INTERNATIONAL ATOMIC ENERGY AGENCY
UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION

INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
34100 TRIESTE (ITALY) - P.O.B. 586 - MIRAMARE - STRADA COSTIERA 11 - TELEPHONE: 2240-1
CABLE: CENTRATOM - TELEX 460392-I

SMR.379/21

COURSE ON BASIC TELECOMMUNICATIONS SCIENCE

9 January - 3 February 1989

Digital Modulation of Carrier Waves

P.A. MATTHEWS

University of Leeds, Dept. of Electrical and Electr. Eng., U.K.

DIGITAL MODULATION OF CARRIER WAVES

PETER A. MATTHEWS

November, 1988

1 INTRODUCTION

In both line and radio systems we transmit information using modulated carrier waves. When the information we wish to transmit is in digital form, e.g. from a computer or p.c.m. speech we must decide how best to modulate a carrier by the digital bit stream. There are many types of digital modulation which have been designed with the aim of providing efficient communication subject to various constraints on the transmission channel. For example some are designed to minimise the bandwidth occupied by the modulated carrier whilst others are designed for use in circuits with non-linear amplifiers.

As in analogue modulation a carrier can be modulated by controlling the amplitude, frequency, or phase either separately or together. If the frequency or phase is controlled the signal has a constant envelope and may be amplified using non-linear amplifiers. If the amplitude is controlled then amplifiers and other circuits with a linear response are need.

As the many types of digital modulation are derived from the simple forms controlling amplitude, frequency, and phase we shall consider these first.

*Professor, Department of Electrical and Electronic Engineering, University of Leeds, Leeds LS2 9JT, UK

2 DIGITAL MODULATION

We shall assume that the bit stream which is to modulate the carrier has a uniform bit rate and that it is a binary signal with logical levels $(0, 1)$. The logical levels can by transmitted at baseband as a uni-polar signal with voltage levels $(0, A)$ or a bi-polar signal with levels $(-A, +A)$. Suppose we use these levels to amplitude modulate a carrier which has an angular frequency ω_c. The modulation is carried out using a multiplicative mixer as in d.s.b.s.c. a.m., Fig.1.

If the modulating signal is $m(t)$ then the output signal from the mixer $s(t)$ is
$s(t) = m(t)\, cos(\omega_c t)$.

The signal $s(t)$ will have two states depending on the voltage levels of $m(t)$ associated with the logical levels $(0, 1)$. If $m(t)$ is a uni-polar signal the two states are

Logic Level	$m(t)$	$s(t)$
0	0	$s(t) = 0$
1	A	$s(t) = A\, cos(\omega_c t)$

The signal is switched on and off. This kind of modulation is called on-off keying or o.o.k. It is a particular form of amplitude shift keying, a.s.k., in which the amplitude is shifted between the levels $(0, A)$.

If $m(t)$ is a bi-polar signal the two states are

Logic Level	$m(t)$	$s(t)$
0	$-A$	$s(t) = -A\, cos(\omega_c t)$
1	A	$s(t) = A\, cos(\omega_c t)$

However, as $-cos\theta = cos(\theta - \pi)$ the two states can also be written

Logic Level	$m(t)$	$s(t)$
0	$-A$	$s(t) = A cos\,(\omega_c t - \pi)$
1	A	$s(t) = A cos\,(\omega_c t)$

In this case the signal has a constant amplitude but the phase is shift by π or reversed. This form of modulation is called phase reversal keying, p.r.k., which is a particular form of phase shift keying or p.s.k.

We see that these two forms of digital modulation are similar to analogue a.m. and p.m. with the modulating signal varying between two levels. We may expect that there will be a form of frequency modulation similar to frequency modulation. This is so, the carrier frequency is decreased for one state and increased for the other. If the shift in angular frequency from the carrier is ω_m the two states are

Logic Level	$m(t)$	$s(t)$
0	$-A$	$s(t) = A\ cos\ (\omega_c - \omega_m t)$
1	A	$s(t) = A\ cos\ (\omega_c + \omega_m t)$

This type of modulation is called frequency shift keying or f.s.k.

Frequency shift keying can be produced in several ways. The simplest perhaps is to apply the modulating signal to a voltage controlled oscillator. With such a simple system it may be difficult to control the frequencies precisely. An alternative is to have two oscillators at the frequencies $(\omega_c - \omega_m)$ and $(\omega_c + \omega_m)$ and to switch between the two outputs. Yet another method is to use a single side band modulator and switch from upper to lower side band.

In simple f.s.k. modulation there is no relation between the bit rate, the carrier frequency or the shift frequency. The switchings may occur at any point is a carrier cycle. There are advantages in switching when the carrier passes through zero. This requires a fixed ratio between the carrier frequency and the bit rate. In the case of f.s.k. the two frequencies should be different multiples of a common frequency which itself is some multiple of the bit rate.

Waveforms for the different types of modulation are shown in Fig.2.

3 SPECTRA OF THE MODULATED CARRIERS

As all three types of modulation can be produced by forms of amplitude modulation we can readily deduce the forms of spectra at the outputs of the modulators. We know that in d.s.b.s.c. a.m. the spectrum of the modulated carrier is centred on the carrier frequency and that the spectrum of the upper sideband has the same shape as that of the baseband spectrum of the modulating signal whilst the lower sideband is the mirror image about the carrier frequency.

To find the general form of the spectra consider a baseband signal which is a random bit stream. For a bi-polar signal the d.c. component is zero. If the bit rate is p b/s and the bit period $\tau = 1/p$ the baseband spectrum will have a *sinc* form with a first zero at a frequency of p Hz. The first zero is at p Hz but the spectrum extends to an infinite frequency. The uni-polar signal has a similar form with the addition of a d.c. component.

The spectra of the modulated waves will therefore have the same *sinc* form. For p.r.k. there is no component at the carrier frequency. The first zero will be at $(\omega_c \pm 2\pi p)$. For o.o.k. there will a carrier component and again the first zeros are at $(\omega_c \pm 2\pi p)$.

If we consider f.s.k. to be produced by o.o.k. of two separate frequencies the two spectra will be centred on the frequencies $(\omega_c \pm \omega_m)$ and will overlap. The amount of overlap depends on the spacing of the two frequencies or the value of ω_m.

The exact form of the spectra will depend on the nature of the modulating bit stream. If there a periodicities in the bit stream there will be lines in the spectra.

The theoretical spectra have an infinite width. This is not acceptable in practice and the spectrum of the signal transmitted over the channel must be band limited. The bandwidth may be limited by passing the signal before or after modulating the carrier through filters. However the effect of filtering is to introduce amplitude variations into the modulating signal waveform and the envelope of the modulated carrier. If the bandwidth limitation is to be maintained the system must have a linear response. If the band limited signal is passed through a non-linear circuit the bandwidth

will be extended.

More complicated forms of modulation have been devised to overcome this problem.

We can express the spectra more precisely using the auto-correlation function of the binary signal.

For a random binary waveform with a bit period T_b and two equally probable levels $(0, A)$ the a.c.f. of the waveform is

$$R(\tau) = \frac{A^2}{4}\Lambda(\frac{\tau}{T_b})$$

in which

$$\begin{aligned} \Lambda(\tfrac{\tau}{T_b}) &= 1 - \tfrac{|\tau|}{T_b}, \quad |\tau| < T_b \\ &= 0, \qquad\quad |\tau| > T_b \end{aligned}$$

Taking the Fourier transform gives the power spectrum

$$S(\omega) = \frac{A^2 T_b}{4} Sinc^2\left(\frac{\omega T_b}{2}\right) .$$

As $Sinc(\frac{\omega T_b}{2})$ has a first minimum when $\frac{\omega T_b}{2} = \pi$ the minimum occurs for $f = \frac{1}{T_b}$. The spectrum has a theoretical bandwidth extending to infinity although most of the energy lies within the range $0 \rightarrow f$.

When the signal with this spectrum modulates the carrier at ω_c the spectrum of the modulated wave for o.o.k. will be

$$S(\omega) = \frac{A^2 T_b}{4} Sinc^2\left(\frac{(\omega_c + \omega)T_b}{2}\right) .$$

In the case of p.r.k. the spectrum will have the same form but the power in the spectrum is doubled because the signal is on for all the time.

In the case of f.s.k. the spectrum will have the form

$$S(\omega) = \frac{A^2 T_b}{4}\left[Sinc^2\left(\frac{(\omega_c + \Omega + \omega T_b)}{2}\right) + Sinc^2\left(\frac{(\omega_c - \Omega + \omega)T_b}{2}\right)\right]$$

The power in this spectrum is the same as the power in the p.r.k. spectrum but the f.s.k. spectrum has two peaks separated by 2Ω.

For the two waveforms in f.s.k. to be orthogonal the peak separation

should be

$$2\Omega = \frac{2}{T_b}$$

For minimum error rate it can be shown that the separation should be

$$2\Omega = \frac{3\pi}{2T_b}$$

When the separation of the two frequencies is that needed for orthogonal waveforms the two parts of the spectrum overlap as shown in Fig.3. We can separate the two waveforms by a process of coherent detection. However this separation is too small to use with an incoherent detection system.

4 DEMODULATION OF O.O.K., F.S.K. AN P.R.K.

Two types of demodulator can be considered, non-coherent and coherent. Non-coherent demodulation of o.o.k. and f.s.k. can be carried out using a simple envelope detector and a frequency discriminator respectively. An exact knowledge of the carrier frequency or phase is not required. This leads to simple receivers for this type of signal. However such receivers are not optimum.

As p.r.k. can be produced by a d.s.b.s.c. a.m. modulator and the information depends on the phase we may expect that a synchronous or coherent detector will be required. In a coherent detector a local reference oscillator at the receiver must be locked in frequency to the received carrier and must have a constant phase relation. For p.r.k. the required signal can be derived from the received signal. The received signal is modulated by changing the phase from 0 to π radians. If this signal is passed through a frequency doubler the phase change is also doubled and becomes 0 to 2π. However 2π cannot be distinguished from 0. There are no phase fluctuations on the frequency at $2\omega_c$. This can then be divided by 2 to give a constant reference at ω_c. The coherent detector then produces the product of the received signal and the reference signal.

For the two states the signals are

Logic Level	$s(t)$	Product
0	$s(t) = Acos(\omega_c t)$	$r(t) = Acos(\omega_c t)cos(\omega_c t)$
1	$s(t) = Acos(\omega_c t + \pi)$	$r(t) = -Acos(\omega_c t)cos(\omega_c t)$

But $cos^2(\omega_c t) = \frac{1}{2}[cos(2\omega_c t) + 1]$ so if the product signal is passed through a low-pass filter the output is $(-A, A)$, the original signal.

Coherent detection can also be used for o.o.k. and f.s.k. and is more efficient when the signal is received with added noise. For o.o.k. the local reference must again be at ω_c. For coherent detection there must be phase continuity between the times when the signal is on. This requires a continuously running oscillator at the transmitter the output of which is switched on and off and a continuously running local reference oscillator at the receiver.

For f.s.k. two local reference oscillators are required at the two frequencies $(\omega_c \pm \omega_m)$. The received signal is split and passed through two coherent detectors, Fig.4. When the frequency $(\omega_c - \omega_m)$ is received the detector using the reference at $(\omega_c - \omega_m)$ will produce a d.c. output, the detector using the reference at $(\omega_c + \omega_m)$ will produce an output frequency of $2\omega_m$. When the signal frequency is shifted the situation will reverse. If the low pass filters in the coherent detectors have a cut-off frequency below $2\omega_m$ the output levels on the two detectors will be $(0, A)$ and $(A, 0)$. The two outputs can be subtracted to give the binary bit stream.

