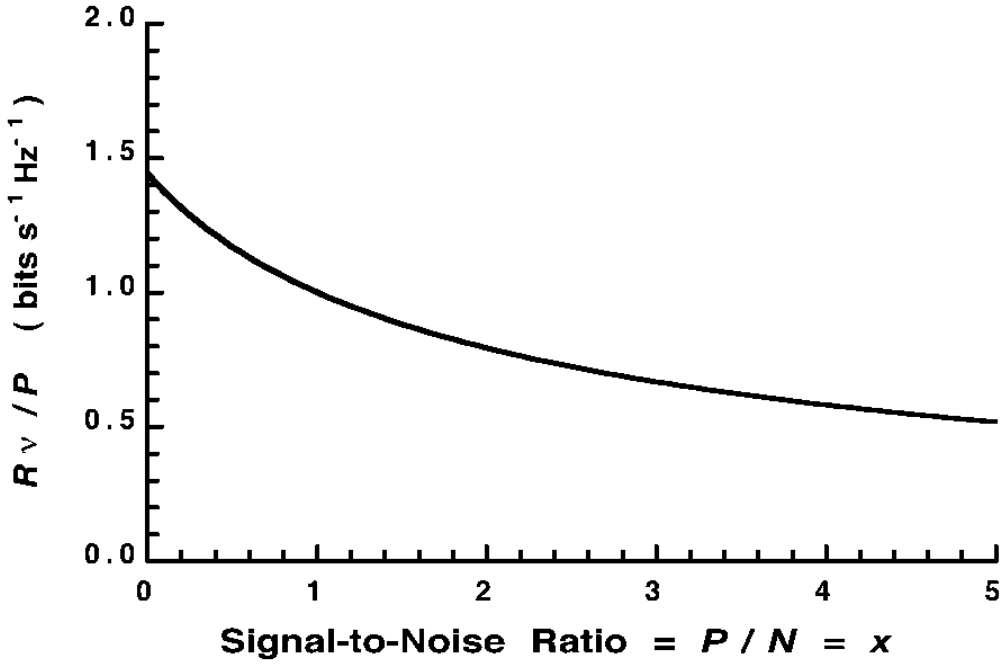

Fig. 4: Channel capacity $R\,\nu\,/\,P$ (bits s^{-1} Hz^{-1}) versus Signal-to-Noise ratio P / N in the Power-limited case. Note that the limiting Channel Capacity as x → 0 is 1.443 bits.

5 Fourier Theory of discontinuous functions

I like to follow Bracewell in defining the Fourier Transform (FT) of a function of time $f(t)$ as $F(s)$, where s is frequency in cycles/unit of time, as

$$F(s) = \int_{-\infty}^{+\infty} f(t) \exp(-i\,2\,\pi\,s\,t)\,dt \qquad (15).$$

I shall need to use two results that follow. The first is the so-called *Shift Theorem*. It states how $F(s)$ is modified if $f(t)$ is shifted along the time axis. If $f(t)$ is retarded by a time T, that is to say shifted to the right on the time axis, then it becomes $f(t - T)$ and its FT becomes $F(s) \exp(-i\,2\,\pi\,s\,T)$. It acquires a linear negative phase slope. The second, which follows from the first, is the *Derivative Theorem*, which I will express inversely in terms of an integral. If $f(t)$ is integrated with respect to t, its FT becomes

$$F(s) / (i\,2\,\pi\,s).$$

Consider the following development. Start with a delta function of time $\delta(t)$. This has value of 0 for all $t \neq 0$ and its integral from $t = -\varepsilon$ to $+\varepsilon$ equals 1

even as $\varepsilon \to 0$. Integrating it with respect to time one obtains the step function shown at the bottom of Fig. 5. Shift this to the left (advance it in time) by amount *T*/2 and shift it to the right (retard it in time) by *T*/2, and subtract the second from the first. We obtain the square pulse or "top hat" function shown one line up. This has width *T*. Repeat the process. The next line up shows the integral of the top hat as a ramp, and the line above again shows the triangular pulse resulting from shifting left and right by *T*/2 and taking the difference. The figure shows the effect of repeating this process twice more.

The process described amounts to repeated convolution by the top hat function

$$\Pi\left(t \,/\, T\right) \;=\; 1 \,/\, T \;\text{ for }\; -T/2 < t < +T/2,$$

and is elsewhere zero. The effect of the successive convolutions is to produce an ever-smoother pulse. The top-hat function has abrupt sides, so it is a discontinuous function. On integration the ramp is continuous but it has discontinuity in its slope or first derivative. The next smoother integral is discontinuous in its second derivative and the top one is only discontinuous in its third derivative.

Now the Fourier Transform of the top hat function is well know to be the sinc function

$$\sin\left(\pi\, s\, T\right) / \,\pi\, s\, T \;,$$

and by the *Convolution Theorem,* which I have not discussed, or from the *Shift Theorem* which I have, we can see that the FT's of the various pulses are

$$\left[\sin\left(\pi\, s\, T\right) / \,\pi\, s\, T\right]^{n} \;,$$

with $n = 1$ for the top hat, $n = 2$ for the triangular pulse, and so on. We see that the envelope of the FT falls off as s^{-n}. So the smoother the pulse the faster the FT falls off. This is a manifestation of a general rule that a function, which is discontinuous in its n^{th} derivative, has an FT with an envelope that falls asymptotically as $s^{-(n+1)}$ in the frequency domain.

If a signal is composed of a string of pulses, as many are, the form of the resultant power spectrum is the square of the magnitude of the FT of one pulse. So if the pulses are discontinuous in their n^{th} derivatives, the resultant power spectrum falls as $s^{-2(n+1)}$. Viewed on log-log scales the spectrum falls as -6 (n + l) dB/octave or -20 (n + l) dB/decade. These results are very germane to the matter of Out-Of-Band emissions (OOBs).

I have plotted the power spectra corresponding to the pulses of Fig. 5 in Fig. 6, but on a dB scale vertically and a linear frequency scale horizontally. The deep nulls occur at frequencies which are multiples of 1/*T*. I have drawn a vertical dashed line at 5/*T* which is 250 % of (2/*T*).

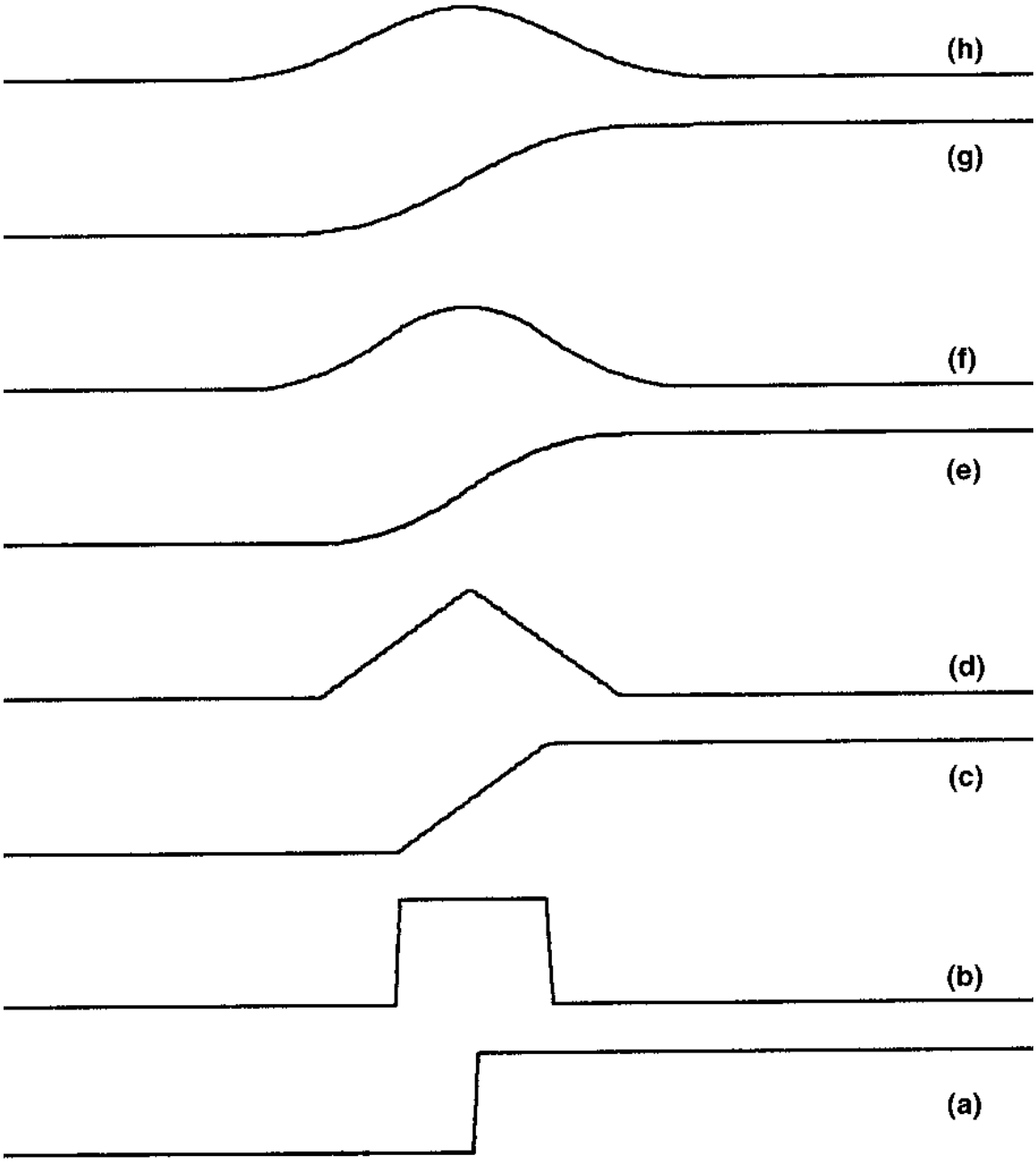

Fig. 5: The successive integration and differencing of a delta function makes successively smoother pulses.

