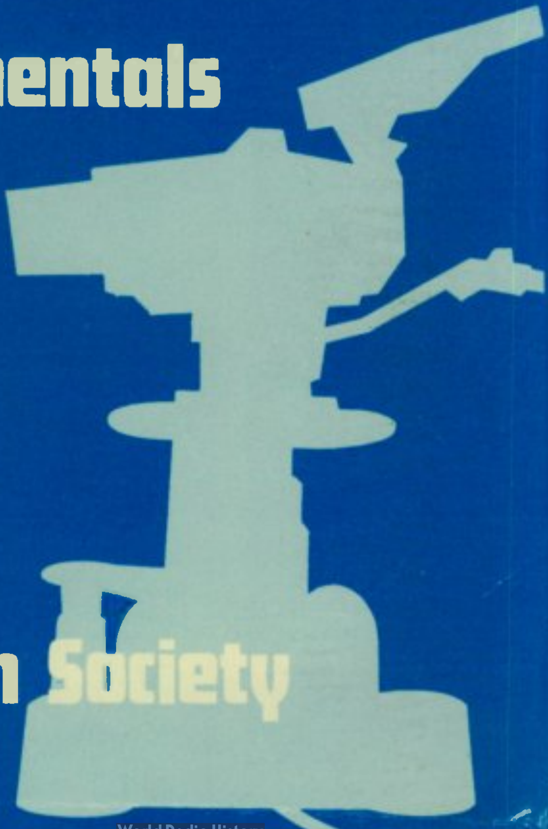
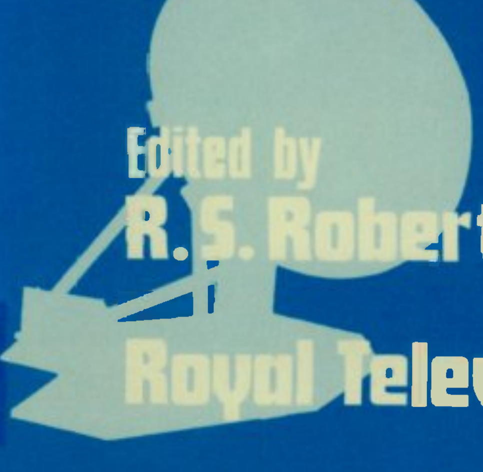
TELEVISION ENGINEERING

Broadcast, Cable and Satellite

Part 1: Fundamentals





Edited by

R. S. Roberts

Royal Television Society

PRINCIPLES OF DIGITAL DATA TRANSMISSION, 2nd edn.

A. P. Clarke, Professor of Telecommunications, Loughborough University

1983 0-7273-1613-3 (cloth), 1614-1 (paper) 308 pp. illustrated.

This second edition provides an up-to-date textbook suitable for presentation as a course on digital communications at either final-year undergraduate or first-year postgraduate level and is a valuable source of reference for practising engineers.

Contents: Introduction; Distortion; Noise; Transmission of Timing Information; Modulation Methods; Ideal Detection Processes; Practical Detection Processes; Transmission Rates; Binary Data-Transmission Systems for use over Telephone Circuits; Multi-Level Data-Transmission Systems for use over Telephone Circuits; Data-Transmission Systems for use over HF Radio Links; Matched-Filter Detection; Rectangular Baseband Signals; Rounded Baseband Signals; Partial-Response Channels; Modulated-Carrier Signals; References; Index.

HIGH PERFORMANCE LOUDSPEAKERS, 3rd edn.

Martin Colloms, Audio and Electronics Consultant

1985 328 pp. illustrated 0-7273-0806-8

Considerable changes have taken place in high performance loudspeaker design since publication of the second edition in 1980. Topics which have been updated or appear for the first time include low mass honeycomb speaker enclosures; open baffle systems; electrostatic speaker theory; time delay spectrometry; new application of shaped impulses for speaker analysis; low distortion magnet systems; new diaphragm materials; computer controlled tests systems; time delay networks; high power pulsed distortion measurements; power handling; listening room design; system acoustic adjustability.

Contents: Introduction; Fundamental principles of drive units and loudspeaker acoustics; The moving coil family; Loading; Systems and crossover networks; High power systems; Measurements, subjective and objective evaluation; Enclosures; Loudspeakers and the environment; The design of a compact system; References; Index.

TRANSMISSION AND DISPLAY OF PICTORIAL INFORMATION

D. E. Pearson, Department of Electrical Engineering Sciences, University of Essex

1975 220 pp. illustrated 0-7273-2001-7

Essential reading for all communication engineers concerned with broadcast and cable television, video-telephone systems, facsimile, computer generated visual displays, scanning electron microscopes, etc. whose job requires a clear understanding of the nature of video signals and their efficient transmission. Lecturers will find the material particularly suitable for use as a course textbook in television systems at advanced undergraduate and post-graduate level.

Contents: The mathematical analysis of images; An introduction to multi-dimensional Fourier theory; Properties of the eye affecting system design; Scanning; Reception and display of monochrome pictures; Transmission of monochrome information; Displays of colour; Transmission of colour information; Subjective assessments of picture quality.

PENTECH PRESS LIMITED

Marketed by JOHN WILEY & SONS LTD., Baffins Lane, Chichester, Sussex PO191UD

TELEVISION ENGINEERING

Technical advances in television engineering have progressed rapidly since a broadcast service commenced in the UK in 1936. Training for the very broad field of television technology is carried out by 'in house' departments in many broadcasting and manufacturing organisations and, of course, at polytechnics and technical colleges but there have been very few publications directly related to training.

These two volumes are based on the Royal Television Society's Television Engineering Course which the Society has run each year for the past four years. The chapters are based on lectures presented by the authors, many of whom are acknowledged experts in their own field and of international standing.

The contents range from basic fundamentals in Part I, to specific engineering applications in Part II. Students of television will need Parts I and II and although practising engineers will be more interested in Part II, they will nevertheless find Part I a valuable reference for fundamentals.

The subject treatment is unique in many cases. For example, Chapter 2 on 'Modulation' does not confine itself to the types of modulation to be found in any current television system. It deals with modulation in a comprehensive manner that will be basic to any further developments that may take place in the future. Chapter 8 on 'The NTSC and PAL systems' is also a comprehensive treatment which is basic to any system which uses a sub carrier. The first sub-carrier colour broadcast system was the US NTSC system which was used to launch transmission in 1955. The NTSC system is dealt with in depth because it is the basis of the later variants such as PAL and SECAM.

Another chapter with special features is Chapter 11 on 'Camera Tubes and the Camera' where the author has brought together a large amount of widely-scattered information while Chapter 16 on 'The Receiver' is probably the first concise treatment of the modern television receiver.

TELEVISION ENGINEERING

Broadcast, Cable and Satellite

Edited by R. S. Roberts,
Royal Television Society

Part 1: Fundamentals

1985 0-7273-21048

Contents:

TELEVISION ENGINEERING

Part 1

Contributing authors

Mrs. G. B. M. Claydon (Chapter 9)
formerly Marconi Communications Systems
now Science Museum, London

J. L. Eaton (Chapter 12)
Consultant, formerly
BBC Research Department

A. Fremont (Chapter 6)
Marconi Communications Systems

I. E. Gibson (Chapter 11)
Marconi Communications Systems

Dr R. D. A. Maurice, OBE (Chapters 2 and 8)
Consultant, formerly
BBC Research Department

R. S. Roberts (Chapters 1, 4, 5 and 7)
Consultant, formerly
Department of Telecommunications
Polytechnic of North London

P. Sproxton (Chapter 3)
Abekas Video Systems, Ltd.

R. M. Stevens (Chapter 10)
Thames Television Ltd

TELEVISION ENGINEERING

Broadcast, Cable and Satellite

Part 1: Fundamentals

Edited by

R. S. Roberts
Royal Television Society

PENTECH PRESS
London

First published 1985 by
Pentech Press Limited
Graham Lodge, Graham Road
London NW4 3DG

British Library Cataloguing in Publication Data

Television engineering.
1. Television
I. Roberts, R.S.
621.388 TK 6630

ISBN ~~0-7273-2104-8 Part 1~~
ISBN 0-7273-2105-6 Part 2

Typeset by Mid-County Press, London SW15
Printed in Great Britain

Preface

This book has its origin in a Course on Television Engineering which has been organised by the Royal Television Society for several years. The Course consists of some 27 or so lectures each Autumn and Winter session and, over the years, succeeding Course programmes have been up-dated and have benefitted from the experience of the preceding Courses. As a result, a framework for the Course has evolved which appears to satisfy several levels of need in Television Engineering.

The book has followed the same general pattern of the Course. It consists of two parts. The first twelve chapters deal with basic fundamentals, and the remaining ten chapters deal with the specialised applications that constitute modern television engineering practice.

The Course has indicated that there are four main types of engineer to whom the book should be of interest. There is the young graduate engineer who is going into television as a career. There is the engineer who has been in television for some time, and who requires some revision and, possibly, a study of areas in television that had not been his main interest. There is the engineer who is engaged on the 'fringe' of television activities (e.g. film, lighting, etc.), and who requires a wider knowledge of the subject. A fourth type of possible interest is the engineer in associated disciplines, such as communications, who wishes to change his main interest into television, or who finds that he cannot avoid a concern with television.

The lecturers are experts in their particular subjects, and many are of international renown. All are experienced 'hands-on' engineers. As a consequence many of the chapters are unique in the sense that the author has pulled together a mass of scattered information, not previously available in a concise form.

Another feature of the specialist authorship for each chapter has resulted in some chapters being authoritative on their particular subject to an extent well beyond the boundaries of television engineering. An example is Chapter 2 on Modulation. The subject has been dealt with fully in terms of all present and possible future types of modulation. It would have been unrealistic to limit this chapter to the three types of modulation used in a typical present-day television channel.

Black-and-white television has been with us since the BBC commenced a regular broadcast service in 1936, but all modern colour television systems have evolved from the US NTSC[1] system, which commenced colour broadcasts in 1955. The PAL[2] system and others are variants of the NTSC system and, therefore, Chapter 8 deals with NTSC in depth, PAL being treated as one of the variants.

Television engineering is dynamic and is the subject of continuous change. It is hoped that this book, in covering present-day techniques will, at the same time, form a useful foundation for understanding techniques likely to be used in the next decade. It will be noted that, in many cases, reference has been made to VHF television and the use of Standard A, 405 lines. These services were discontinued in the UK at the end of 1984 and the former Bands I and III are to be used for various mobile communication services. However, the bands are still in use for television in other countries—not using Standard A.

Grateful acknowledgement is made to the following for their support:

Ampex Ltd
Antiference Ltd
BBC
British Telecom
Crow of Reading
Marconi Communications Systems
Michael Cox Electronics
Mullard Ltd
Rank Cintel
Thames Television

R. S. Roberts

[1] National Television Standards Committee (US).
[2] Phase Alternating Lines.

Contents: Part 1

Chapter 1

Survey of basic principles

R. S. Roberts

Many readers will have some knowledge of the basic principles underlying the engineering of a television system. However, many of the principles are so basic and have such an importance, that it is useful to gather together the most important.

1.1 SEEING

The eye, as a visual communication system, 'sees' a large amount of visual detail simultaneously, by virtue of the fact that it has several million communication channels operating at all times. The electrical signals generated by the millions of sensors in the eye are partly processed in the retina at the back of the eye, and further processed in the brain to provide the familiar human experience of normal vision. The mass of detail forming the visual scene consists of variations in light and shade, colour and, because we have two eyes, perspective. Chapter 5 will deal with the visual experience in more detail.

Picture transmission, using electronic means to convey picture information, cannot be carried out as a simultaneous process, embracing the total field of view, at any one instant. Any telecommunications system can only handle a single bit of information at any instant and, hence, transmission of the information contained in any visual scene, must be analysed in some fashion whereby the complete scene is transmitted as separate bits of electrical information. At the receiver, the individual bits of transmitted information are recovered and processed for display by some form of display unit (usually a cathode-ray tube) into the complete visual scene as viewed at the transmitter.

1.2 SCANNING

The visual scene is explored by examining the small areas of detail that make it up, and the process is termed *scanning*. It is a similar process that we use when reading a page of a book. It is not possible for even the human eye to look at a book page and take in all of the information at a glance. It is necessary to 'scan' the page line-by-line to extract the total visual information. Electronic scanning carries out a similar line-by-line scan process. At the transmitter, the detail encountered during the scan is translated by the camera (or similar scanning device) into voltages that can be used to modulate a radio transmitter. At the receiver, the received signals are demodulated and used to modulate the beam current(s) of a display tube, the beam of which is sweeping in synchronism with the transmitter scanning beam, across the face of the tube.

What constraints does the scanning process impose on any electronic system? One of these which is not too obvious is the need to put a 'frame' round the field of view that is to be transmitted. In the human 'seeing' process the eye is quite unrestricted in its movements and it roams freely over about 180° which, with head movements, provides an unlimited range of view. In the scanning process, a finite limitation must be imposed on the dimensions of the field of view, within which the picture can be analysed line by line.

Many forms of picture transmission and reception have been used for many years. Still pictures are communicated regularly using telecommunications links, by means of a system termed *Facsimile* (in USA, 'FAX'). The picture is wrapped round a drum, and scanned line by line as the drum is rotated and advanced with each turn. At the receiver, a photo-sensitive paper, wrapped round a similar drum, is synchronously rotated past a light beam which is modulated by the received signals. A high-quality picture takes about 12–15 minutes to transmit a 250 × 200 mm picture over a voice communication circuit by this means.

The only difference between transmission of a still or moving picture is one of time. A still picture can take as much time as we wish, but a moving field of view must be totally scanned in a time that is very short, compared with the time being taken by any movements in the field of view. In other words, complete scanning must be so fast that we are concerned with what is, virtually, a still picture.

1.3 MOVING PICTURES

The display of moving pictures has been carried out for many years by the cinema and, more recently, by television. The basic method used is common to both the cinema and television, and is probably well known to most people.

The human eye exhibits a number of properties that we shall be considering at length, but one of these, termed 'persistence of vision' enables display of a moving scene. When the image of a still picture is impressed on the eye, removal of the visual stimulus does not result in an immediate cessation of the signals passed to the brain. An exponential lag takes place with a relatively long time required for total decay. The cinema has exploited this effect by presenting to the eye a succession of still pictures (termed 'frames'), one after the other, each frame differing only by the re-positioning of any moving objects in the field. The presentation of one frame after another does not allow time for the decay to become obvious and providing that the presentation is sufficiently rapid and not too bright an illusion of continuous movement can be sustained. Cinema film presentation consists of a succession of still pictures projected on to a screen. A frame is drawn into position with the light cut off by means of a rotating shutter. As the shutter opens, the frame is stationary and the projected light is exposed through the frame. The shutter cuts off the light, the next frame is drawn into position, the light is re-exposed through the frame, and so on in a continuous sequence.

The cinema industry has been around for very many years and the industry has established a large number of basic principles concerning the visual presentation of a moving scene. The newcomers to the art are the television engineers who, very wisely, have adopted many of the practices evolved in film presentation. Many of these we must consider in detail.

At what rate is it necessary to present a succession of still pictures to the eye, so that the illusion of movement is complete? This was a basic film problem. If the picture rate is too low, the eye will see each individual picture, but if the rate is too fast, a large amount of costly film would be wasted.

The eye has another property that affects the projection rate. If the projection rate is just high enough to satisfy the illusion of movement, the eye will still be aware that the light source is being interrupted. The eye will see 'flicker', and awareness of this effect depends on the brightness.

Let us now examine the extent to which the television engineer has drawn on the wide experience of the film industry, for his standards.

1.4 ASPECT RATIO

What size and shape should the frame be? The film industry has experimented with this problem throughout its existence. A rectangular frame has an 'aspect ratio' which is the term for the ratio of width to height. A standard aspect ratio is, clearly, of fundamental importance. If a system has a standard ratio at both the transmitter and the receiver, picture size is of no importance. The relative dimensions of the objects in the field will be correct. The film industry has used many aspect ratios over the years, but finally came up with an optimum ratio of 4×3. This ratio is still in general use for the main products of the film industry, despite the use of various 'wide screen' and other ratios.

The first television engineers concerned with the need to establish standards could see no reason to depart from the 4×3 film aspect ratio, particularly as it was realised that film would constitute a large proportion of the television programmes. These engineers were the British team that established standards for the World's first broadcast service of regular television transmissions in 1936. The 4×3 ratio has been adopted by all the World's television systems and, only recently, has this ratio been considered for possible change.

Figure 1.1 shows a simple 6-line television picture consisting of a black bar on a white background and, underneath, the voltage output from the scanning device during a 1-line scan, e.g. the camera voltage output. The scanning spot sweeps across the picture from left to right along line 1, immediately returns to the left, is displaced vertically by one line, sweeps line 2, and so on. The voltage output may be as shown with maximum voltage indicating peak white, and minimum voltage corresponding to black. In a practical case the polarity may

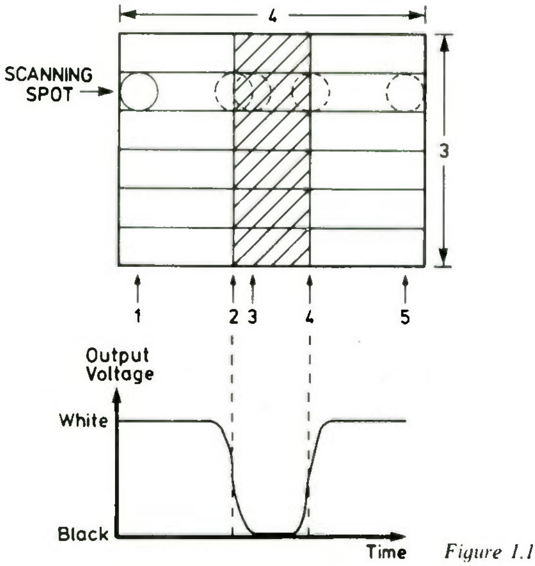


Figure 1.1

be inverted, i.e. minimum voltage equal to white and maximum to black. This, however, depends on the camera tube or other signal source and is of no real importance, because the polarity can be inverted at any point in the system by using an amplifier stage with a gain of 1. When line 6 has been completed, a rapid return to the top of the picture results in a repeat of the scanning cycle. This method of scanning is termed 'sequential scanning'.

1.5 APERTURE DISTORTION

Figure 1.1 shows a basic technical problem that exists with the scanning process. The picture has a sharp transition at the edges of the bar where the voltage should change instantly from the white to the black value, but it does not. As the spot reaches the bar the voltage changes and reaches a half of the black value in position 2. It only reaches the full black value when in position 3. and the overall voltage for the full line scan is as shown. This voltage is used to modulate the transmitter and, when demodulated at the receiver and applied to control the beam current of the display tube, cannot show the sharp transition at the edge of the bar, even if the receiver circuitry is perfect. The edge will be broadened and 'smeared'. The signal that is generated by the scanning process is distorted by what is termed 'aperture distortion' and it is always present.

In a practical television system, the picture quality will depend on the ability of the system to reproduce at the display tube all sharp edges and fine detail. In order to minimise aperture distortion, the spot size must be reduced and this will require more lines for a complete picture scan. Thus, a small scanning spot with a consequent high number of lines is required for an acceptable picture quality.

Some electronic processing can reduce the effects of aperture distortion.

1.6 PICTURE RATE

We have still not fully answered the question 'What should the picture presentation rate be?' A great deal of early work with film revealed that, for most people, a picture rate as low as 10–12/s is adequate to present the illusion of movement. However, the eye is extremely sensitive to the interruption of light at this rate, and would be very aware of 'flicker'.

The early standard film projection rate was 16 frames/s, well above that necessary for movement and, with the low-level illuminants of those early days, adequate to minimise the awareness of flicker. The eye sensitivity to flicker is a function of picture brightness and interruption rate, the flicker rate needing to be increased if brightness is increased. With the improvements in projector lamps over the years, flicker became a problem. Raising the projection rate would reduce flicker, but would result in a larger quantity of expensive film being required.

An ingenious solution of this problem was evolved, and is a standard feature of all film projection systems. The principle has been adopted for all television systems since it was first used in System A in 1936. The film is drawn into position with the light cut off by the shutter, as previously explained. The shutter exposes the light through the frame, cuts it off, re-exposes the light, cuts it off and only then is the next frame drawn into position for the next double exposure. The light is exposed and cut off twice during each frame. For a picture exposure rate of 16 frames/s, the interruption frequency is raised to 32/s, and the visibility of flicker is very much reduced without doubling the length of film.

As time passed, better illuminants came into use and flicker re-appeared. This problem was overcome by solving another problem concerned with a sound track that films now required. Film was not passing through fast enough for good sound quality. The frame rate was raised from 16 to 24 frames/s, the present film standard. This increased the length of sound track from 12 to 18 in/s, and raised the interruption frequency to 48/s.

How all this affects television picture-rate standards will be discussed in the following sections.

1.7 INTERLACED SCANNING

Firstly, a scanning system was developed which provided a similar effect to the double-shuttering used in film projection. Referring to Fig. 1.1, television scanning is *not* carried out sequentially, as shown. The picture is scanned by one field scan using lines 1, 3, 5 and a second scan filling the gaps by scanning with lines 2, 4, 6. This is termed 'interlaced scanning' and constitutes two sweeps of each field for each complete picture. This has all the advantages of the double-shuttering used in film projection, and is used in all standard television systems.

1.8 TELEVISION PICTURE FREQUENCY

The second important film standard that influenced television standards is the frame projection rate, now 24/s. The use of film in television is extensive and films should be projected at the same rate as in the cinema. The engineers who had the task of establishing standards for the first system (System A) were concerned at that time with the possibility that any residual 50 Hz or 100 Hz power supply ripple in the receiver may modulate the tube beam current; the mains power supply, having a frequency of 50 Hz, might result in a 'beat' becoming visible if film was projected at 24 pictures/s. For this reason, it was decided that the picture rate would be 25/s, and film used for television would be operated at 25 frames/s instead of the cinema rate of 24 frames/s. The effect on picture and sound quality is not serious and, in the presence of subharmonic interfering modulation of the display-tube beam current, would result in a stationary bar across the picture. It was considered that the stationary bar would be less offensive than an interfering bar that would sweep vertically down the picture at the beat frequency of 1/s.

The concern that prompted the adoption of a 25/s picture rate has, with time, resulted in the consideration that it was a bad choice and picture frequency could have been higher, with advantage. In the event, visual interference from receiver power supply systems is negligible. Indeed, as will be seen in Chapter 8, the modern colour receiver has a picture frequency which is nominally 25/s, but it is not related to the power supply frequency and the modern receiver does not display any interference effects.

1.9 FLICKER

The interruption rate, using interlaced scanning, is 50/s and this imposes a limitation on the brightness level at which the display tube can be operated before flicker becomes visible. This can be observed on any UK television receiver by increasing the brightness and observing the screen from the side of the eye. Viewing with this part of the eye is termed 'peripheral vision' and is particularly sensitive to interrupted light.

The relationship between flicker and brightness is expressed by the Ferry–Porter law, which is given by:

$$f_c = F + 12.6 \log_{10} B$$

where f_c is the critical frequency below which flicker is observed, F is a constant related to viewing conditions and B is the luminance of picture highlights.

Tests on the viewing of television pictures have suggested a value of about 37 for F and, using 50 for f_c (as in all European systems), a picture highlight value of 10 foot-lamberts is obtained.

When the USA decided on their television standards, they adopted the same general principles that were used in the UK. The picture rate would be related to the power-supply frequency which, in the US, is 60 Hz, resulting in a picture rate of 30/s and a light-interruption rate of 60/s. This increase of interruption frequency, compared with the European rate, results in a permissible high-light value increased by 6.8 times. American television pictures are very much brighter than European pictures can be!

It is of note that the difference in picture frequency between European and US television is too great for direct scanning of film as in the UK. As a result,

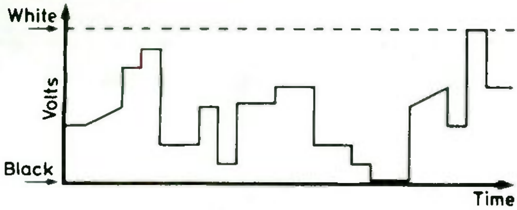


Figure 1.2

US film scanners use a complicated scanning system whereby the film is projected at 24 frames/s but produces 30 television pictures/s.

1.10 THE VIDEO SIGNAL

We will now examine the 'video' signal as it is derived by the camera or other scanning device, for a black-and-white system. It will very rarely be as straightforward as shown in Fig. 1.1. In a practical case, the field of view will consist of random black, white and images in shades of grey, and any single line scan will result in a totally random variation in output voltage, such as shown in Fig. 1.2. Two important conclusions arise from consideration of this type of output voltage:

(1) Voltage variations will generally consist of 'step' changes from one value to another. Smooth transitions from peak white to black, or from black to white, will be very rare.
(2) AC voltage variations are extremely unlikely, and their rare appearance would arise from a scan across regular bars such as the black and white sections of a Test Card.
(3) A third important fact is derived from Fig. 1.3(a) and (b)—each show the same magnitude of signal voltage variation during a line scan. In Fig. 1.3(a) a

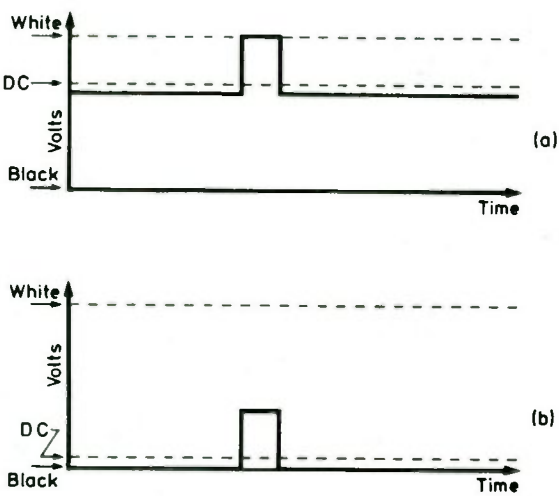


Figure 1.3

white bar is shown on a grey background and in (b) the same output voltage variation shows a grey bar on a black background. The difference between the two identical voltage variations is due to there being in each, an average d.c. voltage component, shown dotted in Fig. 1.3. In Fig. 1.3(a) the d.c. voltage is higher than in Fig. 1.3(b). In practical terms, the d.c. voltage level determines brightness.

1.11 AMPLITUDE MODULATION BY VIDEO

The video output signals are processed in the transmission equipment and, finally, are used to amplitude-modulate the transmitter carrier frequency output. All standard broadcast transmitters use a.m., but f.m. is used for certain link systems and for use with satellite systems.

Amplitude modulation has been used since well before sound broadcasting commenced in 1920. The principles are dealt with in Chapter 2, but there are some important differences between the use of sound or video as modulating signals. Figure 1.4(a) shows an a.c. current variation that might be measured in the antenna system of a sound transmitter. Over the period A the carrier is unmodulated. During B, one cycle of an audio modulating tone is applied. The familiar features of this process are:

(1) The unmodulated carrier is radiated.
(2) The carrier peak level is varied at the modulating frequency. The l.f. variations in carrier peak values during modulation is termed the 'envelope'.
(3) A limit exists in the modulating process whereby the carrier peak must not exceed twice the unmodulated carrier peak level if distortion is to be avoided.

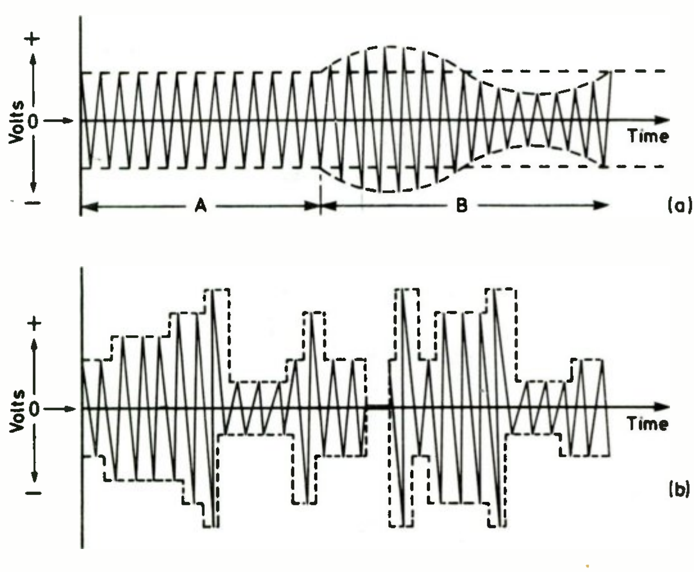


Figure 1.4

(4) The average carrier level remains constant, whether it is modulated or not. This is an important distinction between audio and video amplitude modulation.

Figure 1.4(b) shows a similar situation to that in Fig. 1.4(a), but it is amplitude modulation by a video signal. The modulation envelope is now of a random character and there is now no mean carrier of constant level during modulation. As shown, when no modulating signal is present, no transmitter output is radiated. The peak modulating signal level, which could correspond to a peak white for example, 'turns on' the r.f. power output to its maximum value. Intermediate values of output will correspond to shades of grey.

If the video modulation shown in Fig. 1.4(b) is such that maximum r.f. output is produced by a peak white signal and zero output by a singla representing black, the carrier is said to be 'positive modulated'. As indicated earlier, some standards use a reversal of this whereby minimum or zero output represents white. Such a system is termed 'negative modulation'. In the UK, System A was positively modulated and System I is negative.

1.12 VIDEO AND CHANNEL BANDWIDTH

We now consider what bandwidth will be required for a video signal, and what the consequent channel bandwidth will be when the video signal is used to modulate an r.f. carrier?

It is well known that when an r.f. carrier is modulated, side frequencies are generated in the modulating process. If a carrier is amplitude modulated with, say, a 1 kHz tone, three frequencies are produced; the carrier, a frequency lower than the carrier by 1 kHz, termed a 'lower side frequency', and a frequency above the carrier by 1 kHz which is the upper side frequency. If the modulating signal is a band of frequencies such as voice, music or video, a band of frequencies is generated each side of the carrier, termed 'sidebands', extending on each side of the carrier out to a limit determined by the highest frequency in the band of modulating frequencies. For example, if an audio modulation includes frequencies up to 15 kHz, the r.f. channel will be centred on the r.f. carrier frequency, and will occupy a total band ± 15 kHz. The channel bandwidth is 30 kHz, and the full title of the process is 'double-sideband amplitude modulation', which is more fully described in the following chapter.

So far as amplitude modulation with a video signal is concerned, it is probably realised that the frequency range of a video signal must be very wide and, consequently, the r.f. channel bandwidth will be wide. If we are producing 25 pictures/s and using a small diameter scanning spot in order to resolve fine detail, we shall require a large number of lines to explore fully the field of view. The scanning spot must, therefore, move very fast and any fine detail in the picture will generate high-frequency signal components.