If there is a phase error in the frequency of the local reference the output of the coherent detector is reduced. If the received signal is

$$r(t) = Acos(\omega_c t)$$

and the local reference is

$$\ell(t) = cos(\omega_c t + \phi)$$

then the product is

$$\begin{aligned} r(t)\ell(t) &= Acos(\omega_c t)cos(\omega_c t + \phi) \\ &= \frac{A}{2}[cos(2\omega_c t + \phi) + cos\phi] \ . \end{aligned}$$

After the low pass filter the output signal is

$$o(t) = \frac{A}{2}cos\phi$$

This has a maximum when $\phi = 0$ but will be zero if the phase error is $\phi = 90^0$. This shows the need for precise control of the local reference. As the frequency and phase of the received signal varies with variations in the transmission channel it is generally essential to lock the local reference to the received signal. Variations in the channel may be due to temperature variations in cables, changes due to switching, or due to multipath propagation and Doppler shifts on radio systems.

When the signal is transmitted over a fluctuating channel it may be difficult to generate the stable reference required at the receiver. One way of overcoming this problem is to use a differential modulation scheme, d.p.s.k.. In this scheme the bit stream is recoded using the rule that if the incoming bit stream contains a 1 the output state remains unchanged, if it contains a 0 there is a change in state of the output. This is illustrated in Fig.5. At the demodulator the signal of the previous bit serves as the reference for the present bit. If there is no change in phase a '1' is output, if there is a change of π a '0' is output.

To operate a differential scheme of this kind the bit rate must be constant. If the phase of the signal is disturbed by noise or distortion in the previous bit an error may occur in the present bit. There is therefore a tendency for errors to occur in pairs. For a given noise level there will be a higher probability of error using d.p.s.k. than when using p.r.k..

5 COMPLEX ENVELOPE OF THE SIGNAL

To give a physical understanding of the different types of modulation it is best to think of the real variations of voltage with time as has been done so far. However, for the theoretical analysis of the signals it is convenient to use the complex envelope concept. The use of the complex envelope of a bandpass signal is restricted to narrow-band bandpass signals, i.e. signals with a bandwidth less than $\omega_c/2$. Strictly this concept cannot be used for

signals of infinite bandwidth such as a carrier modulated by a rectangular pulse. In practice the concept is used and in practice the signals are band limited.

For o.o.k. using complex notation and the subscripts $(0,1)$ to indicate the two states of the signal.

$$\begin{aligned} s_0(t) &= 0 \\ s_1(t) &= \Re\{Aexp[j\omega_c t]\} \end{aligned}$$

The complex envelope then has two values

$$\begin{aligned} u_0(t) &= 0 \\ u_1(t) &= A \end{aligned}$$

For f.s.k. the signals are

$$\begin{aligned} s_0(t) &= \Re\{Aexp[j(\omega_c t + \pi)]\} \\ &= \Re\{Aexp[j\pi]\, exp[j\omega_c t]\} \\ s_1(t) &= \Re\{Aexp[j\omega_c t]\} \end{aligned}$$

The corresponding complex envelope signals are

$$\begin{aligned} u_0(t) &= Aexp[j\pi] \\ u_1(t) &= A \end{aligned}$$

For the f.s.k. the signals are

$$\begin{aligned} s_0(t) &= \Re\{Aexp[j(\omega_c - \Omega)t]\} \\ &= \Re\{Aexp[-j\Omega t]\, exp[j\omega_c t]\} \\ s_1(t) &= \Re\{Aexp[j\Omega t]\, exp[j\omega_c t]\} \end{aligned}$$

and

$$\begin{aligned} u_0(t) &= Aexp[-j\Omega t] \\ u_1(t) &= Aexp[j\Omega t] \end{aligned}$$

We can use the complex envelope to show the different modulations on phase plane plots, Fig.6. O.o.k. is shown in Fig.6(a). The signal has two positions at $(0,0)$, and $(A,0)$. For p.r.k., Fig.6(b), the two positions are at $(-A,0)$ and $(A,0)$. For f.s.k., Fig.6(c), the points rotate on a circle radius A, clockwise for logical 0 and anti-clockwise for logical 1.

6 MULTI-LEVEL AND MULTI-PHASE MODULATION

In case of o.o.k., p.r.k., and f.s.k. which we have considered so far there is a one-to-one relation between a bit and the waveform. However, we can group sequences of bits into symbols and use methods of modulation which transmit signals representing the different symbols. A simple case is that in which we take the bits in the bit stream in pairs. Then each symbol represents a pair of bits. As there are four possible ways in which pairs of bits can occur the modulation must be able to take up four positions on the phase plane. This could be four positions along the real axis but it is better to use four positions around a circle centred on the origin. Then for each symbol the amplitude is the same and the phase of the signal represents the symbol. This is shown in Fig.7. Two sets of phase positions are convenient to use. In the first the logical levels in the pairs of bits control the phase of the in-phase and quadrature carriers leading to the four positions lying on the two lines at 45^0 to the axes. Alternatively we can rotate this pattern so that the four positions lie on the axes. This type of modulation is called quadrature phase shift keying or q.p.s.k..

We can continue this process further by making up symbols containing more bits. Fig.8 shows the case in which each symbol contains four bits. With four bits in the symbol there are 16 different symbols. Each symbol modulates the carrier to one of 16 different positions on the phase plane. These positions can lie on a circle or a grid as shown. Clearly we can extend this process for symbols containing more bits.

We must note that with 16 positions on a circle the phase difference between adjacent positions is only 22.5^0. If the amplitude of the signal is A the distance between the points is $0.39A$.

When the points all lie on a circle the modulation is called M-ary p.s.k.,

in this case 16-p.s.k.. When the points lie on a grid the modulation is called quadrature amplitude modulation or q.a.m.. For 16 points it would be called 16-q.a.m.. Up to 256 points are in use in systems, 256-q.a.m..

When the points corresponding to the different symbols lie on a circle the signal has a constant amplitude. If we vary amplitude and phase we can position the points on a grid as shown in Fig.8. For sixteen symbols a 4 x 4 grid is needed. If the maximum amplitude of the signal is A the spacing between individual points is $\frac{2A}{3\sqrt{2}} = 0.47A$. For a peak power limited system the points are spaced further apart than for 16-p.s.k..

When we group the bits into symbols the symbol rate is less than the bit rate. The symbol rate in the examples above is a half or a quarter of the bit rate. The symbol rate is also called the baud rate. The bandwidth of the modulated carrier signal depends on the baud rate of the modulating signal. We can reduce the r.f. bandwidth occupied by the signal by using a multi-level modulation.

We shall discuss the ways in which we group bits into symbols in the lectures on coding. We shall find that by choosing the correct method we can improve the resistance of the signals to noise.

7 OPTIMUM DETECTION IN NOISE

We can show that we can maximise the ratio of energy to noise spectral density for a finite energy signal by using a matched filter. If each symbol in the digitally modulated carrier is independent and is detected independently the appropriate matched filter to use is that matched to the signal transmitted during the symbol. This signal, in the forms of modulation described so far, is a constant amplitude sinusoid.

If the waveform during the symbol $s(t)$ and the symbol period is T_s the response of the matched filter is

$$h(t) = s(T_s - t).$$

For a noise free input the output of the matched filter is

$$\begin{aligned} y(t) &= s(t) \otimes h(t) \\ &= R_s(T_s - t) \end{aligned}$$

where $R_s(t)$ is the a.c.f. of the signal $s(t)$. As the a.c.f. has a peak value when $(T_s - t) = 0$ the optimum time to sample the output to detect the signal is at $t = T_s$, the end of the symbol.

As the symbol is of constant amplitude A and duration T_s the energy in the symbol is

$$E = \frac{A^2 T_s}{2}$$

At the output of the matched filter the mean square noise output voltage is

$$\overline{n_0^2(t)} = \frac{E_\eta}{2}$$

where $\eta/2$ is the double sided noise spectral density.
This gives

$$\overline{n_0^2(t)} = \frac{A^2 T_s}{4}\eta$$

As the noise is normally distributed the output of the matched filter at the optimum sampling time has a mean value

$$m = \frac{A^2 T_s}{2} = E$$

and a variance

$$\sigma^2 = \frac{A^2 T_s \eta}{4} = \frac{E_\eta}{2}$$

In the detection process we have to set a threshold to decide whether we consider that a symbol has been transmitted. For the p.r.k. or f.s.k. the threshold is at zero. For o.o.k. the threshold is at half the mean output level of the detector when logical 1 is transmitted.

For o.o.k. or p.r.k. there is only one matched filter in the detector, Fig.9(a), and the variance of the noise is that given above. In the case of coherent detection of f.s.k. there are two matched filters, one for each frequency, Fig.9(b), and the outputs are added. As there are two paths for noise through the detector the noise at the output has a variance double that above.

Hence, for the three simple cases of o.o.k., p.r.k., and f.s.k. the mean signal outputs, noise outputs, and threshold levels are as given in Table 1.

	OOK	PRK	FSK
$Mean, 1$	$\frac{A^2T_s}{2}$	$\frac{A^2T_s}{2}$	$\frac{A^2T_s}{2}$
$Mean, 0$	0	$\frac{-A^2T_s}{2}$	$\frac{-A^2T_s}{2}$
$Variance, \sigma^2$	$\frac{A^2T_s\eta}{4}$	$\frac{A^2T_s\eta}{4}$	$\frac{A^2T_s\eta}{2}$
P_E	$Erfc\sqrt{\frac{E}{2\eta}}$	$Erfc\sqrt{\frac{2E}{\eta}}$	$Erfc\sqrt{\frac{E}{\eta}}$
	$Erfc\sqrt{\frac{S}{2N}}$	$Erfc\sqrt{\frac{S}{N}}$	$Erfc\sqrt{\frac{S}{N}}$

As the noise has a Gaussian p.d.f.,

$$p(y) = \frac{1}{\sigma\sqrt{2\pi}}\, exp\left(\frac{-(y-m)^2}{2\sigma^2}\right)$$

for which the probability of exceeding a level is μ is

$$P_E = Erfc\left(\frac{\mu-m}{\sigma}\right)$$

The probabilities of error for each type of modulation are given in the table. The probabilities of error are given both in terms of the ratio of energy per symbol to noise spectral density and also mean signal power to

noise power on the assumption that the noise bandwidth is the inverse of the symbol period. We must remember that the bandwidth in the detector for f.s.k. is twice that for o.o.k. and p.r.k. and that the mean power for o.o.k. is half the peak power as on the average the signal is on for only half the time.

We have found these probabilities as the symbol error rates at the output of the matched filter coherent detection system when each symbol is independent. For o.o.k., p.r.k., and f.s.k. a symbol corresponds to a bit so that the bit error rate is the same as the symbol error rate.

The probabilities or error can be calculated for different types of modulation and are shown for o.o.k., f.s.k., p.r.k., and d.p.s.k. in Fig.10.

8 BANDWIDTH LIMITING TECHNIQUES

We have seen that an unfiltered binary signal has in theory an infinite bandwidth. This signal may be filtered before modulation to limit the modulating signal bandwidth but then a linear system must be used in the amplifiers and mixers after modulation. The wide bandwidth of the modulated signal arises because of the rapid changes of amplitude and phase. If the steps in amplitude and phase can be limited then the bandwidth of the significant components of the modulated signal can be reduced.