It follows from this result that any signal which is infinitely differentiable will have a power spectrum that falls off faster than any negative power of frequency. The key example is of course the pure sine wave. Its power spectrum is a vanishingly thin delta function. We are back to our boring 60 Hz power line!

A Gaussian shaped pulse is also infinitely differentiable. It has the pleasing property that its FT is also a Gaussian and, indeed, viewed on a log-log scale its power spectrum has no asymptotic rate of fall-off. It falls ever faster as the frequency is increased. But it can't be used for communication because it has a very small but infinite precursor in time. It starts infinitely far back in time. As soon as it is modified so that the precursor is chopped off, one no longer has an infinitely differentiable function and the corresponding power spectrum has an asymptotic rate of fall-off.

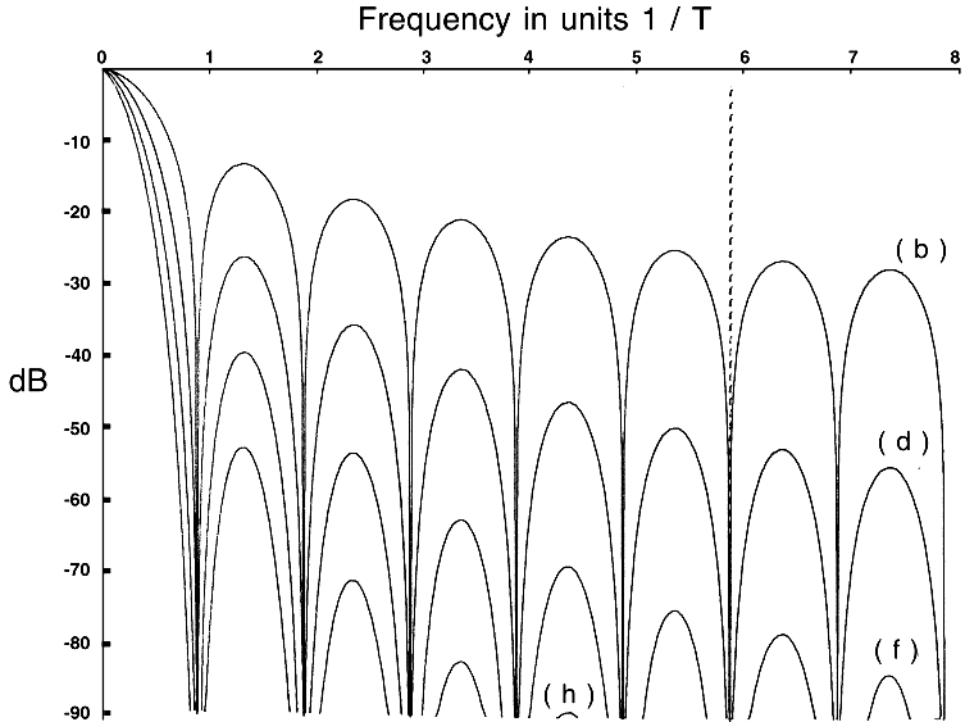

Fig. 6: Power spectra in dB corresponding to the pulse shapes in Fig. 5.

Because an information conveying system must transmit unpredictable messages, the symbols cannot have infinitely long precursors. Thus I maintain all information bearing transmissions must have power spectra that fall asymptotically no faster than some negative power of frequency.

Everything that has been said here about smoothness of pulses and the fall-off of the power spectrum has been treated as if we were concerned with single-sided spectra going down to zero frequency. But everything remains the same if the pulse shapes are the envelopes of a high frequency carrier. Then the rates of fall-off are measured from the carrier frequency.

6 Filters

Presented with a transmitter whose Out-Of-Band (OOBs) are unacceptable, it is natural to suggest that an output filter should be added. If an effective classical electrical filter can be fitted, a filter made up of a number of coupled resonators (Fig. 7), then one must ask what it does to the signal.

Every filter has a frequency response, let us call it $\mathtt{F}\ (\mathtt{s})$, which is in general a *complex* function of frequency. Every frequency component of an applied signal gets changed in amplitude and in phase. There is not much scope in design for independent control of the amplitude and phase responses. In fact, for every given amplitude response there is an inherent minimum lagging phase response. One talks of *minimum-phase networks* and most filters are of this type. The constraint stems from the fact that a filter is a *causal* system, it is not clairvoyant, it cannot possibly respond to an impulse before it occurs. Since in an impulse $\delta\,(t)$, all frequencies are

present with equal amplitude, the *impulse response* $I(t)$ of a filter is simply the Fourier Transform of its *complex* frequency response (see Fig. 8)

$$I(t) = \int_{-\infty}^{+\infty} \mathrm{F}(s) \exp(+i\,2\,\pi\,s\,t)\,ds \qquad (16).$$

Since $I(t) \neq 0$ only for $t > 0$, and recalling the *Shift Theorem*, one is not astonished that there has to be at least a certain minimum negative phase slope associated with $\mathrm{F}(s)$.

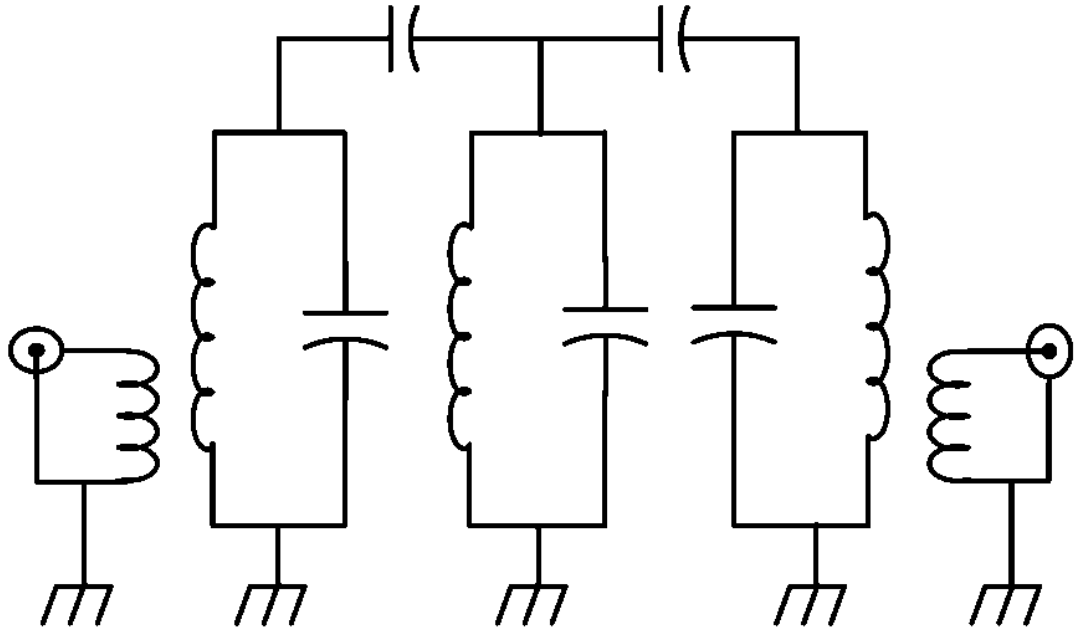

Fig. 7: Typical Band-Pass-Filter.

A filter is certainly a device that changes the amplitude and the phase of every frequency component of the applied signal. And it is easy to fall into the way of thinking that it somehow does a Fourier Analysis of the signal, changes the amplitudes and phase of each component, and then reassembles them to form the output signal. That is quite a task for a few interconnected resonators. One may think that way but it isn't a physically correct description of how a humble filter actually works. This multiplication by the frequency response in the frequency domain is in reality achieved by a convolution in the time domain. How a filter really filters is by convolving the input signal with the filter impulse response,

$$V(t)_{out} = V(t)_{in} * I(t) \qquad (17).$$

The output waveform is the input waveform convolved with the impulse response. Here the * denotes convolution. One may consider the input signal as being subdivided into a succession of elementary contiguous impulses of varying amplitudes. Each one excites the filter's impulse response. The output signal is the superposition of all the elementary impulse responses.