There are several ways whereby the highest-frequency component in the video signal can be determined and, thereby, the channel bandwidth. One of these is indicated in Fig. 1.5, which represents at (a) the top left-hand corner of

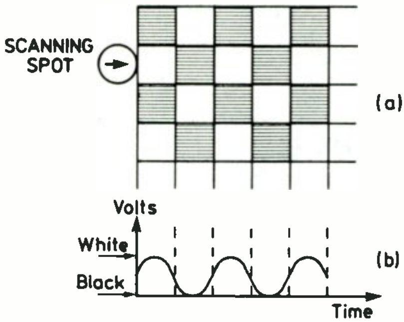


Figure 1.5

a picture consisting of a regular pattern of alternate black and white squares. The sides of the square are equal in length to the scanning spot diameter and, as a consequence, the signal generated will be as shown in Fig. 1.5(b). It is a sine-wave signal and the frequency is the highest that will be generated at full amplitude. Any smaller detail will not generate maximum voltage, because the scanning spot would more than cover the area of detail.

Some arithmetic will enable us to calculate the highest frequency generated in this manner. A return to Fig. 1.1 shows the simple 6-line picture considered earlier. Suppose that this picture consisted of alternate black and white squares, as shown in Fig. 1.5. The total number of squares would be $(6 \times 6)4/3$ (to account for the aspect ratio), i.e. 48 squares. If these squares are to be scanned in $\frac{1}{25}$ s for a complete picture scan, then $48 \times 25 = 1{,}200$ squares are scanned each second. The resulting a.c. cycle would correspond to a scan over one black and one white square. Thus, the highest modulating frequency would be 600 Hz, and the two sidebands would produce an r.f. channel bandwidth of 1.2 kHz. We have seen that the resulting picture quality would be poor and severely lacking in the resolution of edges and fine detail. The picture quality can be improved by reducing the spot size and increasing the number of lines, but the arithmetic shows where this leads. The price to be paid for improved resolution of small detail is bandwidth. For example, if we halved the spot size of our Fig. 1.1 model, we would have to double the number of lines to scan the whole picture, and the arithmetic would look like this:

$$\text{Total number of squares} = (12 \times 12)4/3 = 192$$

$$\text{Number of squares scanned in } 1\text{ s} = 192 \times 25 = 4{,}800$$

resulting in a channel bandwidth of 4.8 kHz. The bandwidth is four times that of the original 6-line system. The bandwidth increases as the square of the number of lines.

It is seen that a standard adopted for a good quality television system has to be a compromise between the need for acceptable definition of fine detail in the picture, and the overall bandwidth of the channel that will concern the generation, transmission, reception and final display of the picture.

1.13 SYSTEM A

It is worthwhile to re-consider the first-ever broadcast standards that were used in 1936. We can now see how the engineers of that time put together their standards which have formed the basis of all today's standards.

The picture frequency was 25/s, interlaced scanning was used and the aspect ratio was 4×3. The compromise on definition and bandwidth was decided on the basis of the scanning spot being of such a size that 405 lines would be required to cover the picture area, and the resulting picture quality would be acceptable. As calculated for the 6-line example of Fig. 1.1, the highest video frequency becomes:

$$(405 \times 405)4/3 \times 25/2 = 2.7 \text{ MHz}$$

The output of detail smaller than 1/405 of picture height results in an output voltage less than maximum. Detail of 3.0 MHz, for example, would be about -3 dB. If the spot diameter is reduced to less than the line width, reproduction of a 3 MHz detail would be fully restored, but gaps would appear between the lines. (Some early 405-line receivers used a 'spot wobble' system to sweep the tube beam vertically over a very small angle to fill the spaces between the lines.)

The total video bandwidth in System A goes up to 3.4 MHz and, as discussed in connection with Fig. 1.3, the lowest video frequency is zero or d.c. The overall r.f. bandwidth is about 6 MHz.

It is of interest to know why this first television broadcast service was situated in the r.f. spectrum on a carrier frequency of 45 MHz. Before 1936, the radio spectrum was fully occupied, from the very low frequencies to about 30 MHz, this high frequency being about the highest frequency usable for long-distance communications. The channel bandwidth for the various services at that time never exceeded about 3 kHz. The demand for spectrum space in which to place a broadcast television channel could only be met by using a carrier frequency higher than 30 MHz. Such ventures would be into relatively unknown regions. Transmitter valves, for example, were available for frequencies up to 30 MHz, but little was known concerning transmission above 30 MHz at high power. The channel frequency of 45 MHz was chosen as being not too far into the unknown above 30 MHz, so that development of the service would not be inhibited. Of course, after the War, which produced a massive development in VHF and UHF, a new technology became available.

The total spectrum space occupied by the first System A channel is shown in Fig. 1.6. The overall video space is about 6 MHz. A separate a.m. audio transmitter is spaced 3.5 MHz from the video carrier, on 41.5 MHz. So far as the receiver is concerned, the television channel included the quite separate sound transmission, and the overall channel space was 6.5 MHz (excluding any 'guard' space at the ends of the channel that might be necessary if any adjacent channel space were to be used).

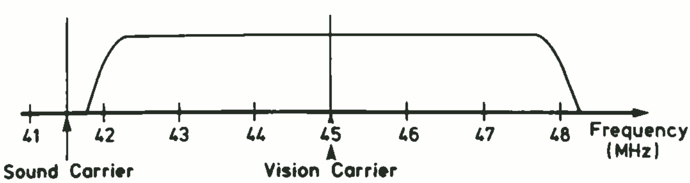


Figure 1.6

1.14 SYNCHRONISM BETWEEN SCANNING SYSTEMS

Referring to Fig. 1.4(b), this could represent the signal voltage modulation of the transmitter during, say, a 1-line scan. At the receiver, the modulating signal is recovered by, possibly, an envelope detector and, after suitable processing, is used to modulate the beam current of the display tube, thus re-creating the transmitted picture. We have omitted, so far, two important functions that are necessary to ensure that the scanning spot at the transmitter, and the beam position on the face of the receiver display tube, occupy identical positions in their 4 × 3 frame. It is necessary to transmit two extra pieces of information from the transmitter; one that 'tells' the receiver the position of the scanning spot in the horizontal plane and another that gives the position in the vertical plane.

It would appear from Fig. 1.4(b) that there is no way in which two types of additional information can be transmitted, but the pioneers who developed System A showed how it could be done, in a manner that forms part of every television standard today.

Figure 1.7 shows how the video signal can be established between two new voltage limits, to represent the voltage changes between black and white. We now have the region between zero volts and black level, into which, we can put extra information. However, as indicated at the beginning of this chapter, the transmitter/receiver system can only deal with one piece of information at one instant of time and if we intend to use the levels between black and zero to provide extra information we must remove the video information at the time when we intend to use the added information. This is achieved as shown in Fig. 1.8. A line 'blanking pulse' blanks out all video information at the start of the line scan, down to black level. A narrow pulse is inserted from the foot of the blanking pulse down to zero and the leading edge of this pulse is used to start the line scan on its traverse from left to right. At the receiver, the leading edge of the demodulated narrow pulse is used to start the sweep across the face of the display tube from left to right. At the end of the line scan, a further blanking and synchronising pulse triggers the return of the spot to the left-hand side and the scanning cycle starts again, this time, slightly displaced vertically to trace a new line path across the picture. This line-by-line repetitive action ensures that each line is started and stopped at exactly the same instants in both the transmitter and the receiver. Figure 1.9 shows the waveform of a line of the UK System I (which is inverted compared with System A), and shows the timing

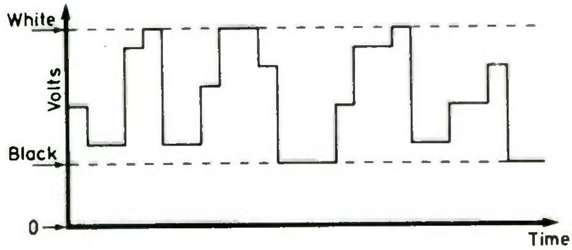


Figure 1.7

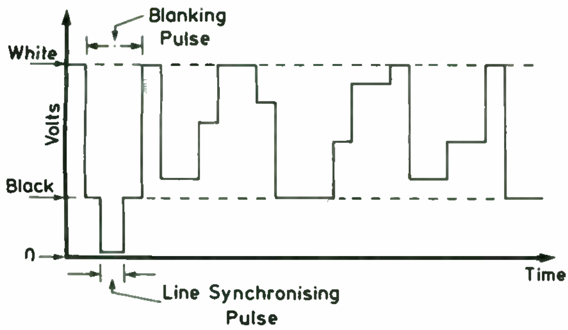


Figure 1.8

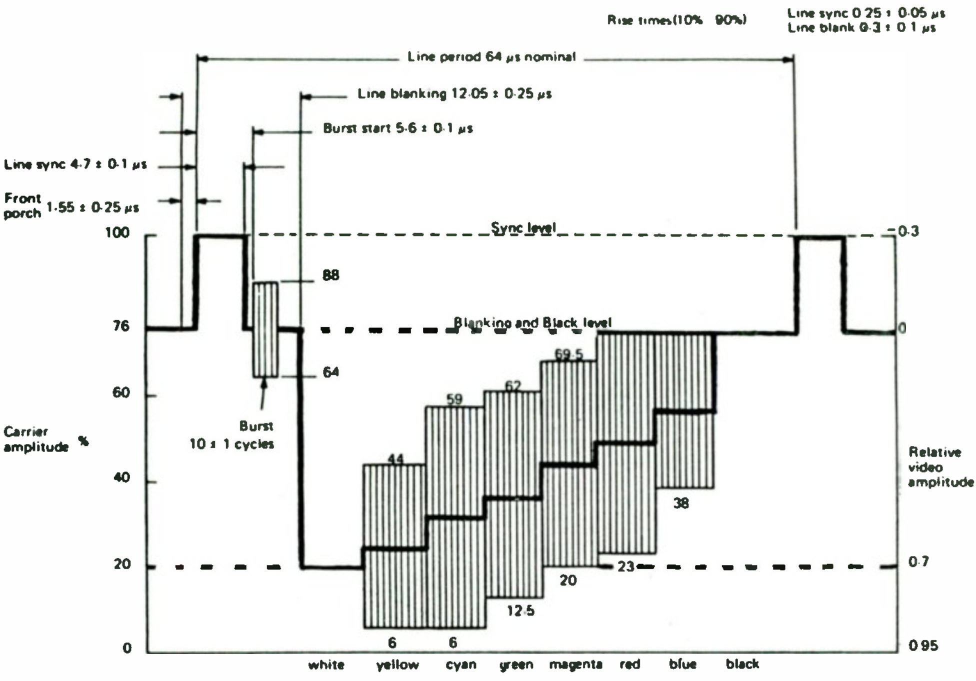


Figure 1.9 Transmitted carrier waveform showing line synchronizing signals, colour synchronizing burst signal and colour bar waveforms

intervals involved with blanking and the line sync pulse. All timing is with reference to the leading edge of the narrow sync pulse.

We see how line synchronism has been achieved between transmitter and receiver, but how is the spot moved vertically? There is only the space between black level and sync-pulse tips, and we appear to have used this space for horizontal synchronising information. However, we can use the same space for a different kind of pulse. The line sync pulses are narrow and relatively infrequent and, if we can use a pulse-type signal with characteristics that differ

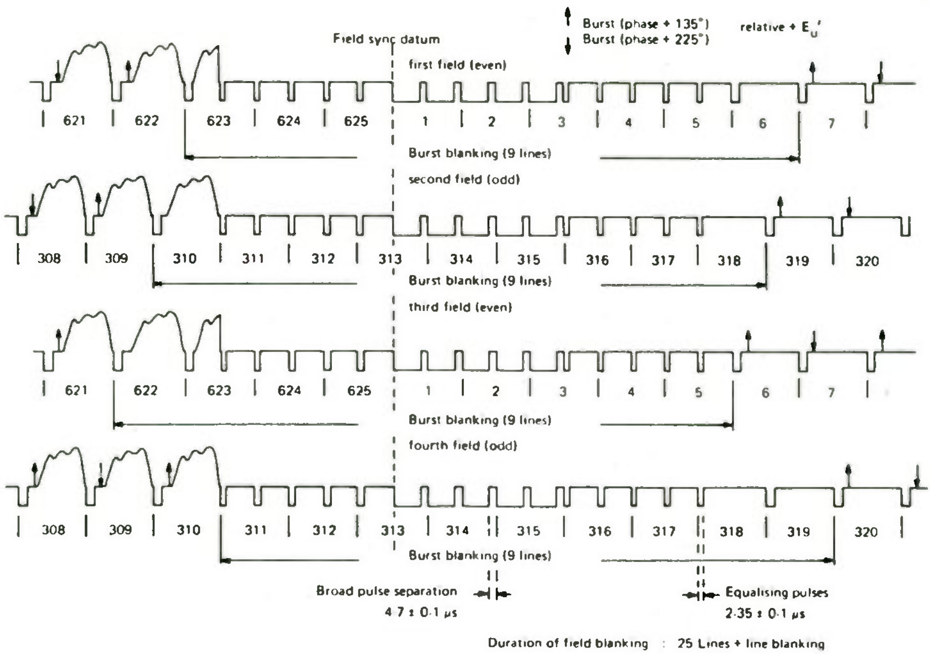


Figure 1.10 Waveforms showing field synchronizing signals, colour synchronizing burst signals and burst blanking

considerably from the line sync pulses, we can send them in the same space as the line sync pulses and separate them at the receiver. There are two types of contrasting pulse that can be used. A single broad pulse would do or a train of broad, frequent pulses.

When the scan has reached the bottom of the picture, the video information is blanked out by a field blanking pulse that has a width covering several lines, thus removing some picture height. A train of broad, frequent pulses is inserted in the blanking time interval, as shown in Fig. 1.10. These broad pulses are extracted at the receiver, processed in a manner that separates them from the line pulse information, and used to trigger the vertical sweep of the display-tube beam.

We now have the three types of information, transmitted in a manner whereby they can be separated at the receiver for individual processing. At this stage we have only considered black and white television—colour signals will require the transmission of even more information, as is shown in Chapter 8.

1.15 PULSE SEPARATION

When the received signals are demodulated, a 'composite' signal is extracted that consists of the complete waveform of the video and pulse signals, as shown in Fig. 1.8. Pulse separation is effected in two stages (often by a single active

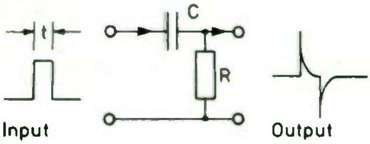


Figure 1.11

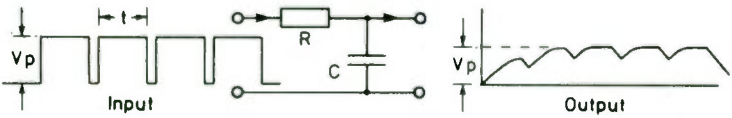


Figure 1.12

circuit). First, the pulse information is extracted from the composite signal by means of an amplitude discrimination system. The two types of pulse are then passed to passive circuitry that separates the functions of the two types of pulse.

Figure 1.11 shows one form of a differentiation circuit in which the time constant *CR* is short compared with *t*. An output is derived from the change of input voltage only and, as shown, a narrow pulse (or 'spike') can be derived from the leading edge of an input pulse and/or from the trailing edge, if required.

Figure 1.12 shows an integrating circuit. The *CR* time constant is long compared with the pulse duration. It is seen that successive pulses build up the voltage on *C* to the value V_p and, at the end of the pulse train, *C* discharges. The output waveform constitutes a single pulse, used to trigger the vertical sweep circuit, which operates at the relatively slow speed of 50 sweeps/s.

The differentiation circuit does not distinguish between line or field pulses, it will generate 'spike' pulse output from the edges of any pulses fed to it. The integration circuit is the one that *does* discriminate between the two types of pulse. The narrow, infrequent line sync pulse will provide no output from the integrator.

1.16 PORCHES

Figure 1.8 shows two other features concerned with the line sync pulse region. It will be noted that the pulse is not inserted at the centre of the blanking interval.

Voltage or current changes cannot take place instantaneously. A video signal change from black to peak white, or white to black, requires time for completion. A black/white edge results in a voltage change as shown in Fig. 1.13. The time of the rise (or fall) is important, but it is not possible to determine exactly where the change starts or stops. As a result, a definition for 'rise time' (or fall time) has been adopted. It is the time required for the change to take place between 10 and 90% of the final value, this time being readily determined from an oscilloscope display.

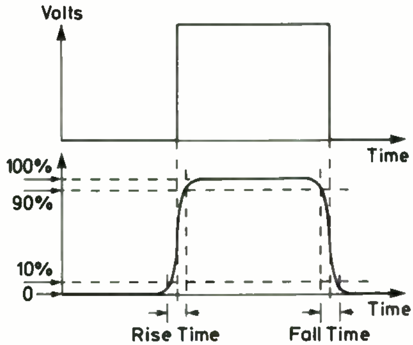


Figure 1.13

Ahead of the pulse in Fig. 1.9 is a narrow plateau, termed the 'front porch'. Its purpose is to allow the video signal of the previous line (which might have been at peak white, for example) to reach black, before the sync pulse starts. Behind the sync pulse is another plateau termed the 'back porch'. The original purpose of the back porch in the System A was to make sure that there was plenty of time for the line scan circuitry to effect complete re-trace of the line scan and for the spot to be in its correct position for commencement of the next video signals. By comparison with today's technology, the scanning circuits of 1936 were crude, ponderous and extravagant of power, and they required a lot of time for the re-trace. Today, the back porch provides plenty of time for re-trace and it serves another purpose by providing the space for colour information to be transmitted.

1.17 VIDEO VOLTAGE

How is the magnitude of a video voltage determined? Figure 1.7 shows the random manner in which the video signal varies. It bears no resemblance to a sine-wave signal, which has a comfortable relationship between its peak value, r.m.s. and average values. The most meaningful way in which a video signal can be expressed is in terms of its magnitude between black and peak-white values, its peak-to-peak value.

The complete signal, including sync pulses, is termed a 'composite signal' and its value is determined by the overall value from peak, through black to the sync tips. The *p-p* voltage value usually used for general handling and processing of video signals is 1 V, of which the ratio of video to pulse might be 0.7 to 0.3.

1.18 SPECTRUM UTILISATION

We have considered the amplitude modulating signals in detail. How is this information established in the very wide channel space that is required for a television transmission?

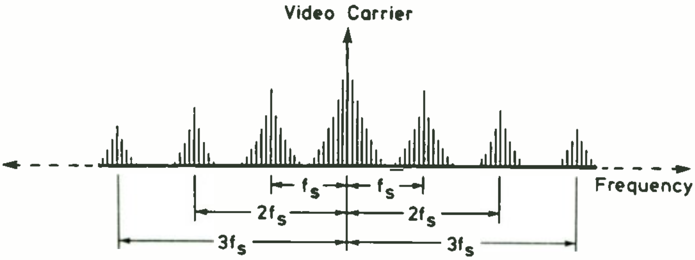


Figure 1.14

Chapter 2 gives most of the answers to this question. The picture frequency is 25/s, and this must be one of the modulating frequencies. Interlaced scanning produces another modulating signal at 50/s. The line scanning rate is another constant frequency (15,625 Hz in System I) and all of these are modulating the carrier, whether video signals are present or not.

The line frequency is predominant, and harmonics of the line frequency are spaced on either side of the carrier frequency, as shown in Fig. 1.14. The line frequency harmonics f_s extend out to the bandwidth limits, and they are of a descending level of energy as they are more remote from the carrier. Around each side-frequency harmonic there are further side-frequencies, spaced above and below by the field and picture frequencies, as shown. It is seen from Fig. 1.14 that, over the channel bandwidth, the energy derived from the scanning process is contained in discrete clusters, spaced at the line-scanning frequency from the carrier and each other. Large regions in the overall channel space contain little or no energy. What happens when video signals are included in the modulating process?

In 1934, two mathematicians, Merz and Gray, studied this question and derived the, perhaps surprising, conclusion that the video energy simply joins the existing clusters round each line scan harmonic, leaving the overall situation very much as shown in Fig. 1.14, complete with its gaps between the clusters. The only differences between one picture and another or whether one is stationary or moving, are relatively small variations in the magnitude of the energy clusters. The low or zero energy gaps between the line frequency harmonics remains with all types of video signal.

The Merz and Gray investigation was based on black and white television systems, which require to transmit three separate types of information, video and sync pulses. It has proved to be of considerable value when considering possible interference effects between transmitters operating on the same nominal channel frequency. It is only necessary to operate one of the transmitters with a carrier frequency 'offset' by half the line frequency, to ensure that any interfering signals from each transmitter would occupy a zero-energy space in the channel space of each.

Of more importance, as seen in Chapter 8, spaces exist in a black and white channel where the extra information required for colour can be placed, with minimum interference effects, between the colour signals and the black and white video signals. The half-line offset mentioned above has to be modified when colour transmission is involved.

Chapter 2

Modulation

R. D. A. Maurice

2.1 INTRODUCTION

Almost any waveform may be modulated; that is, have another waveform impressed upon it in some way or another. Usually it is a sinewave (cosinewave) that is modulated by another wave containing or representing the information that it is desired to transmit over a distance. Electromagnetic radiation will propagate through space only if it is varying in intensity. A direct voltage or current will pass along a wire, but will not propagate in space. It can be modulated by switching it on and off as in Morse code keying; but the frequencies of the pulse thus produced are so low that the transmitting and receiving aerials required for wireless transmission would need to be impractically long. A sinewave modulated with the information to be transmitted can be given any frequency that is suitable for the transmission band allocated for a particular service, provided that the information is itself contained within a frequency band that is smaller than the sinewave carrier's own frequency. Note the introduction of the word carrier. We shall be dealing with pulses later on, but our first considerations will be confined to sinewaves. Although we shall, at first, confine ourselves to sinewave carriers, the modulating information—a sound radio programme or a television video signal—may well not have a sinusoidal form. Thanks to Fourier we know that any waveform, whatever its shape, can be expressed as a sum of sinewaves. If the waveform is periodic; that is, repeats itself, the sinewaves will have frequencies that are integer multiples of the lowest frequency of the series; they are harmonics of the fundamental. If the waveform is not periodic, then the number of sinewaves that constitute it will be infinite, each one will have an infinitesimal amplitude and adjacent waves will have frequencies that are infinitely close to each other. This is the Fourier integral. The periodic waveform, from Fourier series, has a line spectrum; that is, each component at fundamental and harmonic frequencies has a finite amplitude. The non-periodic waveform has a continuous spectrum; that is, a smooth curve whose ordinate values are densities such as volts per Hertz or amps per Hertz.

It must be clearly understood that a spectrum, in the sense meant by communications engineers, is not a function of time. It consists of the amplitudes of the terms of the Fourier series or, in the case of a non-periodic or transient time function, the density values of the oscillating components that constitute the Fourier integral. Sidebands, with which we shall be dealing shortly, are functions of time and can appear and disappear on the screen of a cathode-ray tube, a behaviour not open to spectra which are concepts rather than physical entities.

To summarize, a periodic function of time with period T may be expressed as the Fourier series

$$f(t)=\tfrac{1}{2}b_0+b_1\cos 2\pi\frac{t}{T}+b_2\cos 4\pi\frac{t}{T}+\ldots+b_n\cos 2\pi n\frac{t}{T}+\ldots$$

$$+a_1\sin 2\pi\frac{t}{T}+a_2\sin 4\pi\frac{t}{T}+\ldots+a_n\sin 2\pi n\frac{t}{T}+\ldots \qquad (2.1)$$

where

$$b_n=\frac{2}{T}\int_{-T/2}^{T/2} f(t)\cos 2\pi n\frac{t}{T}\,dt \qquad (2.2)$$

$$a_n=\frac{2}{T}\int_{-T/2}^{T/2} f(t)\sin 2\pi n\frac{t}{T}\,dt \qquad (2.3)$$

$\frac{1}{2}b_0$ is the average value or d.c. component of the periodic function. A non-periodic, or transient, function of time may be expressed as the Fourier integral

$$f(t)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(u)\,e^{j2\pi f(t-u)}\,du\,df \qquad (2.4)$$

$$=\int_{-\infty}^{\infty} f(u)\,e^{-j2\pi fu}\,du\int_{-\infty}^{\infty} e^{j2\pi tf}\,df \qquad (2.4a)$$

In Equation 2.4a, the left-hand integral is a function of f and not t. It is therefore the counterpart to the coefficients b and a of Equations 2.2 and 2.3, which are also not functions of t, but only of T and n, the harmonic number. If we give a name to the left-hand integral in Equation 2.4a—the Fourier transform $f[f]$* of the time function $f(t)$—we can re-write Equation 2.4(a) as

$$f(t)=\int_{-\infty}^{\infty} f[f]\,e^{j2\pi tf}\,df \qquad (2.4b)$$

where

$$f[f]=\int_{-\infty}^{\infty} f(t)\,e^{-j2\pi ft}\,dt \qquad (2.5)$$

Equation 2.4b expresses $f(t)$ as the inverse Fourier transform of $f[f]$.

Now the foregoing mathematics enable us to distinguish clearly between spectra and time functions. For the Fourier series, which is a time function, the

* Here, square brackets are used to indicate a frequency function, whilst parentheses are reserved for time functions. Of course, $f(t)$ and $f[f]$ are quite different functions. They could have been written as $f(t)$ and $g(f)$.

coefficients a_n and b_n represent the heights or values of the spectral lines. For the Fourier integral, which is a time function (Equations 2.4, 2.4a), $f[f]$ represents the spectral density as a function of frequency.

2.2 AMPLITUDE MODULATION OF A SINUSOIDAL CARRIER

2.2.1 Double sideband suppressed carrier amplitude modulation

Since any modulating waveform may be regarded as a sum (infinite or finite) of sinewaves we can study the phenomenon with adequate precision by considering the modulation to consist of a single sinewave at, for example, audio or video frequency f_a. The carrier wave will have a radio frequency $f_r \gg f_a$. We let the carrier be $\sin 2\pi f_r t$ and the modulation be $m \sin 2\pi f_a t$. The process of modulating the amplitude of the carrier is simple multiplication, thus the resultant wave is

$$R(t) = m \sin 2\pi f_a t \sin 2\pi f_r t \tag{2.6}$$

and this is shown in Fig. 2.1(a). Note that the points at which zeros of the two sinewaves occur together are not, in fact, sharp transitions, because there can be no frequencies in the spectrum of $R(f)$ that exceed the sum $f_a + f_r$.

Figure 2.2 shows the product $\sin\theta \sin(4.3\theta)$ where the zeros of the two waves do not coincide. It can be seen that at $\theta = 180°$ when $\sin\theta = 0$ the curve of the product is quite smooth.

$$R(t) = \frac{m}{2}\cos 2\pi(f_r - f_a)t - \frac{m}{2}\cos 2\pi(f_r + f_a)t \tag{2.6a}$$

The two terms in the right-hand side are sidebands. There is no carrier; that would be a term at frequency f_r. Figure 2.1(b) is a vector or phasor diagram in which the non-existent carrier is shown as a dashed line. If we imagine the whole diagram to be rotating anticlockwise at an angular velocity $2\pi f_r$ then the directions and relative angular velocities of the two sidebands are as shown.

To detect such a signal without distortion it is necessary to generate, at the receiver, a carrier of frequency f_r and to adjust its phase angle ϕ for maximum output. The process is, once again, multiplication, thus

$$(m \sin 2\pi f_a t \sin 2\pi f_r t)\sin(2\pi f_r t + \phi)$$
$$= m \sin 2\pi f_a t\{\tfrac{1}{2}\cos\phi - \tfrac{1}{2}\cos[2(2\pi f_r t) + \phi]\} \tag{2.7}$$

The term $\frac{1}{2}\cos[4\pi f_r t + \phi]$ when multiplied by $m \sin 2\pi f_a t$ will produce, once again, two sidebands and the lower frequency one will have a frequency of $2f_r - f_a$. A low pass filter cutting off at a frequency just below $2f_r - f_a$ will, therefore, eliminate the second term of the right-hand side of Equation 2.7. If ϕ is adjusted for maximum output, that is, $\phi = 0$ and $\cos\phi = 1$, the detected signal will be $(m/2)\sin 2\pi f_a t$.

Suppressed carrier modulation as described in Section 2.2.1 is used in NTSC and PAL colour television systems for carrying the chrominance signal. It is also used in the pilot-carrier system of stereophonic broadcasting for carrying

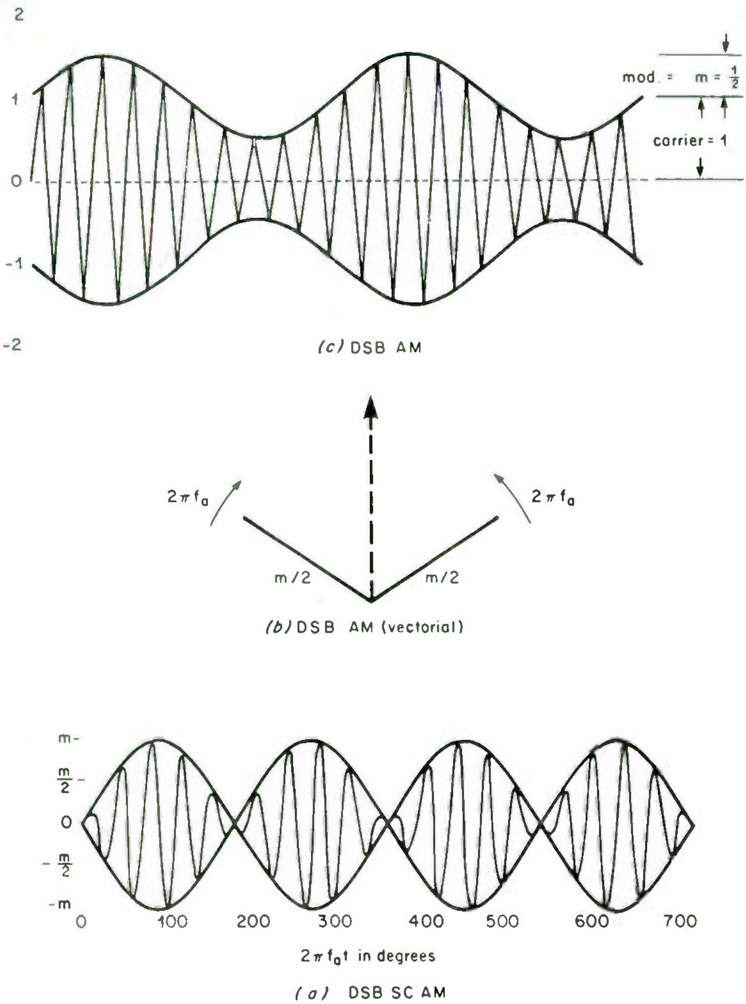


Figure 2.1 Amplitude modulation

the difference signal between left-hand and right-hand microphones. In all the above-mentioned systems the carrier that has to be generated in the receiver for multiplying with the signal to be detected is synchronised to a separate signal. It is then transmitted either in a different time slot or at a different frequency from the position of the suppressed or non-existent carrier in the signal itself.