Using q.p.s.k., in which both in-phase and quadrature components are switched at the same time, the steps in phase may be $\frac{\pi}{2}$ or π depending on whether one or both components are switched. If the two modulating signals are staggered in time only transitions of $\pm\frac{\pi}{2}$ will occur. If a single bit stream is switched alternately to the two branches of the modulator, Fig.11, each branch will be switched at half the bit rate and the switching points will be staggered by one bit. This type of modulation is called offset q.p.s.k. or o.q.p.s.k..

This idea can be extended to various continuous phase or frequency shift modulation schemes.

We can write a phase modulated signal

$$s(t) = Acos\left(\omega_c t + \theta(t)\right)$$

We wish to control the phase $\theta(t)$ so that it is a continuous function of time. In frequency shift keying the frequency is controlled so that

$$s(t) = A\cos(\omega_c t \pm \Delta\omega t + \theta(0))$$

over the bit period T. In this expression $\theta(0)$ is the phase at $t = 0$. If the two transmitted frequencies are ω_1 and ω_2 we have

$$\omega_c = \frac{\omega_1+\omega_2}{2}$$
$$\Delta\omega = \frac{\omega_1-\omega_2}{2}$$

As a function of time the phase $\theta(t)$ is

$$\theta(t) = \pm\Delta\omega t + \theta(0)$$

For the two transmitted signals to be orthogonal there must be an integer number of cycles in each transmitted bit interval. For this $2\Delta\omega.T = \eta\pi$. For the minimum shift

$$\Delta\omega.T = \frac{\pi}{2}$$

or $$2\Delta f = \frac{1}{2T}$$

With this frequency shift

$$\theta(t) = \pm\frac{\pi t}{2T} + \theta(0)$$

The value of $\theta(0)$ depends on the past history of the signal and can be chosen to be zero. Over any bit period the phase is advanced or retarded by exactly $\frac{\pi}{2}$ with respect to the reference carrier.

If the modulating bit stream of data is a bipolar signal with values $a_k = \pm 1$ and θ_k is a phase constant during the kth data interval, i.e. for $kT \leq t \leq (k+)T$ the modulated signal can be written for m.s.k.

$$s(t) = \cos\left(\omega_c t + \frac{\pi a_k}{2T}.t + \theta_k\right)$$

The values of the constants θ_k depend on the need for the waveform to be continuous at the bit transition times. For this to be true

$$\theta_k = \theta_{k-1} + (a_{k-1} - a_k)\frac{\pi k}{2}$$

Now at some particular time θ_{k-1} can be set to zero and then

$$\theta_k = (a_{k-1} - a_k)\frac{\pi k}{2}$$

As $a_k = \pm 1$ the term in brackets is zero or two and so $\theta_k = 0$ or $\pi mod 2\pi$. We can define $\theta(t)$ to be

$$\theta(t) = \theta_k + \left(\frac{\pi a_k}{2T}\right)t.$$

Then $\theta(t)$ is a piece-wise linear function added to the linearly increasing carrier phase, $\omega_c t$. The phase function for a particular data stream takes a particular path through the phase trellis, Fig.12. At each increment of T the phase of the m.s.k. waveform is advanced or retarded by 90^0 with respect to the carrier phase.

The expression above for the signal $s(t)$ can be expanded trigonometrically and as $\theta_k = 0$, $\pi mod 2\pi$ we are left with

$$s(t) = \cos(\theta_k)c(t)\cos(\omega_c t) - a_k\cos(\theta_k)d(t)\sin(\omega_c t)$$

where $c(t) = \cos\left(\frac{\pi t}{2T}\right)$ and $d(t) = \sin\left(\frac{\pi t}{2T}\right)$

This shows that the m.s.k. can be produced by the modulation of in-phase and quadrature carriers. The in-phase channel gives the signal

$$s_I(t) = \cos(\theta_k)c(t)\cos(\omega_c t)$$

and the quadrature channel

$$s_Q(t) = a_k\cos(\theta_k)s(t)\sin(\omega_c t)$$

The two signals are weighted in amplitude during the bit period by the

functions $c(t)$ and $d(t)$. These are half sinusoidal over a time of $2T$ or two bit periods and each will fall to zero on alternate values of T. As the value of a_k can change every bit period it would at first seem that $cos(\theta_k)$ and $a_k cos(\theta_k)$ will vary at the end of each bit. However, it can be shown that $cos(\theta_k)$ only changes at the zero crossings of $d(t)$. In each channel the transitions occur at intervals of $2T$. This is shown in Fig.13.

As we have designed a continuous phase signal we would expect it to occupy a smaller bandwidth than q.p.s.k. or o.q.p.s.k.. The spectra have been calculated for random bit streams and are shown in Fig.14.

The bit error rates of m.s.k. and o.q.p.s.k. can be calculated for an ideal channel corrupted by additive white gaussion noise and using perfect coherent detection. Both m.s.k. and q.p.s.k. can then be considered to be orthogonal channels with antipodal signals. The probability of error is given by

$$P_e = Erfc\left(\sqrt{\frac{2E_b}{N_0}}\right) = \int_{\lambda}^{\infty} \frac{1}{\sqrt{2\pi}} exp\left[\frac{-x^2}{2}\right] dx$$

where $\lambda = \sqrt{\frac{2E_b}{N_0}}$, E = signal energy per bit, and $\frac{N_0}{2}$ is the double side spectral density of the AWGN. This probability of error can be compared with that for coherently detected orthogonal f.s.k. with a tone spacing of $\frac{1}{T}$. For this modulation

$$P_e = Erfc\left(\sqrt{\frac{E_b}{N_0}}\right)$$

We see that m.s.k. has a 3dB advantage.

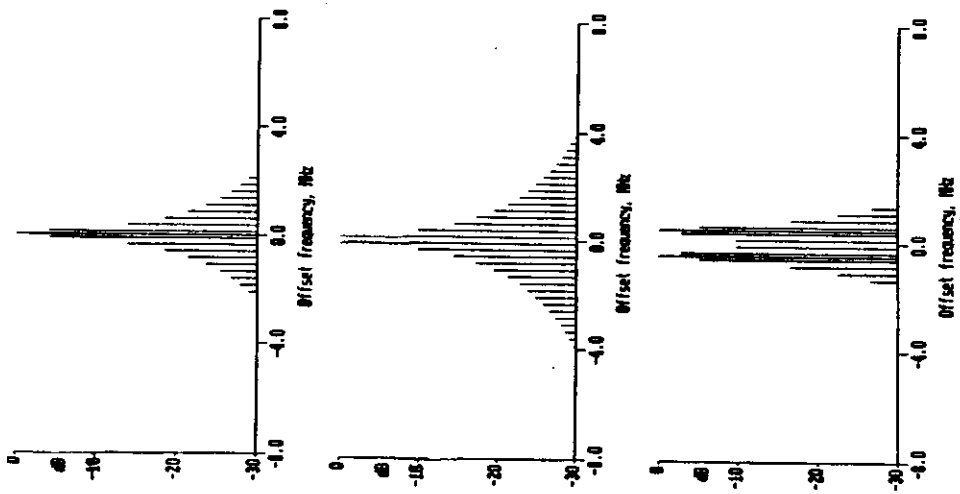
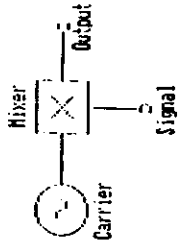


Fig.3. Spectra for (a) o.o.k., (b) p.s.k., (c) f.s.k. about the carrier frequency. Note that the spectra for o.o.k. and p.s.k. are the same as those for unipolar and b-polar baseband signals offset to the centre frequency of the carrier.



Multiplicative mixer for double sideband suppressed carrier amplitude modulation.

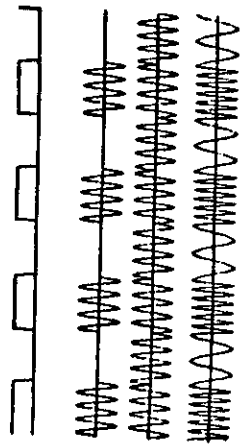


Fig.2. Waveforms for (a) o.o.k., (b) p.s.k., (c) f.s.k..

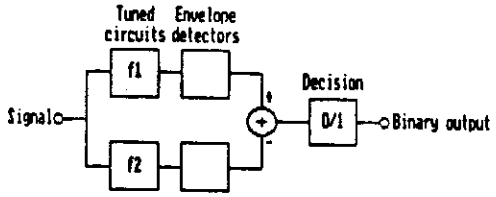


Fig.4. F.s.k. detector block diagram.

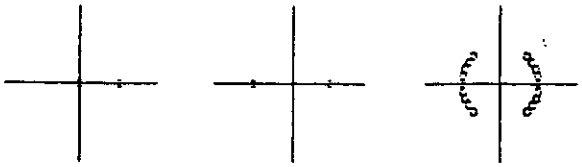


Fig.6. Phase plane plots, (a) o.o.k., (b) p.r.k., (c) f.s.k..

Code input 1 0 1 1 0 1 0 0 1
Differential code 1 1 0 0 0 1 1 0 1 1

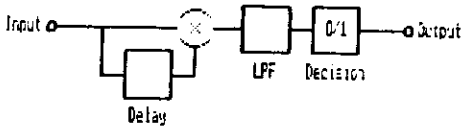


Fig.5. Differential p.s.k., (a) coding, (b) receiver

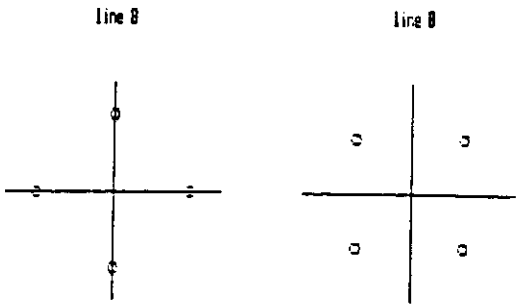


Fig.7. Phase plane plots for q.p.s.k.

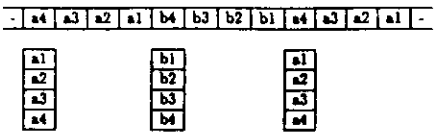


Fig.8. Conversion of serial bit stream to four bit symbols. Note that the symbol rate is one quarter the bit rate.

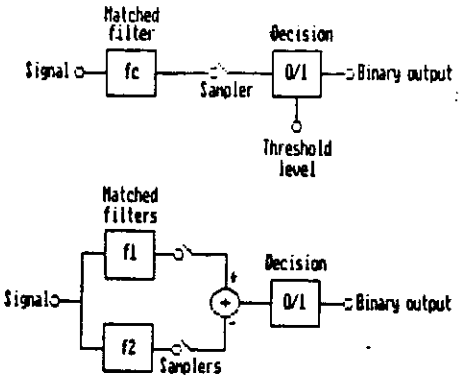


Fig.10. Matched filter detection of (a) o.o.k. and p.s.k., (b) f.s.k..

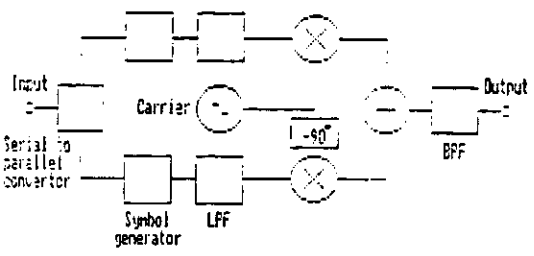


Fig.9. Quadrature amplitude modulation.

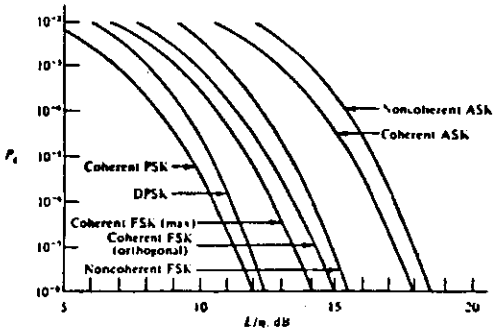


Fig.11. Probability of error for binary modulations.