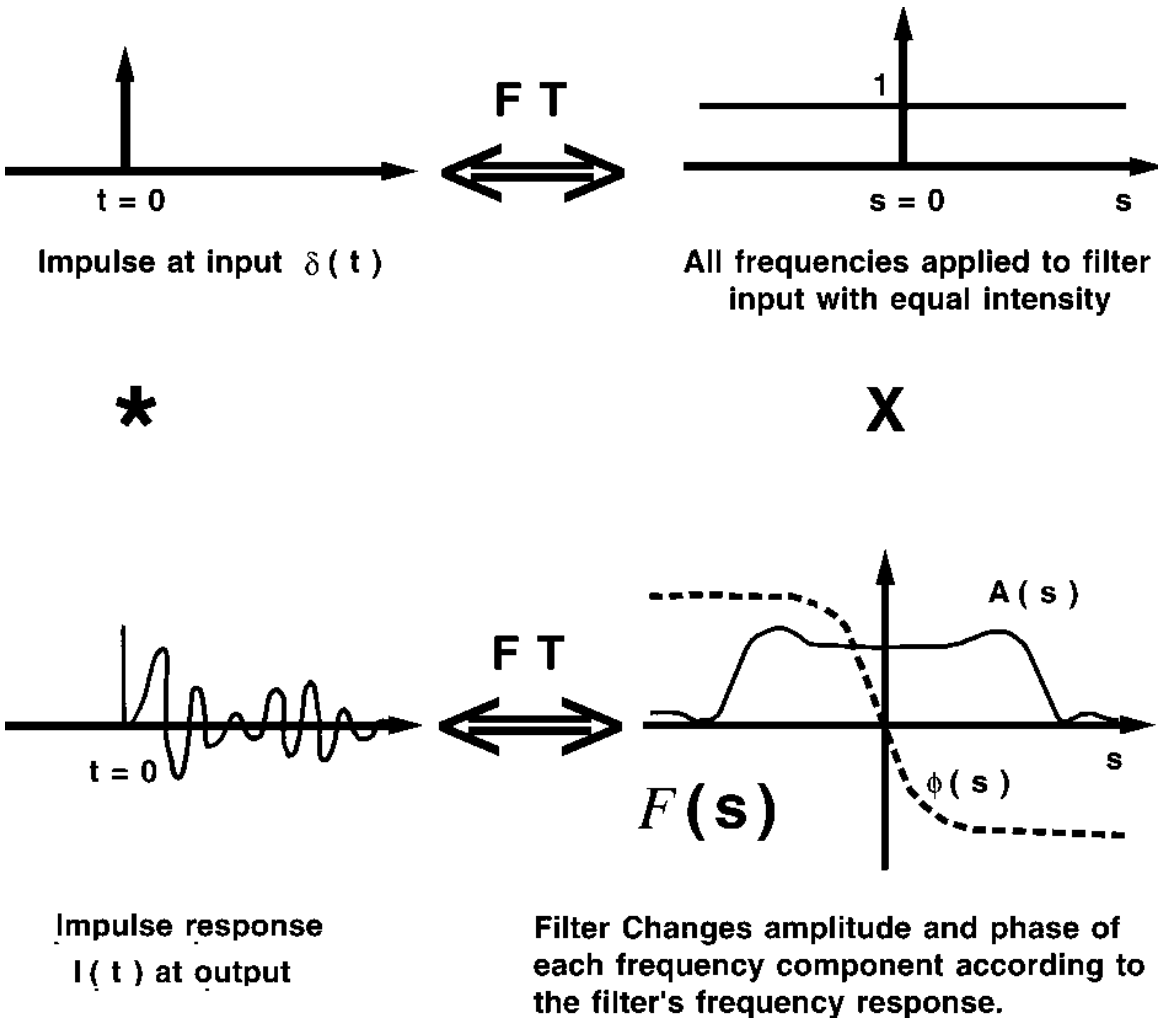

Fig. 8: Fourier Transform of the Impulse and the Impulse-Response of a filter.

6.1 Filtering of constant amplitude signals

We shall come later to various forms of modulation, but at this point we need only say that there are forms of modulation which only alter the phase of the signal and leave the amplitude constant. Frequency modulation (FM), as used for audio broadcasting, is one such form. The modulated signal may be Fourier analyzed into a carrier and a lot of sidebands. It is easy to understand that it requires a very intricate and specific inter-relationship between the amplitudes and phases of all these constant frequency components to make them all add up to form a signal of constant amplitude but varying frequency. The condition is in fact that the autocorrelation function of the FT of the signal should be a delta function

$$\delta(s) = \int F(s_1)\, F^*(s_1 - s)\, ds_1 \qquad (18).$$

Any tampering with the amplitudes and phases of the components will cause departure from this condition and give rise to a signal which is no longer of constant amplitude.

So passing an FM signal through a filter, even one that looks as if its passband is wide enough to pass all the major frequency components, will result in an output signal that is no longer of constant amplitude. The same happens with any constant amplitude signal phase modulated in some way.

A common form of modulation is Binary Phase Shift Keying (BPSK). This is a digital modulation scheme where the 1's and 0's of the data stream are represented by two alternate versions of the RF carrier mutually 180° in phase. An elementary way of modulating such a signal is with a switch as shown on the LHS of Fig. 9. The abrupt phase reversals are discontinuities in the zeroth derivative of the amplitude of the signal and generate sidebands that fall off as s^{-2} in power, where s is now the frequency offset from the carrier. Only the designers of the *GPS* and *GLONASS* systems thought this was an acceptable form of signal to transmit. Generally, some effort is made to reduce the amplitudes of the *unwanted* (an ITU technical term) sidebands. A suitable filter will do it. How it does it is by imposing continuity on a greater number of the signal's derivatives. There is no other way. But modifying the amplitudes and phases of the signal components inevitably causes the signal amplitude to vary. One gets PM to AM conversion: *Phase Modulation* into *Amplitude Modulation.*

This is of no great consequence provided subsequent handling of the signal is entirely linear, as it would be if the filter was between the final power amplifier and the antenna. However, there are good reasons for not wanting to put a narrow band filter there. They have to be made of high-Q resonators which magnify the applied voltages. They are inevitably lossy, and at high power subject to voltage breakdown. High-Q high-power filters are possible, but are very much to be avoided, not least because they will be big, heavy, and expensive.

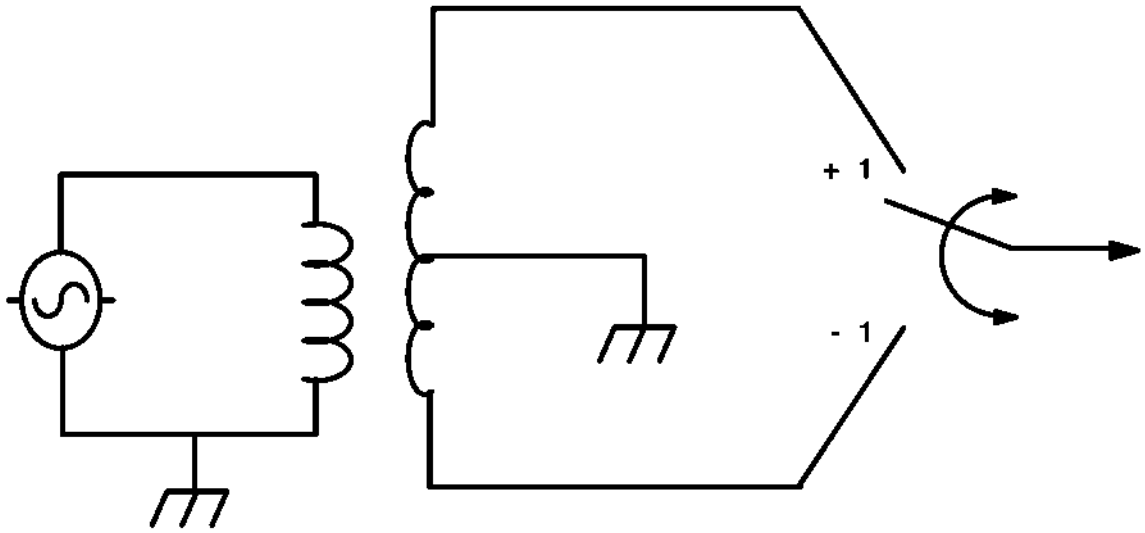

Fig. 9: Reversing switch used to achieve Binary Phase Shift Keying (BPSK).

6.2 Sideband Recovery - Spectral Regrowth

An alternative place to put the filter is before the final Power Amplifier (PA). Here the power level is less, insertion loss is of less consequence, and it would be good

provided the final PA was linear. However one of the attractions of these constant amplitude modulation schemes is that the final PA can be run at high efficiency at its maximum power level. That means the amplifier is driven into saturation and in turn means that the amplitude fluctuations at the input don't appear at the output. This reassertion of the constant amplitude condition further modifies the amplitudes and phases of the signal components. It has the effect of undoing some of the good work of the filter. One has the phenomenon of *Sideband Recovery* alias *Spectral Regrowth*: see Fig. 10. I don't know if it is known whether a long chain of filters and saturating amplifiers would eventually produce a signal with both low sidebands and of constant amplitude, it is an interesting academic speculation, but certainly at present people seem simply to accept that they cannot get rid of their unwanted sidebands, and so far as I can see the ITU writes rules which provide no great incentive to find a way around the problem.