For some point-to-point circuits such as those used by the BBC External Services for feeding sound programmes to distant relay stations, a system based on DSB SC AM is used. In fact, the carrier strength is reduced rather than completely suppressed so that a very narrow-band circuit in the receiver can select the reduced carrier level without the sideband(s) and thus effect

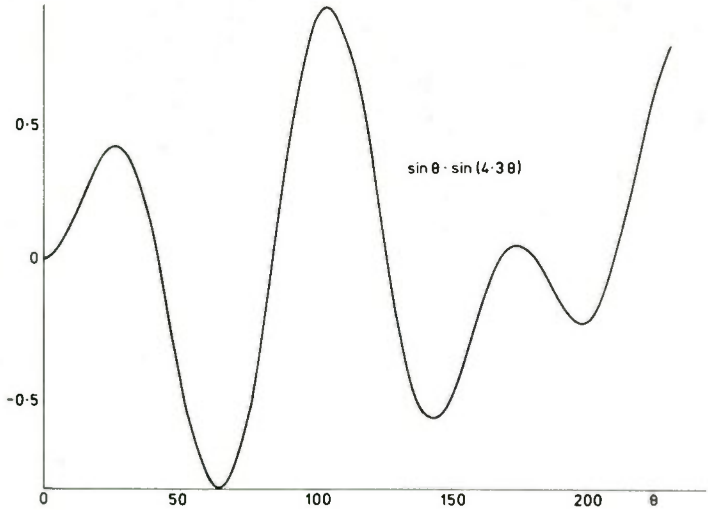


Figure 2.2 DSB SC AM wave zero crossings, not coincident

detection by multiplication. Also, one of the two sidebands is suppressed, so the signal becomes

$$R(t)=\frac{m}{2}\cos 2\pi(f_r-f_a)t \tag{2.8}$$

If $R(t)$ be multiplied by $\cos 2\pi f_r t$ and the product passed through a low-pass filter as described earlier, the result is

$$\frac{m}{4}\cos 2\pi f_a t$$

2.2.2 Double sideband amplitude modulation

This was, after Morse code transmissions, the first system used for sound broadcasting. Since the frequency range of speech and music does not go lower than some tens of Hertz, a sound transmitter unlike a television transmitter, can transmit a continuous, uninterrupted, carrier which is unaffected by the modulating signal. Although much more power has to be radiated if a carrier is always present, the detection process in the receiver can be much simpler and more reliable than the multiplication process that we have been discussing. In order to ensure the continuous presence of a carrier wave we alter the modulating wave from that discussed in Section 22.1 by adding a constant number to the modulating sinewave, thus $1+m\sin 2\pi f_a t$. In order to avoid distortion in the receiving detector we ensure that $m<1$. The radiated signal now becomes

$$R(t)=(1+m\sin 2\pi f_a t)\sin 2\pi f_r t \tag{2.9}$$

$$=\sin 2\pi f_r t+\frac{m}{2}\cos 2\pi(f_r-f_a)t-\frac{m}{2}\cos 2\pi(f_r+f_a)t \tag{2.9a}$$

$R(t)$ is shown in Fig. 2.1(c) in which, to save effort, the actual high frequency modulated wave is plotted as triangular although it is really curved as in Fig. 2.1(a). The frequencies of carrier and modulation are the same in the two figures. It is simple to see that a simple diode detector with a time constant long compared with the carrier period $1/f_r$ and short compared with the modulation period $1/f_a$ will act as a peak detector to the high frequency, but will respond faithfully to the modulation. The diode current will consist of a d.c. with the modulation riding upon it; that is

$$1+m \sin 2\pi f_a t$$

The vector diagram will be similar to Fig. 2.1(b) except that the carrier vector should now be a solid line and twice as long as either of the sideband vectors.

Equation 2.9 is called the envelope equation, Equation 2.9a is the sideband equation and by combining the two cosines in Equation 2.9a we arrive at the modulation vector, Equation 2.9b

$$R(t)=\sin 2\pi f_r t + m \sin 2\pi f_a t \sin 2\pi f_r t \qquad (2.9b)$$

which is the same as multiplying through in Equation 2.9. The vector diagram for the modulation vector (Equation 2.9b) is shown in Fig. 2.3 in which the carrier has amplitude 1 and $m=\frac{1}{2}$. The modulation vector is in phase with the carrier and is therefore added on to it in one straight line.

2.2.3 Energy

Now wave motion carries energy with it and the transmitted energy per cycle of modulation is proportional to the integral with respect to time of the square of $R(t)$, thus

$$E_a \propto f_a \int_0^{1/f_A} R^2(t)\, dt \qquad (2.10)$$

The expression for E_a is a lengthy one* that depends upon the ratio of f_r to f_a, but if we assume that $f_r \gg f_a$ and, furthermore, that this ratio is an integer, so

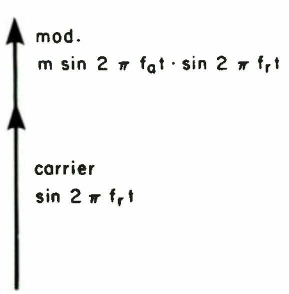


Figure 2.3 DSB AM modulation vector

* The exact formula is

$$E_a \propto \frac{1}{2}\left(1-\frac{\sin 4\pi K}{4\pi K}\right)+\frac{m^2}{8}\left[2+\frac{\sin 4\pi(K-1)}{4\pi(K-1)}+\frac{\sin 4\pi(K+1)}{4\pi(K+1)}\right]$$

where $K=f_r/f_a$

that there is an exact number of cycles of radio frequency carrier in each modulation cycle, we can simplify to

$$E_a \propto \frac{1}{2} + \frac{m^2}{8} + \frac{m^2}{8} \tag{2.10a}$$

A simple way to establish Equation 2.10a is to note that the energy in a cycle of sinewave is proportioned to

$$\frac{1}{2\pi}\int_0^{2\pi} \sin^2 \theta \, d\theta = \frac{1}{2\pi}\left|\frac{1}{2}\theta - \frac{1}{4}\sin 2\theta\right|_0^{2\pi}$$

$$= \frac{1}{2\pi}\left(\frac{2\pi}{2} - \frac{1}{4}\sin 4\pi - \frac{0}{2} - \frac{1}{4}\sin 0\right)$$

$$= \frac{1}{2} \tag{2.11}$$

If we now apply this to each of the terms in the right-hand side of Equation 2.9a, we have

$$E_a \propto \frac{1}{2} + \frac{1}{2}\frac{m^2}{4} + \frac{1}{2}\frac{m^2}{4}$$

The first term represents carrier energy and the remaining two terms represent sideband energy. If we assume 100% modulation, $m = 1$, the percentage energy in each wave is

Carrier energy: 66.6%
Each sideband energy: 16.6%

So single sideband suppressed carrier modulation requires, in theory, only 16.6% of the energy required for full double sideband amplitude modulation. The receiver bandwidth required for it is half that needed for double sideband plus carrier, so the received signal-to-noise ratio would be +3 dB (less noise) − 6 dB (half the signal strength because only one sideband) or −3 dB relative to the signal-to-noise ratio for double sideband amplitude modulation (with carrier) or DSB AM. Of course, the actual transmitter power could be doubled for the single sideband suppressed carrier case (SSB SC), in which case the received signal-to-noise ratio would be 3 dB greater than that achieved with DSB AM with carrier and for a radiated power only one-third of the DSB AM transmission.

It will have been observed that the carrier carries no information, but is a necessary catalyst for simple diode detection.

2.2.4 Signal-to-noise ratio

The performance of receivers is often rated by a quantity called the noise figure. This can be defined in more than one way, but a usefully simple one is the ratio of the output signal-to-noise ratio to the input signal-to-noise ratio. These ratios are expressed in terms of available signal powers and available noise powers. If the receiver input impedance is a resistance R and the source of signal is matched to that resistance by having, itself, an internal resistance R, for example, the radiation resistance of the aerial—75 Ω if a simple dipole—then the available input power is $E^2/4R$ if E is the source e.m.f. The available noise power due to the source resistance only is kTB where k is Boltzmann's constant (1.3805×10^{-23} Joules), T is the absolute temperature ($273.15 + \theta$°C) and B is the effective bandwidth of the receiver. B is usually the total effective intermediate frequency bandwidth if the transmission is DSB AM or SSB AM or vestigial sideband AM (VSB AM) as used for television. The noise figure F is thus

$$F = (E^2/4RkTB)/(S_0/N_0)^2 \tag{2.12}$$

If, as is always the case, the output signal-to-noise ratio is lower than that at the input, due to noise generated within the receiver including its input resistance R, then F can never be less than 2. This cannot be seen from Equation 2.12, but it may be guessed from the fact that, if matching is assumed, there are two resistances R, each producing available noise power $4kTB$, whilst the definition of F admits noise from only the source as 'legitimate'. If we assume that the receiver temperature including that of the source resistance is 16.85°C so that $T = 290°$ Kelvin, or absolute, we find from Equation 2.12

$$E(\mathrm{dB}/\mu\mathrm{V}) = -78 + 10\log_{10} R + 10\log_{10} B + F_{\mathrm{dB}} + \left(\frac{S_0}{N_0}\right)_{\mathrm{dB}} \tag{2.13}$$

where R is in ohms and B is in Hz.

As an example, let us find the required receiver aerial e.m.f. E to give an output signal-to-noise ratio of 40 dB for a SSB SC overseas broadcast relay circuit. Let $R = 100\,\Omega$, $B = 10$ kHz and receiver noise factor 10 dB, Equation 2.13 yields

$$\begin{aligned} E(\mathrm{dB}/\mu\mathrm{V}) &= -78 + 10\log_{10} 100 + 10\log_{10} 10{,}000 + 10 + 40 \\ &= 32\ \mathrm{dB}/\mu\mathrm{V} = 40.3\ \mu\mathrm{V} \end{aligned}$$

If such a signal were achieved by fully loading a given transmitter, then if the system were changed to DSB AM with carrier, the aerial e.m.f. at the receiver would be 3 dB less (37.3 μV) and so would be the output signal-to-noise ratio (37 dB).

2.3 FREQUENCY AND PHASE MODULATION OF A SINUSOIDAL CARRIER

A sinewave has three attributes or parameters each of which is susceptible of being modulated

$$A \sin(\omega t + \Phi)$$

where A is the amplitude, $\omega = 2\pi f$ is the angular frequency and Φ is the phase angle.

2.3.1 Frequency modulation

For this we modulate ω or f and we may, without loss of generality, put $\Phi = 0$.

Again, we let the modulating signal be sin $2\pi f_a t$ or for simplicity sin $\omega_a t$. There are, however, some remarks to be made before proceeding. The modulation must now have the dimensions of angular frequency: let us multiply by, say, $\Delta\omega$, some angular frequency or other. If we write, letting $A = 1$ for simplicity,

$$R(t) = \sin[(\Delta\omega \sin \omega_a t)t]$$

we should have a frequency swing from $-\Delta\omega$ through zero to $+\Delta\omega$ which is useless; so the first thing to do is add the frequency equivalent to the d.c. term required to establish the carrier in DSB AM. Thus the modulation becomes

$$\omega_0 + \Delta\omega \sin \omega_a t \tag{2.14}$$

Frequency can be defined as the derivative of phase with respect to time, thus

$$\omega = d\phi/dt$$

or angular frequency equals rate of change of phase or the corollary

$$\phi = \int \omega \, dt \tag{2.15}$$

This is a logical statement that works in practice. Now our final modulated waveform can be written with complete generality

$$\begin{aligned} R(t) &= \sin \phi(t) = \sin \int \omega \, dt \\ &= \sin \int (\omega_0 + \Delta\omega \sin \omega_a t) \, dt \\ &= \sin\left(\omega_0 t - \frac{\Delta\omega}{\omega_a} \cos \omega_a t\right) \end{aligned} \tag{2.16}$$

Equation 2.16 could be called the envelope equation. The signal has constant amplitude and a phase angle that has a linear increase with time with a cosinusoidal wobble superimposed upon it. The quantity $\Delta\omega/\omega_a$ or $\Delta f/f_a$ is termed the modulation index and has some similarity to the percentage modulation, $100m$, in amplitude modulation. It is not restricted, however, as is $m < 1$.

We must now try to establish the modulation-vector and the sideband equations. These will introduce Bessel functions[1] whose values can be 'looked up' in tables, just as for sines and cosines. Take Equation 2.16 and expand:

$$\begin{aligned} \sin(\omega_0 t - m \cos \omega_a t) = \sin \omega_0 t \cos(m \cos \omega_a t) \\ - \cos \omega_0 t \sin(m \cos \omega_a t) \end{aligned}$$

where $m = \Delta\omega/\omega_a = \Delta f/f_a$

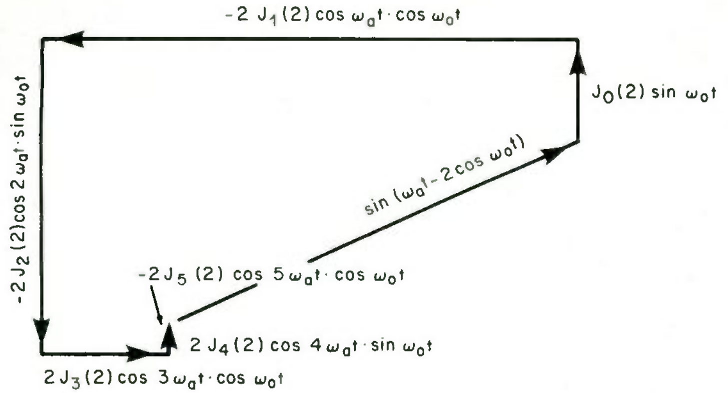


Figure 2.4 FM modulation vectors

$$\begin{aligned}\sin(\omega_0 t - m\cos\omega_a t) &= \sin\omega_0 t\{J_0(m) - 2[J_2(m)\cos 2\omega_a t - J_4(m)\cos 4\omega_a t \\ &\qquad + J_6(m)\cos 6\omega_a t - \ldots]\} \\ &\quad - \cos\omega_0 t\{2[J_1(m)\cos\omega_a t - J_3(m)\cos 3\omega_a t \\ &\qquad + J_5(m)\cos 5\omega_a t - \ldots]\}\end{aligned} \tag{2.17}$$

This may be re-written*

$$\begin{aligned}\sin(\omega_0 t - m\cos\omega_a t) = J_0^* \sin\omega_0 t &- 2J_2\cos 2\omega_a t\sin\omega_0 t \\ &+ 2J_4\cos 4\omega_a t\sin\omega_0 t \\ &- 2J_6\cos 6\omega_a t\sin\omega_0 t + \ldots \\ &- 2J_1\cos\omega_a t\cos\omega_0 t \\ &+ 2J_3\cos 3\omega_a t\cos\omega_0 t \\ &- 2J_5\cos 5\omega_a t\cos\omega_0 t + \ldots\end{aligned} \tag{2.18}$$

The vector diagram is shown in Fig. 2.4. The carrier has an amplitude $J_0(m)$; a Bessel function of the first kind and of zero order. If we add, vectorially, all the individual modulation vectors the sum is, of course, $\sin(\omega_0 t - m\cos\omega_a t)$. Figure 2.4 was drawn with $m=2$.

Happily, the higher-order Bessel functions, J_1, J_2, J_3, etc., diminish rapidly. There would have been many more component vectors of significant values if m had been much larger, say 5 or 10. In studying the figure it should be noted that the positive sine direction is up the figure and the positive cosine direction is across the paper to the right.

The sideband equation is easily derived from the modulation-vector (Equation 2.18). Apart from the carrier $J_0 \sin\omega_0 t$, each term gives rise to a couple of sidebands, one below and one above the carrier frequency $\omega_0/2\pi$.

* The (m) is omitted after each Bessel function $J_n(m)$.

Thus

$$\begin{aligned}
&\sin(\omega_0 t - m\cos\omega_a t)\\
&\quad = J_0(m)\sin\omega_0 t - J_1(m)[\cos(\omega_0+\omega_a)t + \cos(\omega_0-\omega_a)t]\\
&\qquad - J_2(m)[\sin(\omega_0+2\omega_a)t + \sin(\omega_0-2\omega_a)t]\\
&\qquad + J_3(m)[\cos(\omega_0+3\omega_a)t + \cos(\omega_0-3\omega_a)t]\\
&\qquad + J_4(m)[\sin(\omega_0+4\omega_a)t + \sin(\omega_0-4\omega_a)t]\\
&\qquad - J_5(m)[\cos(\omega_0+5\omega_a)t + \cos(\omega_0-5\omega_a)t]\\
&\qquad - J_6(m)[\sin(\omega_0+6\omega_a)t + \sin(\omega_0-6\omega_a)t]\\
&\qquad + J_7\\
&\qquad + J_8\\
&\qquad - J_9\\
&\qquad \text{etc.}
\end{aligned} \tag{2.19}$$

Figure 2.5 shows Equation 2.19 for $m=2$.

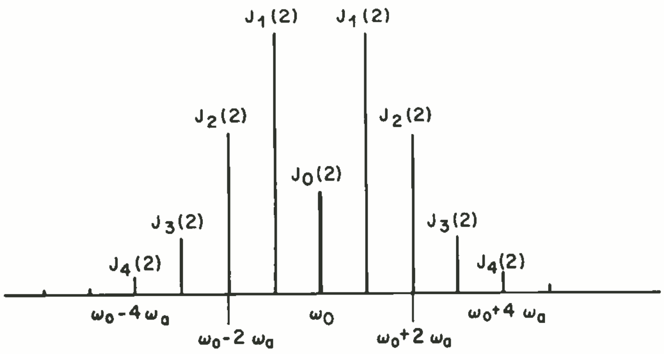


Figure 2.5 FM sidebands

It can be seen that frequency modulation requires a wider bandwidth than DSB AM. It is usual to adopt Carson's rule that sidebands contributing less than about 1% of the total energy can be neglected.

Energy However the phase of a sinewave may vary, its energy per cycle remains proportional to a half. Thus the energy represented by Equation 2.19 is proportional to one-half, that is, the sum of the squares of the amplitudes of the 13 terms in Equation 2.19 should be one, apart from the neglected higher-order sidebands, because the equation must represent the same amount of energy per cycle on each side of the equals sign.

Demodulation of a frequency modulated signal How do we detect $R(t)$ in Equation 2.16?

$$R(t)=\sin\left(\omega_0 t-\frac{\Delta\omega}{\omega_a}\cos\omega_a t\right)$$

It can be seen that if we have a circuit with a sloping characteristic of amplitude as a function of frequency, the cosinusoidal variations of phase can be considered as sinusoidal variations of frequency because, as we have said earlier $\omega=d\phi/dt$, and so these variations will give rise to variations of amplitude. Mathematically, a sloping (linear) frequency characteristic is the equivalent of a time derivative: take a time derivative of Equation 2.4b

$$\frac{d}{dt}f(t)=\frac{d}{dt}\int_{-\infty}^{\infty}f[f]\,e^{j2\pi ft}\,df=\int_{-\infty}^{\infty}f[f]j2\pi f\,e^{j2\pi ft}\,df$$

$f[f]$, a given frequency characteristic, has now become $j2\pi f\,.\,f[f]$. If $f[f]=1$, for example, then it now becomes $j2\pi f=j\omega$, proportional[2] to f or ω.

$$\frac{d}{dt}R(t)=(\omega_0+\Delta\omega\sin\omega_a t)\cos\left(\omega_0 t-\frac{\Delta\omega}{\omega_a}\cos\omega_a t\right) \tag{2.20}$$

We now have a carrier of amplitude ω_0 and a modulation vector of amplitude $\Delta\omega$. An envelope detector which follows the sloping frequency characteristic will complete the job. The complete device is called a discriminator. It is usual to precede the discriminator with an amplitude limiter so as to eliminate all spurious amplitude variations from $R(t)$ before converting it to an amplitude-modulated signal.

Conventional discriminators use the composition of two vectors at right angles when thee is no frequency deviation and when the frequency departs from its state of rest the angle between the vectors departs from 90°. Figure 2.6(a) shows an elementary balanced discriminator. The capacitors C tune the inductances to resonance at the centre frequency of the intermediate frequency band of the receiver. The resistors R represent the series resistances of the inductances. At resonance the voltage V_1 will be in quadrature with the voltages in each half of the secondary winding. The voltage V_3 will be the vector sum of V_1 and V_2, and V_4 will be $\vec{V}_1-\vec{V}_2$. If, to simplify, we let $V_1=V_2=V$ we have (Fig. 2.6(b))*

$$V_3=V\sqrt{2(1+\cos\phi)}$$

$$V_4=V\sqrt{2(1-\cos\phi)}$$

The detected voltage across AB will be the algebraic difference.

$$V_3-V_4=2V(\cos\phi/2-\sin\phi/2) \tag{2.21}$$

Equation 2.21 is shown as curve a in Fig. 2.7. The scales are A and D (or ϕ). The dashed straight line shows that $(V_3-V_4)/2V$ is reasonably linear when $45°<\phi<135°$. Now, if the primary and secondary circuits of Fig. 2.6(a) each

* Note that, in the figure $V_1\neq V_2$.

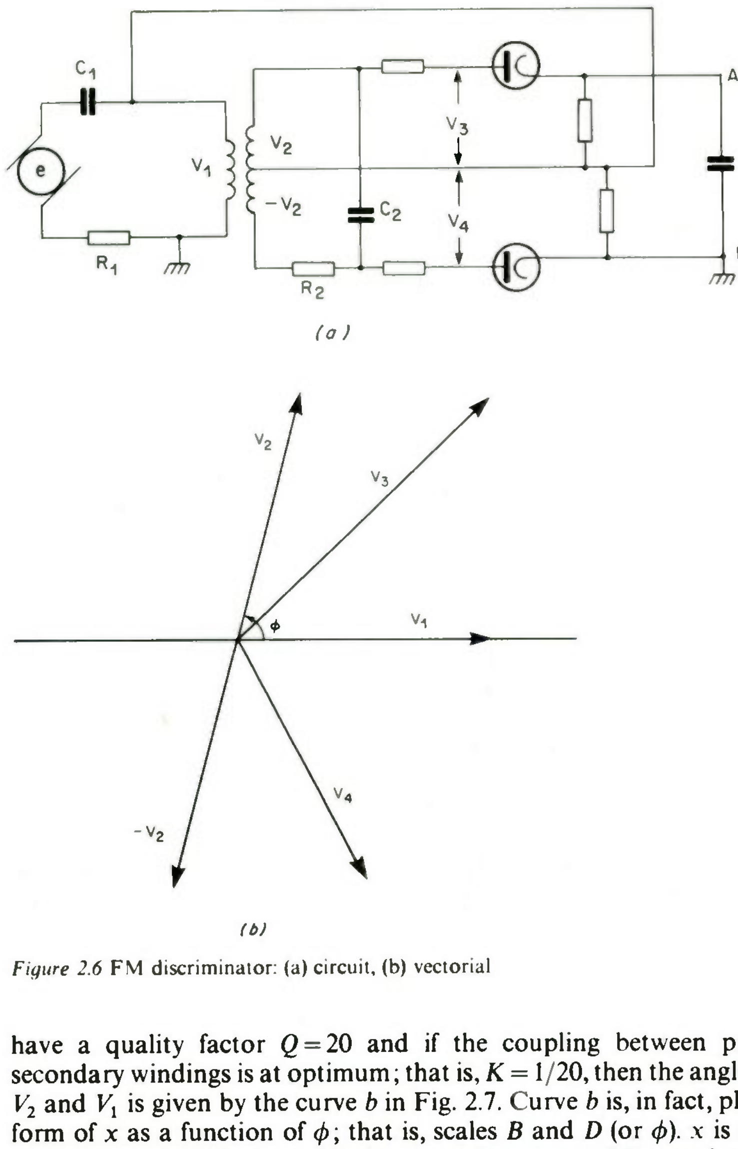


Figure 2.6 FM discriminator: (a) circuit, (b) vectorial

have a quality factor $Q=20$ and if the coupling between primary and secondary windings is at optimum; that is, $K=1/20$, then the angle ϕ between V_2 and V_1 is given by the curve *b* in Fig. 2.7. Curve *b* is, in fact, plotted in the form of x as a function of ϕ; that is, scales *B* and *D* (or ϕ). x is a frequency variable and is f/f_i where f_i is the centre frequency of the receiver's i.f. band We may write $f=f_i+\Delta f$ where Δf will represent a departure from f_i or the frequency deviation of the received carrier due to frequency modulation. x is thus $1+\Delta f/f_i$. Again, by referring to the dashed straight line we see that ϕ is an almost linear function of x and therefore of f when $0.99<x<1.01$. Finally, curve *c* shows the output voltage across *AB* in Fig. 2.6(a), namely $(V_3-V_4)/2V$ (scale *C*), as a function of Δf (scale *E*), based on the assumption that Fig. 2.6(a) is tuned to an i.f. centre frequency $f_i=10$ MHz. Once again, the dashed straight line shows reasonable linearity when $-75\text{ kHz}<\Delta f<75\text{ kHz}$. The

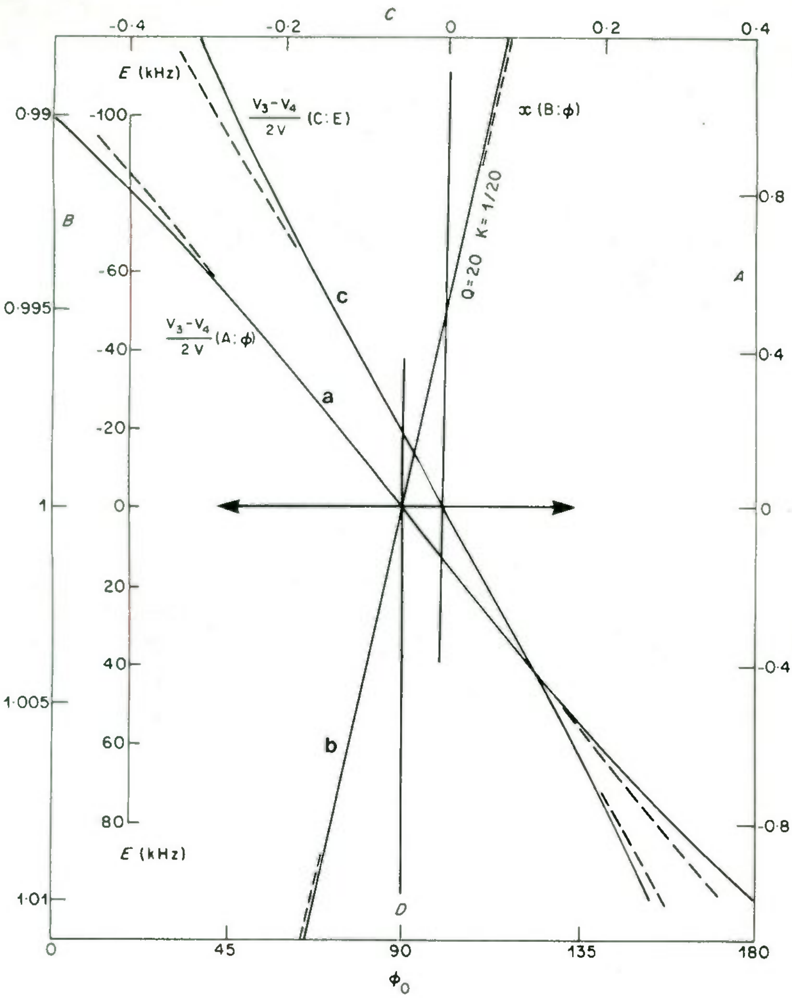


Figure 2.7 FM discriminator theory

foregoing approximate calculations are intended to give an insight into the working of a discriminator.

Some modern discriminators use a counter to count the zero crossings of the frequency modulated wave, Equation 2.16. This is an accurate method of demodulation.

Measurement of deviation A rather precise method relies on checking those values of the modulation index $m = \Delta f/f_a$ that give rise to zeros in the carrier amplitude. Equation 2.17 shows the carrier to be

$$J_0(\Delta f/f_a) \sin \omega_0 t$$

and

$$J_0(\Delta f/f_a)=0$$

when $(\Delta f/f_a)=2.4048$, 5.5201, 8.6537, 11.7915, 14.9304, 18.0711, etc.

Signal-to-noise ratio[2] We shall compare the f.m. signal-to-noise ratio with that for DSB AM with carrier. When we know the FM/AM improvement factor we can then always calculate the required aerial e.m.f. to give a given signal-to-noise ratio in a FM receiver by first calculating the aerial e.m.f. for an AM (DSB with carrier) receiver having the same gain and same noise factor and then multiplying by the reciprocal of the FM/AM improvement factor or subtracting the decibel figure of it. The simplest method of obtaining this factor is to assume the simplest possible discriminator; that is, a differentiator as shown by Equation 2.20. The FM receiver's discriminator characteristic then becomes that shown in Fig. 2.8(a); it is a straight line proportional to $f/\Delta f$ the relative deviation. We now assume that the magnitude of the receiver noise at maximum (or any other fixed value), deviation is the same as that for the comparable AM receiver at 100% modulation (or any other fixed percentage). We next assume the noise spectrum to be uniform over the i.f. band of each receiver. The AM noise power at baseband will be, if N is the constant noise voltage/Hertz,

$$P_{\mathrm{AM}} \propto \frac{1}{2f_a}\int_{-f_a}^{f_a} N^2\, df \propto N^2 \tag{2.22}$$

where f_a is the maximum audio or modulation frequency. The FM noise power at baseband will be (see Fig. 2.8(b))

$$P_{\mathrm{FM}} \propto \frac{1}{2f_a}\int_{-f_a}^{f_a} \left(N\frac{f}{\Delta f_{\max}}\right)^2 df \propto N^2\frac{f_a^2}{3(\Delta f_{\max})^2} \tag{2.23}$$

The ratio of FM baseband noise to that of AM noise is thus $[20\log_{10}(f_a/\Delta f_{\max})-4.72]$ decibels for equal modulation output from each type of receiver. If the two systems AM and FM are assumed to have the same audio band of, say, 15 kHz and the FM system has a maximum deviation of 75 kHz we have

$$\text{Signal/noise: } \frac{\text{FM}}{\text{AM}} = -[20\log_{10}(15/75)-4.77] = 18.75 \text{ dB} \tag{2.24}$$

a worthwhile improvement at the expense of occupying a bandwidth of five times that required for DSB AM and ten times that required for SSB AM.