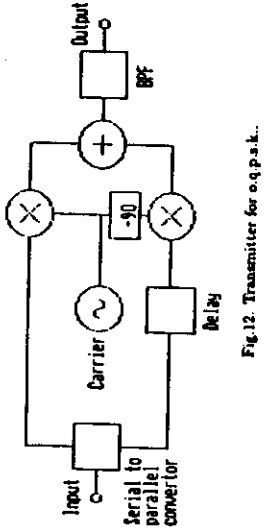


Fig.12. Transmitter for o.q.p.s.k.

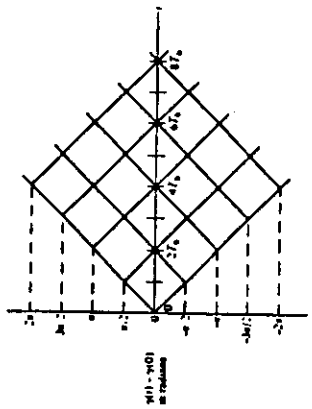


Fig.13 Phase trellis for m.s.k.

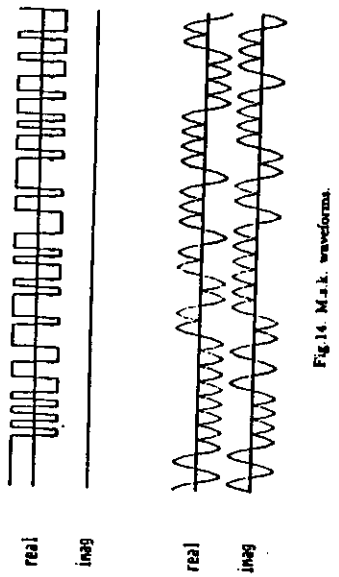


Fig.14. M.s.k. waveforms.

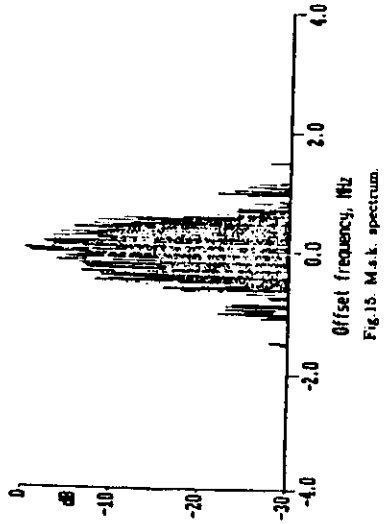


Fig.15. M.s.k. spectrum.

APPENDIX: ANTIPODAL AND ORTHOGONAL SIGNALS

In a digital transmission system we seek the optimum method of transmitting and demodulating signals in the presence of noise. In general an input digital bit stream will be converted to symbols and a M-ary transmission system used. The bit stream is divided into blocks of k bits and there are $2^k = M$ distinct blocks used to transmit the information.

The possible transmitted signal waveforms are $\{s_m(t)\}, m = 1, 2, \ldots, M$. If these signals are bandpass signals they can be represented by

$$s_m(t) = \Re\{u_m(t)\exp(j\omega_c t)\}, \quad m = 1, 2, \ldots, M$$

where $\{u_m(t)\}$ is the set of equivalent low-pass waveforms.

The energy in any one of the M signals is

$$\begin{aligned} \mathcal{E}_m &= \int_0^T s_m^2(t)\,dt \\ &= \frac{1}{2}\int_0^T |u_m(t)|^2\,dt, \quad m = 1, 2, \ldots, M \end{aligned}$$

and the complex cross-correlation coefficient between any pair of the signals is

$$\rho_{jm} = \frac{1}{2\sqrt{\mathcal{E}_m \mathcal{E}_j}} \int_0^T u_m(t) u_j^*(t)\,dt\,.$$

For the bandpass signal waveforms

$$\Re\{\rho_{jm}\} = \frac{1}{\sqrt{\mathcal{E}_m \mathcal{E}_j}} \int_0^T s_m(t) s_j(t)\,dt\,.$$

It will be assumed that the signal is transmitted through a channel with no bandwidth limitation and that the signal suffers the same attenuation

and the same time delay at all frequencies, i.e. the channel is distortionless. However noise is added to the signal. The noise is additive white Gaussian noise, AWGN.

At the receiver the signal is

$$e(t) = \Re\{[\alpha u_m(t-t_0)\exp(-j\omega_c t_0) + z(t)]\exp(j\omega_c t)\}$$

where α is the attenuation, t_0 is the time delay, and $z(t)$ is the equivalent low-pass added noise. The added noise is a zero-mean Gaussian stationary random process with an auto-correlation function

$$\Phi_{zz} = \frac{1}{2}E[z^*(t)z(t+\tau)]\,.$$

At the receiver the low-pass equivalent signal can be written

$$r(t) = \alpha\exp(-j\phi)u_m(t) + z(t)$$

in which ϕ shows the dependence on the phase of the carrier at the receiver. The explict dependence on the time delay t_0 has been suppressed and it is assumed that this time delay can be estimated at least approximately at the receiver. At the receiver the carrier phase is sensitive to the time delay. In some receivers the carrier phase is known or can be estimated exactly and used in the demodulation process. The demodulation process is then called coherent demodulation. If the carrier phase is not used in the demodulation process it is called non-coherent.

At the receiver the problem is to decide which of the M signals was transmitted. Because of noise in the system there is some probability of error in making this decision. The receiver must estimate and maximise the *a posteriori* probability that a signal m from the possible set of signals was transmitted.

The *a posteriori* probability for each of the M possible transmitted signals can be expressed in terms of the received signal $r(t)$. If each of the M transmitted signals is equally probable and the noise is AWGN noise then the optimum receiver should estimate the M decision variables, [1],

$$U_m = \Re\left\{\exp(j\phi)\int_0^T r(t)u_m^*(t)\,dt\right\} - \alpha\mathcal{E}_m,\quad m = 1,2,\ldots,M$$

and select the signal which produces the highest value for U_m. In this expression the terms αE_m are bias terms depending on the energies in the different transmitted signals. If all the signals are transmitted with the same energy then all these terms are the same at the receiver and can be ignored in the decision process. It is seen that the carrier phase enters into the decision process. This may be estimated by observation of the received signal over a period of time.

It can be seen that the process which the receiver carries out within the integral corresponds to a process of cross-correlation or a process of matched filtering. These two processes are equivalent. They are shown in block diagram form in fig.A1.

Binary signalling

In a binary system only two signalling waveforms are used, $\{s_m(t)\}$, $m = 1,2$ with equal energies. Their complex cross-correlation coefficient is

$$\rho = \rho_R + j\rho_Q = \frac{1}{2\mathcal{E}}\int_0^T u_1(t)u_2^*(t)\,dt\,.$$

The optimum demodulator forms the decision variables

$$U_m = \Re\left\{\exp(j\phi)\int_0^T r(t)u_m^*(t)dt\right\},\quad m = 1,2$$

and selects the signal which yields the largest value of U.

Suppose the signal $s_1(t)$ is transmitted during the bit period $0 \le t \le T$. The equivalent low-pass received signal is

$$r(t) = \alpha\exp(-j\phi)u_1(t) + z(t)$$

and so the two decision variables are

$$\begin{aligned} U_1 &= \Re\{2\alpha\mathcal{E} + N_1\} \\ &= 2\alpha\mathcal{E} + N_{1R} \\ U_2 &= \Re\{2\alpha\mathcal{E}\rho + N_2\} \\ &= N_{2R} \end{aligned}$$

in which

$$N_m = \exp(j\phi) \int_o^T z(t) u_m^*(t)\, dt$$

represents the noise and $N_{mR} = \Re\{N_m\}$.

The probability of error is the probability that $U_2 > U_1$. Now

$$P(U_2 > U_1) = P(U_2 - U_1 > 0) = P(U_1 - U_2 < 0)$$

so it is mathematically convenient to consider

$$\begin{aligned} V &= U_1 - U_2 \\ &= 2\alpha\mathcal{E}(1-\rho_R) + N_{1R} - N_{2R}\,. \end{aligned}$$

As each of the noise terms is a zero-mean Gaussian process so is their difference and the variable V has a Gaussian distribution with a mean value

$$\mu_V = E(V) = 2\alpha\mathcal{E}\,(1-\rho_R)$$

and a variance

$$\begin{aligned} \sigma_V^2 &= E[(N_{1R} - N_{2R})^2] \\ &= E(N_{1R}^2) - 2E(n_{1R}N_{2R}) + E(N_{2R}^2) \\ &= 2\mathcal{E}\, N_0(1-\rho_R) \end{aligned}$$

where N_0 is the power spectral density of the noise $z(t)$.

The probability of error is then

$$\begin{aligned} P(V<0) &= \int_{-\infty}^{0} p(v)dv \\ &= \frac{1}{\sigma_V\sqrt{2\pi}} \int_{-\infty}^{0} \exp\left(\frac{-(v-\mu_V)^2}{2\sigma_V^2}\right) dv \\ &= \frac{1}{2}\, erfc\left(\sqrt{\frac{\alpha^2 E}{2N_0}(1-\rho_R)}\right)\,. \end{aligned}$$

As the probability of $s_2(t)$ being transmitted is the same as the probability of $s_1(t)$ being transmitted the probability of error in detecting $s_2(t)$ is the same as the probability of error in detecting $s_1(t)$. This probability of error depends on the correlation between the two signals.

Two important cases are (a) when the signals are un-correlated, $\rho = 0$, and (b) when the signals are 'antipodal', i.e. when $\rho = -1$ or $s_1(t) = -s_2(t)$ or $u_1(t) = -u_2(t)$.

When the signals are antipodal the probability of bit error is

$$\begin{aligned} P_b &= \frac{1}{2}\, erfc(\sqrt{\frac{\alpha^2\mathcal{E}}{N_0}}) \\ &= \frac{1}{2}\, erfc(\sqrt{\gamma_b}) \\ &= Erfc\sqrt{2gamma_b} \end{aligned}$$

where $\gamma_b = \alpha^2\mathcal{E}_b/N_0$ is the signal to noise ratio per information bit.

When the signals are un-correlated, e.g. orthogonal signals, then the probability of error is

$$P_b = \frac{1}{2}\, erfc\left(\sqrt{\frac{\gamma_b}{2}}\right)$$

which gives a 3dB worse performance than with antipodal signals.

A particular form of antipodal modulation is p.s.k. A particular form of orthogonal signalling is binary f.s.k. The probability of error expected for these two types of modulation when optimum detection is used is shown in fig.A2.