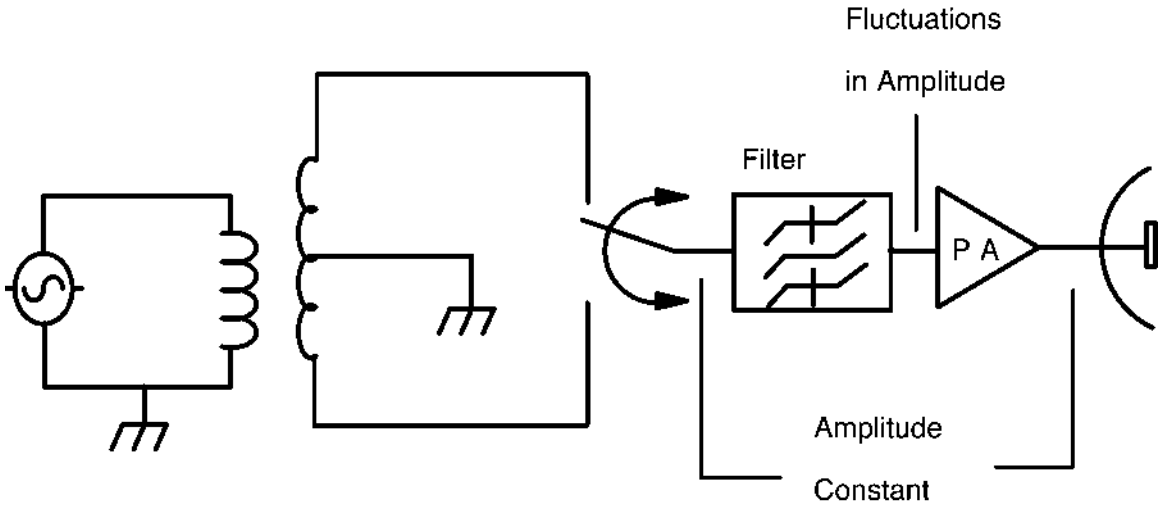

Fig. 10: Sideband Recovery – Spectral Regrowth. The modulator in the first unit introduces sudden phase jumps into a constant amplitude signal, whose spectrum then has far-flung sidebands. Passing this signal to a filter rejects the sidebands, but causes the amplitude to fluctuate. Saturating the power amplifier reimposes a constant amplitude, but sidebands partially return.

In my view it is a mistake to generate sidebands and then try to remove them. I think it is much better to modulate the signal in such a way that the unwanted sidebands are not generated in the first place. I have built a *proof-of-concept* QPSK modulator that does just that. Instead of switching the phase of the signal abruptly from one phase state to another, it is guided slowly in a controlled way from one state to the next. This in an example of what I later learnt was already known as CPM, *Continuous Phase Modulation*. The amplitude of the signal is constant, the phase changes are gradual, the sidebands are inherently low and there is no need for a filter. *Sideband Recovery / Spectral Regrowth* is then not an issue.

7 Modulation

I have already mentioned certain types of modulation. For completeness' sake we must mention the various classical modulations. The earliest type was just ON-OFF keying of a carrier and this was used with Morse Code for *Wireless Telegraphy*. This is still practiced by radio amateurs and it is a minor art form when done well. It was

noticed long ago that it was important not to have the transmitter come on too abruptly when the Morse key was pressed, as it caused *Key Clicks* audible on adjacent channels. We now know why.

7.1 AM and SSB

For broadcasting *Amplitude Modulation* (AM) was first adopted. Here the envelope of the carrier is made to vary according to the waveform of the audio signal carried: see Fig. 11. Its chief merit is that it is easy to make a simple detector for recovering the audio signal and this was important in the days when radio sets had very few active components. However, it is a wretchedly inefficient scheme from two points of view. Firstly the amplitude of the carrier has to exceed twice the maximum value of the sum of all the sidebands. Thus most of the RF power goes into radiating a pure monotonous sinewave that is as boring as the 60 Hz on the power lines. So it is inefficient power-wise. It is also by any definition inefficient from the point of view of use of the spectrum. The reason is that the sidebands are generated in mirror-image pairs, an *upper sideband* and *lower sideband*, and each one alone carries the audio signal. So it occupies at least twice the spectrum that it needs. Despite these inefficiencies it is still much used.

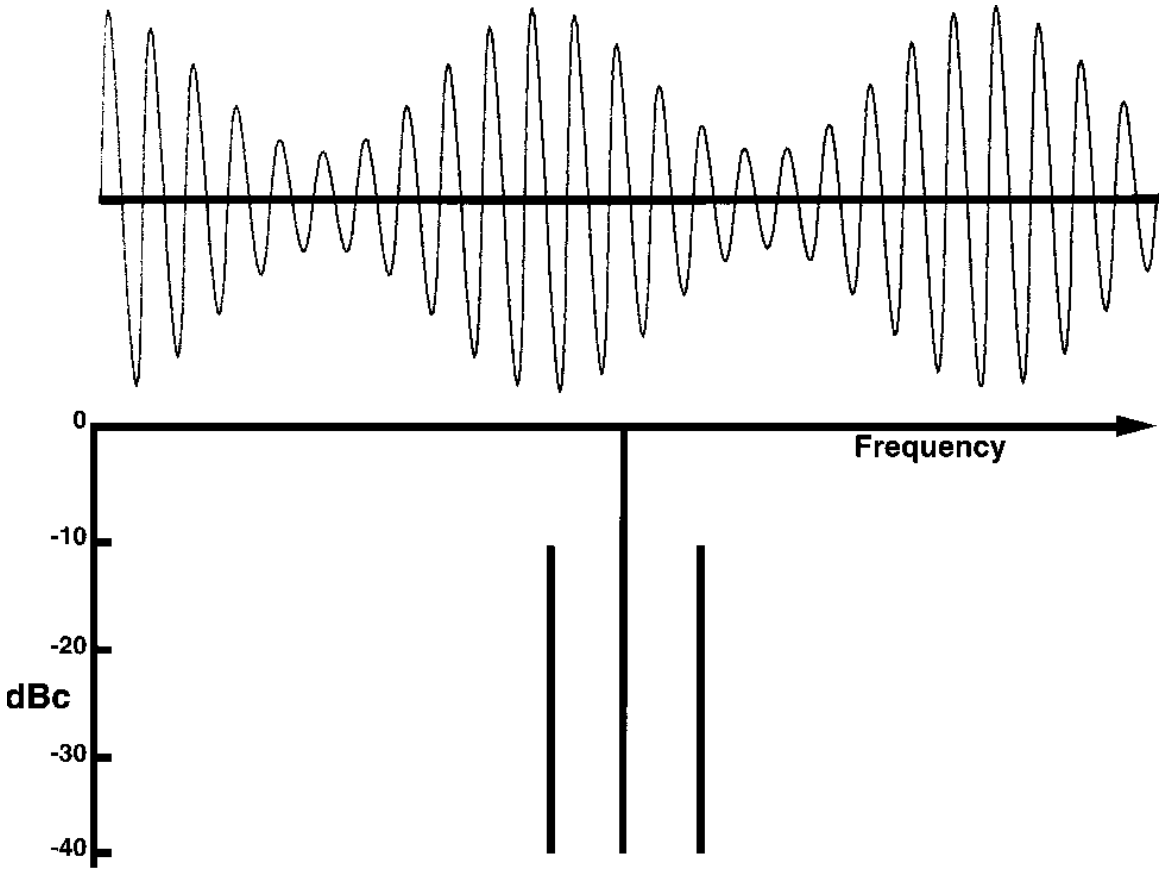

Fig. 11: Amplitude modulation with a modulation index m = 0.6. Each sideband is at -10.45 dBc.

An improvement is *Single Side Band* (SSB). As the name implies only one side band is transmitted. Generally the carrier is also suppressed so the power efficiency is much improved. However the audio signal is now carried in both the

amplitude and the phase of the resulting signal, demodulation is more complicated, and the transmitter PA must be strictly linear to carry the signal. There is a move afoot to change over to SSB for shortwave broadcasting, but it will be a very long time before it wholly replaces AM.

7.2 FM

I have already discussed FM. All I need add is that it exists in two forms. Wide deviation FM is used for audio broadcast and narrow deviation FM is used for such things as marine VHF communications. It is interesting that provided the RF signal-to-noise ratio exceeds a certain threshold, the signal-to-noise ratio of the audio at the demodulator output is much better than the RF signal-to-noise ratio at the input.

7.3 Digital modulation

More interesting for us are the various forms of digital modulations. There is a whole class in which the signal moves from one to another discrete phase state. In principle these changes can be at the Nyquist rate for the given bandwidth. These phase states can be represented on an Argand diagram and may be described as "Constellation diagrams". Four such schemes are shown in Fig. 12 and their theoretical limiting performance are tabulated in Table 1.