Because the FM baseband noise spectrum is triangular there is some advantage in pre-emphasizing the higher audio frequencies at the sending end and de-emphasizing them (to maintain a uniform audio-frequency characteristic) at the receiver. A pre- and de-emphasis characteristic is shown in two parts in Fig. 2.9. The upper circuit represents a 50 μs de-emphasis circuit and its response is given in decibels by the graph below. The lower circuit represents 50 μs pre-emphasis and it can be made to approximate very closely to the reciprocal of the response of the de-emphasis circuit if $R' \ll R$. Its gain

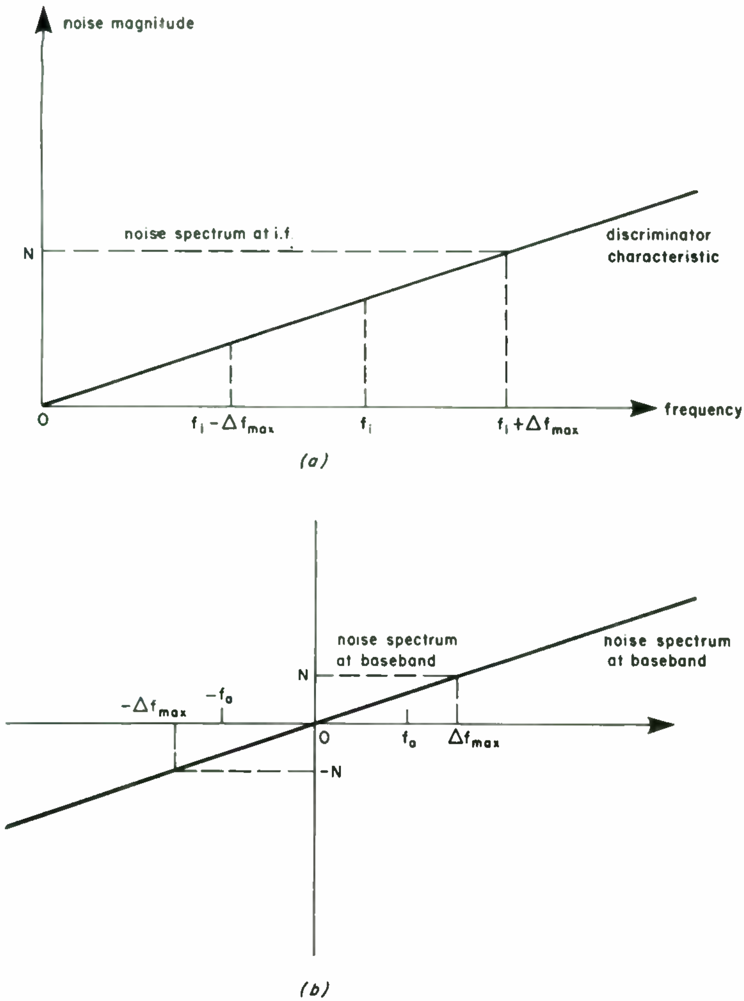


Figure 2.8 FM signal-to-noise ratio

curve is thus the same as that of de-emphasis circuit, only the decibel figures will be positive. The modulation is, theoretically, unaffected by pre- and de-emphasis and thus there should be a further worthwhile gain in signal-to-noise ratio. Unfortunately, pre-emphasis tends to overmodulation (over deviation), and it is necessary to reduce the actual modulation by several decibels to avoid this. The result is that pre- and de-emphasis result, overall, in only about 2 dB increase in signal-to-noise ratio. We can now revise our FM/AM improvement factor by adding 2 dB to it, thus reaching a figure of 21 dB.

The FM/AM improvement threshold The improvement in FM signal-to-noise ratio over that of AM, $\Delta f\sqrt{3}/f_a$, suffers a threshold at which any

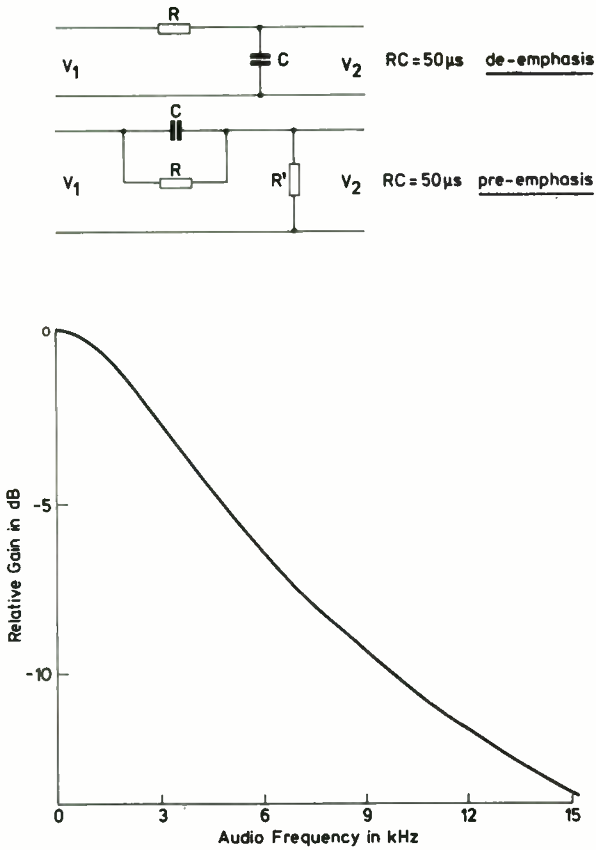


Figure 2.9 FM pre- and de-emphasis

decrease in signal-to-noise ratio or in signal-to-interference ratio in the r.f. or i.f. portion of the receiver gives rise to a catastrophic decrease in output signal-to-noise (or interference) ratio. What happens to the threshold is that the noise spectrum at baseband (after demodulation by the discriminator) becomes more or less uniform instead of maintaining the triangular form shown in Fig. 2.8(b). Figure 2.10 shows at (a) a case in which the noise or interference, represented as a vector rotating around the (unmodulated) carrier, is less in magnitude than the carrier. If the interference is, itself, an unmodulated carrier, its angle with respect to the wanted carrier will be αt where α is the difference in frequency between those of the wanted and interfering carriers. If the interference is random noise, then the interference vector in Fig. 2.10(a) will represent the unit impulse or Dirac function response of the receiver r.f./i.f. circuits; that is, a more or less damped sinewave at i.f. centre frequency modulated by a wave whose frequency is related to the i.f. half-bandwidth. The locus of such a vector is no longer the circle depicted in Fig. 2.10, but is a sausage-shaped figure. Moreover, its phase with respect to the wanted carrier

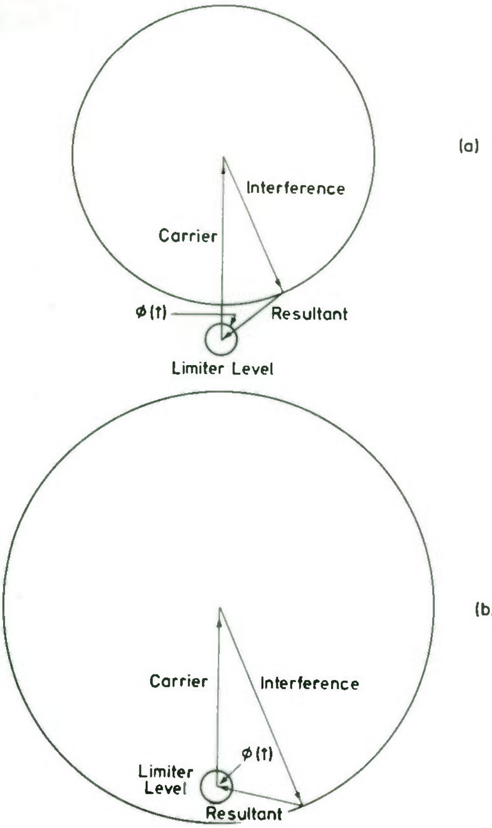


Figure 2.10 FM noise threshold

is a complicated function of time which starts at an arbitrary angle that depends upon the time of arrival of the noise pulse. The above discussion applied equally to impulsive noise like motor car ignition interference or random noise, except that for random noise account must be taken of its Gaussian statistical frequency distribution. The threshold phenomenon is not changed by the type of interference, although its onset value does depend upon it.

Now, in Fig. 2.10(a), it can be seen that the angle $\phi(t)$ between resultant and carrier varies in a more or less symmetrical manner around zero. If we assume, for simplicity, that the starting angle between interference and wanted carrier is zero, then $\phi(t)$ starts at zero, rises to a positive maximum, sinks to zero, 'rises' to a negative 'maximum' and sinks to zero. If the starting angle of the interference is, say, at right angles to the carrier, then $\phi(t)$ starts at some angle other than zero, sinks to zero, 'rises' to a negative 'maximum', sinks to zero and rises again to its initial angle. In the first example there were two zeros and a zero crossing; in the second example there were two zero crossings. The frequency modulation of the resultant of carrier plus interference, with respect to the carrier, is $d\phi(t)/dt$ and for every lobe of $\phi(t)$ its derivative will have one positive lobe and one negative lobe. Now the spectrum of a single pulse having just the one lobe may be approximated by putting the unit impulse or Dirac function $\delta(t)$ into Equation 2.5. First, $\delta(t)$ has infinite height and zero time duration with unit area—it is a spike! Thus

$$f[f] = \int_{-\infty}^{\infty} \delta(t)\, e^{-j2\pi ft}\, dt \tag{2.25}$$

If we arrange that the spike $\delta(t)$ occurs at $t=0$, then the integral will be zero outside of the very small period when $\delta(t)$ is other than zero so we can assume that $e^{-j2\pi ft}$ has not had time to 'get going' and is $e^{-j0}=1$. Equation 2.25 thus becomes

$$f[f] = \int_{-\infty}^{\infty} \delta(t)\, dt = 1 \tag{2.25a}$$

since, by definition, the area under $\delta(t)$ is one. Thus the spectrum of a single spike is uniform. If the spike becomes rounded and has a finite height and a finite duration—still with unit area, its spectrum will be more or less uniform at the lower frequencies, but it will tail-off at the higher frequencies just as if it had been band limited. If we, therefore, take the simplest possible view of the phenomenon, we consider $\phi(t)$ as a single pulse of phase whose frequency is $d\phi(t)/dt$. Now the pulse has a more or less uniform spectrum so its derivative with respect to time will have a more or less triangular spectrum as we showed in the section on Demodulation of a Frequency Modulated Signal with the aid of Equation 2.4(b).

Now consider Fig. 2.10(b). It is easy to see that whatever the form of the interference vector, $\phi(t)$ will undergo a rotation of 2π radians and it will not return to zero. Its time derivative will consist of a single pulse whose spectrum will be more or less uniform. The uniform spectrum will contain a lot more energy than the triangular spectrum resulting from Fig. 2.10(a). Thus, when the interference magnitude exceeds the amplitude of the wanted carrier the FM signal-to-noise ratio is much reduced. That is the FM threshold.

If the interference is a continuous or a modulated carrier wave, Fig. 2.10(b) shows that it will be the wanted carrier vector that rotates around the interference vector and the limiter circle will operate upon the interference, whilst the wanted carrier becomes, in effect, the interference. This is a feature of the threshold phenomenon and it is referred to as the capture effect.

2.3.2 Phase modulation

Let the modulating wave be proportional to $\Phi_0 \sin \omega_a t$ and let this wave modulate the phase, not directly the frequency, of the radio frequency wave $A \sin(\omega_0 t + \Phi)$. Thus, the transmitted signal is

$$R(t) = \sin(\omega_0 t + \Phi_0 \sin \omega_a t) \tag{2.26}$$

This is completely analogous to the FM signal except that $m = \Delta\omega/\omega_a$ now becomes a simple phase angle Φ_0. Equations 2.17–2.19 are valid if Φ_0 be substituted for m. A conventional FM discriminator will demodulate this wave, but the time derivative of $R(t)$ above will yield an audio wave $\omega_a \Phi_0 \cos \omega_a t$ which rises at 6 dB per octave with ω_a. A uniform baseband signal (as a function of frequency) can then be obtained by inserting a circuit whose response falls by 6 dB per octave.

2.4 PULSE CODE MODULATION (PCM)

We now enter the digital field and we shall restrict our discussion to binary digital systems. Instead of sending a signal one of whose parameters is made to vary in accordance with the strength of the modulation, that is to say, by analogue means, we can sample the modulation at regular intervals, measure the height of each sample and convert the height into a group of digits representing a binary number. For example, if a sample height was 2.3 V we should have a digit group of:

2^2	2^2	2^1	2^0	2^{-1}	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}	2^{-9}
		1	0	0	1	0	0	1	1	0	0	1

The row of negative powers of 2 approximate 0.3, in fact, the total comes to 2.29883. We can choose a range of sample values, say 0 to 127, and seven digits would cover this range. A row of seven zeros would mean zero sample height; a single one in the left-hand column would mean $2^6 = 64$. A row of ones would mean $2^6 + 2^5 + 2^4 + 2^3 + 2^2 + 2^1 + 2^0 = 127$, etc. A seven binary digit (bit) signal could be sent over seven separate wires or communication channels or, at seven times the bandwidth of one of the channels, the signal could be 'serialised' and sent in a single channel. In PCM a 'one' is represented by a short pulse and a zero by no signal. The duration of the 'one' pulses must be determined by the number of digits to be sent per unit time. Let us take a television example. There are 625 lines transmitted in $\frac{1}{25}$ s so the duration of a television line is 64 μs. About 12 μs of this are taken up by the line-blanking pulse, leaving 52 μs for picture information. The video band is 5.5 MHz and thus a true picture element has a duration of $1/5.5 = 182$ ns. (The figure of $1/(2 \times 5.5)$ is sometimes taken and this is the time duration of a black-to-white or white-to-black transition, whereas the time required to produce a white dot on a black background or vice versa will be twice that figure or $1/5.5$.) Thus each sample should not last longer than 182 ns and the number of picture elements per line is $52/0.182 = 286$. It has been found by trial and subjective testing that an eight-bit digital signal provides enough levels to represent a television video signal with adequate precision, so we have to fit a possible eight digit pulses in each 182 ns of time. The reciprocal of $182 \times 10^{-9}/8$ is about 44 MHz. However, a further consideration is that of the Nyquist sampling theorem[2]. This states that if a signal is to be sampled without distortion due to aliasing* the sampling rate must not be less than twice the highest frequency in the signal spectrum. Our figure of 44 MHz now becomes $2 \times 5.5 \times 10^6 \times 8 = 88$ MHz. We thus require 16 times the video bandwidth that is required for analogue television. The advantages of digital transmission are

* See Section 2.82.

several, but one of the principal ones is the gain in signal-to-noise ratio. Before discussing that aspect, we must consider sound transmission. For high quality sound, the Nyquist sampling theorem must be rigidly obeyed and the usual sampling frequency is 32 kHz or, for television sound-in-synch, twice the line-scan frequency, namely 31.25 kHz. It has been found from subjective tests that 13 bits or 8192 levels are required for high quality audio, but if suitable companding is used the figure may be reduced to 10 bits (1024 levels). This is the figure used for the sound-in-synchs system. For VHF radio, however, 13 bits has been adopted without companding. Such a system requires a bandwidth of about 416 kHz, precisely 416K bit/s. This figure takes no account of additional bits for error detection and other purposes.

2.4.1 Noise

The receiver has only to recognize the existence of a 'one' pulse or its absence as a 'nought'. If the noise level is such that the receiver always correctly registers the 'ones' and 'noughts', then there is no transmission noise in the output. There is, however, another form of noise known as quantising noise. This results from the inaccurate description of a sample height which falls between two quantum levels. This indeterminacy results in more or less uniform-spectrum 'white' noise in audio transmissions and appears as 'contour' lines in areas of very slowly changing luminance in television images. The presence of quantising noise indicates an insufficient number of quantum levels and therefore of bits. It can be mitigated, however, by the addition of a low level of random noise or of a low level of pseudo-random noise. In the latter case, this may be entirely eliminated in the receiver if the latter is designed to generate an equal pseudo-random noise in synchronism with the received version. Such signals are referred to as 'jitter' signals.

Quantising noise is easily calculated. Consider one quantum step q. The maximum indeterminacy of measurement of a sample of the original analogue signal is $q/2$. The mean square error is therefore

$$\bar{E}^2 = \frac{1}{q/2}\int_0^{q/2} x^2\,dx = q^2/12 \tag{2.27}$$

or

$$\sqrt{\bar{E}^2} = \frac{q}{2\sqrt{3}} \tag{2.27a}$$

The maximum value of the received signal is, of course, $q \times 2^n$ where n is the number of bits. The signal-to-r.m.s. quantising noise ratio is therefore $2^{n+1}\sqrt{3}$. For a 13-bit audio system the figure is 28,378 or 89 dB. For an 8-bit television system the figure is 59 dB.

Since there is no transmission noise in a received digitally coded signal, the concept of signal-to-noise ratio must undergo a radical revision. In digital systems it is 'error rates' that have replaced signal-to-noise ratio. If we regard random or continuous noise as a succession of voltage or current spikes having random times of arrival and random magnitudes, then every so often a noise spike will be large enough to persuade the receiving decoder that there is not a nought but rather a one. How often will such an error occur? Random noise due to the thermal agitation of electrons in circuits or to the behaviour of electrons emitted from hot cathodes in vacuum tubes is said to have a Gaussian statistical frequency distribution: the probability of occurrence of a noise voltage x in the range $x-\frac{1}{2}\,dx<x<x+\frac{1}{2}\,dx$, is

$$dP=\frac{1}{\sigma\sqrt{2\pi}}\,e^{-x^2/2\sigma^2}\,dx \tag{2.28}$$

and the probability of occurrence of a voltage exceeding α times the standard deviation or r.m.s. noise voltage σ is

$$P'=\frac{1}{2}-\frac{1}{\sqrt{\pi}}\int_0^{\alpha/\sqrt{2}} e^{-u^2}\,du \tag{2.29}$$

This assumes that only positive-going noise voltages can cause trouble. If, for example, digital 'ones' were represented by positive voltage pulses and digital 'noughts' by negative voltage pulses it would be necessary to double P', but, of course, the digital voltage pulses might have been twice the magnitude of the one-sided pulse transmission system. Figure 2.11 shows P' as a function of α. For example, when $\alpha=0$, that is, nought times the r.m.s. (standard deviation) noise, the value of P' is $\frac{1}{2}$: there is one chance in any two noise pulses that there will be a positive one. When $\alpha=4\frac{3}{4}$, $P'=10^{-6}$ or one in a million (on average) noise pulses will exceed $4\frac{3}{4}\,\sigma$. Now $4\frac{3}{4}=13.5$ dB and so an error rate of 10^{-6} will occur in a digital system in which the digit pulses exceed the r.m.s. noise by 13.5 dB. This assumes that the receiver decoder is so good that a noise pulse must actually equal a digit pulse for an error to occur. It is frequently assumed that the digit pulse must exceed the noise pulse by 6 dB if conventional

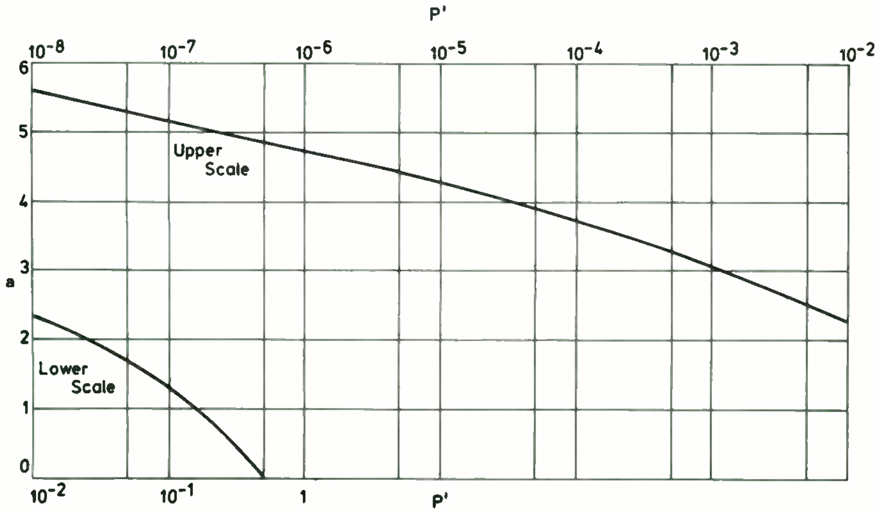


Figure 2.11 PCM error rate

decoding circuits are used. In this case the 10^{-6} error rate requires a digit pulse magnitude that exceeds the r.m.s. noise by 19.5 dB. Assuming the two-to-one magnitude ratio we find from Fig. 2.11 that an error rate of 10^{-8} requires a digit-pulse-to-r.m.s. noise ratio of $2 \times 5.58\ \sigma$, or 21 dB. For television transmissions, a signal-to-r.m.s. noise ratio of 19.5 dB for digital transmission compares very favourably with approximately 40 dB required for analogue transmission. However, the bandwidth required is, theoretically, 88 MHz compared with 6.25 MHz for analogue transmission. The bandwidth ratio is, therefore, about 14 which represents an increased noise power of 11.5 dB; so the digit pulse magnitude will have to be 19.5 dB greater than an r.m.s. noise power which itself will be 11.5 dB greater than that occurring in a 6.25 MHz band used for analogue transmission—all for an error rate of 10^{-6}.

There are error recognition, error concealment and error correction techniques all of which require the addition of further bits to the digital signal. The use of error recognition and concealment can, for both television and sound, permit a higher noise level than that for an error rate of 10^{-6}. Error rates of 10^{-4} or 10^{-5} are being considered as the maximum allowable for high quality broadcasting.

2.4.2 Bit rate reduction

Research is in train with the object of reducing the number of bits required adequately to represent a (colour) television signal. Some success has been achieved in reducing the frequency of sampling from 2×5.5 MHz to 2×4.43 MHz$=8.86$ MHz; that is, twice the frequency of the colour sub-carrier. Certain techniques are used to avoid aliasing[2] which the Nyquist theorem predicts. A further means of reducing the bit rate is to use differential pulse code modulation[3] or DPCM. In this system it is assumed that there is some correlation between adjacent picture elements both along the television line and down the picture from line to line.

The transmitted digital information consists of the digital representation of the difference between an actual sample of the television waveform and a predicted value for that sample plus any error resulting from the quantiser which may be non-linear in the sense that it quantises small differences more accurately than large ones. The use of non-linear quantisation assumes that large differences between sample and predicted value of it are less frequent than small differences. This enables some bit saving to be achieved. The receiver has an identical predictor to that used at the sending end and so predicts in an identical manner. All it needs, therefore, to reproduce the real actual sample is the difference between the sample and its predicted value and it is just this information that is transmitted—non-linear quantising errors included inevitably. The receiver decoder must start with the same sample value in its predictor, as is in the transmitter predictor and this setting-up operation must be repeated at intervals in case a transmitted error has intervened. This re-setting can, for example, be repeated during each line-blanking interval.

2.5 PHASE MODULATION WITH PULSES

So far, we have treated a PCM signal as a baseband signal and, of course, it could be transmitted in that form through a transmission line. For microwave transmission, the complete PCM pulse signal could be amplitude modulated or frequency modulated on to a microwave carrier and, in fact, this technique is used. It does not seem to require any further discussion here since we have already studied AM and FM. There are, however, further methods of transmitting a signal consisting of a sequence of pulses such as PCM. One of them is phase modulation. Another is phase-shift keying (PSK). The first is simply to pulse the phase of a sinusoidal carrier in accordance with a digit 'one' at constant carrier amplitude. This is genuine phase modulation and it may, therefore, occupy more bandwidth than the second method which is to allow the carrier amplitude to fade to zero during phase changes. First let us try to find the spectrum of a constant amplitude carrier undergoing a pulsed phase change. This is not a simple problem. Probably the easiest phase-change waveform to adopt is

$$\Phi_0 \cos 2\pi F t \qquad \text{with} \qquad -1/4F < t < 1/4F \tag{2.30}$$

That is the positive lobe of a cosine wave around $t=0$. The complete carrier may be written as

$$\cos[2\pi f_0 t + \Phi_0 \cos 2\pi F t] \tag{2.31}$$

but with the restriction, expression 2.30, on t applying only to the cosinusoidal phase term. If we expand expression 2.31 as we did in Equation 2.18 we have

$$\begin{aligned}\cos[2\pi f_0 t + \Phi_0 \cos\theta] \\ &= \cos(2\pi f_0 t)[J_0(\Phi_0) - 2J_2 \cos 2\theta + 2J_4 \cos 4\theta - \ldots] \\ &\quad - \sin(2\pi f_0 t)[2J_1(\Phi_0)\cos\theta - 2J_3 \cos 3\theta + \ldots]\end{aligned} \tag{2.32}$$

where $\theta = 2\pi F t$.

Each of the two terms of the right-hand side of Equation 2.32 is a product of two functions of time. Now the spectrum of the product of two time functions is the 'convolution' product[2.4] of their spectra, so if we find the spectra of each of the functions of θ (therefore with $-1/4F < t < 1/4F$) in the brackets of Equation 2.32 and then the spectra of $\cos 2\pi f_0 t$ and $\sin 2\pi f_0 t$, with $-\infty < t < \infty$ we can obtain our final spectrum by convolving the latter spectra with their appropriate coefficient spectra from the bracketed functions.

Now the spectrum of terms such as $J_n(\Phi_0)\cos n\theta$ is, from Equation 2.5,

$$\frac{J_n(\Phi_0)}{2\pi F}\int_{-\pi/2}^{\pi/2} \cos n\theta \, e^{-j\alpha\theta}\, d\theta \tag{2.33}$$

where $\alpha = f/F$. Expression 2.33 is readily calculated to be

$$\frac{J_n(\Phi_0)}{2\pi F}\left[\frac{\sin(n+\alpha)\frac{\pi}{2}}{(n+\alpha)\frac{\pi}{2}} + \frac{\sin(n-\alpha)\frac{\pi}{2}}{(n-\alpha)\frac{\pi}{2}}\right]\frac{\pi}{2} \tag{2.33a}$$

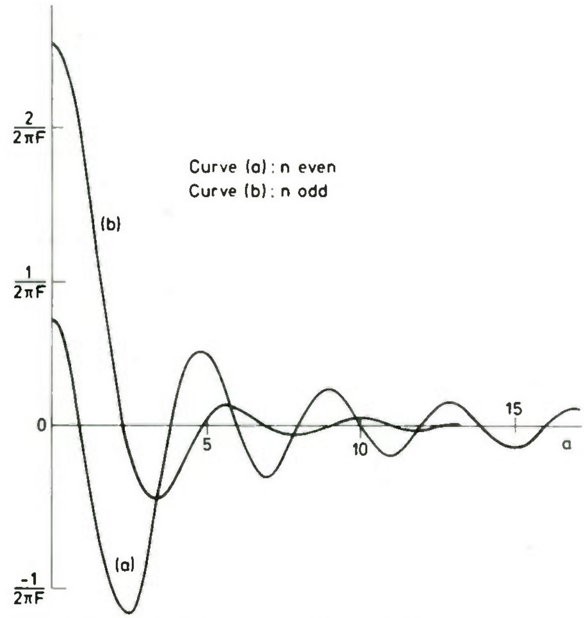


Figure 2.12 Pulsed phase modulation spectra.
Curve (a) spectrum of $[J_0(\phi_0)-2J_2(\phi_0)\cos 2\theta-4J_4(\phi_0)\cos 4\theta\ldots]$
Curve (b) spectrum of $[2J_1(\phi_0)\cos\theta-2J_3(\phi_0)\cos 3\theta+2J_5(\phi_0)\cos 5\theta\ldots]$

So the spectrum of $J_0(\Phi_0)$ is obtained from 2.33a with $n=0$. Similarly, the spectrum of $-2J_2(\Phi_0)\cos 2\theta$ (see expression 2.32) is obtained from 2.33a with $n=2$ and the result to be multiplied by -2 and so on for each of the terms in the brackets of Equation 2.32. The method is to form the appropriate sums of terms with $n=0, 2, 4$, etc. and $n=1, 3, 5$, etc. all for one value of $\alpha=f/F$ and so on. Thus, for example, letting $\Phi_0=2$ the spectrum of the bracketed coefficient of $\cos 2\pi f_0 t$ in Equation 2.32 is the ordinate values of curve a in Fig. 2.12. The spectrum of the bracketed coefficient of $-\sin 2\pi f_0 t$ in Equation 2.32 is the ordinate values of curve b in Fig. 2.12. Figure 2.12 shows half of each spectrum. The other halves are symmetrical with curves a and b about the zero-frequency ordinate. If we assume $F=1$ Hz. Figure 2.12 has the ordinate dimension of volts (say) per cycle per second of bandwidth. If $F=1$ MHz, the dimension will be volts per MHz. (The phase shift used for Fig. 2.12, namely $\Phi_0=2$ rad or 114.6° instead of $\Phi_0=\pi/2$ rad or 90°, was due to the non-availability of higher order Bessel functions of 1.5708.)

We now consider the spectra of $\cos 2\pi f_0 t$ and $\sin 2\pi f_0 t$; this time with t unrestricted, that is, $-\infty<t<\infty$. To find the spectrum of $\cos 2\pi f_0 t$ and $\sin 2\pi f_0 t$ use is made of the Dirac function or unit impulse. This function, written $\delta(x)$, has infinite height for zero width and unit area[2]. Thus

$$\int_{-\infty}^{\infty}\delta(x)\,dx=1 \qquad \text{at } x=0 \tag{2.34}$$

and it only exists when $x=0$. Similarly,

$$\int_{-\infty}^{\infty} \delta(x \pm a)\, dx = 1 \qquad \text{at } x = \mp a$$

Now, working backwards as it were, let us find the time function whose spectrum is $\delta[f-f_0]$.

$$g(t) = \int_{-\infty}^{\infty} \delta[f-f_0]\, e^{j2\pi tf}\, df \tag{2.35}$$

(Equation 2.4b).

Now, since

$$\int_{-\infty}^{\infty} \delta[f-f_0]\, df = 1$$

when $f=f_0$ and 0 ($f \neq f_0$), Equation 2.35 becomes

$$g(t) = e^{j2\pi f_0 t} \int_{-\infty}^{\infty} \delta[f-f_0]\, df \tag{2.35a}$$

because $e^{j2\pi tf}$ becomes a constant equal to $e^{j2\pi f_0 t}$.