Reference

1. PROAKIS, J.G.: 'Digital communications', 1983, McGraw-Hill, Inc.

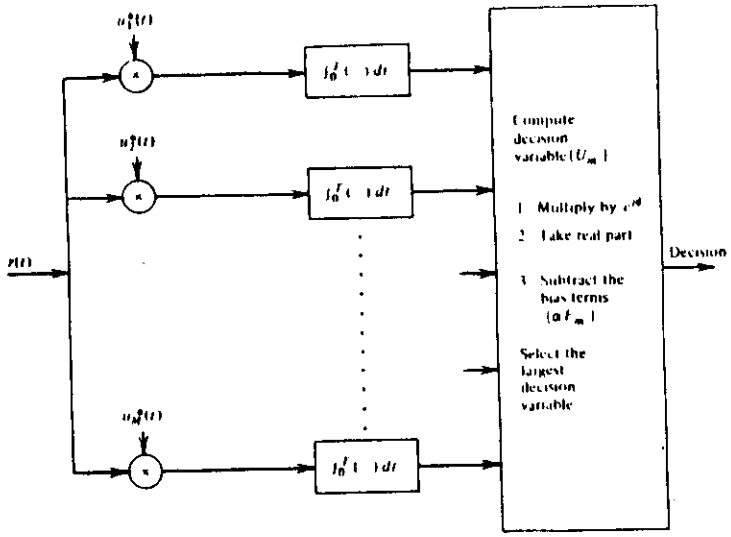


Fig.A1(a) Cross-correlation demodulator

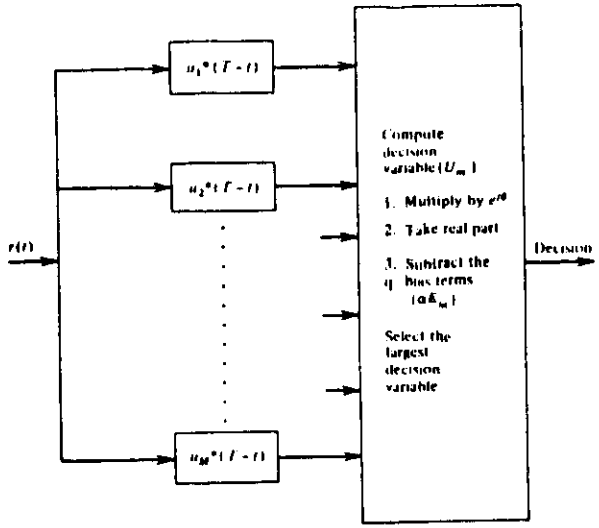


Fig.A1(b) Matched filter demodulator

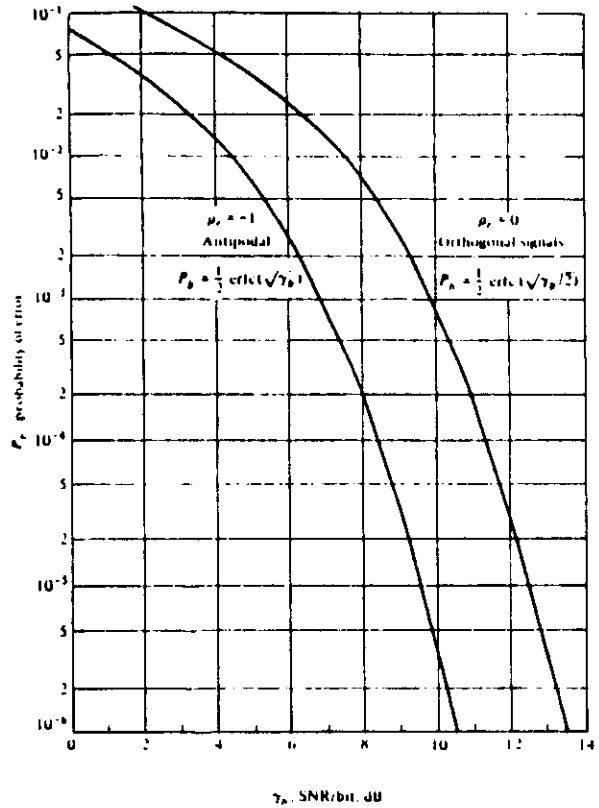


Fig.A2. Probability of error for antipodal and orthogonal binary signals.

M-ARY ORTHOGONAL SIGNALLING

PETER A. MATTHEWS

November 1988

1 INTRODUCTION

Orthogonal signals have the property that their cross-correlation coefficients are zero. An example of orthogonal signals are the pairs of frequencies used for binary f.s.k. For binary f.s.k. and phase coherent detection the probability of bit error is

$$P_b = \frac{1}{2}\, erfc\left(\sqrt{\frac{\gamma_b}{2}}\right)$$

in which γ_b is the received s.n.r. per information bit.

In a binary system there are only two orthogonal states, however we can have M-ary systems with M mutually orthogonal states. For example we can use M different frequencies all of which are mutually orthogonal. At the receiver we use M matched filters or correlators to detect which signal was transmitted. If the signal is received in AWGN the outputs of all the filters except one will be noise. The one filter will produce an output due to the signal.

2 NOISE IN M-ARY SYSTEMS

At the receiver in the absence of noise we receive a signal for the mth state

*Professor, Department of Electrical and Electronic Engineering, University of Leeds, Leeds LS2 9JT, UK

$$s_m(t) = \Re\left\{u_m(t)\, exp\left[j\omega_c t\right]\right\}$$

where $u_m(t)$ is the low pass equivalent signal to $s_m(t)$.

In practice we receive the signal with noise. The low pass equivalent received signal is

$$r_m(t) = u_m(t) + n(t)$$

in which $n(t)$ is the low pass equivalent noise. This noise is independent in each matched filter. The matched filters produce outputs over the symbol period T given by

$$U_m = \Re\left\{\int_0^T r(t)u_m(t)\, dt\right\} \text{ for } m = 1, ..., M.$$

If the signal transmitted is that for $m = 1$ then the output of the corresponding matched filter is

$$U_1 = 2E + N_1.$$

where E is the energy due to the received symbol and N_1 is due to the noise output.

For the other matched filters the outputs are

$$U_m = N_m, \quad m = 2, ..., M$$

All the noise outputs have zero mean and have the same variance even though they are statistically independent. All the distribution are Gaussian are so the probabilities of a given level occurring are

$$P(U_1) = \frac{1}{\sigma\sqrt{2\pi}}\, exp\left[-(U_1 - 2E)^2/2\sigma^2\right]$$

$$P(U_m) = \frac{1}{\sigma\sqrt{2\pi}}\, exp\left[-U_m^2/2\sigma^2\right], \quad m = 2, ..., M$$

in which the variance $\sigma^2 = 2EN_0$.

For the receiver to make a correct decision U_1 must be greater than all

the other U_m. The probability of this occurring is

$$P_c = \int_{+\infty}^{-\infty} P(U_2 < U_1,\ U_3 < U_1, ...U_m < U_1|U_1)p(U_1)du_1.$$

In this expression $P(U_2 < U_1,\ U_3 < U_1, ...U_m < U_1|U_1)$ is the joint probability that $U_2, U_3, ..., U_m$ are all less than U_1 given U_1. As all the U_m are independent the joint probability is the product of the separate probabilities. Each probability has the form

$$\begin{aligned} P(U_m < U_2|U_1) &= \int_{-\infty}^{U_1} p(U_m)dU_m \quad m = 2, ..., M \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{U_1/\sigma} exp\left[\frac{-x^2}{2}\right] dx \end{aligned}$$

hence the probability of a correct decision is

$$P_c = \int_{-\infty}^{+\infty} \left(\frac{1}{\sqrt{2\pi}}\int_{\infty}^{U_1/\sigma} exp\left[\frac{-x^2}{2}\right] dx\right)^{M-1} p(U_1)dU_1.$$

The probability of a symbol error is $P_M = 1 - P_c$.

We must remember that each symbol corresponds to k-bits. Therefore, a symbol error leads to one or more bit errors.

The probability of error finally reduces to

$$P_M = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} \left\{1 - \left[1 - \frac{1}{2}\, erfc\,(\frac{\gamma}{\sqrt{2}})\right]^{M-1}\right\} x\ exp\left[-(\gamma - \sqrt{2\gamma})^2/2\right] dy$$

in which $\gamma = \frac{E}{N_0}$ is the energy per symbol/noise power/ Hz.

There are $M = 2^k$ waveforms and each symbol corresponds to k bits. Hence, $\gamma = k\gamma_b$ where γ_b is the s.n.r. per bit. Also as the symbol errors are equally probable we can put

$$\frac{P_M}{M-1} = \frac{P_M}{2^k - 1}$$

The number of ways in which n bits out of k can be in error is $\binom{k}{n}$.

Therefore the average number of bit errors per k bit symbol is

$$\sum_{n=1}^{k} n \binom{k}{n} \frac{P_M}{2^k - 1} = k\,\frac{2^{k-1}}{2^k - 1}.\ P_M$$

Hence the average bit error rate is

$$P_b = \frac{2^{k-1}}{2^k - 1}.\ P_M$$

This gives the probability of bit error as a function of γ shown in Fig.1. This shows the advantage of increasing the number of waveforms. However, we must remember that increasing the number of waveforms will generally increase the overall system bandwidth.

As we increase the number of waveforms s.n.r. required for a given error rate decreases but the function has a steeper slope. In the limit as $M \to \infty$ the function becomes a vertical line. A simple upper bound occurs at $y = 1.39$ or $1.42dB$. Above this s.n.r. the probability of error would be zero. This gives us a limit beyond which we cannot expect to improve the system by increasing M.

3 M-ARY BI-ORTHOGONAL SIGNALLING

For each of a set of M orthogonal signals there is a set of M negative signals. The matched filter for the detection of the signal $s_m(t)$ will give a negative output for $-s_m(t)$. Signals from a set of orthogonal signals and their negatives we call bi-orthogonal. From $\frac{M}{2}$ orthogonal signals and their negatives we have an M-ary set of bi-orthogonal signals. An advantage of using bi-orthogonal signals is that we only need half the number of matched filters and half the band width compared with orthogonal signals. However, as we might expect there is some disadvantage because a higher s.n.r. is needed for the same probability of error.

For bi-orthogonal signals the probability of symbol error is, [1],

$$P_m = 1 - \frac{1}{\sqrt{2\pi}}\int_{-\sqrt{2\gamma}}^{\infty} exp\left[\frac{-u^2}{2}\right]\ du \left(\frac{1}{\sqrt{2\pi}}\int_{v+\sqrt{2\gamma}}^{v+\sqrt{2\gamma}} exp(\frac{-x^2}{2})dx\right)^{m/2-1}$$

This is shown in Fig.2. We can note that the cases of M=2 and M=4 correspond to p.s.k. and q.p.s.k. Also we must note that this figure is for a symbol error rate whilst Fig.1.is for the bit error rate.

4 Reference

1. Proakis. J.G.:"Digital Communications", Mc-Graw Hill Book Co.

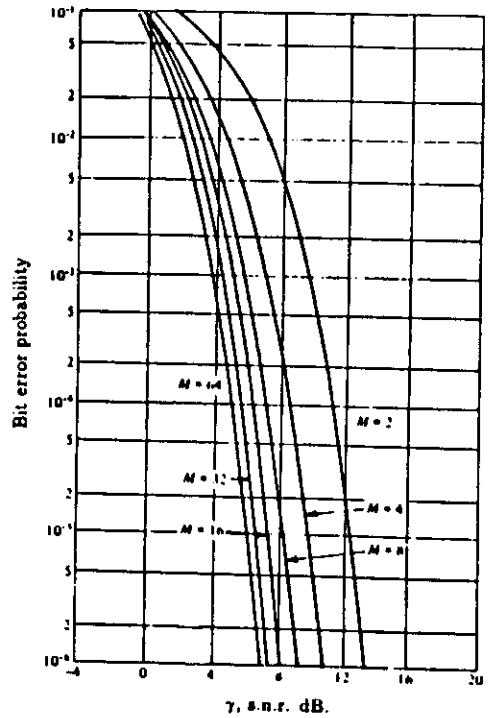
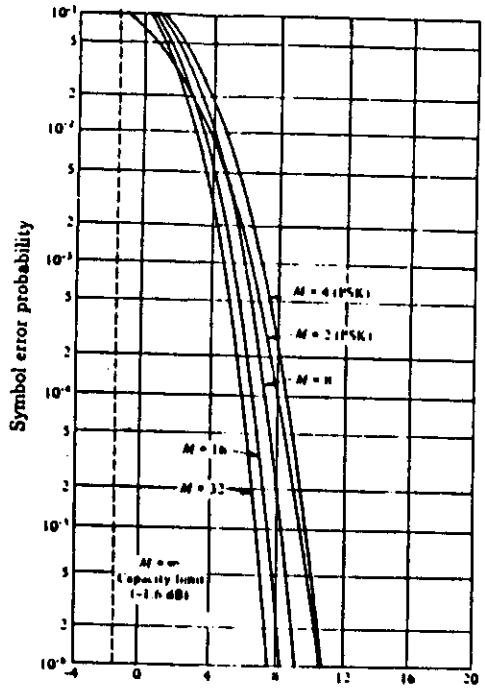


Fig.1. Probability of bit error for coherent detection of orthogonal signals.