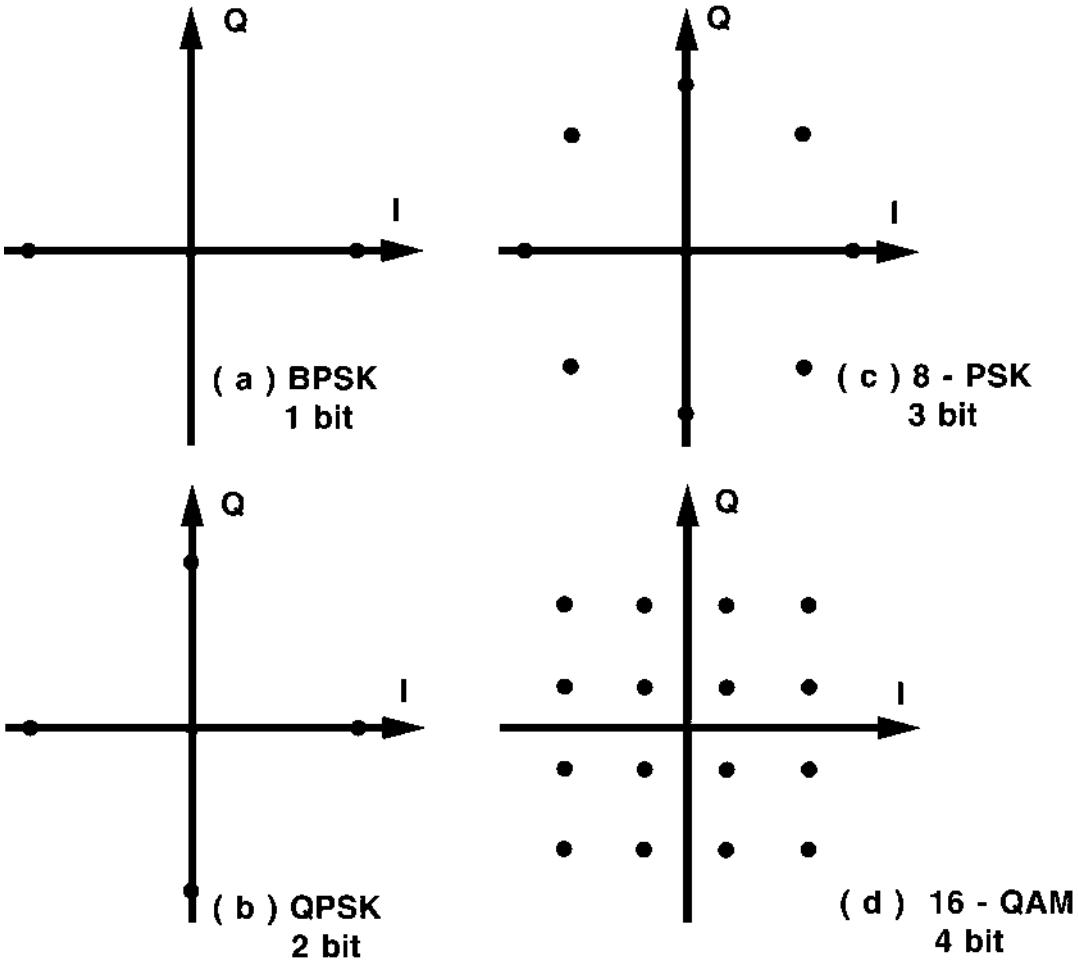

Fig. 12: "Constellation" Diagrams for various digital modulators.

		No. of Phase States	max R bits/sec/Hz
Binary Phase Shift Keying	BPSK	2	2
Quadrature Phase Shift Keying	QPSK	4	4
	8-PSK	8	6
	16-QAM	16	8

Table 1

It will be noticed that the first three operate in principle with a signal of constant amplitude. The last requires the signal to have three distinguishable amplitudes. For this the PA need not be highly linear. Each one listed has variants. For instance with QPSK the x-wise transitions may or may not coincide with the times of the y-wise transitions.

The given bit rates are based on the assumption that the signal can unambiguously adopt a new phase state every $1/(2B)$, at the Nyquist rate. In reality there may be ambiguity because the impulse response of the filters causes some residue of one phase state to be carried over to the next, and this causes *inter symbol interference* (isi). Of course additive noise makes the states indistinct and introduces errors in the bit stream. By adopting error correcting codes, errors can be tolerated, but at the expense of adding "overhead" bits, which then reduce the capacity for the main "payload" bits. There are an enormous number of variants. From the point of view of Spectrum Management, however, the key thing is that all these schemes are conceived in principle as making instantaneous changes of state, and all therefore are inherently prone to emitting a sinc-squared form of power spectrum. Any filter added to reduce the OOBs is certain to add to the isi problem, and therefore users of these schemes cherish their supposedly unwanted emissions.

7.4 Coded Orthogonal Frequency Division Multiplex: COFDM

An interesting and relatively new form of modulation is *Coded Orthogonal Frequency Division Multiplex* (COFDM). It is used for the new Digital TV and Digital Audio Broadcasting (DAB). It is conceived as a large number of very closely spaced carriers each of which is QPSK modulated at some very low rate. From our point of view it has two very interesting properties. One is that it occupies a well-defined band with virtually uniform power spectral density and the power spectrum falls very fast at the edge of the band. The other is that the superposition of the large number of unit amplitude carriers results, by virtue of the *Central Limit Theorem,* in an emitted waveform which very much looks like Gaussian noise. Recall Shannon's Channel Capacity Theorem. I don't know how closely it approaches the theoretical limit, but perhaps it is a move in the right direction. But it does have a down side. That is that the effective peak to rms ratio for Gaussian noise is such that the PA has to be linear and operate in Class-A, which is inherently inefficient from the power point of view.

7.5 Spread Spectrum

There are three types of spread spectrum. All were invented either to provide cryptographically secure communication or to allow covert communication. The difference is that cryptographically secure means that there is no secret that there is a transmission, it is just that it is scrambled in such a fashion that no unauthorized interceptor is able to read the message. Covert however seeks to conceal the very existence of the transmission.

There are three types of spread spectrum. The first is frequency hopping. If a communication system has a large number of otherwise conventional channels, a transmission can be very effectively jumbled up by jumping channels several times per second, possibly at the phoneme rate, and jumping in a prearranged but apparently random fashion between the channels. Naturally such a transmission in effect uses a band as wide as the total spread of the channels. I think this is very straightforward and there is nothing more to be said, except that so far as I know only the military use frequency hopping.

The second type is time-hopping or burst transmission. The presumably digitized message is stored up, and at a prearranged moment transmitted at an enormous data rate, so that the whole event is finished before a would be interceptor has time to get set up to receive it. Again I think this is exclusively a military technique.

The third type, which is what one normally thinks of as *spread spectrum,* is in more general use. For reasons that I don't quite understand, it is called *Direct Sequence Spread Spectrum* (DSSS). The scheme is outlined in Fig. 13. An RF carrier is modulated with BPSK, which simply reverses the phase very fast, according to some prearranged pseudo-random code. The consequence of this in the frequency domain is to spread the energy rather thinly over a wide band. It can be spread so thin that for some receivers it is below the noise level, and thus its very presence is hard to discern. However a receiver "in the know" and provided with an identical generator of the pseudo-random code, and having a similar reversing switch, undoes the effect of the switch in the transmitter. So the sinewave signal is reconstituted. As described

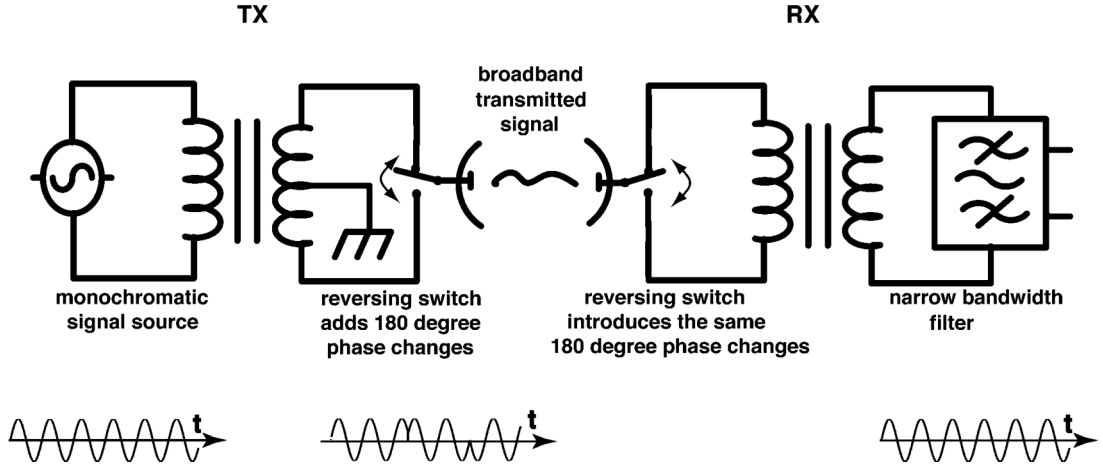

Fig. 13: Direct Sequence Spread Spectrum (DSSS) using BPSK modulation, so the first reversing switch introduces 180 degree phase reversals according to a pseudo-random code, while the second introduces the same reversals to reconstitute the original, narrow-band signal. The output is a “recompressed” narrow-band signal.

there is no communication. But as well as the fast pseudo-random code the outgoing signal can be phase reversal modulated a second time at a much lower data rate at the transmitter, and this second modulation is of course not stripped by the switch in the receiver. The de-spread signal has a much smaller bandwidth than the spread signal, and so in its own bandwidth the signal can be well above noise. Any other transmission entering the receiver that might cause interference gets chopped up by the receiver switch and its energy is spread out and then rejected by internal filters. So this type of transmission may be covert, is cryptographically secure, and is robust against interference whether inadvertent or deliberate.