Thus we have

$$g(t) = e^{j2\pi f_0 t}$$

since the integral is unity and the spectrum of $\cos 2\pi f_0 t + j \sin 2\pi f_0 t$ is $\delta[f-f_0]$. Similarly, the spectrum of $\cos 2\pi f_0 t - j \sin 2\pi f_0 t$ is $\delta[f+f_0]$. If we add the two expressions we find that the spectrum of $\cos 2\pi f_0 t$ is

$$\tfrac{1}{2}\{\delta[f-f_0] + \delta[f+f_0]\}$$

and, by subtracting, the spectrum of $\sin 2\pi f_0 t$ is[4]

$$-\tfrac{1}{2}j\{\delta[f-f_0] - \delta[f+f_0]\}$$

We now have to convolve

$$\tfrac{1}{2}\delta[f-f_0] + \tfrac{1}{2}\delta[f+f_0] \quad \text{and} \quad -\tfrac{1}{2}j\delta[f-f_0] + \tfrac{1}{2}j\delta[f+f_0] \tag{2.36}$$

with expression 2.33a expanded in accordance with expression 2.32 (brackets only, of course).

Convolution of a function by the delta function $\delta(x \pm a)$ is explained in Reference 2, page 51 and further on. Since the value or area of the delta function is unity, it amounts to a single ordinate of value 'one' at an abscissa of $\pm a$ and convolution with it amounts simply to a shift or translation by $\pm a$ of the abscissa values of the function being convolved. All we have to do, therefore, is to shift the spectra of Fig. 2.12 (after including their negative-frequency halves) to $f \pm f_0$ with signs respected and noting that the factor $j = \sqrt{-1}$ represents a rotation of 90°. The final spectrum is, therefore, three dimensional—frequency on one axis, real terms (cosines) on another and imaginary terms (sines) on the third. Figure 2.13 shows, omitting the half factors in expression 2.36, half (for positive frequencies only) the spectrum that will appear around each frequency $\pm f_0$. Note that, taking account of the

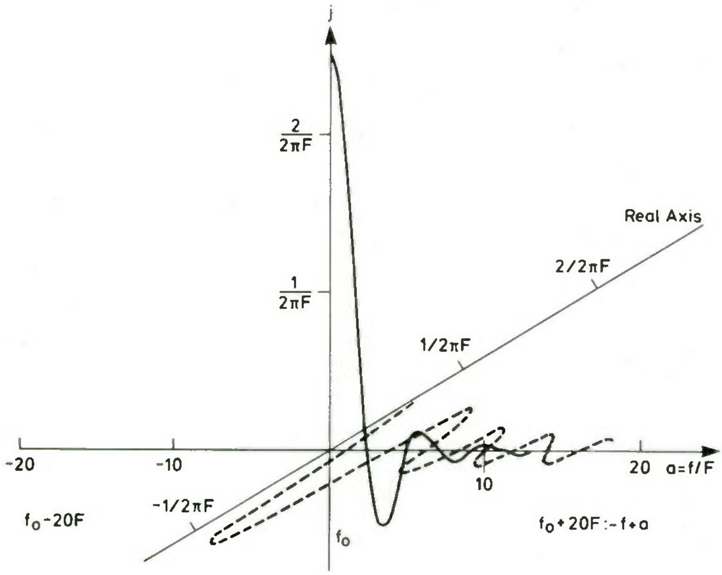


Figure 2.13 Spectrum of pulsed phase modulation

negative sign before sin $2\pi f_0 t$ in Equation 2.32 and the negative and positive signs before the two imaginary terms in expression 2.36, the imaginary spectrum around f_0 will be as shown in Fig. 2.13 whereas the imaginary spectrum around $-f_0$ will be the negative of that shown in Fig. 2.13. Figure 2.14 shows in relative terms the sideband power spectrum around f_0 (or around $-f_0$).

Note, from Fig. 2.14, that the power takes $\alpha=8$ or $f=8F$ for it to decrease to just less than 1% of its maximum value at $\alpha=0$. Since the pulse duration is $1/2F$ we can say that the period of the frequency at which power has decreased to just less than 1% is two-eighths of the pulse period. Of course, all the foregoing remarks apply to a phase pulse of 2 rad or 114.6°. Other values would result from a lesser or greater phase pulse.

2.6 PHASE-SHIFT KEYING (PSK)

With this system a sine- or cosine-wave generator produces two identical outputs. One of them is passed through a phase-changing network so that the two outputs are, say, cos $2\pi f_0 t$ and $\cos(2\pi f_0 t+\Phi_0)$. Often Φ_0 is 90° or 180°, but in our example we shall keep the value $\Phi_0=2$ rad or 114.6° so that we can compare PSK with true phase modulation.

By amplitude modulation, the wave cos $2\pi f_0 t$ is pulsed to zero and at the same time the wave $\cos(2\pi f_0 t+\Phi_0)$ is pulsed from zero to 1. Let us use, again as the pulse waveform

$$K(t)=\cos 2\pi Ft \qquad \text{with} \qquad -\frac{1}{4F}<t<\frac{1}{4F} \tag{2.37}$$

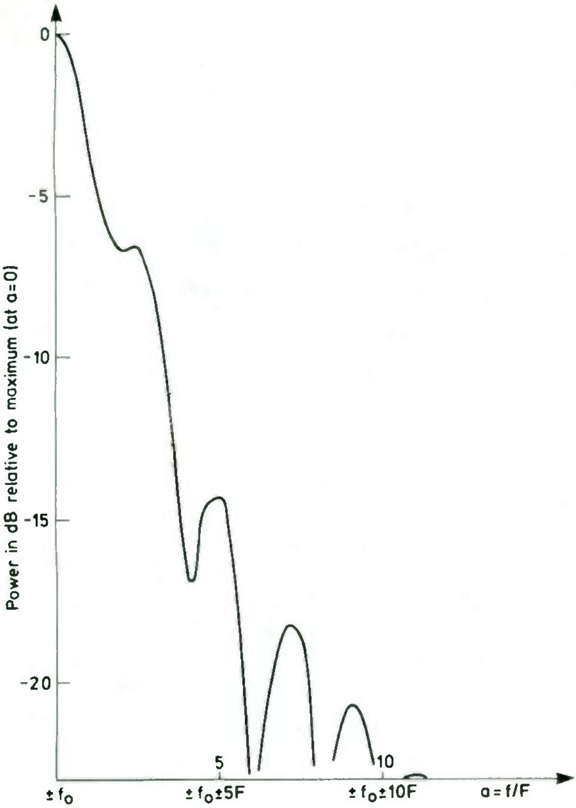


Figure 2.14 Relative power spectrum of pulsed phase modulation

The transmitted signal will be

$$R(t)=\cos 2\pi f_0 t - K(t)\cos 2\pi f_0 t + K(t)\cos(2\pi f_0 t + \Phi_0) \qquad (2.38)$$

or, to distinguish between the $-\infty < t < \infty$ for the signal waves $\cos 2\pi f_0 t$ and $\cos(2\pi f_0 t + \Phi_0)$ on the one hand, and the $-1/4F < t < 1/4F$ of the pulsing waveform on the other hand, we write

$$R(t,\theta)=\cos 2\pi f_0 t - K(\theta)\cos 2\pi f_0 t + K(\theta)\cos(2\pi f_0 t + \Phi_0) \qquad (2.38a)$$

where $K(\theta)=\cos\theta$, $-\pi/2<\theta<\pi/2$.

The spectrum $R[f]$ will consist of the spectrum of $\cos 2\pi f_0 t$ minus the convolution product of the spectra of $K(\theta)$ and of $\cos 2\pi f_0 t$ plus the convolution product of the spectra of $K(\theta)$ and $\cos(2\pi f_0 t + \Phi_0)$.

The spectrum of $K(\theta)$ is

$$K[f]=\int_{-\pi/2}^{\pi/2} \cos\theta \, e^{-j\alpha\theta}\, d\theta/2\pi F \qquad (2.39)$$

$$=\frac{\pi}{2}\left[\frac{\sin(\alpha+1)\frac{\pi}{2}}{(\alpha+1)\frac{\pi}{2}}+\frac{\sin(\alpha-1)\frac{\pi}{2}}{(\alpha-1)\frac{\pi}{2}}\right]\Big/2\pi F \tag{2.40}$$

For clarity we let

$$\delta[f-f_0]=H \qquad \delta[f+f_0]=J \qquad K[f]=K$$

and the final spectrum, after expanding $\cos(2\pi f_0 t+\Phi_0)$, is

$$R[f]=\tfrac{1}{2}\{H+J-(H+J)*K+(H+J)*K\cos\Phi_0$$
$$+j(H-J)*K\sin\Phi_0\} \tag{2.41}$$

where the asterisk $*$ means 'convolved with'. Collecting together the spectral terms around $f=f_0$, that is, H, and those around $f=-f_0$; that is, J, we have

$$R[f]=\tfrac{1}{2}\{H-H*K+H*K\cos\Phi_0+jH*K\sin\Phi_0$$
$$+J-J*K+J*K\cos\Phi_0-jJ*K\sin\Phi_0\} \tag{2.41a}$$

The H and the J all by themselves must be given the value unity, but only when $\alpha=0$, that is $f=\pm f_0$. They are spectral lines. The greater part of the complete spectrum is continuous—not a line spectrum, but one in which the ordinates are given in volts (say) per Hertz. This is shown in Fig. 2.15. The spectrum around $-f_0$ is the same as that shown in Fig. 2.15 except that the imaginary terms will then be negative. As before, only half the spectra is shown since they are symmetrical around $\pm f_0$.

Figure 2.16 shows the relative power spectrum around f_0 (or around $-f_0$). For pulsed PSK the spectrum is considerably narrower than for pulsed phase

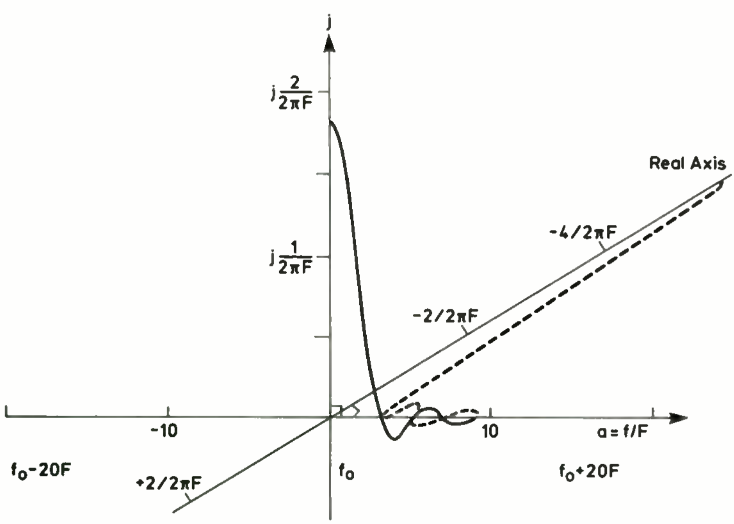


Figure 2.15 Spectrum of pulsed PSK

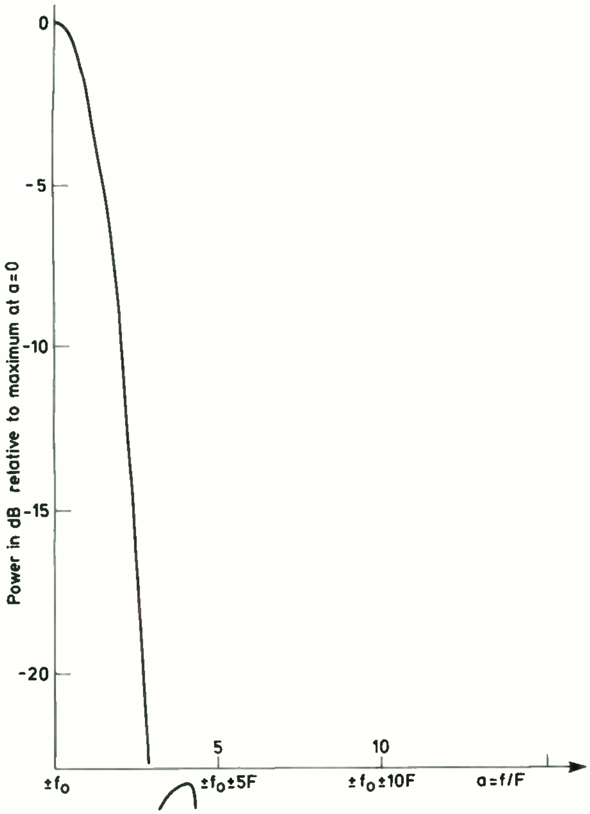


Figure 2.16 Relative power spectrum of pulsed PSK

modulation. The power decreases to less than 1% of its maximum before α reaches 3.

2.7 A SINGLE RECTANGULAR PULSE APPLIED FIRST AS PHASE MODULATION AND SECONDLY AS PSK

2.7.1 Phase modulation by a single rectangular pulse

Consider a wave

$$R(t)=\cos[2\pi f_0 t+K(t)] \quad \text{where } K(t)=\Phi_0 \quad \text{when } -\frac{T}{4}<t<\frac{T}{4}\ldots \tag{2.42}$$

We can write

$$R(t)=\cos 2\pi f_0 t \cos K(t)-\sin 2\pi f_0 t \sin K(t) \tag{2.42a}$$

Since $K(t)$, when it exists, is Φ_0, we have

$$R(t)=\cos 2\pi f_0 t \cos \Phi_0-\sin 2\pi f_0 t \sin \Phi_0 \tag{2.42b}$$

Now $\cos \Phi_0$ and $\sin \Phi_0$ are rectangular pulses occurring when $-T/4<t<T/4$, so we must find the spectrum of a rectangular pulse

$$U(\theta)=1 \qquad \text{when} \qquad -\frac{\pi}{2}<\theta<\frac{\pi}{2} \tag{2.43}$$

From Equation 2.5 we have

$$U[f]=\frac{1}{2\pi F}\int_{-\pi/2}^{\pi/2} 1\, e^{j\alpha\theta}\, d\theta \qquad \text{with} \qquad \alpha=\frac{f}{F}=fT=2f\,\frac{T}{2} \tag{2.44}$$

$$=\pi\,\frac{\sin \alpha\dfrac{\pi}{2}}{\alpha\dfrac{\pi}{2}}\Bigg/2\pi F \tag{2.44a}$$

So our final spectrum is

$$R[f]=\frac{1}{2}\left\{[\delta[f-f_0]+\delta[f+f_0]]*\pi\cos\Phi_0\,\frac{\sin \alpha\dfrac{\pi}{2}}{\alpha\dfrac{\pi}{2}}\right.$$

$$\left.-j[\delta[f-f_0]-\delta[f+f_0]]*\pi\sin\Phi_0\,\frac{\sin \alpha\dfrac{\pi}{2}}{\alpha\dfrac{\pi}{2}}\right\}\Bigg/2\pi F \tag{2.45}$$

With, as before, $H=\delta[f-f_0]$ and $J=\delta[f+f_0]$, we have

$$R[f]=\frac{\pi}{2}\left\{H*[\cos\Phi_0-j\sin\Phi_0]\,\frac{\sin \alpha\dfrac{\pi}{2}}{\alpha\dfrac{\pi}{2}}\right.$$

$$\left.+J*[\cos\Phi_0+j\sin\Phi_0]\,\frac{\sin \alpha\dfrac{\pi}{2}}{\alpha\dfrac{\pi}{2}}\right\}\Bigg/2\pi F \tag{2.45a}$$

or

$$R[f]=\frac{\pi}{2}\left\{H*\frac{\sin \alpha\dfrac{\pi}{2}}{\alpha\dfrac{\pi}{2}}\,e^{-j\Phi_0}+J*\frac{\sin \alpha\dfrac{\pi}{2}}{\alpha\dfrac{\pi}{2}}\,e^{j\Phi_0}\right\}\Bigg/2\pi F \tag{2.45b}$$

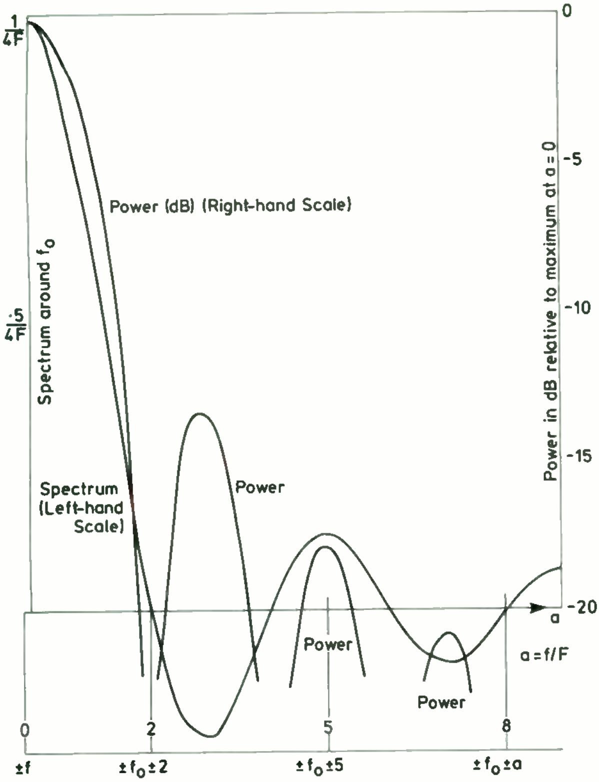


Figure 2.17 Signal and power spectra. Phase modulation, rectangular pulse

Equation 2.45(b) shows that the final spectra at f_0 and $-f_0$ can be plotted in two dimensions instead of three; that is, at a fixed angle of $-\Phi_0$ and $+\Phi_0$ at f_0 and $-f_0$ respectively. Figure 2.17 shows (for $\Phi_0=2$) either of the two terms of Equation 2.45b. The plane of the paper in which Fig. 2.17 has been drawn is assumed to be at either $-\Phi_0$ or $+\Phi_0$ from the plane formed by the real axis and the frequency axis. The power has fallen to less than 1% of its maximum when $\alpha=5.4$; that is, at a frequency whose period is 2/5.4 of the pulse duration.

2.7.2 A single rectangular pulse as amplitude modulation or PSK

Taking $U(\theta)$, as in Equation 2.43, we can write

$$R(t)=\cos 2\pi f_0 t - U(\theta)\cos 2\pi f_0 t + U(\theta)\cos(2\pi f_0 t+\Phi_0) \quad (2.46)$$

The second term of the right-hand side of Equation 2.46 'kills' the first term for the period $-\pi/2<\theta<\pi/2$ or $-T/4<t<T/4$, whilst the third term introduces the second carrier at phase Φ_0 for the same period. If we expand the third term of Equation 2.46, we have

$$R(t)=[1-U(\theta)+\cos\Phi_0 U(\theta)]\cos 2\pi f_0 t - \sin\Phi_0 U(\theta)\sin 2\pi f_0 t \qquad (2.46a)$$

or

$$R[f]=\tfrac{1}{2}\{(H+J)+(H+J)*(\cos\Phi_0-1)U[f] + j\sin\Phi_0 U[f]*(H-J)\} \tag{2.47}$$

Finally,

$$R[f]=\frac{\pi}{2}\left\{H+H*(\cos\Phi_0-1)\frac{\sin\alpha\frac{\pi}{2}}{\alpha\frac{\pi}{2}}+jH*\sin\Phi_0\frac{\sin\alpha\frac{\pi}{2}}{\alpha\frac{\pi}{2}}\right.$$

$$\left.+J+J*(\cos\Phi_0-1)\frac{\sin\alpha\frac{\pi}{2}}{\alpha\frac{\pi}{2}}-jJ*\sin\Phi_0\frac{\sin\alpha\frac{\pi}{2}}{\alpha\frac{\pi}{2}}\right\}\Big/2\pi F \tag{2.47a}$$

where, as before, $\alpha=f/F=2f(T/2)$. This time, again, we have an H and a J all by themselves, so when we come to evaluate Equation 2.47a we let $H=1$ only when $\alpha=0$; that is when we are situated at $f=f_0$. Similarly, $J=1$ only when $\alpha=0$ because we are then at $f=-f_0$. Thus once again (refer back to Equation 2.41a) we have a composite spectrum consisting partly of lines and mainly of a continuous curve; but we shall take account only of the continuous portion. The power spectrum is the same as that shown in Fig. 2.17 apart from the factor $2(1-\cos\Phi_0)=2.83$ for $\Phi_0=2$. Figure 2.18 shows the real and imaginary spectral components around $-f_0$. The imaginary component is reversed in sign (that is, positive) for $+f_0$.

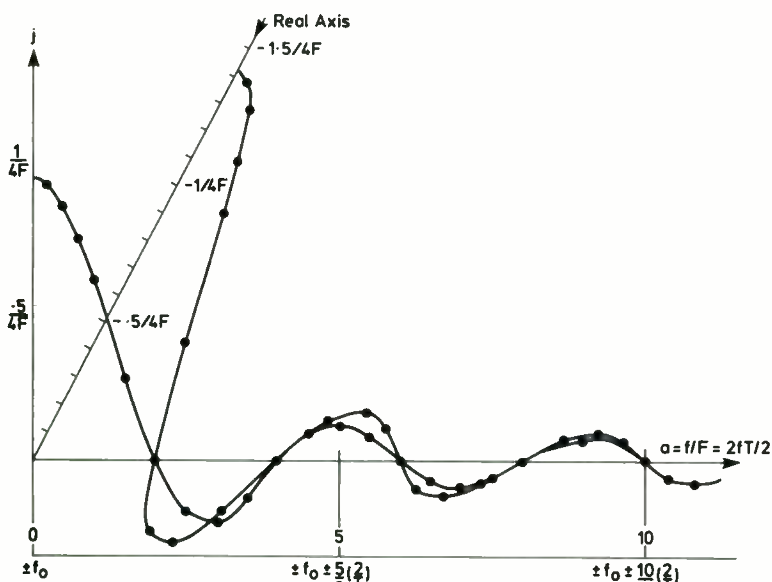


Figure 2.18 Signal spectrum PSK rectangular pulse

The studies described in Sections 2.5–2.7 raise some interesting points. The statement was made early in Section 2.5 that 'genuine phase modulation at constant carrier amplitude *may* occupy more bandwidth than the achievement of a phase change by amplitude modulation methods'. In Section 2.6, the power spectrum of the AM method (PSK) was compared to the power spectrum of true phase modulation and it was found that with a *half-cosine pulse* PSK occupied about three-eighths of the bandwidth required by true phase modulation. This result is what might be expected from experience with AM and FM in sound and television technology. Section 2.7, however, has shown that, for a *rectangular pulse*, both true phase modulation and the amplitude modulation system, PSK, require precisely the same bandwidth. This rather surprising fact is mentioned in Bennett and Davey[5] in connection with frequency shift keying compared with pulsed frequency modulation (at constant amplitude). It can be seen that when the baseband signal is a rectangular pulse, Equations 2.42a and 2.42b show that the phase modulation turns into amplitude modulation, in effect.

Another point of interest is the difference between a line spectrum (for a repetitive function) and a continuous spectrum (for a transient function). In truth, they cannot both be plotted on the same piece of paper, because they have different dimensions. The frequency axis is common to both spectra, but the heights or ordinate values in a line spectrum represent, or rather have the same dimensions as, the original time-function signal; for example, volts or amps or volts/metre, etc. in the case of continuous spectrum the ordinates have the dimensions of a density; for example, volts/hertz or amps/hertz or volts/metre/hertz, etc. This must be so, since in a continuous spectrum the magnitude of each single component is infinitesimal; for example $f[f]\,df$. If the dimensions of $f[f]$ are, say, volts/Hertz, then a finite spectral component at frequency f is

$$\int_{f-\Delta f}^{f+\Delta f} f[f]\,df$$

where Δf is as small as you like, but not infinitesimal as is df; Equation 2.22 in Section 2.3.1 is an example.

There is a great deal more to be said about the modulation of phase with digit pulses and the reader is recommended to refer to References 2 and 5.

2.8 PSK: SYMBOL ERROR RATE

2.8.1 Symbols and bits[5,6]

We now consider an N-phase system of phase-shift keying in which the circle of pre-determined phase angles of the transmitted carrier is divided into N sectors: the phase angle that the carrier is constrained to adopt may occupy the centre of any of the sectors. Each carrier phase angle represents a certain number of bits (binary digits). In this study the word 'symbol' means a carrier vector occupying a specific sector, that is, having a pre-determined phase angle and representing a certain number of bits of information. For example, if only

Table 2.1 QPSK

Sector number	*Phase angle*	*Digits represented*
1	0	00
2	90°	01
3	180°	10
4	270°	11

two sectors of the circle of phase angles are used, then the upper half-circle could represent a 'one' digit and the lower half-circle a 'zero' digit. That would be a one-bit system ($2'$ states). A two-bit system would have $2^2=4$ states represented by four phase angles, say 0, 90°, 180° and 270°. Each phase angle would have to represent two digits, as for example in Table 2.1 which refers to quadrature PSK or QPSK. Synchronous detection in the receiver would reveal which phase-angle sector of the circle was being occupied by the transmitted carrier-pulse at each clock-pulse interval of time.

If the receiver fails to measure correctly the phase of a carrier pulse, a symbol error will occur. The translation of the symbol error into erroneous bits depends upon the coding system being used. For a one-bit system a symbol error would produce only a single bit error. In a two-bit system such as that shown in Table 2.1 a single symbol error could give rise to one or two bit errors. For example, the sectors (of the circle) in which each symbol should appear, would be as in Table 2.2.

Table 2.3 shows the more probable symbol errors and their resulting bit errors. Errors resulting from a 180° rotation of the carrier from its transmitted phase are significantly less probable than errors resulting from a distorting phase rotation of 90°. In general terms, error phase shifts from the correct sector to either of its adjacent neighbours are most likely.

Table 2.2 SECTOR BOUNDARIES

Sector number	*Sector boundaries*	*Phase angle at centre of sector*
1	$\pm 45°$	0°
2	45° to 135°	90°
3	$\pm 135°$	180°
4	$-45°$ to $-135°$	270°

Table 2.3 BIT ERRORS

Erroneously detected carrier pulse in sector	*Sector that carrier pulse should have occupied*	*Number of bits in error*
2	1	1
4	1	2
3	2	2
4	2 or 3	1

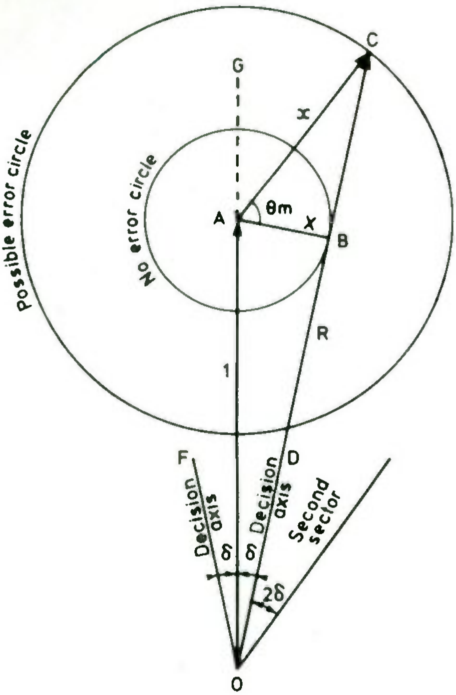


Figure 2.19 Geometry of carrier and noise

2.8.2 Symbol errors[5–8]

This study will be confined to the examination of symbol errors due to interference from random fluctuation noise whose amplitude has a Rayleigh statistical frequency distribution.

In Fig. 2.19 $\vec{OA}$ is a carrier vector of unit amplitude appearing at the input to the demodulator of the receiver. $\vec{AC}$ of amplitude x and phase angle θ_m with respect to the line $AB = X$ (normal to the tangent OC), represents a noise pulse of random amplitude in accordance with the Rayleigh distribution and random phase angle $\angle GAC$ having a rectangular distribution between 0 and 2π rad. In the present study it is the angle θ_m that is of interest, rather than $\angle GAC$ to which θ_m is related by the expression

$$\angle GAC = \theta_m - (\pi/2 + \delta)$$

The vector $\vec{OC}$ is the resultant R of the unit carrier $\vec{OA}$ and the noise $\vec{AC}$. The angular width of a symbol sector is 2δ and we shall examine the state of affairs when resultant R turns up in the wrong (unintended) sector. In Fig. 2.19 the intended sector is FOD and the wrong sector is labelled 'second sector'. From triangle OAB we have

$$\sin\delta = X \qquad \text{or} \qquad X = \sin(\pi/N) \tag{2.48}$$

since a sector angular width is $2\delta = 2\pi/N$ where N is the number of symbol sectors. From triangle ABC we have

$$\theta_m = \arccos(X/x) \tag{2.49}$$

Now from Fig. 2.19 it can be seen that if $x < X$ no symbol error can occur whatever random phase angle θ may turn up. So $x < X$ is a no-error condition. On the other hand, if $x > X$, as is the case in Fig. 2.19 symbol errors will occur if $\theta < \theta_m$, since in that situation resultant $\vec{OC}$ will form an angle with the centre-line OG of the wanted sector that exceeds δ and therefore involves the decision axis OD. If $\vec{AC}$ forms with OA an angle $GAC > GAB$ we have a symmetrical situation, so we may restrict our calculations to cases of the type shown in Fig. 2.19. Furthermore, for noise pulses whose vector representations would appear to the left of OG, the considerations would be identical to those for the right-hand side. From this discussion we may write that for any given value of $x > X$ the probability that an error-creating angle will occur will be $\theta_m/(\pi/2)$, because for $X < x < \infty$ we have $0 < \theta_m < \pi/2$. Thus if p_θ represents that probability, we write

$$p_\theta = (2/\pi)\arccos(X/x) \tag{2.50}$$

The probability that x will lie within the range $x - \frac{1}{2}\,dx < x < x + \frac{1}{2}\,dx$ is, from the Rayleigh statistical distribution

$$dP = (x/\sigma^2)\, e^{-x^2/2\sigma^2}\, dx \tag{2.51}$$

and the probability for a symbol error to occur is, of course, the product

$$dp(x, \theta_m) = p_\theta\, dP$$

whence

$$p(X) = (2/\pi)\int_X^\infty (x/\sigma^2)\arccos(X/x)\, e^{-x^2/2\sigma^2}\, dx \tag{2.52}$$

If we let C be the peak-carrier-to-r.m.s.-noise ratio, $C = 1/\sigma$, and $u^2 = x^2/2\sigma^2 = C^2x^2/2$ and $a = CX/\sqrt{2} = C\sin(\pi/N)/\sqrt{2}$ then we have

$$p(a) = (4/\pi)\int_a^\infty u\arccos(a/u)\, e^{-u^2}\, du \tag{2.52a}$$

Figure 2.20 shows the symbol error probability $p(a)$ as a function of a and also $p(C)$, the error probability as a function of peak-carrier-to-r.m.s.-noise ratio with N, the number of phases, as parameter. Table 2.4 shows, for various values of N, the values of C at various error probabilities. The case of $N = 6$ is somewhat hypothetical since $6 = 2^{2.585}$; that is, a non-integer number of binary digits.