INFORMATION THEORY AND COMMUNICATION ENGINEERING

PETER A. MATTHEWS

December 1988

1 INTRODUCTION

Engineering is the pursuit of the ideal, a goal rarely if ever achieved. In this pursuit we need to know what the goal is else we may waste our efforts in trying to attain an impossible objective. Information theory gives us the goal at which to aim in communications engineering but it dose not tell us how to reach that goal.

In engineering we deal with things that can be measured. In communications we must be able to measure the information being transmitted through a system. The amount of information we gain on receiving a message depends on the probability of our knowing the content of the message before it is received. If we receive a message each morning that the sun has risen we gain little information because we are pretty sure that the sun will rise. If we receive a message that there has been an eclipse of the sun we may gain little information because we can accurately predict when eclipses will occur. However if we a driving a car we must watch out for traffic signals because we cannot predict when they will change from green to red. If we miss the information given by the signal there may be a nasty crash.

*Professor, Department of Electrical and Electronic Engineering, University of Leeds, Leeds LS2 9JT, UK

The measure of information depends on our prior knowledge of the message before we receive it. We should note that in a technical sense the written words of the numbers making up a message are not important. It is the change in probability of our knowledge which is important.

2 A DIGITAL COMMUNICATION SYSTEM

Information theory can be applied to both analogue and digital communication systems but it is easier to explain the ideas in relation to digital systems. A block diagram of a simple system is shown in fig.1. The data source produces a data stream which passes through the communication channel to the data receiver. In the channel the data stream is distorted and noise is added to the signal. The job of the receiver is to detect the incoming signal and to produce an output which should reproduce the signal from the source. We would like to know how much information we can pass through the channel and whether we can receive the data without error. We need to define what we mean by information, rate of transmission and the capacity of the channel.

2.1 Rate of transmission

Suppose that the data stream is a stream of multi-level symbols transmitted at a constant rate, for example the output of a p.c.m. coder. If the number of levels in each symbol is n then for two symbols the number of possible combinations is n^2. For r symbols the number of combinations is n^r. If the symbols are transmitted at a rate of r symbols/sec and the message lasts for T seconds the total number of combinations is n^{rT}.

If we have a message which lasts twice as long we expect to be able to transmit twice as much information. If we use a logarithmic measure we get the result we expect. Then the information transmitted is given by

$$Information \propto rT \log n \, .$$

If we use logarithms to the base two we have

$$Information = rT \log_2 n$$

or the rate of transmission is

$$R = r \log_2 n \, .$$

Now $\log_2 n$, where n is the number of levels, is just the number of bits required to express the level as a binary number, e.g. if $n = 16$ then the number of bits required is four. Therefore the information rate is expressed in bits/sec is

$$R = r \log_2 n \, .$$

Note that so far we have not said anything about what limits the rate of transmission R or anything about the possibility of errors occurring in transmision.

2.2 Capacity of a channel

Information when it is transmitted through a channel may not be detected correctly at the receiver. There are errors in the received signal. To overcome these errors we can code the signal at the receiver and then we have to decode the signal at the receiver. This coding usually reduces the rate at which information can be passed through the channel but it reduces the probability of error. In a fixed time with coding more information can be passed through the channel without errors than if coding is not used. There will be a maximum rate of transmission R which the channel will pass without error. This rate is called the capacity of the channel C. The capacity is measured in bits/second.

If the information rate R is less than or equal to the channel capacity, $R \leq C$, then Shannon, [1], showed that the information can be transmitted with an arbitrarily low probability of error. If $R > C$ then the error rate cannot be reduced to zero.

3 INFORMATION CONTENT OF A MESSAGE

In communications engineering we are not concerned with the meaning, the semantics, of a message when we speak of the information content of

a message. What we are concerned with is the probability of being able to predict the symbols making up the message. If at the receiver we can predict the symbols of a message before the message is received then no information is gained on receiving the message. If we cannot predict the symbols then a maximum of information is received.

3.1 Information in a binary message

For simplicity suppose that only two symbols can be transmitted, 0 and 1. If the probability of 0 being transmitted is p then the probability of 1 being transmitted, q, is $q = (1 - p)$. We define the information on transmitting 0 to be $-\log_2(p)$ and the information on transmitting 1 to be $-\log_2(q)$. If the symbol transmission rate is r symbols/sec on the average rp $0's$ and rq $1's$ will be transmitted per second. The average information rate is

$$\begin{aligned} H(r) &= -rp \log_2(p) - rq \log_2(q) \\ &= r\,(p \log_2(1/p) + (1-p) \log_2(1/(1-p)) \end{aligned}$$

The average information per symbol measured in bits is

$$H = p \log_2(1/p) + (1-p) \log_2(1/(1-p)) \, .$$

When $p = 0$ or $p = 1$ then $H = 0$. We know the message before it is sent so no information is gained on receiving the message. If $p = 1/2$ then $H = 1$ and H is a maximum. The information per symbol is one bit which is exactly what the message is as we started by saying the symbols in this example were 0 or 1. The variation in H with p is shown in fig.2.

3.2 Information in a message containing random variables

We can generalise this concept of information for messages containing any number of random variables, [2]. Suppose we are dealing with a random variable X which can take up values from the set of values $a_1, a_2, \ldots, a_k$. We say that X has a sample space $\Omega_X = \{a_1, a_2, \ldots, a_K\}$. The *self-information* when X takes the value a_m is defined to be

$$h(a_m) = -\log_2(p_X(a_m))$$

where $p_x(a_m)$ is the probability that the value a_m will occur. If $p_x(a_m)$ is small then $h(a_m)$ is large as we expect. With this definition the self-information is positive, again as we expect.

We expect that if we have two independent events the total gain in information from the two events should be the sum of the information in the two events taken separately. If the two random variables are X and Y and $\Omega_Y = \{b_1, b_2, \ldots, b_N\}$ the self-information resulting from the observation of the two variables is, as the two variables are independent

$$\begin{aligned} h(a_m, b_n) &= -\log_2(p_{X,Y}(a_m, b_n)) \\ &= -\log_2(p_X(a_m)) - \log_2(p_Y(b_n)) \\ &= h(a_m) + h(b_n) \end{aligned}$$

If we have a sampled and quantized system we can convey the information about the state of the system at the sampling times by binary symbols. If there are only two quantizing levels we need one bit $(0, 1)$, to convey the level. With two bits we can convey four levels, with three bits, eight levels and so on. By using $\log_2$ we have a direct relation between the measure of information and the number of bits needed to transmit that information. Even when the information is in the form of a continuous variable we can still use $\log_2$ to measure the information in bits.

When we send a long message made up of many symbols the probabilities of the separate symbols occurring may differ. However we can take an average to get the average information conveyed by the symbols in the message. The average is

$$\begin{aligned} H(X) &= E[-\log_2(p_X(X))] \\ &= -\sum_{x \epsilon \Omega_X} pX(x) \log_2 p_X(x) \end{aligned}$$

The quantity H is called the entropy in X. It is the average information obtained on receiving the message *or* the average uncertainty about X before the message has been received *or* the average uncertainty removed by receiving X. This average information is measured in *bits*.

4 CHANNEL CAPACITY AND NOISE

For a noise free channel the capacity is

$$C = r \log_2 n$$

If the channel is bandlimited with a bandwidth BHz the rate at which symbols can be transmitted is

$$r = B$$

so that

$$C = B \log_2 n$$

If noise is added to the signal in transmission and the noise has a uniform spectral density across the bandwidth of the channel Shannon showed that with suitable coding the capacity of the channel measured in bits/sec is

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

where S is the average signal power and N is the average noise power.

If the signal being transmitted is a binary signal and the transmission rate is R bits/sec, $R < C$ then the probability of error is bounded by

$$P_e \leq 2^{-E(C,R)T}$$

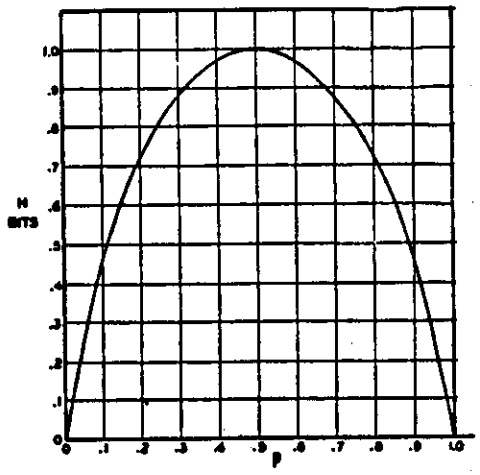
where $E(C, R)$ is a monotonically decreasing function and T is the time taken to transmit the encoded message. If R and C are fixed the probability of error can be decreased by increasing T.

5 REFERENCES

1. SHANNON, C.E. and WEAVER, W.: "The mathematical theory of communication", 1949, University of Illinois Press, Urbana, USA.

2. LEE,E.A. and MESSERSCHMITT, D.G.: "Digital communication", 1988, Kluwer Academic Publishers, Boston, USA.

Figure 1: Block diagram of a simple communication system.



RADIO RELAY SYSTEMS

PETER A. MATTHEWS

December 1988

1 INTRODUCTION

The term 'radio relay system' is used to describe point-to-point communication systems operating at microwave frequencies. At these frequencies high gain antennas with narrow beamwidths can be used and transmitter powers of a few watts are sufficient to provide reliable communication over link distances up to about 50Km. At these frequencies a line-of-sight path is needed between transmitter and receiver. To cover long distances a number of links in series are used. At the ends of the system and at some intermediate stations there will be connections between the microwave system and the local telephone network. This requires modulators and demodulators to transfer the baseband signals to the microwave carrier. At other intermediate relay stations there is no need to connect to a local network. For comparative purposes hypothetical reference circuits have been defined by CCIR for systems using analogue and digital modulation techniquees.

Before the introduction of digital transmission in telecommunication networks most radio relay systems used frequency modulation with a small frequency deviation. Frequency modulation used in this way does not give a significant overall improvement in s.n.r. on demodulation but it has the advantage that non-linearities in the system produce less distortion than if a.m. is used. As the bandwidth available for radio relay systems laid down in international agreements is limited the bandwidth occupied must be

*Professor, Department of Electrical and Electronic Engineering, University of Leeds, Leeds LS2 9JT, UK

minimised if systems are to carry a large number of telephone or television channels.

When f.m. is used the signal at a repeater station is normally converted to an intermediate frequency, amplified, and re-transmitted on a different carrier frequency to the next station along the system. In this process noise is added to the signal. At the next station this added noise appears to be part of the signal and is amplified whilst more noise is added. As the signal is transmitted through the system the noise level builds up.