As well as these properties it may serve another function. The despreading only occurs if the receiver switch is in step with the phase reversals on the incoming signal. If the receiver switch drifts out of synchronization, the despreading is lost. The timing has to be precise to within a fraction of the reciprocal of the so-called *chip rate,* which is just the bit rate of the pseudo-random pulses. It is this feature of DSSS, which makes it the key to the *GPS* and *GLONASS* navigation satellite systems. In *GPS* the civilian so-called C/A code has a chip rate of 1.023 MHz (Note: $2^{10} - 1 = 1023$). A timing shift of ~1 μs is enough for the codes to get completely uncorrelated. But the receiver adjusts its code generator to track the incoming code, using a delay-lock loop, and synchronization is maintained to about 10 ns. By this means the one-way propagation delay from the distant satellite is measured with an error corresponding to a distance error of only a few metres. This is the key to the system's wonderful positional precision. The *GLONASS* system is in effect identical except that its C/A code chip rate is 0.511 MHz (this rate was chosen because $511 = 2^9 - 1$).

This matching of the pseudo-random codes is highly specific and works like a lock matching a key. With *GPS* all the satellites transmit on the same carrier frequency, and therefore the signals from all visible satellites are superimposed at the input to the receiver. But all are below the noise, and even the aggregate of all the signals barely increases the total noise. Independent despreading switches, each controlled by its own pseudo-random code generator, can simultaneously extract and track the signal from its own satellite.

DSSS can of course be used for covert communication, and it is often said that such systems can be "overlaid" across otherwise occupied bands without the users of those bands being aware of it. I believe there was once such a system in Europe that used to overlay the entire broadcasting "Medium Wave Band". Maybe that can be done in a broadcasting band, but it surely can't be done in a radio astronomy band! It would be noticed very quickly.

To my mind such covert overlaying is rather like the old reprehensible practice of coin clipping. In the days when coins were made of precious metal, certain people filed a little off the edge of every coin, thinking the recipient wouldn't notice the loss but little by little they would get rich. I fear this very thing is now being allowed to happen under the name of *Ultra Wide Band*. Perhaps we should consider this a form of spread spectrum. It is certainly in the spirit of coin clipping. It is like stealing from supermarkets. The shoplifter says, "I take so little, they won't notice the loss".

Direct sequence spread spectrum is used with mobile phones under the name *Code Division Multiple Access* or CDMA. Here an integrated system chooses to

reuse the same carrier, just like *GPS*, and each user uses his own pseudo-random code. It certainly provides privacy but I can't recall the precise justification for its adoption. From our point of view the worry of it is that the sidebands generated by the spreading process will fall far outside the system's allocated band.

7.6 No Modulation

The ITU treats the Radio Astronomy Service as a sort of "pretend" communication service. Yet there are profound differences between all forms of remote sensing, both active and passive, and communications.

In a communication system there are:

1. Agreed frequencies, modulation scheme, symbols, codes, ciphers, etc.
2. A distant agent is seeking to "inform" an idea in the recipient's mind. Both share a common "universe of discourse".

A radio astronomer in making an observation is not receiving a communication from a distant galaxy. Nature is not sending messages to astronomers any more than the White Cliffs of Dover send messages to a ship's radar. In remote sensing there are:

1. No agreements.
2. No sending agent.
3. There is no rate of transmission in bits/sec.
4. The observer is alone, trying to "inform" his own mind.

Remote Sensing is NOT communication. One would hope that the ITU would come to understand this.

8 On Spectrum Efficiency

I do not know of a satisfactory definition of "spectrum efficiency". Generally in science or engineering efficiency is defined as

$$\frac{\text{What you get out}}{\text{What you have to put in}} \quad \text{or} \quad \frac{\text{What you get}}{\text{What you have to pay for}}$$

and the numerator and denominator are expressed in the same or equivalent units. Such a definition is applicable for instance to a Heat Engine:

(Work out / Heat in)

or a radio transmitter:

(RF power out / DC power in)

With such definitions it is quite clear that efficiency can be expressed as a percentage, and that the very best is 100 %.

But there is no such definition for *efficiency* of use of the radio spectrum, though that doesn't stop people talking about it. They mouth platitudes about the importance of using the spectrum efficiently. But what stops them in their tracks is asking what they mean! One only has to ask oneself what scenario would constitute 100 % efficient use of the spectrum, to see that the expression is devoid of meaning.

There are however many definitions of spectrum efficiency, more than I know. Each may have validity in its own limited context. However I don't believe there is any single universally satisfactory definition. What I will do is list the considerations that I think should enter into a satisfactory definition, and then perhaps we could collectively invent a definition that embraces them all.

1. To some it seems obvious that the spectrum is being well used if lots of information is being communicated. If the spectrum is well occupied. This idea suggests that services that operate only occasionally are inefficient users of the spectrum. But is one to say that Emergency services and bands allocated to Search and Rescue represent inefficient use of the spectrum? Is a marine radar using the spectrum inefficiently when it receives no echoes? Absence of a signal is good news!

   An analogy may help. The bureaucrats who manage our universities view lecture halls as "plant" which should be used "efficiently". It is inefficient they say to have plant unused and standing idle. They are keen that the plant, provided of course at great expense, be used to full capacity. But they are not consistent in this industrial view. It is not the view they adopt for the provision of rest rooms. No one suggests it is inefficient if the rest rooms aren't used to capacity! Everyone agrees the important thing is that the capacity should be available to meet anticipated demand. Telephone engineers install channel capacity on that basis without it being thought inefficient. Telephone systems are designed so that there is *nearly always* considerable unused capacity. *Quality of service*, which includes finding an unused line available on demand, is the decisive criterion. Why should it be different for lecture halls and the radio spectrum?

2. One important aspect is the volume of space or the area of ground in which one user of the spectrum denies its use to another. Clearly the smaller the space occupied, the more often the same frequencies can be used. The ultimate in this regard is the telephone system. Every pair of wires can reuse the same range of frequencies! This effect is recognized in measures of spectrum efficiency of cellular systems. In an urban environment where the propagation losses are high, the same frequency can be reused close by. In open country, where losses are low, the cells have to be much bigger and the "reuse distance" gets greater.

3. The idea that use of spectrum may deny its use to others should also be applied in "frequency space". If the Out-Of-Band emissions (OOBs) of a broadcasting satellite, say, prevent an adjacent band from being used for its intended purpose, then it is using frequency space that is not properly its own, and this "loss of amenity" for its neighbour in frequency space needs to be included in any comprehensive measure of the efficiency with which it uses the spectrum. I know of no measure of spectrum efficiency that

includes this trespass.

4. How should the spectrum usage of the Radio Astronomy Service be measured? Is it using the spectrum inefficiently if every radio observatory is not observing on every band allocated to radio astronomy all the time? Or if it isn't looking in every direction! Is it efficient that the various allocations to the RAS should be like rest rooms, usually unused but available "on demand"?

5. Is it efficient if a band is occupied? Remember that information is conveyed by improbable messages. How improbable are the signals conveyed by a TV signal? All those sync pulses are absolutely predictable and therefore convey no information, in Shannon's sense! One often overhears people speaking into a mobile phone. How often has one overheard something momentous being said? Is it efficient to clutter the spectrum with inconsequential babble? It may be efficient from the point of view of a mobile phone company, but it is not spectrum efficient.

What they get: Income Revenue

What is paid for: others (RA) give up their use of the spectrum

That is called "externalizing your costs"!

9 Transmitters

For our purposes, stripped of inessentials, a radio transmitter is made up of a high power amplifier (PA), which is supplied by a signal source and a source of generally DC power, and is followed by an output filter, a transmission line, and an antenna.

The term *high power* is strictly relative. In a mobile phone it may be less than 1 Watt. On the Arecibo radar it may be 1 MW. But it is nearly always the dominant power consuming part of any radio installation. The power efficiency of a PA is always a matter of concern. At one end it may be because the power determines how long the batteries last, at the other end it is because the electricity bill becomes considerable. On a spacecraft power is always limited.

At low frequencies, one can make one's amplifiers linear by negative feedback, as with the common op-amp, but this is only possible if the amplifying device has a good deal more intrinsic gain than one really needs. At RF gain is not so easily come by and linearity cannot be assured by negative feedback. So one is forced to accept the inherent non-linearity of the amplifying device. It is a great Universal Truth that

ALL AMPLIFIERS ARE NON-LINEAR

For small signals they may be regarded as linear, but when they are pushed to the point that a reasonable fraction of the DC power input gets transformed into RF output power, they manifest non-linearity. The nonlinearity may be manifest not only

in a lack of proportionality between the output and input levels, but also as a change in the phase relationship between the input and output. This is particularly true of transistor PAs.