2.9 FREQUENCY MODULATION WITH PULSES

It would seem that pulsed frequency modulation is less likely to be used for digital sound or digital television than pulsed phase modulation and so this section will be confined to a reference back to Sections 2.5–2.7. Just one brief example may be given: Thus consider a single rectangular pulse of frequency,

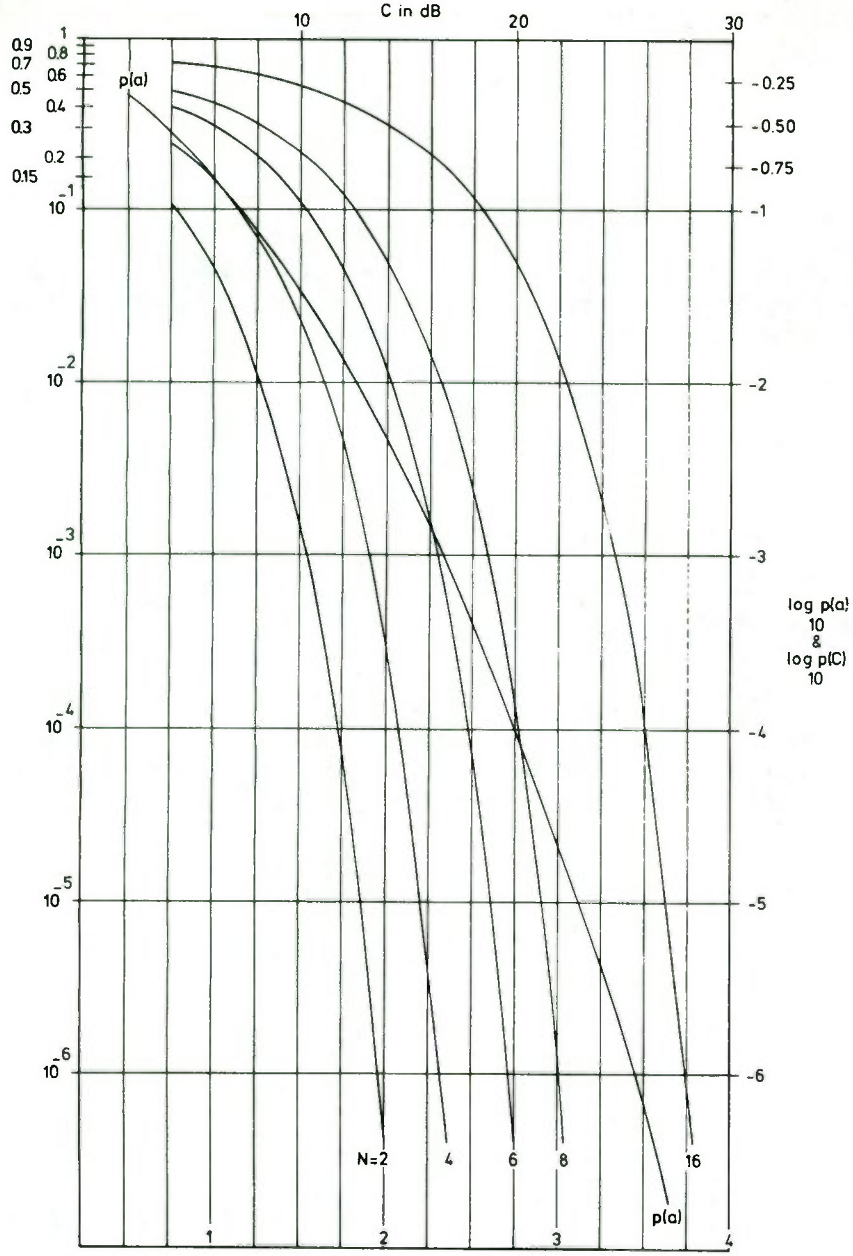

C in dB
10
20
30
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.15
p(a)
-0.25
-0.50
-0.75
10^{-1}
-1
10^{-2}
-2
10^{-3}
-3
$\log_{10} p(a)$
&
$\log_{10} p(C)$
10^{-4}
-4
10^{-5}
-5
10^{-6}
-6
N=2
4
6
8
16
p(a)
1
2
3
4

Figure 2.20

Table 2.4

N	10^{-6}	10^{-5}	10^{-4}	10^{-3}	10^{-2}	10^{-1}
16	28	27.2	26	24.5	22.5	18.4
8	22.1	21.3	20.2	18.7	17	12.7
6	19.7	19	17.9	16.3	14.2	10.3
4	16.7	15.8	14.7	13.3	11.2	6.9
2	13.8	12.9	11.7	10.5	8.5	4

of magnitude Δf and duration $T/2$. Call it $\Delta f K(t)$, with $-T/4<t<T/4$. If the carrier wave has a frequency f_0, the total frequency will be $f_0+\Delta f K(t)$. The total phase will be (Equation 2.16)

$$\phi(t)=2\pi\int_0^t [f_0+\Delta f K(t)]\,dt \tag{2.53}$$

$$=2\pi f_0 t+2\pi\Delta f t \tag{2.53a}$$

because the integral or area of the square wave is $\Delta f t$. When $t=T/4$ it becomes $\Delta f T/2$. So our carrier wave may be written

$$R(t)=\cos(2\pi f_0 t+2\pi\Delta f t) \tag{2.54}$$

with the restriction on the t that multiplies the $2\pi\Delta f$: $-T/4<t<T/4$.

We may also expand the cosine, obtaining

$$R(t)=\cos 2\pi f_0 t \cos 2\pi\Delta f t-\sin 2\pi f_0 t \sin 2\pi\Delta f t \tag{2.55}$$

We have the algebraic sum of the products of two time functions so we may, once again, convolve their spectra to obtain the final spectrum, $R[f]$. The spectra of $\cos 2\pi\Delta f t$ and $\sin 2\pi\Delta$ are respectively,

$$\int_{-T/4}^{T/4} \cos(2\pi\Delta f t)\, e^{-j2\pi f t}\,dt$$

and (2.56)

$$\int_{-T/4}^{T/4} \sin(2\pi\Delta f t)\, e^{-j2\pi f t}\,dt$$

The first is

$$\frac{\sin\frac{\pi}{2}(f+\Delta f)T}{\frac{\pi}{2}(f+\Delta f)T}\frac{T}{4}+\frac{\sin\frac{\pi}{2}(f-\Delta f)T}{\frac{\pi}{2}(f-\Delta f)T}\frac{T}{4}$$

and the second is (2.57)

$$j\left[\frac{\sin\frac{\pi}{2}(f+\Delta f)T}{\frac{\pi}{2}(f+\Delta f)T}\frac{T}{4}-\frac{\sin\frac{\pi}{2}(f-\Delta f)T}{\frac{\pi}{2}(f-\Delta f)T}\frac{T}{4}\right]$$

So, with our previous notation, we have, if $F=1/T$ and $m=\Delta f/F$

$$R[f]=\frac{1}{8F}H*\left\{\left[\frac{\sin\frac{\pi}{2}(m+\alpha)}{\frac{\pi}{2}(m+\alpha)}+\frac{\sin\frac{\pi}{2}(m-\alpha)}{\frac{\pi}{2}(m-\alpha)}\right]\right.$$

$$\left.+j\left[\frac{\sin\frac{\pi}{2}(m+\alpha)}{\frac{\pi}{2}(m+\alpha)}-\frac{\sin\frac{\pi}{2}(m-\alpha)}{\frac{\pi}{2}(m-\alpha)}\right]\right\}$$

$$+\frac{1}{8F}J*\left\{\frac{\sin\frac{\pi}{2}(m+\alpha)}{\frac{\pi}{2}(m+\alpha)}+\frac{\sin\frac{\pi}{2}(m-\alpha)}{\frac{\pi}{2}(m-\alpha)}\right.$$

$$\left.-j\left[\frac{\sin\frac{\pi}{2}(m+\alpha)}{\frac{\pi}{2}(m+\alpha)}-\frac{\sin\frac{\pi}{2}(m-\alpha)}{\frac{\pi}{2}(m-\alpha)}\right]\right\} \tag{2.58}$$

The similarity to the treatment for phase modulation is evident, but this is a simple example and there are considerations such as continuity or discontinuity of phase to be considered. The reader is referred to Bennett and Davey[5].

2.10 MODULATION OF BASEBAND PULSES

2.10.1 General remarks[2]

In order to sample an analogue waveform, for example, an audio or video signal, it is only necessary to multiply it by a repetitive wave of short pulses. If the pulses are short enough their shape becomes unimportant. In order to see clearly the phenomenon we consider an infinite sequence of delta functions

$$s(t)=\sum_{n=-\infty}^{\infty} A\delta(t-nT) \tag{2.59}$$

The 'weight' A of the delta functions will have the dimensions, volts × seconds, for example. Each delta function is zero when $t\neq nT$ and becomes effective when $t=nT$. The Fourier series expression[2] for $s(t)$ (Equation 2.1) is

$$s(t)=\frac{A}{T}\sum_{n=-\infty}^{\infty}\cos\frac{2\pi n}{T}t \tag{2.60}$$

which, from our previous knowledge (see remarks before expression 2.36), has the spectrum

$$s[f]=\frac{A}{2T}\sum_{n=-\infty}^{\infty}\left\{\delta\left[f-\frac{n}{T}\right]+\delta\left[f+\frac{n}{T}\right]\right\} \tag{2.61}$$

This spectrum is uniform in the sense that the delta functions do not 'fade away' as the frequency f increases. If the time sequence of pulses consisted of, say, rectangular pulses of finite width, the spectrum would cease to maintain uniformity with frequency and it would decrease with increasing frequency in either a monotonic or an oscillatory manner. Now, if the sequence of delta functions, Equations 2.59 and 2.60, be used to sample, that is, multiply, a sinusoidal signal, the total signal would obviously be

$$R(t)=s(t)\sin\omega_a t \tag{2.62}$$

$$=\frac{A}{T}\sum_{n=-\infty}^{\infty}\cos 2\pi nFt\sin 2\pi f_a t \tag{2.62a}$$

$$=\frac{AF}{2}\sum_{n=-\infty}^{\infty}[\sin 2\pi(nF+f_a)-\sin 2\pi(nF-f_a)] \tag{2.63}$$

where $f_a=\omega_a/2\pi$, $F=1/T$.

The result of the sampling has been to produce double sideband amplitude modulation upon a series of harmonically related suppressed carriers. The spectrum consists of lines at $\pm f_a$, $\pm F\pm f_a$, $\pm 2F\pm f_a$, $\pm 3F\pm f_a$, $\pm 4F\pm f_a\ldots$ when $n=0,1,2,3,\ldots$. A low-pass filter cutting off at a frequency between f_a and $F-f_a$ will reestablish the original modulation $\sin 2\pi f_a t$. If $F-f_a<f_a$ there will be interference from the lower sideband of the fundamental term $F\pm f_a$. This is called 'aliasing' and to avoid it $F-f_a$ must exceed f_a, whence we must have $F>2f_a$. This is one of Nyquist's many rules.

If a d.c. term is added to the modulation, which now becomes

$$B+\sin 2\pi f_a t \qquad \text{with} \qquad B\leqslant 1 \tag{2.64}$$

then the spectrum has, in addition to the lines at $\pm nF\pm f_a$, lines at $\pm nF$ including one representing d.c. when $n=0$.

If $F<2f_a$ and aliasing occurs the spectrum can be found by situating the spectrum of the modulation, now assumed to have a certain bandwidth, centrally on $f=0$, $f=\pm F$, $f=\pm 2F$, and adding, algebraically, the spectral components that occur at each frequency. Figure 2.21 illustrates this*; it shows the spectral components around each harmonic nF of the sampling frequency F. The rectangular spectra, which are those of the example modulation that is being sampled, are the spectrum of the response of an ideal low-pass filter of bandwidth W (that is $-W<f<W$) to a delta function excitation. That response is

$$2AW\frac{\sin 2\pi Wt}{2\pi Wt} \tag{2.65}$$

and, of course, its spectrum or Fourier transform (Equation 2.5) is a rectangle,

* Reference 2: Fig. 10.15, page 119.

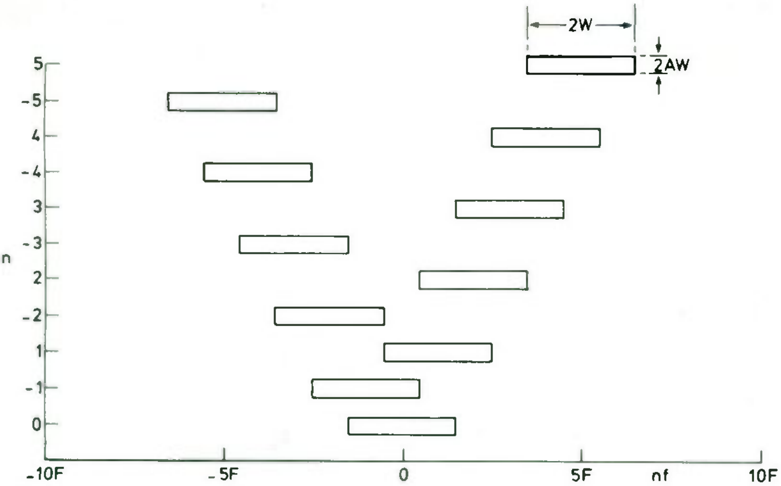


Figure 2.21 Aliasing spectrum

since that is the response-versus-frequency characteristic of the ideal low-pass filter excited by a delta function, or unit impulse that has a spectrum that is uniform for $-\infty < f < \infty$. It therefore 'fills' completely the filter response characteristic.

The final spectrum to be obtained from Fig. 2.21 is $3 \times 2AW$ over all nF with $-\infty < n < \infty$, because at any given frequency nF, there are at least a part of each one of three rectangular spectra to be added. If the spectral lines in the component spectra have phase angles that change with frequency, then the addition must be vectorial.

The above remarks cover what is termed pulse amplitude modulation (PAM).

2.10.2 Pulse width modulation (PWM)

Consider a repetitive waveform of rectangular pulses of width 2τ and of periodicity T, Fig. 2.22. If $r(t)$ is the Fourier series expansion of Fig. 2.22 we have from Equations 2.2 and 2.3:

$$b_n = \frac{2}{T}\int_{-\tau}^{\tau} 1 \cos 2\pi n \frac{t}{T}\, dt = 4\frac{\tau}{T}\frac{\sin 2\pi n\tau/T}{2\pi n\tau/T} \tag{2.66}$$

$$a_n = \frac{2}{T}\int_{-\tau}^{\tau} 1 \sin 2\pi n \frac{t}{T}\, dt = 0 \tag{2.67}$$

whence

$$r(t) = 4\frac{\tau}{T}\left[1 + \frac{\sin 2\pi\tau/T}{2\pi\tau/T}\cos 2\pi\frac{t}{T} + \frac{\sin 2.2\pi\tau/T}{2.2\pi\tau/T}\cos\left(2.2\pi\frac{t}{T}\right) + \ldots\right] \tag{2.68}$$

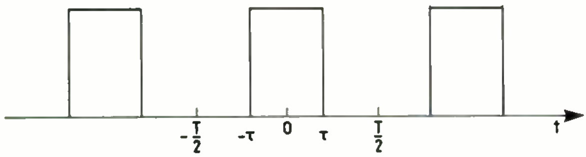


Figure 2.22 Pulse width modulation

If, now, we modulate τ with an audio or video sinewave $m \sin 2\pi f_a t$ with $m<1$, we have for our PWM signal:

$$R(t)=\frac{4\tau}{T}(1+m\sin 2\pi f_a t)\left\{1+\frac{\sin\left[\frac{2\pi}{T}\tau(1+m\sin 2\pi f_a t)\right]}{\frac{2\pi}{T}\tau(1+m\sin 2\pi f_a t)}\cos\frac{2\pi}{T}t\right.$$
$$\left.+\frac{\sin\left[\frac{2.2\pi}{T}\tau(1+m\sin 2\pi f_a t)\right]}{\frac{2.2\pi}{T}\tau(1+m\sin 2\pi f_a t)}\cos\frac{2.2\pi}{T}t+\dots\right\}\tag{2.69}$$

This is a very complicated frequency or phase modulated signal full of Bessel functions which, in the actual technique[10] employed, is eliminated by a bandpass filter that attenuates the d.c. component $4\tau/T$ and also attenuates all the terms in the braces that have $\cos n2\pi t/T$ as a factor. In other words, PWM is a way of transmitting a sinewave modulation by constant amplitude pulses that have a desirable resistance to noise.

2.10.3 Pulse position modulation (PPM)

In this system the pulse width is constant, but its position in time is varied in accordance with the modulation. Since the pulse width is constant, we can take a sequence of delta functions as our carrier wave. The sequence of delta functions

$$s(t)=\sum_{n=-\infty}^{\infty} A\delta(t-nT)\tag{2.59}$$

represents short-duration pulses occurring at intervals T.

The Fourier series expansion for $s(t)$ is

$$s(t)=\frac{A}{T}\sum_{n=-\infty}^{\infty}\cos\frac{2\pi n}{T}t\tag{2.60}$$

Now let us advance and delay the time t by an audio or video sinewave

$$Y\sin 2\pi f_a t\tag{2.70}$$

Of course, to obviate overlapping of pulses we must ensure that $Y<T/2$.

Putting expression 2.70 into Equation 2.60 we have

$$R(t)=\frac{A}{T}\sum_{n=-\infty}^{\infty}\cos\left[\frac{2\pi n}{T}(t+Y\sin 2\pi f_a t)\right] \tag{2.71}$$

If $Y=KT$, $K<\frac{1}{2}$, we can expand Equation 2.71, thus

$$R(t)=\frac{A}{T}\sum_{n=-\infty}^{\infty}\left[\cos(2\pi nK\sin 2\pi f_a t)\cos\frac{2\pi n}{T}t - \sin(2\pi nK\sin 2\pi f_a t)\sin\frac{2\pi n}{T}t\right] \tag{2.71a}$$

Letting $2\pi K=Z$ we expand the first factors in the two terms of the right-hand side of Equation 2.71a:

$$\begin{aligned}R(t)=\frac{A}{T}\sum_{n=-\infty}^{\infty}\Big\{&[J_0(nz)+2J_2(nz)\cos 2\omega_a t+2J_4(nz)\cos 4\omega_a t\\ &+2J_6(nz)\cos 6\omega_a t+\ldots]\cos n\omega t\\ &-[2J_1(nz)\sin\omega_a t+2J_3(nz)\sin 3\omega_a t\\ &+2J_5(nz)\sin 5\omega_a t+\ldots]\sin n\omega t\Big\}\end{aligned} \tag{2.71b}$$

where $\omega_a=2\pi f_a$ and $\omega=2\pi/T$.

Now $J_m(-nz)=(-1)^m J_m(nz)$ and $\sin(-m\omega_a t)=-\sin(m\omega_a t)$ whence we can change $\sum_{n=-\infty}^{\infty}$ to $\sum_{n=0}^{\infty}$ and write

$$\begin{aligned}R(t)=\frac{2A}{T}\sum_{n=0}^{\infty}\{&J_0(nz)\cos n\omega t+J_2(nz)\cos(n\omega\pm 2\omega_a)t\\ &+J_4(nz)\cos(n\omega\pm 4\omega_a)t+J_6(nz)\cos(n\omega\pm 6\omega_a)t+\ldots\\ &+J_1(nz)\cos(n\omega\pm\omega_a)t+J_3(nz)\cos(n\omega\pm 3\omega_a)t\\ &+J_5(nz)\cos(n\omega\pm 5\omega_a)t+\ldots\}\end{aligned} \tag{2.71c}$$

wherein the odd order terms of lower sidebands have the negative sign in front.

If we now expand Equation 2.71c giving $n=0,1,2,3$, etc. we arrive at

$$\begin{aligned}R(t)=\frac{2A}{T}\{1&+J_0(z)\cos\omega t+J_1(z)\cos(\omega\pm\omega_a)t\\ &+J_2(z)\cos(\omega\pm 2\omega_a)t+J_3(z)\cos(\omega\pm 3\omega_a)t+\ldots \quad \Big]\, n=1\\ &+J_0(2z)\cos 2\omega t+J_1(2z)\cos(2\omega\pm\omega_a)t\\ &+J_2(2z)\cos(2\omega\pm 2\omega_a)t+J_3(2z)\cos(2\omega\pm 3\omega_a)t \quad \Big]\, n=2\\ &+J_4(2z)\cos(2\omega\pm 4\omega_a)t+\ldots\\ &+J_0(3z)\cos 3\omega t+J_1(3z)\cos(3\omega\pm\omega_a)t\\ &+J_2(3z)\cos(3\omega\pm 2\omega_a)t+J_3(3z)\cos(3\omega\pm 3\omega_a)t \quad \Big]\, n=3\\ &+J_4(3z)\cos(3\omega\pm 4\omega_a)t+\ldots\}\end{aligned} \tag{2.71d}$$

Now, if our carrier wave of short pulses (delta functions) were changed to a sequence of rectangular pulses of width 2τ, repeated at intervals T, we could modify Equation 2.71d by multiplying each term by

$$\frac{\sin nx}{nx} \qquad \text{with} \qquad x = 2\pi\tau/T \tag{2.72}$$

(See Equation 2.68.)

As an example, let $2\tau = T/3$ and $Y = 0.159T$ or $K = 0.159$. Then we have $z = 2\pi K = 1$ and Table 2.5 gives the relevant values of $J_m(nz)$ with $x = 2\pi\tau/T = 1.047$. The curious value of $Y = 0.159T$ or $K = 0.159$ was the nearest figure to 1/6 that made $z = 1$ for which adequate Bessel-function values could easily be obtained from printed tables.

Figure 2.23 shows some of the spectrum, omitting the d.c. value of 1. The first 'chunk' of spectrum, around W is not unlike DSB AM. The usual

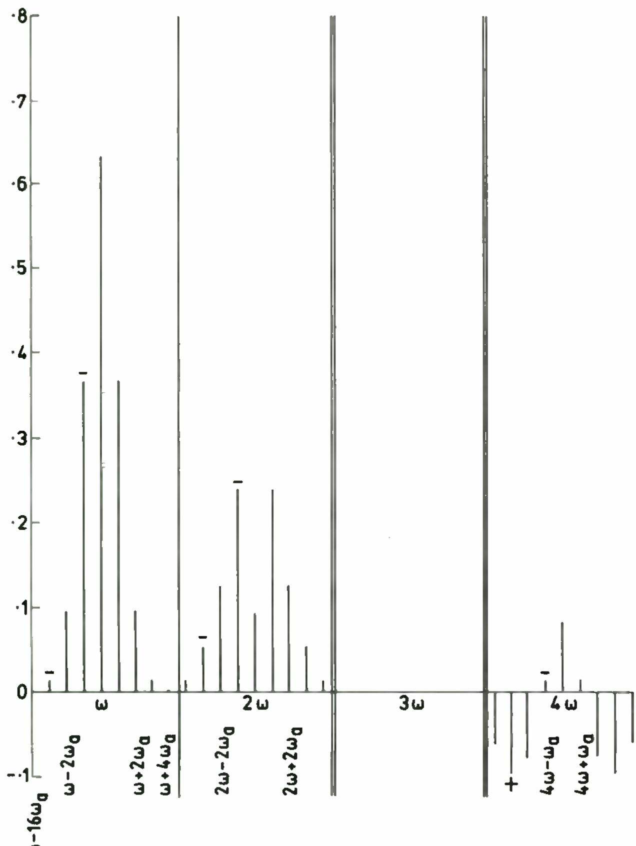


Figure 2.23 Spectrum of PPM

Table 2.5 COEFFICIENTS FOR EQUATION 2.17d

n	J_0	J_1	J_2	J_3	J_4	J_5	$\dfrac{\sin nx}{nx}$
0	1	0	0	0	0	0	1
1	0.7652	0.4401	0.1149	0.0196	0.0025	0.0002	0.8271
2	0.2239	0.5767	0.3528	0.1289	0.034	0.0070	0.4137
3	0.2601	0.3391	0.4861	0.3091	0.132	0.0430	0.000189
4	−0.3971	−0.066	0.3641	0.4302	0.2311	0.1321	−0.2067

technique employing PPM is, however, to convert it on reception either to PWM or to PAM and then use a low-pass filter to extract the original modulation.

REFERENCES

1. McLachlan, N. W., *Bessel Functions for Engineers*, p. 43, Clarendon Press, Oxford (1934)
2. Maurice, R. D. A., *Convolution and Fourier Transforms for Communications Engineers*, Pentech Press, London (1976)
3. Devereux, V. G. and Jones, A. H., 'Pulse code modulation of video signals: codes specifically designed for PAL', *Proc. IEE*, Vol. 125, No. 6 (1978)
4. Campbell, G. A. and Foster, R. M. *Fourier Integrals for Practical Applications*, p. 39, pair 203; p. 148, pairs 983, 984, Van Nostrand, Amsterdam (1967)
5. Bennett, W. R. and Davey, J. R., *Data Transmission*, p. 79, McGraw-Hill, New York (1965)
6. Spilker, J. R., *Digital Communications by Satellite*, Prentice-Hall, New Jersey (1977)
7. Montgomery, G. F., 'A comparison of amplitude and angle modulation for narrow band communication of binary coded messages in fluctuation noise', *Proc. IRE*, pp. 447–454, February (1954)
8. Feher, K., *Digital Communications: Microwave Applications*, Prentice-Hall, New Jersey (1981) Arnold, London (1973)
9. Edwards, J. R., 'A comparison of modulation schemes for binary data transmission'. *Radio and Electronic Engineering*, Vol. 43, No. 9, September (1973)
10. Connor, F. R. *Modulation. Introductory Topics in Electronics and Telecommunication*, Edward Arnold, London (1973)

Chapter 3

An introduction to digital techniques

P. Sproxton

3.1 INTRODUCTION

The purpose of this chapter is to describe some basic techniques and components used in television equipment incorporating digital electronics. Digital techniques, as used by the television industry (or in telecommunications generally), can be separated into two categories, i.e.

(1) Control of signal path and analogue processing equipment.
(2) Conversion of analogue signals to digital form and processing of digital signals in both converted form and generic digital form.

The first category is illustrated by the hypothetical system in Fig. 3.1 in which digital control signals are sent to the analogue processing elements to alter the parameters of those elements.

The emphasis of this chapter, however, will be upon the second category, particularly the actual operation of analogue/digital and digital/analogue conversion. The operation of analogue to digital conversion involves the replacement of the analogue signal, as a continuous function of time, by a sequence of numerical values in binary form. A simple example of a system incorporating digital techniques is shown in Fig. 3.2. A digital signal must deliver the entire information content of the original analogue signal.

3.1.1 Advantages of digital systems

Replacing existing analogue equipment with its digital counterpart can involve considerable extra cost and complexity, so why bother? This can best be answered by considering the following advantages of digital equipment:

(1) High resistance to noise and circuit distortion. The digital signal exists in binary form, therefore only two amplitude levels need to be discerned.
(2) Complex production effects become possible due to the ease with which digital signals can be stored and manipulated.

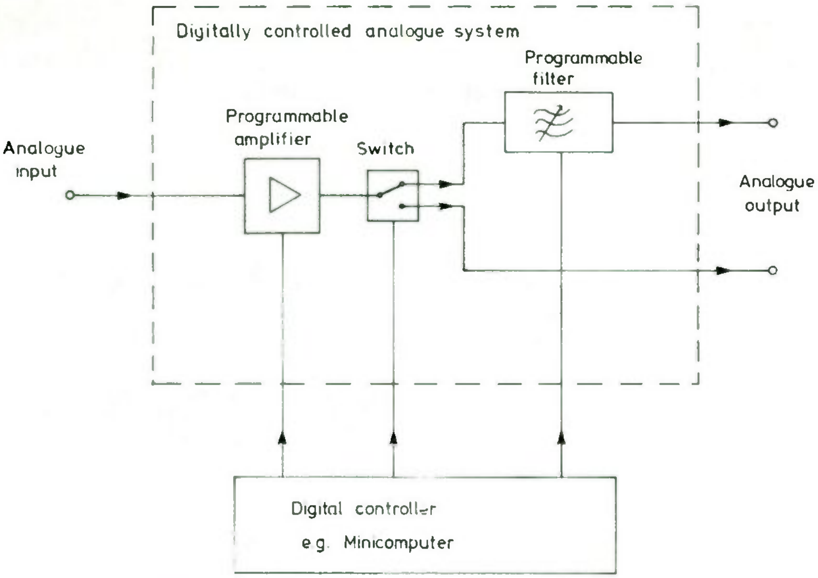


Figure 3.1 Example of digital equipment used for control purposes

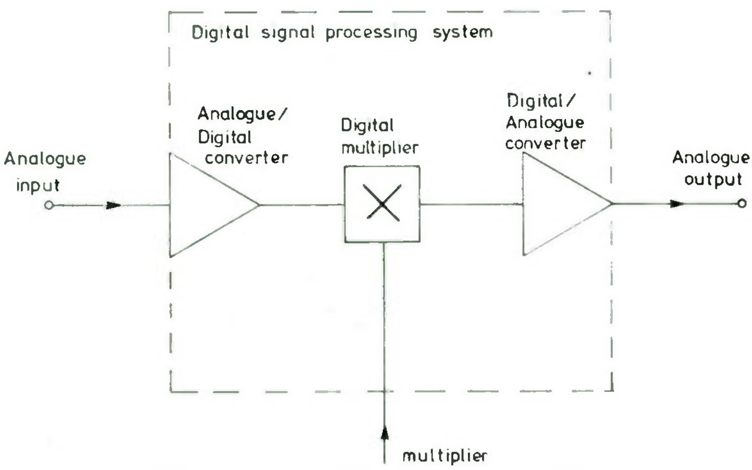


Figure 3.2 Example of digital equipment used for signal conversion and processing

(3) Powerful signal conditioning equipment can be manufactured, again due to the simplicity of digital circuits.
(4) Digital equipment is inherently stable. The effects of temperature, ageing and tolerance variation in digital equipment are much reduced.

The above advantages have been sufficient to prompt engineers to re-design existing analogue electronics.

3.1.2 A disadvantage of digital systems

The main disadvantage of digital signal generation and processing is the considerably greater bandwidth occupied in comparison with the equivalent analogue signal. This results in larger transmission bandwidths and very high speed logic devices needed to deal with real time signal processing.

3.1.3 Digital signal processing equipment

To give some perspective to the techniques and components described in the following sections, it is worth mentioning some of the digital equipment currently available to the television industry. Only a brief outline is given as many of the underlying principles will be discussed in later chapters.