Using digital modulation techniques the signal at relay stations can be regenerated. In a regenerator the digital signal is reformed before it is re-transmitted. Although there may be some errors in this process the noise level does not add up through the system. The overall error rate can be kept very small for a large proportion of the time. However for a small proportion of the time errors will occur because of distortion in the system. In digital systems distortion is more of a problem than noise.

2 DIGITAL TRANSMISSION SYSTEMS

In digital transmission systems there is a hierarchy of bit rates used in transmission. These rates are internationally agreed but there is more than on set in use. They are shown in table I. Because of the need to conserve r.f. bandwidth it is common to use some form of multi-level modulation. This may be q.p.s.k. in simpler systems to 64-QAM in more advanced systems. The most modern systems are designed to use up to 256-QAM. Such systems require the use of linear amplifiers and transmitters to preserve the amplitude and phase relations in the signal. Whilst errors due to added noise can be reduced by increasing the transmitter power this does not reduce the errors due to distortion.

3 SOURCES OF DISTORTION IN RADIO RELAY SYSTEMS

Distortion may occur in the equipment of a system, e.g. the amplifiers, filters, and waveguides, it may also occur in transmission between stations.

For distortionless transmission the requirements are that all frequency components in the signal should suffer the same attenuation and all frequency components should be subject to the same time delay. The latter condition means that velocity of propagation should be independent of frequency and that the phase shift should be proportional to the frequency. If the phase shift is not proportional to frequency the signal spectrum at the input to the demodulators in the system will be distorted and so will be the output.

It is common to express the phase distortion in terms of the group delay through the system. The group delay depends on the rate of change of phase with frequency or the slope of the phase characteristic. If the phase characteristic as a function of frequency is

$$\phi = f(\omega)$$

then the group delay characteristic is

$$\tau = \frac{d\phi}{dt} = \frac{d\,f(\omega)}{dt}\,.$$

Components in the system which have a non-linear phase characteristic are in particular filters. However filters can be designed to have a largely linear phase response and it is possible to include in the design sections to equalise the group delay across the band of interest. In many systems there are long waveguide feeds between the equipment and the antennas. If there are reflections in the feed system the signal at the antenna or the receiver is the sum of a number of components. As the different components arrive with different time delays depending on the number of times they have been reflected up and down the guide the phasor sum will vary with frequency. Careful design and installation will reduce the amount and effect of these reflections.

A similar effect is multipath propagation which may occur on the radio path between transmitter and receiver. This multipath propagation occurs because of conditions in the atmosphere which lead to abnormal changes in the vertical structure of the refractive index profile. The refractive index in the lower atmosphere normally falls with height and near the ground the fall is approximately linear with height. However under some atmospheric conditions there can be sharp changes in the refractive index. When such changes occur the signal can be refracted or reflected from the boundary

and the result is that the signal arrives at the receiver by more than one path. The effect on the frequency transfer function of the path is to cut a notch in the amplitude response and associated with the notch is a change in the group delay characteristic. fig.1. Because this effect depends on the conditions in the atmosphere it changes as the atmosphere changes. The distortion is time varying and un-predictable. Although the atmospheric conditions leading to this kind of distortion occur infrequently on most radio relay paths there are some paths, particularly those near the sea and in hot climates which are most liable to problems. This is because the variations in the refractive index depend to a large extent on the humidity in the atmosphere.

To overcome the effects of atmospheric multipath propagation adaptive equalisers may be used in the system. Another method of overcoming the effects of multipath is to use some method of diversity reception. In diversity reception the aim is to receive two or more signals for which the effects producing distortion are un-correlated. The methods which may be used are frequency diversity, space diversity, and angle diversity. To show the effect of multipath propagation consider a two path model with a time delay of say 5nsec between the two paths. With this diference in time there will be minima in the frequency transfer characteristic spaced 200MHz apart in the frequency domain. As typical system r.f. bandwidths are less than 50MHz if carier frequencies are spaced say 100MHz apart then if one signal is in a null in the frequency response the second will be at a maximum. The diversity system may switch between the two signals or they may be combined with a weighting system which takes advantage of the better signal received on one of the channels. A disadvantage of a frequency diversity system is the extra bandwidth required. It is better if the two independent signals can be obtained form a single transmitted signal. This is possible using spaced receiving antennas. If the antennas are more than a few wavelengths apart the phasor sum of the multipath signals received on one antenna will be different to that on the second antenna. Another way in which independent signals can be obtained is by having receiving antennas which collect their signals from different parts of the atmosphere. This can be achieved by using a dual feed system within a single receiving antenna or by using two antennas with different beamwidths, [1].

4 PERFORMANCE OBJECTIVES FOR RADIO RELAY SYSTEMS

To compare different systems and to ensure satisfactory performance of systems which may form part of an international system CCIR have produced recommendations and reports for both analogue and digital radio relay systems, [2]. These recommendations give values for the s.n.r. or b.e.r. which should not be exceded in a hypothetical reference circuit and in working systems. They are revised and up-dated every three years.

5 REFERENCES

1. LIN, S.H., LEE, T.C., GARDINA, M.F.: 'Diversity protections for digital radio-Summary of ten-year experiments and studies', IEEE Communications Magazine, 1988, 26, (2), p.51-63.

2. Recommendations and Reports of the CCIR, Vol.IX. Fixed services using radio-relay systems. ITU, Geneva.

BROADCASTING

PETER A. MATTHEWS

December 1988

1 INTRODUCTION

Broadcasting systems are different from other types of radio system in that one transmitter serves many receivers. Because the receivers are distributed amongst the general population there is a need to design the system so that the receivers will be inexpensive. Because of the small number of transmitters compared with the large number of receivers it is better to put costs into the transmitters rather than the receivers.

In the past to make low cost receivers the receivers were made simple in terms of the electronic circuits. This is necessary when circuits are made from elementary components, resistors, capacitors, inductors and transistors but as more and more parts of the receiver can be made from integrated circuits costs no longer depend on the number of circuit elements. Costs depend on the number of i.c.s. used, the development costs of the i.c.s. and the number of receivers produced using similar i.c.s.. Nowadays it is possible to use complicated circuits provided they are used in large enough quantities.

Another factor which controls the introduction of new systems, e.g. those using new methods of modulation or new frequency bands, is the need for compatibility between new systems and old. Radio receivers can last many years and people expect the system not change. Whilst it is relatively easy to introduce new systems for new services it is an unpopular move to alter a system so that existing receivers have to be replaced. This factor can introduce a large time constant into the alteration of existing services, e.g. the move of t.v. services from v.h.f. to u.h.f. or the adoption of s.s.b. in sound broadcasting.

*Professor, Department of Electrical and Electronic Engineering, University of Leeds, Leeds LS2 9JT, UK

2 SOUND BROADCASTING

Sound broadcasting services for general information and entertainment have been used since the 1920's. These services are provided in the l.f., m.f., h.f., and v.h.f. bands. The characteristics of the radio propagation path will be discussed in the second part of this course. Briefly l.f. is used for national and international services at distances of up to about 1000 km. M.f. is used for regional services at distances up to about 200 km. H.f. is used mainly for long distance international broadcasting throughout the world. V.h.f. is used for local broadcasting. The propagation characteristics of the h.f. bands and at m.f. depend on the properties of the ionosphere and are very variable with diurnal, annual and sunspot cycles. The characteristics of l.f. propagation are much less affected by the variations in the ionosphere. The propagation characteristics at v.h.f. do not depend on the ionosphere but do depend on meteorological conditions in the lower atmosphere, the troposphere.

Sound broadcasting is used for the transmission of both speech and music. Whilst intelligible speech can be transmitted as an analogue signal in a bandwidth of 3 kHz a wider bandwidth is required for high fidelity reproduction of speech and music.

If we consider the total bandwidths available for broadcasting, for example at l.f. where the band is 150 - 250 kHz or m.f. where the band is from 500 - 1500 kHz, then using double sideband a.m. with a signal bandwidth of 15 kHz only 3 transmissions could be accomodated at l.f. and 33 at m.f.. This does not allow any guard band between adjacent transmissions. Such a small use of the bands is not internationally acceptable and in practice in these bands transmissions are spaced at 9 KHz intervals. Transmitters may operate on the same frequency provided they are far enough apart but at l.f. this implies inter-continental distances and at m.f. inter-national distances.

At h.f. there is more bandwidth available but propagation conditions may be such that the signals are receivable at any part of the world. There is need for the international coordination of the use of frequencies to minimise interference. At h.f. because of the variable nature of the ionosphere the bandwidth which can usefully be used is limited. At these frequencies because of multi-path propagation the transfer function of the radio path can vary sufficiently across a bandwidth of a few kiloHertz to cause severe time varying distortion. The bandwidth used at h.f. is limited because of these factors. There is an advantage in using s.s.b. both to reduce distortion and to increase the number of transmission which can be accomodated in a given bandwidth.

At v.h.f. much more bandwidth is available. In most countries the band between 88 and 108 MHz is used for sound broadcasting, a total bandwidth of 20 MHz which is comparable with the whole of the h.f. band and 20 times the bandwidth available at m.f.. Because the service area of a transmitter is limited to near line-of-sight paths frequencies can be reused at distances of less than 200 km although care has to be taken in planning the use of frequencies to avoid interference during abnormal propagation conditions. Because a wide band is available wideband f.m. transmissions can be used to provide high-fidelity stereo sound services.

3 TELEVISION BROADCASTING

Because television entertainment broadcasting services require radio bandwidths of 6-8 MHz they can only be provided at v.h.f. and u.h.f.. Whilst t.v. broadcasting started at low v.h.f. these services are now provided at high v.h.f. or u.h.f.. At u.h.f. most of the band from 450 - 850 MHz is used for t.v. broadcasting. In this band a single receiver tuner can be used to cover the whole band. The area covered by one transmitter is again confined to near line-of-sight paths and frequencies can be reused over relatively short distances. However, a large number of transmitters is needed to provide an adequate service over a large percentage of the population.

By using satellite transmissions much larger areas can be covered by a single satellite transmitter. Satellite transmissions are used for the distribution of programmes and for direct broadcasting to domestic receivers.

For d.b.s. high power satellite transmitters are needed and high microwave frequencies are used. The use of high powers and high frequencies allows the use of relatively small receiving antennas. These services can best be provided using new modulation techniques which are different to those used a v.h.f. and u.h.f..

Because wide bandwidths are available using satellite transmissions it becomes possible to use new high definition systems. Such services can provide a much better quality picture and setting new standards allows the introduction of stereo sound and multi-lingual transmissions without the need to maintain compatability with existing systems.

4 RECENT ADVANCES IN BROADCASTING

The recent advances in broadcasting are largely inaudible or invisible to the casual listener or viewer. Because of the need for compatibility the opportunities for improving the operation of receivers are limited to extracting the best response from receivers. However, there is a wide range of qualities in receivers and few receivers make the best use of the available signals.

In studios and transmitters however, there are advances concerned with improving the performance and improvement of the system. Many of the improvements have become possible by the introduction of digital techniques in the transmission chain between studio and transmitter and in the control and monitoring of systems. Using p.c.m. for signal transmission maintains the quality of the signal both in transmission around studio and in distribution networks. Standards have been established for the transmission of both sound and television signals. In setting up such standards there is a need for compatability with standards for sound and video recording.

Using digital techniques the creation of many effects becomes possible. These effects are widely used by advertisers and can be very eye-catching. Similar techniques allow the rapid preparation of material for services such as news and sport. Photographs can be stored digitally and rapidly recalled for incorporation into programmes.

When digital transmission is used it becomes easier to measure the quality of transmission. This particularly important in large networks. The

quality can be measured and reported automatically from remote transmitters. Staff no longer need be employed to monitor performance at remote stations and service teams can be deployed more economically from regional service stations.