One must understand the effect of amplifier nonlinearity on the signal modulation. It is certain to be distorted. There is only one class of modulation that can be passed through a non-linear amplifier without change of form, and that is a signal of identically constant amplitude. Such a signal must be modulated only in its phase. We have already discussed such signals.

The effect of non-linearity is to add frequency components that were not present in the input signal. The effect of amplifier saturation on an AM signal is shown in Fig. 14. The clipping of the high peaks can be regarded as achieved by the negative addition of the missing portions. In the figure one sees that the negatively added peaks form a series of short pulses and these of course have harmonics which are multiples of the original modulation frequency. In general amplifier saturation leads to intermodulation which is to say the generation of side-bands at sum and difference frequencies of the intended sidebands. These intermodulation components can fall outside the allocated frequency band and constitute OOBs (Out-Of-Band) emissions.

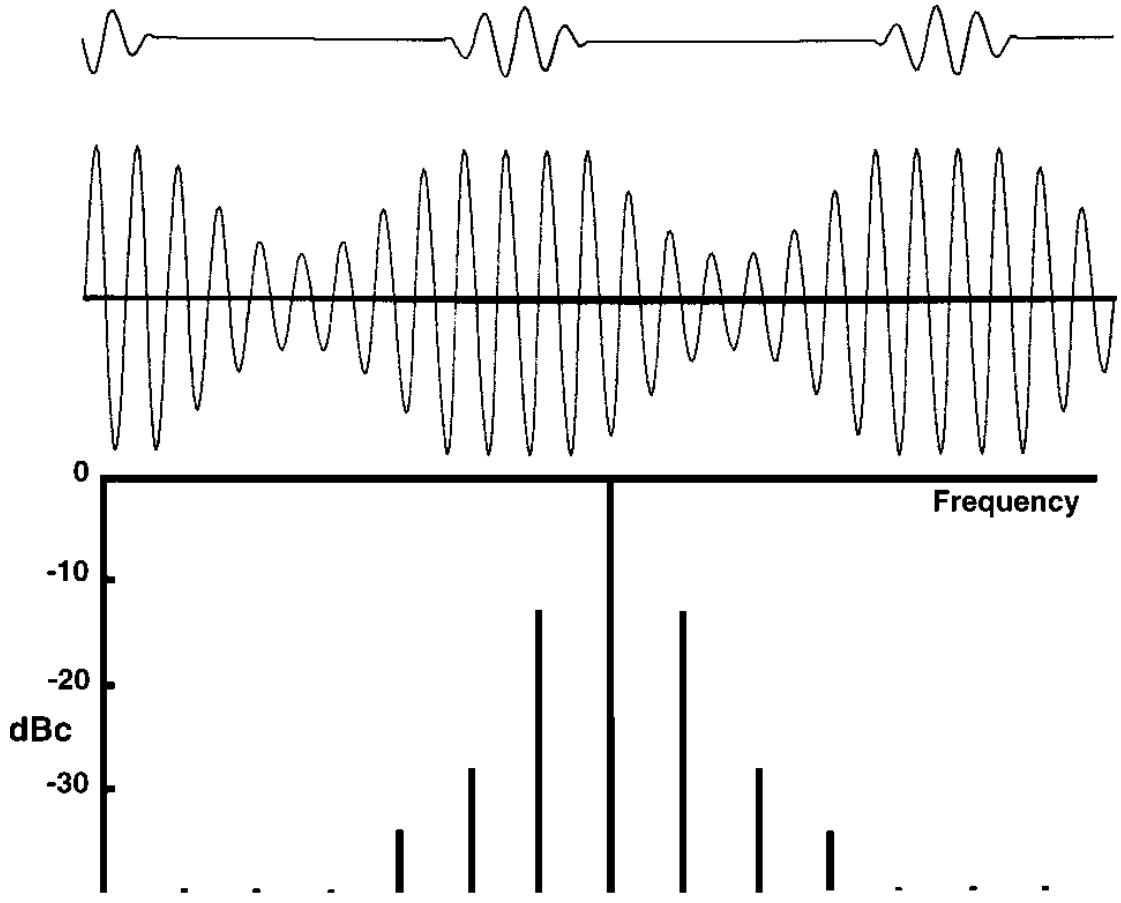

Fig. 14: Intermodulation due to transmitter saturation. Amplitude modulation with m = 0.6 is clipped at +2.27 dBc. In effect the pulses of the first line have been added to the unclipped signal. These pulses are composed of an extended set of sidebands.

This is precisely the problem with the *IRIDIUM* satellites. The PAs have to

handle a number of independent carriers simultaneously. The aggregate sum of these must look very like Gaussian noise. Occasionally these add up to a spike which drives the amplifiers into saturation and out-of-band emission is generated. The aggregate signal level of course depends on the number of carriers, so the problem is "traffic" loading dependent.

10 Antennae

An antenna is a passive reciprocal coupling element, ideally loss-free that couples a guided wave, on a transmission line or wave-guide, to an unguided or free-space wave. It can be used to transmit or to receive. Its key parameters are its far-field pattern, the way its sensitivity varies with direction, its bandwidth, and its input impedance. There are many types of antenna.

10.1 Conventional antennae

It is convenient to start by considering an antenna in transmitting mode. The high frequency current in a radiating element, perhaps a dipole or a monopole in a waveguide, generates an electromagnetic field that carries the power away. The EM field can be conceived as divided into three zones. Very close to the radiating element there is an *induction field*. This is largely confined to within about one wavelength of the radiating element. It has field components which fall with distance r as r^{-2} and as r^{-3}, and these rapidly become insignificant compared to the *radiation field* which varies as r^{-1}. For highly directive antennae the *radiation field* itself is separated into the *near-field* or *Fresnel region* (Augustin Jean Fresnel 1788 - 1827,

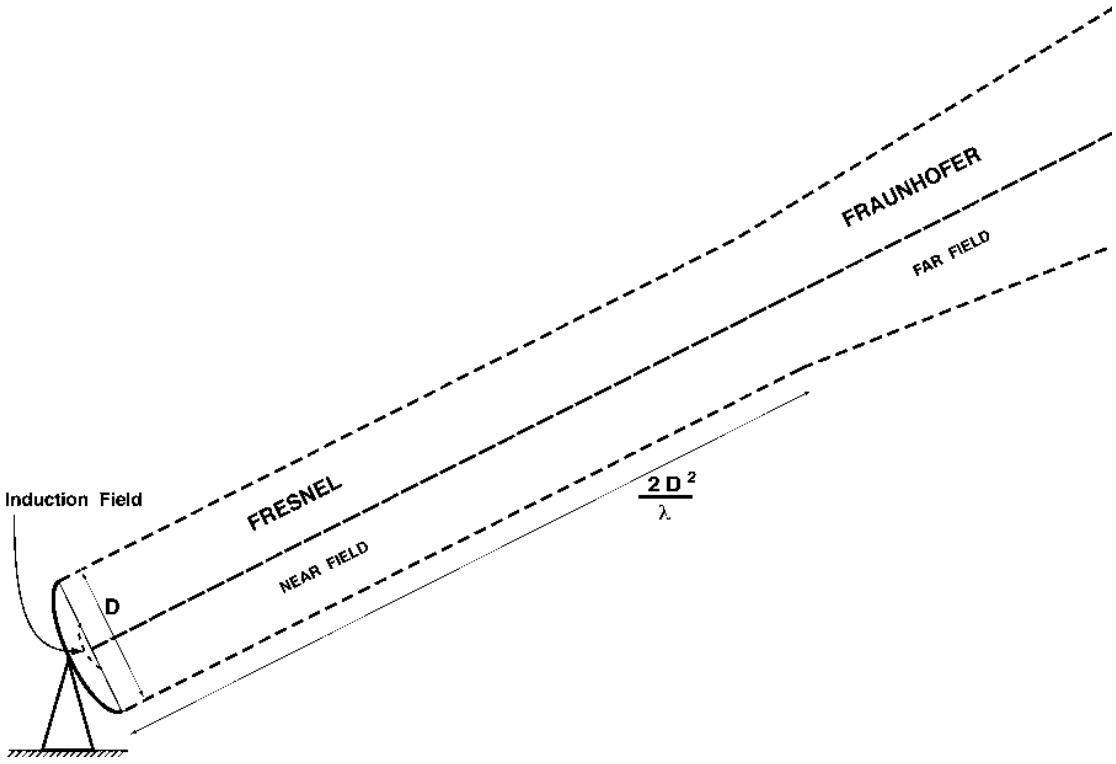

Fig. 15: Field regions associated with a narrow beam antenna in transmitting mode.