Time base correctors This equipment conditions the output from a video tape recorder to remove errors due to elastic distortion in magnetic tape and other timing errors due to mechanical wow and flutter.

Frame synchronisers In a television studio it is essential that video signals from different sources have concurrent timing information. If this is not so, then unpleasant visual effects will occur during a mix between two separate video sources. Frame synchronisers remove these effects by 'locking' source timing information to a common studio sync pulse generator.

Standards converters Digital techniques have made it possible to construct equipment, to convert video signals between two standards (e.g. British PAL to American NTSC), using solid state electronics throughout.

Digital stills stores Conversion of television signals to digital form has made it possible to store high quality 'frozen frame' stills on a computer type disk pack.

Digital production effects equipment This equipment enables a television production team to evolve a wide range of special effects to improve program presentation quality. The ability of the equipment to 'squeeze' or 'zoom' is probably one of the better known characteristics of this equipment.

Digital paint box Using a computer type graphics tablet an artist or production team can literally paint a scene on a video monitor. An operator could also use this equipment to 'touch up' sections of a pre-recorded scene.

An important common factor with all these devices is the ability to store parts, or complete frames of video information. Storage of this kind becomes very easy if the video signal is converted into digital form.

3.2 BASIC LOGIC GATES

This section, although covering basic material, assumes the reader to have some knowledge of logic devices and their operation. The emphasis is upon the types of logic most likely to be encountered in the field of digital television, and for this reason particular emphasis is given to high speed devices. Before going into any detailed description of particular logic gates some very basic terminology will be mentioned—recognised logic circuit symbols are illustrated in Fig. 3.3.

The term 'positive' logic signifies that the voltage level assigned to the '1' state is more positive than the level assigned to the '0' state. Negative logic is the converse and is indicated on a logid diagram by a small circle adjacent to the logic element. Negative logic circuit connections are often referred to as being 'active low'.

The symbol for an inverting buffer is shown in Fig. 3.3(a). The circle indicates the logic polarity being inverted or negated, i.e. Y equals NOT A, or Y is present when A is not present. Figure 3.3(b) shows the non-inverting buffer for both positive and negative logic. The non-inverting buffer is usually used to improve the logic 'fanout'. Fanout is the number of logic inputs a logic device can drive.

Figures 3.3(c) and (d) illustrate a two input AND gate and a two input OR gate respectively. The logical functions of the AND and OR gates are described by their 'truth tables' written next to their symbols in Figs. 3.3(c) and (d).

Figures 3.3(e) and (f) illustrate the symbols for AND/OR functions whose outputs have been negated to form NOT AND, shortened to NAND, and NOT OR, shortened to NOR. Observing the truth table in Table 3.1 it can be seen that the positive logic NAND function is the same as that for the negative

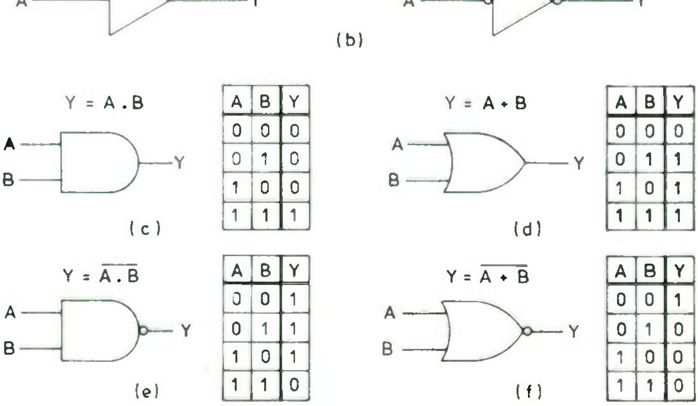


Figure 3.3 Fundamental logic symbols

Table 3.1 TRUTH TABLE FOR NAND AND NEGATIVE LOGIC OR FUNCTIONS

A	B	$\overline{A}$	$\overline{B}$	$Y=\overline{A.B}$	$Y=\overline{A}+\overline{B}$
0	0	1	1	1	1
0	1	1	0	1	1
1	0	0	1	1	1
1	1	0	0	0	0

logic OR function. This is a practical realisation of De Morgan's theorem, written here in conventional Boolean algebra form, i.e.

$$\overline{A.B}=\overline{A}+\overline{B}$$

where $\bar{}$ represents negation, . represents logical AND and + represents logical OR. The same situation holds for the positive NOR gate and negative NAND gate, i.e.

$$\overline{A+B}=\overline{A}.\overline{B}$$

Thus it can be seen that by using De Morgan's theorem a logic designer can implement AND functions with OR gates and vice versa. This enables a design engineer to optimise the number of gates used on a particular circuit card. More powerful logic minimisation techniques are available, e.g. Karnaugh mapping, but these methods are beyond the scope of this book.

3.3 PRACTICAL LOGIC GATES

From the many families of logic gates produced there are three categories of particular importance to digital television applications:

(1) TTL (Transistor Transistor Logic)
(2) ECL (Emitter Coupled Logic)
(3) MOS (Metal Oxide Surface)

3.3.1 The TTL gate

TTL logic has several variants, the properties of some of which are summarised in Table 3.2. Standard series TTL is the oldest type, well established in many systems where speed and power consumption requirements are not critical. Schottky TTL and the recent Texas Instruments Advanced Schottky TTL constitute some of the fastest TTL devices available.

Table 3.2 COMPARATIVE SPEED POWER FIGURES FOR SCHOTTKY TTL

	Standard	*Low power*	*Advanced*	*Advanced low power*
Speed	3 ns	9 ns	1.5 ns	5 ns
Power	19 mW	2 mW	19 mW	<2 mW

The speed of a logic device is the length of time taken for a logical change at the input to propagate through the device to the output. In many digital television applications, logic devices require propagation times (delays) approaching 1 ns. Low power variants are available for applications where low power consumption is important. These low power devices, however, exercise their properties at the expense of a longer propagation delay.

Schottky TTL is described in greater detail in the following section.

3.3.2 Schottky TTL

The essential factor in the operation of the Schottky gate is that, unlike other TTL variants, the switching transistors are not allowed to saturate. This is achieved by using Schottky diode clamped transistors. Figure 3.4 shows a single Schottky TTL NAND gate with clamping diodes shown between collector and base of the relevant transistors. When the transistor is turned

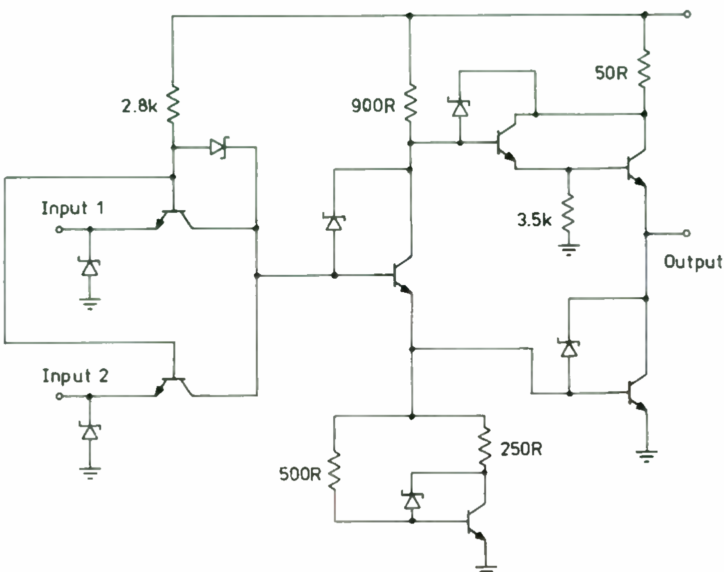


Figure 3.4 Two-input Schottky diode clamped TTL NAND gate

hard on, the diode conducts and excess base current is bypassed to the collector. The low voltage drop of the Schottky diode ensures a small voltage between collector and emitter and the transistor is thus prevented from saturating. Important gate speed and power characteristics for Schottky logic are summarised in Table 3.2.

3.3.3 MOS logic

MOS logic is based upon the insulated gate field effect transistor (IGFET). The technology for MOS logic gates allows for a relatively high density silicon dice operating at a much lower power consumption than bipolar TTL logic. These properties result in MOS logic being used for high density memory devices which will be described in the next section. Until recently MOS devices suffered from being very slow in operation; this is no longer so. Some of the more complex real time processing circuit elements are now implemented using a variant of MOS.

3.3.4 Emitter coupled logic

The basic ECL OR/NOR gate, along with its logic symbol, is shown in Fig. 3.5. Operation of the circuit is based upon current steering in emitter coupled pairs of bipolar transistors. The transistors are non-saturating and are thus inherently very fast.

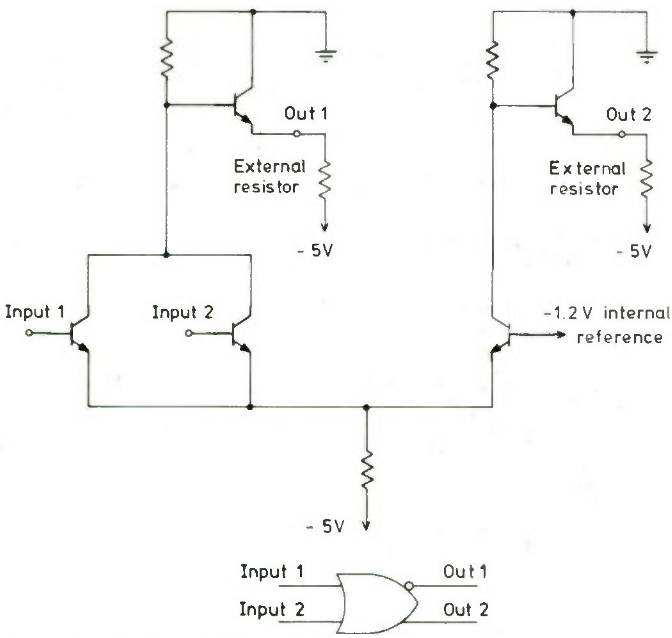


Figure 3.5 Basic ECL OR/NOR gate

A change of input level causes the emitter tail current to be steered to one collector or the other. As a consequence of this design a function and its complement appear simultaneously at the device outputs. The advantages of complementary outputs are:

(1) Reduced circuit package count by eliminating inverters.
(2) Eliminates differential timing problems that could be introduced by inverters.

The differential pair arrangement also has the advantages of reduced power

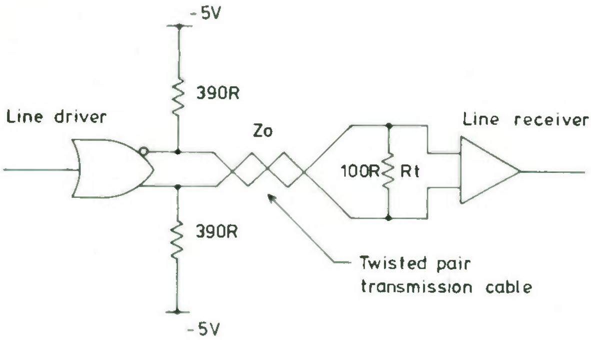


Figure 3.6 Twisted pair transmission link using ECL logic

supply noise and balanced line driving capability. Because the total current demanded by the gate is the same, no matter which state it is in, the power supply rails suffer little disturbance. Complementary outputs along with open emitter facility enable easy implementation of balanced line data transfer. One of the more popular ways to implement an ECL transmission line (e.g. circuit card interconnection) is to use the twisted pair method. This method is illustrated in Fig. 3.6. Here R_t is used to terminate the transmission line. The 1 to 1.5 V common mode noise rejection of the line receiver afford good noise and crosstalk immunity, permitting multiple twisted pairs to be tied into cables. These types of transmission line can work effectively at lengths greater than 300 m although line attenuation will limit bandwidth, degrading edge speeds when long line runs are made.

Important a.c. and d.c. parameters for ECL devices are summarised in Table 3.3.

3.3.5 Tristate TTL gates

Another TTL gate of particular importance is the 'tristate' gate. In addition to a low impedance 'high' and 'low' output state, a third condition causes the output to go high impedance. When in this third state the gate is effectively removed from circuit.

One application where the tristate gate is much used is within micro-processor or general computer equipment. In this example many LSI circuits transmit data to and from a common 'data bus'. Correct operation of the data bus relies on only one device passing data to it at any one time. This is achieved

Table 3.3 TYPICAL CHARACTERISTICS FOR ECL LOGIC

Gate propagation delay	1.5 ns
Output edge speed	2.5 ns
Gate power	25 mW
Speed power product	37 pJ
Gate power supply	−5 V

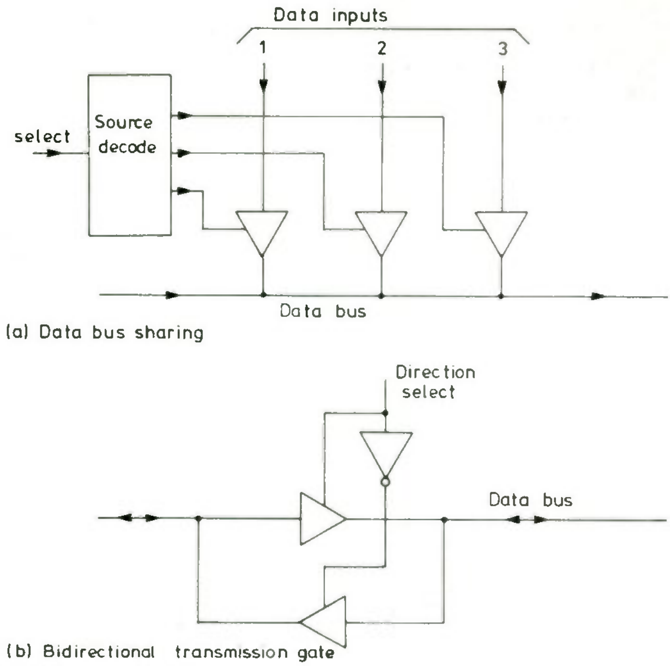


Figure 3.7 Tristate gate implementation

by setting all but the required transmission source to their tristate conditions. A simple example of a tristate data bus is shown in Fig. 3.7(a).

By using a combination of tristate gates we can form a 'bi-directional transmission gate'. This arrangement, shown in Fig. 3.7(b), allows a device to both send and receive data to and from a single data bus.

3.4 MEMORY DEVICES

3.4.1 Introduction

The importance of the memory device in digital television equipment was briefly mentioned in Section 3.1; the purpose of this section is to describe the basic principles employed in their manufacture.

Memory devices are organised as an array of memory 'cells', where each cell stores a unit of binary information, i.e. a '1' or a '0'. The number of cells present in any one memory device can vary between 16 and 256,000 although larger densities are now available.

Semiconductor memory can be divided into two basic categories, i.e. read/write and read only. The read/write, or Random Access Memory (RAM) allows binary information to be stored, and retained so long as power is applied to the device. Information stored in RAM can be subsequently read out and rewritten when necessary. RAM is used in large quantities in the equipment described in Section 3.1, where a complete frame of a digitised video signal is stored for processing. This type of store is periodically overwritten as each video frame becomes available to the memory. In computer applications RAM is used to store a program which is not permanently required.

The Read Only Memory (ROM) is a device which has its binary contents defined at the beginning of its life. ROM contents are set by a photographic mask during production of the silicon dice. Loss of power to the ROM does not affect its contents. ROM can be used to store data or a program which is a permanent feature of the host system. An example of a system using data stored in ROM would be a digital waveform synthesiser. In this equipment a complex analogue waveform is permanently stored in digital form. In computer applications, programs which are permanently in use are often stored in ROM.

The following sections outline, in further detail, the operation of RAM, RAM organisation and two devices which have properties of both ROM and RAM.

3.4.2 Static and dynamic RAM

The more common RAM devices are those implemented using MOS technology. MOS semiconductors can be configured to form two distinct types of memory, these are the 'static' storage cell and the 'dynamic' storage cell.

Static storage cells Figure 3.8(a) depicts the static storage cell implemented using *p* channel MOS technology. A pair of cross coupled inverters, T1 and T2, with MOS load resistors, T3 and T4, will remain in one of two possible states with one 'on' and the other 'off'. The initial state is determined by forcing the voltage level on the gate of either T1 or T2. This state will remain until either T1 or T2 are forced into a different state, or power is removed from the device.

This type of storage element suffers from the disadvantage that the silicon slice cell size is large resulting in a relatively low density memory device.

Dynamic storage cells Optimum use of space on a silicon slice can be

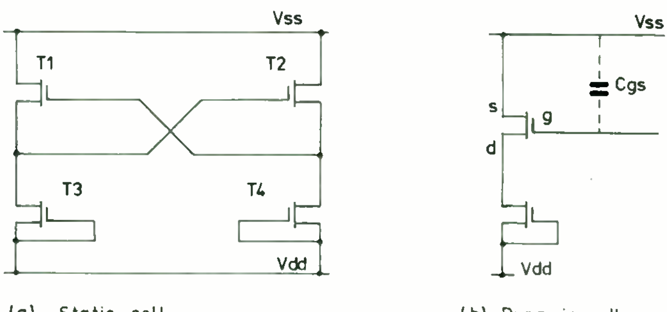


Figure 3.8 Static and dynamic storage cells

achieved by making use of the extremely high impedance of the MOS transistor (about $10^9\ \Omega$).

Figure 3.8(b) shows a single dynamic memory cells where its logic state is determined by whether or not a charge exists across the gate to source capacitance. The period of time the charge is maintained is determined by circuit leakage which is in turn dependent on the temperature of the device. This quasi bistable can form the basis of a useful storage element provided the gate-source charge is 'refreshed' at regular intervals. In practice, for 70°C, many manufacturers will quote a refresh period of 2 ms. For many applications, particularly digital video signal processing, refresh occurs automatically as a result of the repetitive nature of the signal. In microprocessor control or data base applications sections of dynamic memory may not be accessed within the required 2 ms, in this event all dynamic memory must be placed under the supervision of a 'refresh controller'. This is often a single LSI device which ensures that all dynamic memory is refreshed within 2 ms.

3.4.3 Memory organisation

For most purposes, the engineer need have no exacting knowledge of cell level organisation. In practice, it is the package level configuration along with relevant a.c. and d.c. characteristics that are important.

Examples of a 1024 × 8 static RAM and a 16,384 × 1 dynamic memory are illustrated in Figs. 3.9 and 3.10 respectively. The static RAM example is capable of storing 1024 'bytes', where 1 byte is an 8-bit wide parallel binary code. It can be seen that the package has eight bi-directional data connections, 10 address inputs, a read/write pin and a chip-enable pin. The bi-directional

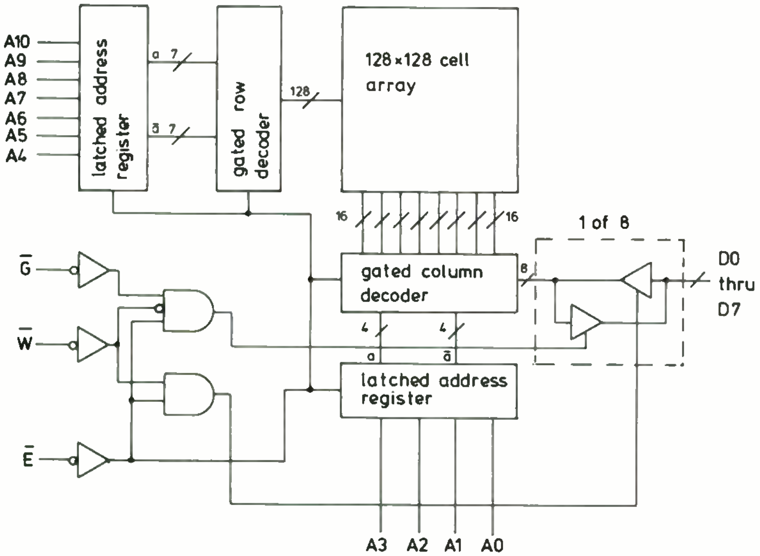


Figure 3.9 Typical internal configuration for a 2048 × 8 static ram

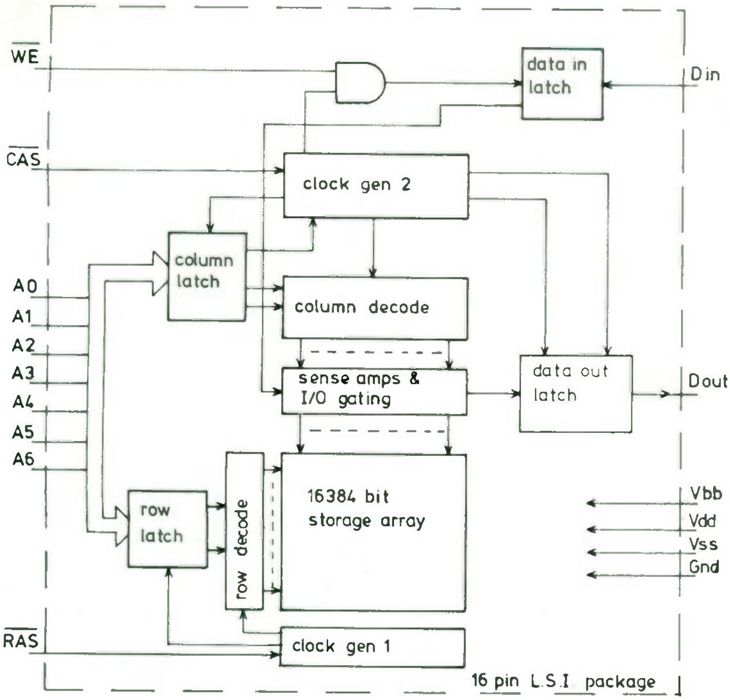


Figure 3.10 Typical internal configuration for a 16 384 × 1 bit MOS dynamic ram

data connections allow a data byte to be either read in, or written out on the same set of pins, depending on the condition of the read/write pin. The condition of the 10 address inputs determine which of the possible 1024 locations the data byte is to be read from, or written to.

There are 1024 combinations for a 10-bit address code therefore each memory location has a unique address code. The chip-enable pin forces any outputs, from the chip, into their tristate conditions.

The block diagram for a typical dynamic RAM is shown in Fig. 3.10. The organisation for this device is 16,384 × 1, therefore only one bi-directional data pin is required. In practice, however, a data input and a data output pin are provided for design flexibility. Only seven address inputs are present, which would seem to imply that unique addressing is limited to only 128 locations. This limitation is imposed by the desire to fit the device into the smaller 16-pin industrial standard DIL package. Two control pins, RAS and CAS, allow full address decode by multiplexing 14 address lines, down to the seven present on the DIL package. This arrangement, where multiplexing is employed, is illustrated in Fig. 3.10 which shows an internal block diagram of a typical dynamic RAM. Here a memory 'column' address of up to 7 bits and a memory 'row' address of up to 7 bits are transferred to the row and column decoders under the control of RAS and CAS. After multiplexing, the total possible number of locations that can be uniquely accessed is 16,384 or number of rows multiplied by the number of columns.

Cell refresh is performed by ensuring that each of the row address lines is energised at least once every 2 ms.

3.4.4 Access and cycle time

The block diagram of the dynamic RAM illustrated in Fig. 3.10 gives an idea of the amount of logic required to implement such a device. The more complex this logic, the greater the overall propagation delay introduced. Internal propagation delays of this nature are a problem in high speed systems and for this reason manufacturers will quote 'access time' and 'cycle time' for their devices. Access time is that time taken for memory data outputs to settle after a read operation. The cycle time is the shortest period in which a read operation can directly follow a write operation.

It is the length of access and cycle time in MOS memories which would, on first inspection, render them unsuitable for use in high speed (real time video processing speeds) equipment. It is becoming common for digital video equipment to incorporate data transmission speeds in excess of 14 Mbit/s, which results in clock periods of less than 70 ns. To capture data at these speeds RAM must operate with access times less than 70 ns, which is almost half the access time of many MOS memories. Considering the relative cost of MOS memory (MOS being the cheapest) and the amount of memory required, it is worth using a multiplex/demultiplex system. This principle is illustrated in Fig. 3.11. By time division demultiplexing a single 14 Mbit/s transmission line down to four separate lines, the maximum data rate can be reduced to 3.5 Mbit/s. The memory capture speed required is then quartered.

3.4.5 Programmable read only memories

The ROM, as discussed in Section 3.4.1, has its contents fixed when it is made; this is a disadvantage if alteration to its data or program has to be made. Two

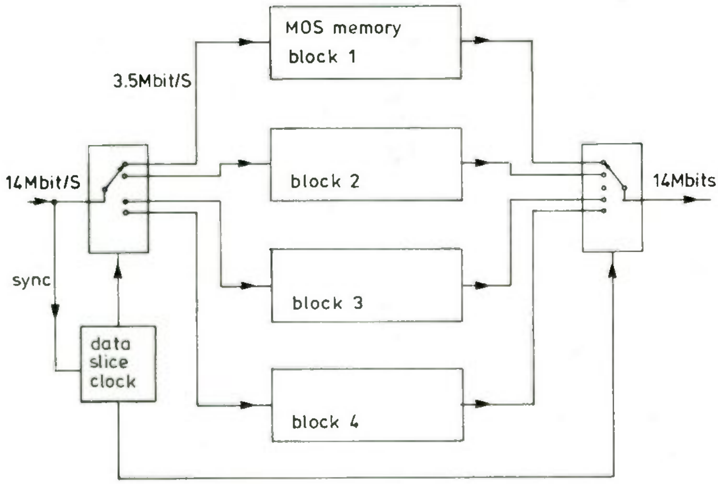


Figure 3.11 Data speed reduction for MOS memory

other types of read only type device, the EPROM and the EAROM allow a degree of flexibility in this respect. The contents of an EPROM (Erasable Programmable Read Only Memory) can be erased by shining ultraviolet light through a quartz lid on the memory package. The EPROM can then be reprogrammed using an EPROM programming circuit.

An EAROM (Electrically Alterable Read Only Memory) is a read only type device which can be reprogrammed in the host circuit (unlike the EPROM). An example of the use of an EAROM is in an auto-dial telephone system. Here the EAROM holds a directory of numbers which can be altered when the need arises. Care must be taken to ensure that a circuit accesses an EAROM only infrequently. This is because the device will be rendered ineffective after a finite number of location accesses.

3.5 SAMPLING THEORY

3.5.1 Signal sampling

Figure 3.12 illustrates a simple circuit for sampling a signal. For a theoretical model of this circuit, it is assumed that the FET has an infinite 'off' resistance, zero 'on' resistance and the capacitor C has zero leakage. An analogue signal

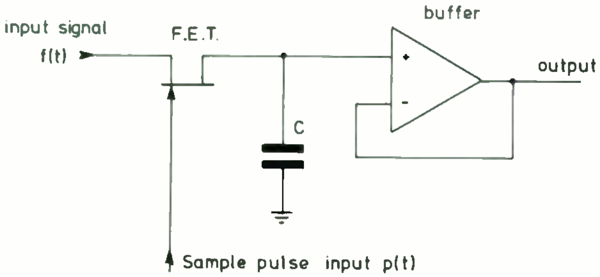


Figure 3.12 Sample and hold circuit

$f(t)$ is applied to the input of the circuit and a periodic narrow sample pulse is applied to the gate of the FET. During any sample pulse, the FET conducts and the input $f(t)$ appears across C. When the FET is turned off, an instantaneous value of $f(t)$ is left stored across C. This instantaneous value of $f(t)$ is presented to the theoretically infinite input impedance of the buffer, which being configured for unity gain, reflects this value at the output. This output is available until a subsequent sample pulse switches on the FET and a new sample of $f(t)$ is taken. This arrangement constitutes a circuit which will sample a signal and then hold the value of that sample, hence it is known as a 'sample and hold' circuit.

The output from the sample and hold circuit can be 'read' by an analogue to digital converter (A/D) to yield a binary code, whose value is determined by the amplitude of the sample. In practice the sample and hold circuit is used for either infrequent sampling, or for slow A/D converters. For continuous video signal sampling the A/D converters used are, relatively speaking, instantaneous in operation.

3.5.2 Theory

It has been determined that the entire information content of an analogue signal can be represented by a periodic series of samples of that signal. The number of samples required, to fully describe the analogue signal, is determined by the analogue signal bandwidth. The following graphical analysis of the spectral components of the sampled and sampling waveforms will yield the relationship between signal bandwidth and the minimum sampling frequency necessary.

A circuit for sampling a signal was described in Section 3.5.1. By considering the analogue signal $f(t)$ and the sampling pulse train $p(t)$ in terms of their spectral components, the minimum sampling frequency necessary can be determined graphically. Figure 3.13(a) shows the analogue signal input as a function varying with time ($f(t)$). Figure 3.13(b) shows the spectral density $F(w)$ (Fourier transform) of the signal. Figure 3.13(b) shows the input signal to occupy a bandwidth f_b. Figure 3.13(c) shows the sample pulse train as a function of time ($p(t)$), with a repetition rate of f_r and a period T where $f_r = 1/T$.

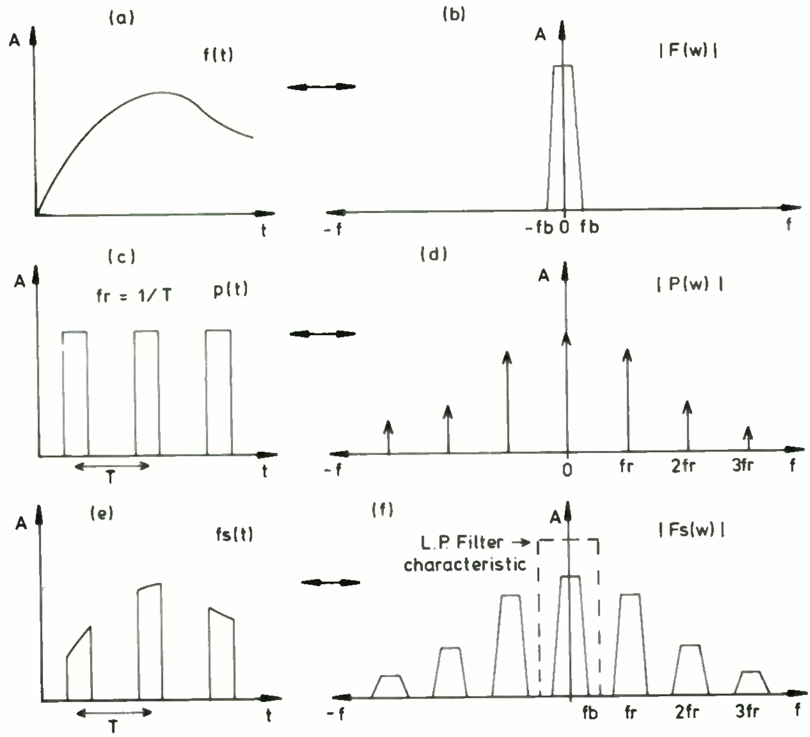


Figure 3.13 Spectra for sampled and sampling waveforms

Figure 3.13(d) shows spectral density $P(w)$ of the sampling pulse train. From this it can be seen that the spectral components exist only at harmonics of the sampling frequency f_r. Figure 3.13(e) shows the sampled waveform $f_s(t)$. Figure 3.13(f) shows the spectral density $F_s(w)$ of this waveform. It can be seen for this spectrum that the original signal spectrum $F(w)$ is reproduced, scaled at harmonics of the sampling frequency f_r.