Using digital techniques additional services can be provided which ride on existing transmissions. Data systems can be provided on sound radio. These systems can be used for the automatic tuning of receivers and for information services, for example about traffic conditions. Data can be transmitted at sub-audio frequencies, stored and displayed whilst normal transmissions are taking place. Other services can take advantage of the bandwidth offered by t.v. transmissions for down loading data during the normally quiet hours of the transmitters. Such added services can provide a valuable financial contribution to the running of broadcast systems. Finally using digital techniques it becomes relatively simple to provide encryption to prevent casual eavesdropping on commercial services and to provide means for charging for extra services.

5 REFERENCES

1. Reports and Recommendations of the CCIR. Vol.X. Broadcasting (Sound). Vol.XI. Broadcasting (Television). ITU, Geneva.

RANDOM VARIABLES AND PROBABILITY

PETER A. MATTHEWS

December 1988

1 INTRODUCTION

In communications systems we have to deal with both deterministic and random or stochastic signals and noise which by its nature is random. We also have to deal with both continuous and discrete valued signals. Random signals and noise are types of random variables. Random variables are described by probability distributions. When the variables are discrete valued we describe them as a set of discrete values. To read the literature on such random variables we must become familiar with the terminology used.

For discrete valued variables the notation of *sets* is used. For example if we carry out an experiment with a six-sided die the *sample space* S of the experiment is the *set* of all possible outcomes

$$s = \{1, 2, 3, 4, 6\} .$$

The brackets { } are used to denote a set.

An *element* of the set S is one member of the set. Thus 3 is an element of the set S. This is written $(3 \epsilon S)$ in which ϵ stands for 'an element of'.

*Professor, Department of Electrical and Electronic Engineering, University of Leeds, Leeds LS2 9JT, UK

An *event* is a *subset* of S which may consist of one or more of the elements in S. For example an event A may be defined by

$$A = \{3,4\}$$

The event A has occurred when an experiment produces the outcomes 3 and 4.

The *complement* of A, denoted by $\overline{A}$, is the set of samples in S which does not include A, in this case

$$\overline{A} = \{1,2,5,6\} .$$

Two events are *mutually exclusive* if no outcomes included in the first set occur in the second set. Thus if $A = \{3,4\}$ and $B = \{1,5,6\}$ then A and B are mutually exclusive.

The *union* of two events is the event which includes the events in the two separate events. Thus if $A = \{3,4\}$ and $C = \{1,3,6\}$ then the union of A and C is the event $D = \{1,3,4,6\}$. This is written

$$D = A \cup C .$$

The union of A and its complement $\overline{A}$ is the entire sample space

$$A \cup \overline{A} = S .$$

The *intersection* of two events is an event consisting of the outcomes which are common to the two events. Thus E the intersection of A and C, which is written

$$E = a \cap C$$

is the event $E = 3$.

When events are mutually exclusive their intersection is the *null set* ϕ. Thus using the values above

$$A \cap B = \phi .$$

By definition $A \cap \overline{A} = \phi$.

The set of *real numbers* is denoted in these notes by $\Re$.

2 RANDOM VARIABLES

A *random variable* is denoted by a capital letter, e.g. X. The *outcome* of a random variable is denoted by a lower case letter, e.g. x. The random variable may be a real or a complex-valued function defined over a *sample space* Ω of all possible outcomes. An *event* E is a set of possible outcomes and may occur with a probability $P[E]$ where $0 \leq P[E] \leq 1$.

Using set notation

$$P[E_1 \cup E_2] = P[E_1] + P[E_2] - P[E_1 \cap E_2]$$

which gives the *union bound*

$$P[E_1 \cup E_2] \leq P[E_1] + P[E_2] .$$

The *cumulative distribution function*, c.d.f., is the probability that $(X \leq x)$ and is denoted by

$$F_X(x) = P[X \leq x] .$$

For a complex-valued random variable Y the c.d.f. is

$$F_Y(y) = P[\Re(Y) \leq \Re(y), \Im(Y) \leq \Im(y)]$$

For a continuous-valued random variable the *probability density function*, p.d.f., $f_x(x)$ is defined so that for any interval $I \subset \Re$

$$P[X \in I] = \int_I f_X(x)dx .$$

For a complex-valued random variable I is a region of the complex plane.

For a real-valued random variable X

$$f_X(x) = \frac{d}{dx} F_X(x) .$$

If the c.d.f. includes a step function the corresponding p.d.f. has a delta function at the step.

For a discrete-valued random variable X the probability of a particular outcome $(x \in \Omega)$ is

$$p_X(x) = P[X = x] .$$

The p.d.f. is

$$f_X(x) = \sum_{y \epsilon \Omega_X} p_X(y)\delta(x-y) .$$

The *expected value* or *mean* of X for a continuous variable is

$$E[X] = \int_{-\infty}^{+\infty} x f_X(x) dx$$

and for a discrete-valued variable it is

$$E[X] = \sum_{x \epsilon \Omega} x p_X(x) .$$

For a complex-valued variable the integration is taken over the complex plane.

If a function $g(\cdot)$ is defined over the sample space X then

$$E[g(x)] = \int_{-\infty}^{+\infty} g(x) f_X(x) dx .$$

The *mean* of X, denoted by μ is

$$\mu = E[X] ,$$

the *variance* of X, the second moment, is denoted by

$$\begin{aligned} \sigma_X^2 &= E[(X - E[X])^2] \\ &= E[X^2] - E[X]^2 . \end{aligned}$$

For a complex-valued random variable

$$\begin{aligned} \sigma_X^2 &= E[|X|^2] - |E[X]^2| \\ &= E[XX^*] - E[X]E[X]^* . \end{aligned}$$

The *joint c.d.f.* of two random variables X and Y is

$$\begin{aligned} E_{X,Y}(x,y) &= P[X \leq x, Y \leq Y] \\ &= \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(p,q) dp dq \end{aligned}$$

where $f_{X,Y}(x,y)$ is the joint p.d.f. Conversely

$$f(x,y) = \frac{\partial}{\partial x \partial y} F(x,y) .$$

The *marginal density*, $f_X(x)$ of the random variable X of the pair of variables X and Y is

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y) dy .$$

The random variables X and Y are *statistically independent* if over the intervals I and J

$$P[X \epsilon I \cap Y \epsilon J] = P[X \epsilon I] P[Y \epsilon J]$$

or

$$f_{X,Y} = f_X(x) f_Y(y)$$

or

$$F_{X,Y} = F_X(x) F_Y(y) .$$

For independent events the *cross-correlation*

$$E[XY] = E[X]E[y]$$

and the events are *uncorrelated.*

The *characteristic function* of X is

$$\begin{aligned} \Phi_X(s) &= E[\exp(sX)] \\ &= \int_{-\infty}^{+\infty} \exp[sX] f_X(x) dx \end{aligned}$$

where s is a complex variable. The characteristic function is the Laplace transform of X at $x = -s$. When s is real-valued the characteristic function is called the *moment generating function.*

If $Z = X + Y$ and X and Y are independent the

$$\Phi_Z(s) = \Phi_x(s) \Phi_Y(s) .$$

The mean of X is

$$E[X] = \frac{\partial}{\partial s}\Phi_X(s)|_{s=0}$$

and

$$E[X^2] = \frac{\partial^2}{\partial s^2}\Phi_X(s)|_{s=0} .$$

The probability that $(X > x)$ is bounded by

$$\begin{aligned} 1 - F_X(x) &= P[X > x] \\ &\leq \exp[-sx]\Phi_X(s) \end{aligned}$$

for any real-valued $s \geq 0$. This is called the *Chernoff bound.*

Also

$$F_X(x) \leq \exp[sx]\Phi_X(-s)$$

for $s \geq 0$. The value of s which makes the bound tightest must satisfy

$$x\Phi_X(s) = \frac{\partial}{\partial s}\Phi_X(-s)$$

and

$$x\Phi_X(-s) = \frac{\partial}{\partial s}\Phi_X(-s) .$$

3 CONDITIONAL PROBABILITIES

The *conditional probability* that a continuous-valued random variable X is in the interval I given that Y is in the interval J is defined, for all values of J for which $P[Y \in J] \neq 0$, by

$$P[(X \in I)|(Y \in J)] = \frac{P[(X \in I) \cap (Y \in J)]}{P[(Y \in J)]}$$

where $P[(Y \in J)]$ is called a marginally probability as it does not take into account any effect that X may have on Y.

If X and Y are independent then

$$P[(X \in I)|(Y \in J)] = P[(X \in I)] .$$

For a continuous-valued variable

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} , \quad f_Y(y) \neq 0 .$$

As $f_{X,Y}(x,y) = f_{Y,X}(y,x)$ it follows that

$$f_{X|Y}(x|y)\, f_Y(x) = f_{Y|X}(y|x)\, f_X(x) .$$

4 GAUSSIAN RANDOM VARIABLES

The *Gaussian* or *normal* p.d.f. is

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-(x-\mu)^2/2\sigma^2]$$

where μ is the mean and σ^2 is the variance of the distribution.

The c.d.f. is

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp[-(\alpha-\mu)^2/2\sigma^2] d\alpha .$$

There is no closed form expression for the c.d.f.. When $\mu = 0$ and $\sigma = 1$ the distribution is called the standard or normalised Gaussian distribution function.

The c.d.f. of the standard Gaussian distribution is called the $Q(x)$ function where

$$\begin{aligned} Q(x) &= P[X > x] = 1 - F_X(x) \\ &= \frac{[}{1}][\sqrt{2\pi} \int_{-\infty}^{x} \exp[-\alpha^2/2] d\alpha . \end{aligned}$$

The Q function is related to the *error function* and the *complimentary error function* by

$$\begin{aligned} Q(x) &= \frac{1}{2} erfc[\frac{x}{\sqrt{2}}] \\ &= \frac{1}{2}\left[1 - erf[\frac{x}{\sqrt{2}}]\right] . \end{aligned}$$

For a Gaussian random variable X with meanμ and variance σ^2

$$P[X > x] = Q\left(\frac{x-\mu}{\sigma}\right) .$$

The momnet generating is given by

$$ln[\Phi_X(s)] = \mu\sigma + \frac{1}{2}\sigma^2 s^2 .$$

The Chernoff bound is

$$1 - F_X(x) \leq \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]$$

and so

$$Q(x) \leq \exp[\frac{-x^2}{2}] .$$

A tighter bound is given by

$$\frac{1}{x\sqrt{2\pi}}(1 - \frac{1}{x^2}\exp[\frac{-x^2}{2}] < Q(x), \frac{1}{x\sqrt{2\pi}}\exp[\frac{-x^2}{2}] .$$

This bound differs little from the true value for $x > 3$.

For two zero-mean Gaussian random variables having the same variance the joint p.d.f. is

$$F_{X,Y}(x,y) = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}}\exp[\frac{-(x^2 - 2\rho xy + y^2)}{2\sigma^2(1-\rho^2)}]$$

where ρ is the correlation coefficient of the two variables given by

$$\rho = \frac{E[X,Y]}{\sigma^2}$$

and $-1 \leq \rho \leq +1$.

When the two variables are uncorrelated $\rho = 0$ and then

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{1}{2\pi\sigma^2}\exp[\frac{-(x^2+y^2)}{2\sigma^2}] \\ &= \frac{1}{2\pi\sigma^2}\exp[\frac{-r^2}{2\sigma^2}] \end{aligned}$$

where $r^2 = x^2 + y^2$. This distribution is the *Rayleigh amplitude p.d.f.*.