French physicist), and the *far-field* or *Fraunhofer region* (Joseph von Fraunhofer 1787-1826, German optician and physicist): note the relationship between their dates. The transition between the two may be defined as at distance r such that $r\lambda = 2D^2$ where λ is the wavelength and D the diameter of the radiating aperture. When a dish antenna is used to transmit, the radiation travels out more or less in a cylinder with the same diameter as the dish, as far as the near/far field transition, and only then spreads out into a conical beam: see Fig. 15. The transition point can be quite a long way away: for example with D = 100 m, λ = 6 cm, $r \approx 330$ km. To efficiently receive a signal from a source at a distance less than this transition distance requires the dish antenna to be refocused. Large optical telescopes need to be refocused to look at the Moon. It is too close to be regarded as at infinity. The same problem may arise when a large radio telescope deliberately receives signals from a LEO (a satellite in Low Earth Orbit).

Consider a transmitting antenna. It radiates with an *angular power flux density* in the far-field $P(\theta, \phi)$ W steradian^{-1}, where θ & ϕ are the spherical coordinates. When multiplied by an arbitrary constant, $P(\theta, \phi)$ is sometimes simply called the *antenna pattern*. It generally consists of a *main beam* confined to a narrow span of directions, and a multitude of smaller *sidelobes* in other directions.

Assuming no ohmic loss, the integral over all directions must be equal to the transmitter power. Thus

$$\text{Transmitter power} \quad P_{tx} = \iint_{4\pi} P(\theta, \phi)\, d\Omega \qquad (19).$$

If the same power were radiated by a hypothetical isotropic or omnidirectional antenna, which incidentally doesn't exist, the *angular power flux density* would be simply $P_{tx}/4\pi$ in all directions. The real antenna concentrates the power more in some directions than in others, but doesn't make or add any to the total. Nevertheless the extent to which an antenna concentrates its power in a given direction is called its *directive gain* $G(\theta, \phi)$, or simply its *gain*

$$G(\theta, \phi) = \frac{P(\theta, \phi)}{\frac{1}{4\pi} \iint_{4\pi} P(\theta, \phi)\, d\Omega} \qquad (20).$$

The fact that this quantity is referred to as a *gain* is the source of much confusion because an antenna doesn't have gain in the same sense as an amplifier has gain. It would have been better if the quantity had been named the *concentration factor,* as that is all it is.

It is easy to see that if all the power were radiated uniformly into a narrow conical beam of angular radius θ radians, the directive gain would be $4\pi / (\pi \theta^2) = 4 / \theta^2$, which is simply the number of times the beam's solid angle goes into 4π.

The concept of antenna *gain* has natural significance in the context of a

transmitting antenna. However when the antenna is used to receive and is exposed to some incident *power flux density* (PFD) W m^{-2}, the received power appears at the output terminals and it is evident dimensionally that the conversion parameter must be an area. For receiving purposes one needs an antenna's *effective area* A_{eff}. For large dish-type antennae, like we see at Green Bank, the effective area is not very much less, possibly between 70 and 80 % of the ordinary physical area of the dish. However for the sort of thin wire antennae that one sees on masts, or for TV antennae, the effective area is vastly greater than the mere physical area of the wires.

Now I think it is obvious that, as a purely passive device an antenna, which when used to transmit concentrates the energy in a narrow range of directions, will likewise be especially sensitive in that direction. Reciprocity applies. So the *directive gain* in a particular direction is proportional to the *effective area* when receiving from the same direction. I won't prove it but there is a universal relationship between *gain* and *effective area*. It is

$$G = 4 \pi A_{eff} / \lambda^2 \qquad (21).$$

I emphasize that here the *gain* is that relative to an isotropic antenna. For a large antenna the *gain* can be a large number and it is usual to express it in *decibels*. It is then 10 $\log_{10}(G)$ dBi, where the i reminds one that it is relative to an isotropic antenna.

For a uniform conical beam

$$G = 4 / \theta^2 = 4 \pi A_{eff} / \lambda^2 \qquad (22),$$

so if we suppose that the aperture efficiency is 100 % for a dish antenna of diameter D, and that $A = \pi D^2 / 4$, we obtain for the angular diameter of the conical beam

$$2\theta = (4 / \pi)(\lambda / D) = 1.27(\lambda / D) \qquad (23).$$

This is very close to the *"half-power beam width"* achieved with a large dish antenna.

Example: What is the gain and the beam width of a 32 m diameter dish at wavelength λ = 21cm?

The physical area is $\pi D^2 / 4 = 256 \pi = 804$ m^2

Typically the *aperture efficiency* $\eta = 0.7$ so

the *effective area* $A_{eff} = 0.7 \times 804 = 563$ m^2

The gain $G = 4 \pi \times 563 / 0.21^2 = 160{,}428 \Rightarrow +52$ dBi

Beamwidth $\approx 1.3\ (0.21 / 32) \times 180 / \pi \approx 0.5°$

The gain is that on the peak of the main lobe. In other directions both the gain and the effective area fall off dramatically. At some point the gain of a large dish becomes no

more than that of an isotropic antenna, which is 1 by definition. At that point its effective area is $\lambda^2/4\pi$, which at 21cm is only 35 cm^2. In reality η might be more or less than 0.7, but the true gain is unlikely to differ from that computed by more than ~ ± 0.6 dB.

It is interesting that the effective area of an isotropic antenna is equal to that of a circle of one wavelength in circumference.

10.2 Active antennae

Despite the usual understanding that an antenna is a passive coupling device, there have been recent developments with integrating radiating and active devices in such a way that there is no distinct interface between the antenna proper and the active amplifier. These integrated devices are referred to as *active antennae.* The antenna proper may be integrated with a low noise amplifier to form an active receiving antenna, or with a PA to form an active transmitting antenna. The *IRIDIUM* satellites' *main mission antennae* are active in both modes and contain embedded T/R switches as well. The development of *active antennae* was not foreseen in the writing of the ITU's RRs. Consequently there is confusion when certain protagonists maintain disingenuously that an *active antenna* is just an *antenna* like any other. It certainly isn't. It presents technical and conceptual difficulties. It is difficult to measure the output power of an active transmitting antenna and the noise factor of an active receiving antenna, and it becomes impossible to specify what may or may not pass on the transmission line between transmitter and antenna, because these isn't one. The absence of a transmission line also makes it impossible to insert a filter.

10.3 Array antennae

Large antennae may be made of arrays of small antennae. At one extreme one has interferometers, which are perhaps to be thought of as sparsely filled arrays, at the other one has things like the *IRIDIUM main mission antennae,* which are composed of 105 individual transmitting and receiving modules laid out on 74" x 34" panels. Each individual "patch antenna" has +4 dBi gain and gives +23.9 dBi for the whole array. The difficulty with arrays in which the elements are close together, is that there is significant interaction between adjacent elements. Consequently the beam doesn't always point in the expected direction, and the element input impedances become functions of beam direction. They are very complicated and I don't think it appropriate to discuss them further.

11 Concluding remarks

I have tried to cover a lot of ground. Each of the topics I've touched on could easily be developed into an extended course. My purpose has been to introduce ideas and subjects with which I believe Spectrum Managers should be familiar. I think the measures of information and channel capacity are especially important, and I have given them prominence not least because I think they may be new to those coming to our subject from a physics background. Possibly the more theoretically inclined with

engineering backgrounds may already be familiar with them. However, I've never been made aware that any of the people I've met at meetings concerned with Spectrum Management have been familiar with them, yet they certainly are relevant to discussions of *necessary bandwidth.* In a certain sense there is perhaps no such thing as *necessary bandwidth,* because in principle any bit rate can be passed through any given bandwidth. No wonder I haven't grasped the ITU's concept of *necessary bandwidth*.

I think the discussion of discontinuous derivatives of signals is simple and cuts through all manner of complication when discussing *Out-Of-Band emissions.* The simple message is that to avoid spilling energy outside an allocated band, the signal must be made smooth to a high degree. There is no other way. It is not clear to me that this is sufficiently widely understood. "Make it smooth" is the message.

I haven't said a great deal about transmitters. The Universal Truth is that ALL AMPLIFIERS ARE NON-LINEAR. It is easy to understand that high efficiency is frequently an operational necessity, but that means operation in nonlinear mode and non-linearity necessarily causes intermodulation. Post PA filters are unwelcome for a number of reasons. My own inclination is to urge the use of *Continuous Phase Modulation,* but I am aware that viewed from the point of view of the *Channel Capacity Theorem* a constant amplitude signal cannot be in every sense optimal.

My discussion of antennae has been an independent departure not well connected with my other themes. But antennae are a vast and difficult subject on their own. I hope I have conveyed the rudiments and provided at least some simple rules of thumb.

12 Bibliography

1. C.E.Shannon, "A Mathematical Theory of Communication," Bell *System Technical Journal, vol.27*, pp 399-423; July: also, pp 623-656; October, 1948.

2. C.E.Shannon, "Communication in the Presence of noise," *Proceedings of the IRE,* vol. 37, pp.10-21; January, 1949.

3. P.M.Woodward, "Probability and Information Theory, with Applications to Radar," *Pergamon Press,* 2nd Ed. 1964.

4. R.N.Bracewell, "The Fourier Transform and Its Applications," McGraw-Hill, 1965.