The original baseband spectrum of the sampled signal ($F(w)$) can be recovered from the sampled waveform spectrum ($F_s(w)$) by using a low-pass filter. This operation is demonstrated by superimposing the low-pass filter characteristic on $F_s(w)$, as shown in Figure 3.13(f).

Explanation of the sampling theory, using a knowledge of the signal spectra, can be extended to describe the effects of sampling at too low a frequency. This will, in turn, lead to the minimum sufficient condition for sampling frequency. Figure 3.14 illustrates how the spectral components of the sampling waveform, $P(w)$, bunch together as the sampling frequency falls. In the extreme case, when the sampling frequency becomes infinitely low, i.e. a single sample pulse occurs, then the spectrum becomes continuous. Figure 3.14 shows the effect of $F_s(w)$ of reducing sampling frequency. It can be seen that the spectra $F(w)$ reproduced, run into one another. The overlap between the repeated spectra represents a region of the signal frequency spectrum that will be distorted due to 'aliasing' component. Figure 3.15(a) shows this effect in more detail by including only the baseband spectrum and the spectrum reproduced at the first harmonic of sampling frequency f_r. Here f_b represents the highest frequency component present in the original signal bandwidth. Figure 3.15(b)

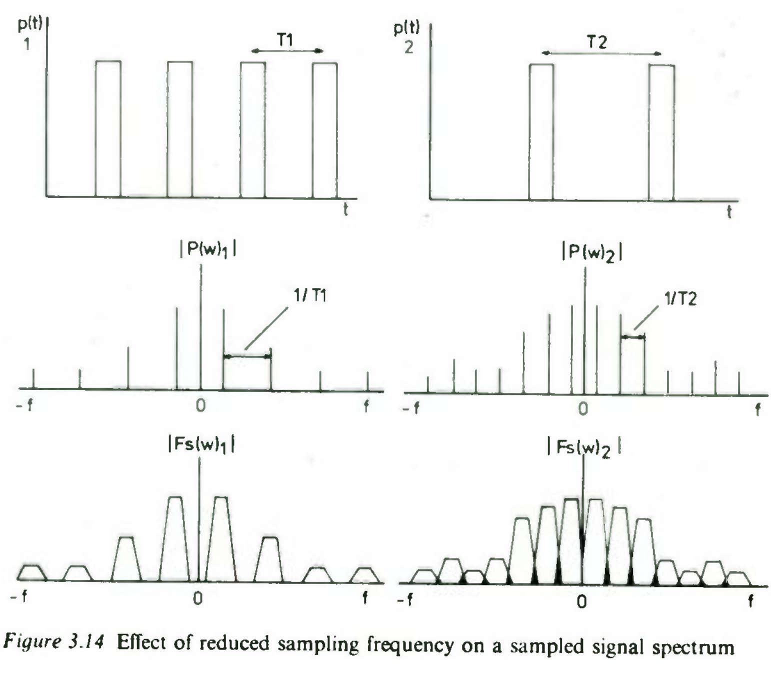


Figure 3.14 Effect of reduced sampling frequency on a sampled signal spectrum

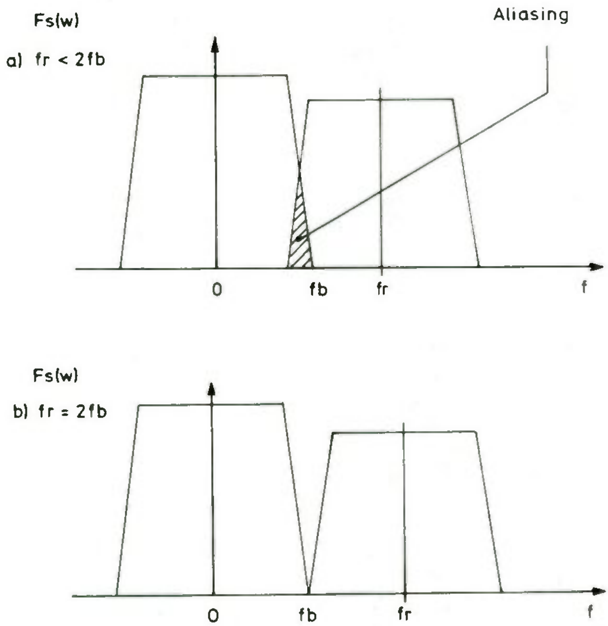


Figure 3.15 Elimination of aliasing components

shows the limiting case where the baseband signal is just free from aliasing components. Thus, due to the symmetrical nature of the spectra, the sampling frequency must be at least twice that of the input signal bandwidth f_b. This situation can be rewritten thus:

$$f_r \geqslant 2f_b$$

where f_r is the sampling frequency and f_b the maximum frequency component, or bandwidth, of the signal to be sampled. The minimum value of f_r to satisfy the above equation is known as the 'Nyquist' sampling frequency.

To ensure that the lowest possible sampling frequency is required, the input signal must be low-pass filtered to remove any unwanted out of band frequency components; components which otherwise introduce aliasing. In practice, sampling frequencies higher than the Nyquist frequency are used. This is because of practical constraints, such as finite roll-off low-pass filters.

Due to the large transmission data rates required for a digitised television signal it is important to take as little samples of the original analogue signal as possible, one such way of satisfying this requirement is to sample at a rate lower than the Nyquist sampling frequency. This is possible due to the discrete nature of the television signal spectrum. A discussion of sub-Nyquist sampling is beyond the scope of this chapter but Barratt and Lucas[1] give an introduction to this technique.

3.6 ANALOGUE/DIGITAL CONVERTERS

3.6.1 Analogue to digital converters

There are several methods of implementing an A to D converter, each with advantages and disadvantages, but the method used for high-speed video conversion is the 'fully parallel' or 'flash' converter. This type of converter is currently being used for conversion rates of over 30×10^6 samples/s. A number of manufacturers are currently marketing high-speed monolithic flash converters, which are described below.

The architecture of a typical flash converter is illustrated in Fig. 3.16. A band-limited video signal is applied, simultaneously, to 255 voltage comparators, thus enabling 256 different voltage levels to be determined (255 levels plus minimum reference). The reference voltage (applied between V_{ref1} and V_{ref2}) for all comparators is derived from a common resistor chain. The analogue signal input range is determined by the reference voltage applied to each end of the resistor chain. This range is so adjusted to give maximum possible resolution over the input voltage range, i.e. for an input signal varying between 0 and 1 V, 1 V is applied to V_{ref1} and V_{ref2} is connected to 0 V. By using a 255 to 8 line encoder an 8-bit code is produced, where the value of that code is a measure of the number of comparators that have been switched. The number of comparators switched is, in turn, a direct measure of the input voltage level.

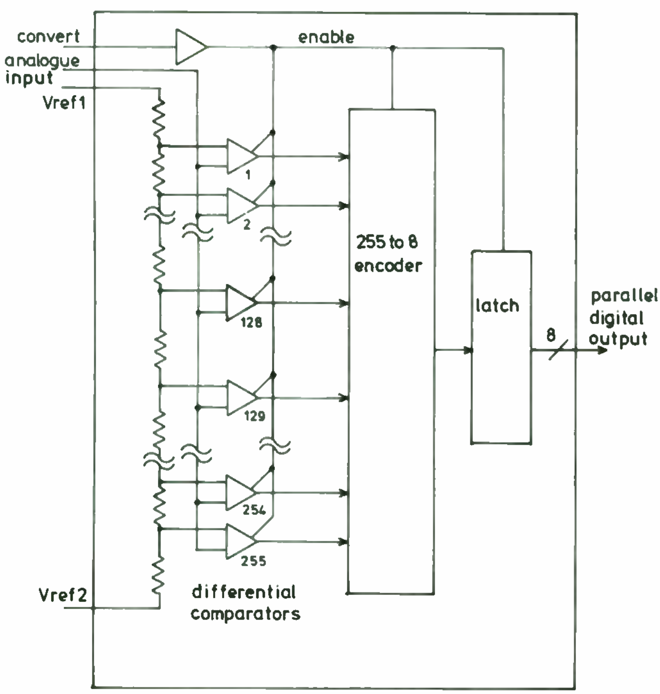


Figure 3.16 Flash analogue to digital convertor

A conversion is made by enabling the comparators, the input to the 255 to 8 line encoder and also the output latch. This enable-line is brought out of the converter package, and is referred to as the convert, sample or strobe pin. An instantaneous sample of the comparator output states (and hence the input voltage) is taken relative to a convert clock edge. Each conversion is held in the output latch until a subsequent sample is ready; in this way propagation delays in the encoder are prevented from affecting the output. Timing for a typical A/D converter is shown in Fig. 3.17. Of particular note here is the length of time, relative to the convert clock period, between a sample being taken and its digital form appearing at the output latch.

Some important definitions associated with A/D converters are listed here:

Quantising error For any A/D converter output code, there is a 1 LSB range of analogue input levels. It is not possible to tell, from the output code, the precise value of the analogue input level, there being a quantising error of $\pm\frac{1}{2}$ LSB. Since all A/D converters have this inherent error it is frequently not quoted by manufacturers.

Resolution The resolution of a converter is the smallest part of a full-scale input voltage that the converter can define, e.g. for a converter with 255

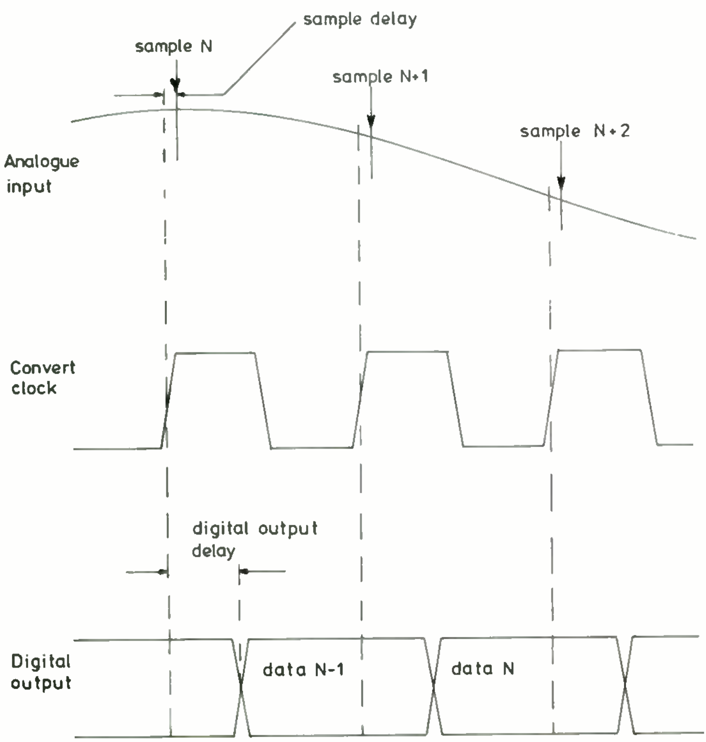


Figure 3.17 High speed flash convertor timing

comparators and full-scale input voltage of 1 V, the resolution is 1/256 or approximately 4 mV. For a converter with 512 comparators the resolution is twice as great, i.e. approximately 2 mV. The number of comparators chosen is always a power of two so there is no redundant output produced by the encoder. For a 6-bit converter, the number of comparators becomes 64 giving a reduced resolution of 1 part in 64, i.e. 16 mV for a 1 V input signal. The resolution implies nothing about the accuracy of the device.

Conversion time This is the time taken for an A/D converter to present a stable binary output after any one sample has been taken. To sample a signal, for example, at a rate of 14×10^6 samples/s the conversion time must be less than 72 ns, i.e. less than the period of the sampling rate.

3.6.2 Digital to analogue converters

High-speed digital to analogue converters are far simpler in design than A/D converters, and correspondingly cheaper to produce. Figure 3.18 shows the basic arrangement for a 3-bit, R-2R converter. Each 2R element is connected

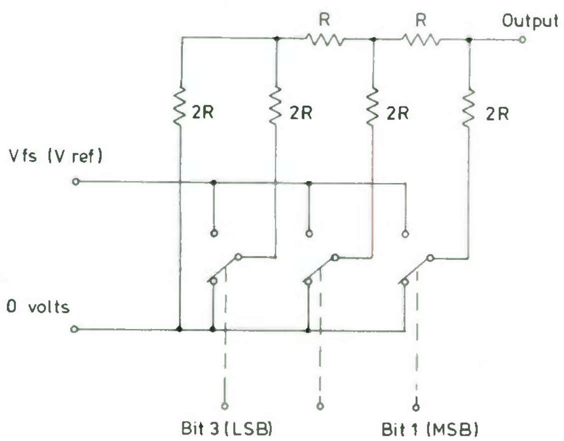


Figure 3.18 3-bit voltage switching DAC

either to 0 V or V_{fs} (V_{ref}) by transistor switches. Binary weighted voltages are produced at the output of the R-2R ladder, the value being proportional to the digital input number. For example, if bit 1 is '1' and bits 2 and 3 are '0' then an output of $V_{fs}/2$ is produced. This is because the resistance of the circuit, in this condition, looking from the output through the first R, is 2R. This forms a 2:1 attenuator with the 2R in series with the MSB switch. Output voltages for other input codes can be similarly determined. The R-2R ladder can be extended to any number of bits, within practical limitations.

The transfer function for the 3-bit D/A converter is shown in Fig. 3.19(a). For the transfer function the converter is assuming that all resistors are perfectly matched and that the transistor switches have zero on resistance. In

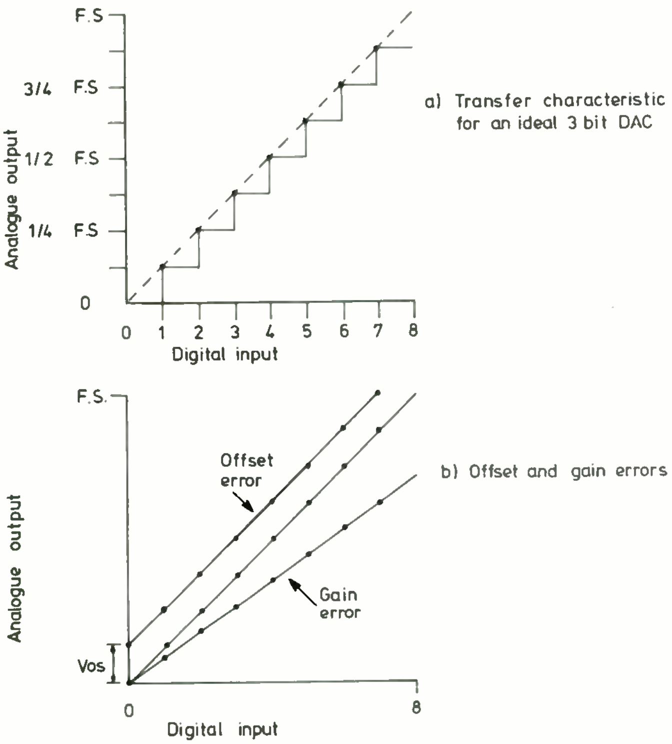


Figure 3.19 DAC characteristics

practice converters suffer from errors; the most significant errors are listed here:

Monotonicity When the input code to a D/A converter is increased in 1 LSB steps, the output should increase with linear step height; if this is so then the converter is monotonic. If, due to variation in resistance values, the step size does not increase in a linear fashion, the the converter is said to be non-monotonic.

Offset (zero error) Due to internal connection resistance and offset voltages in the switches an output offset error may exist. This has the effect of shifting the converter transfer function so that it no longer passes through zero. Offset error is illustrated in Fig. 3.19(b), where V_{os} represents the offset error.

Gain error Assuming the reference voltage to be within specification, the transfer function staircase should follow the ideal straight line. If imperfections in the converter lead to a divergence from the ideal then gain error is present. This is illustrated in Fig. 3.19(b), where the gain error is the measure of difference between the slope of the actual transfer function and that of the ideal transfer function.

Resolution The resolution of a D/A converter is the smallest part of full scale (V_{fs}) that can be produced at the output. The resolution for a D/A converter is normally quoted as the number of bit inputs, e.g. for a converter of 10-bit resolution and a reference voltage V_{fs} of 1 V, the smallest output volt increment is $1/2^{10}$, i.e. approximately 1 mV.

Settling time This is the time taken, after a change in input code, for the output of the converter to settle to within $\pm\frac{1}{2}$ LSB of its final value. This time varies, depending upon which bits are being changed. Settling time may be specified for a change of 1 LSB but this generally gives the most optimistic figure. A more useful measure is that for the worst case transition, i.e. the MSB changes in the opposite direction to that of all other bits (100 to 011 for a 3-bit converter) or from zero to full scale (000 to 111 for a 3-bit converter). To convert a digital input at, for example, a rate of 14×10^6 samples/s, then the output must settle in less than 72 ns, i.e. less than the period of the sampling rate.

Glitching A glitch, in data conversion context, is a transient excursion occurring as a D/A converter switches between two different output levels. For example, when changing from input code 011 to 100 not all bits will switch simultaneously. In the worst case there will be a time T when the code 111 will appear at the output. This period T is very short and gives rise to a very narrow spike, i.e. the glitch. Glitches can generate noise components all through the passband of the host system.

3.7 APPLICATION OF DIGITAL TECHNIQUES WITH PARTICULAR REFERENCE TO FRAME STORAGE

3.7.1 The television waveform

The purpose of this section is to give a physical appreciation of the principles outlined in Sections 3.2 to 3.6. The frame store has been chosen because of its fundamental importance within digital video processing equipment. It is, however, beyond the scope of this chapter to describe how the frame store is actually used in this equipment. This section is aimed at building a simple model of a frame store from first principles. Before continuing the discussion, a rudimentary description of the analogue television signal must be given. To keep this as simple as possible (most aspects of the television signal will be given in later chapters), material covered here describes a monochrome signal only.

A television picture is produced by scanning 312.5 equally spaced lines (British system) across a CRT every 1/50 s, this operation produces a television 'field'. Two such fields are interlaced to form a complete 'frame', which constitutes a complete 'still' picture. Twenty-five such frames are transmitted every second to produce continuous moving pictures. Figure 3.20 shows a simplified section of a transmitted television waveform, sufficient to form a complete video frame. Of particular note, here, are the line and field blanking periods.

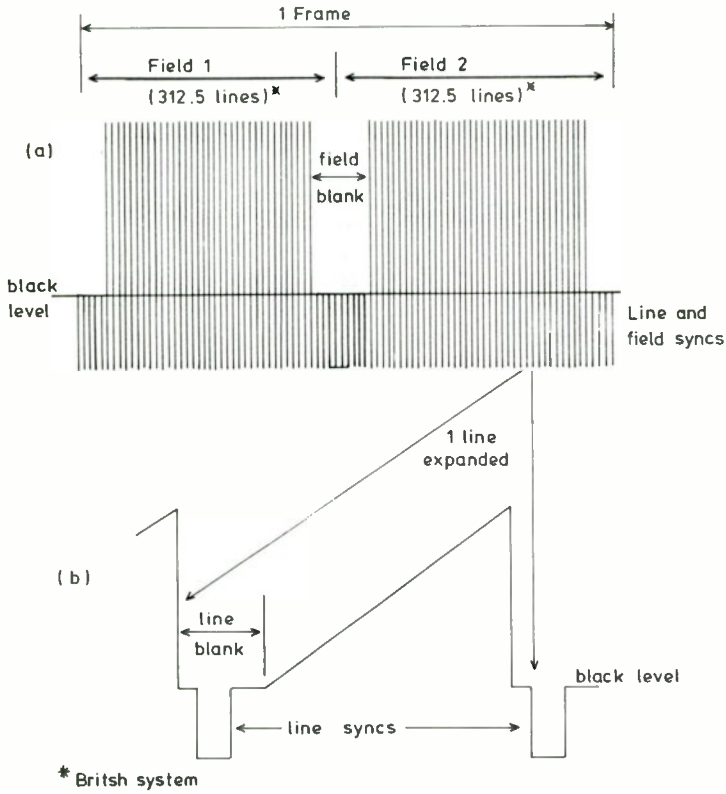


Figure 3.20 Simplified television waveform

3.7.2 Picture elements

The frame store, as currently implemented, is a large array of MOS memory devices, capable of storing a digital representation of a complete frame of video information. A store of this nature is still an expensive item, so it is important to determine the minimum size necessary. To do this the 'resolution' and 'number of grey levels' must be known. The subject of grey levels is covered in Section 3.7.3, resolution is the subject of this subsection.

The resolution of a television picture, and hence the frame, is a measure of the ability to reproduce fine picture detail. The quantitative measure of resolution is the number of 'picture elements' a complete frame can reproduce; a picture element being the smallest, clearly discernible point in the picture.

The resolution of the television frame is primarily determined by the signal transmission bandwidth, for the British system the bandwidth is 5.5 MHz. The reciprocal of the bandwidth gives the period of the picture element, i.e. 91 ns. Taking into account line blanking, the active line period is about 52 μs, therefore the number of picture elements per line becomes 52 μs/91 ns which gives 572. Field blanking leaves about 580 active lines, therefore the total number of picture elements is about 340,000. This means the frame store has to contain the digital representation of 340,000 picture elements, where each picture element has its level (between black and peak white) defined by a binary code. Picture element coding is discussed in the next section.

3.7.3 Picture element coding

Conversion of a television line (such as in Fig. 3.20(a)) into digital form is performed using an analogue to digital converter of the type described in Section 3.6.1, i.e. the flash converter. Here, the A/D converter employed must be adequate to resolve the level of the television signal to a degree fine enough not to produce an unpleasant 'grainy' picture when reconverted into analogue form. To illustrate this point, consider the application of the waveform in Fig. 3.20(b) to an A/D converter of only 4-bit resolution. For only 4 bits the A/D contains only 15 comparators, thus for the 1 V peak to peak signal applied to it only 16 amplitude levels can be resolved. This gives a figure of 64 mV, which means a quantisation error ($\pm\frac{1}{2}$ LSB) of ± 32 mV. Figure 3.21(a) shows the effect on the ramp when it is reconverted into analogue form, it can be seen here that the linear ramp assumes a staircase form. Each step of the staircase represents a discrete 'grey level' observable on the monochrome monitor; for a 4-bit converter 16 grey levels are produced (in fact one being black and another peak white). This is illustrated in Fig. 3.21(a) where only signal information

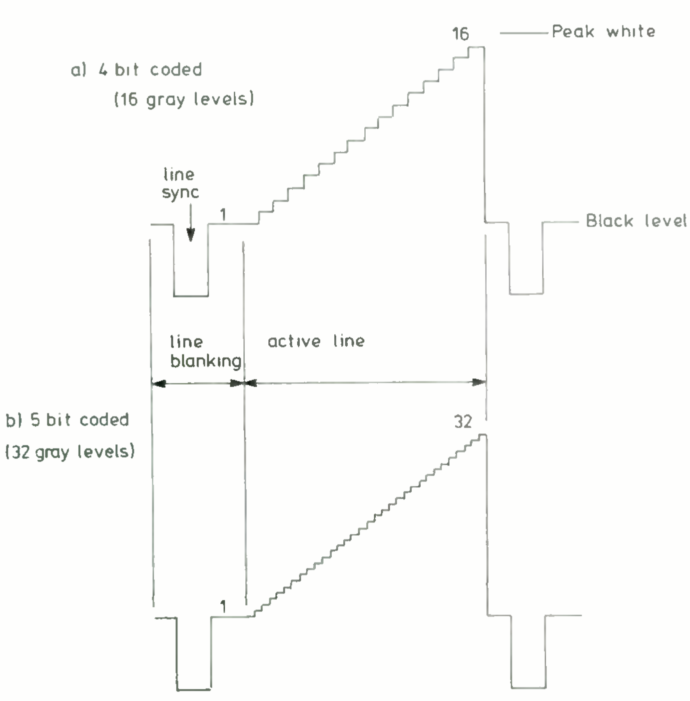


Figure 3.21 Line ramp digitisation

above black level has been digitised; in practice the signal may be digitised down to sync tips to include colour information which extends below black level. As the resolution of the converter (i.e. the number of bits) is increased, the quantisation error decreases and the staircase form approximates more closely to a linear ramp; this is illustrated in Fig. 3.21(b). It has been determined, by early research, that an 8-bit converter is sufficient to sample such a ramp, which when reconverted to analogue form appears to the human eye as a linear ramp. In other words, the staircasing effect is not observable. Converters of 9-bit resolution and greater have been used, but equipment costs rise steeply for every extra bit of resolution.

In conclusion to the above paragraph every picture element requires coding with 8 bits to produce adequate picture quality. For the 8-bit converter 256 levels are discernible.

3.7.4 Frame store organisation

Sections 3.7.2 and 3.7.3 have determined the necessary information to calculate the size of store required for one frame. For a resolution of 340,000 picture elements per frame and a grey level coding of 8 bits per picture element, the total size of store required is 2.72×10^6 bits.

Section 3.7.2 described the picture element period as being the reciprocal of twice the bandwidth; this is due to the Nyquist criterion. This says that the signal must be sampled at least twice that of the maximum frequency component, i.e. for 5.5 MHz the sampling rate must be at least 11 MHz (this is to avoid aliasing). To satisfy this requirement the frame must, in effect, store two samples per picture element (still 8 bits per sample). The minimum store requirement then becomes 2.72×10^6 bits. In physical terms, if this store was implemented in dynamic RAM of the type described in Section 3.4.3, then 167 16-pin DIL packages would be required.

The essential components of the frame store are outlined in Fig. 3.22. A video signal to be stored is applied both to a sync separator and a low-pass filter. The sync separator strips off the timing information carried on the video signal (see Fig. 3.20). Line and field blanking information from the sync separator is used to enable and disable the store, so insuring only active picture information is stored.

The video signal is band limited by the low-pass filter, this removes any unwanted, out of band, signal components that would cause aliasing. Filtering video is then presented to the input of the A/D converter. An analogue to digital conversion is made when the converter receives a strobe pulse from the sample clock. The sample clock, for this system, will run at 11 MHz (twice bandwidth or greater) and can be turned on or off by the sync separator (for blanking periods). At the output of the A/D converter, an 8-bit code is produced for every sample made. The operation of writing to the store is achieved by the following steps:

(1) A sample pulse is generated by the sample clock, and increments the address counter. This opens a new memory location to the digitised signal.
(2) The sample pulse causes the A/D converter to perform a conversion and present its result to the input of the store.

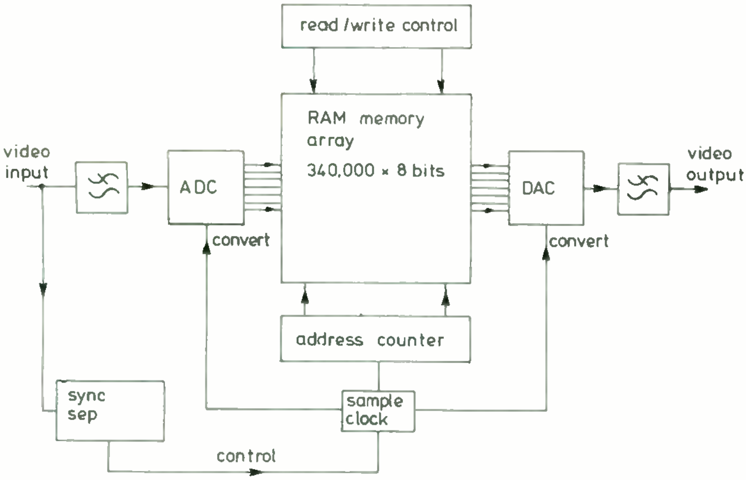


Figure 3.22 Simplified framestore organisation

(3) The store-write strobe is pulsed to transfer conversion result to the memory location selected by the address counter.

This process happens continuously, with the exception of blanking intervals.

Frame store contents can be reconverted into analogue form by sequentially reading the contents back out through a D/A converter. The store-read operation is as follows:

(1) The address counter is clocked by the sample clock and the contents of the memory location selected, presented at the store output.
(2) The D/A converter is then strobed to convert the code present at its input. Video output from the converter is then low-pass filtered to remove any spectral components above 5.5 MHz.

Section 3.2 discussed some of the higher-speed logic families available, equipment such as the frame store use some, if not all these devices. It is common to find the outputs of A/D converter and the inputs of D/A converters implemented using ECL. Apart from the logic speeds involved, the converters are often on separate circuit cards which are connected to the store cards via twisted pair transmission lines. This ensures data integrity and little supply disturbance.

The store itself is usually implemented using MOS memory with the demultiplexing system as described in Section 3.4.4. As the MOS memory described is TTL compatible it is common to find Schottky and low-power Schottky logic performing multiplexing and demultiplexing. The sample clock and address counters are likely to be implemented in Schottky TTL or possibly ECL.

This explanation of the frame store has been kept very basic in order to describe its fundamental parts and organisation with relevance to the previous sections. Utilisation of the frame store in the types of equipment described in

Section 3.1 is certainly beyond the scope of this book, but it is left to the reader to consider the implications of the following effects on the output picture when the store described in Fig. 3.22 is affected in two ways:

(1) Altering the store read timing relative to the store write timing.
(2) Altering store address connections during a store read.

REFERENCE

1. Barratt, K. H. and Lucas, K., 'An introduction to sub-Nyquist sampling', *IBA Technical Review*, No. 12 (1979)

Chapter 4

Standards

R. S. Roberts

4.1 BANDWIDTH

It was shown in Chapter 1 that picture quality is determined by the size of the scanning spot used in the scanning process. A small spot requires a high number of lines in the vertical 4 × 3 dimension, to fully scan the field of view. The number of lines can thus be used to indicate the bandwidth of the video channel, and a compromise must be made between a fully acceptable picture quality and the channel bandwidth, bearing in mind that the demand for bandwidth increases as the square of the number of scanning lines.

It was shown in Fig. 1.6 that, in System A (the first to be standardised, and without field experience), 405 lines was adopted as a compromise between adequate definition and a minimum bandwidth. The resulting complete channel space, including the double-sideband video, a sound channel, and guard space at each end of the channel (necessary to avoid interference with adjacent channels), was an overall channel width of 7 MHz. This was the only UK television channel when broadcasting was re-started on System A in 1945.

Meanwhile, the US had commenced broadcasting a television service with standards based largely on the UK field experience. After a false start on 441 lines, they finalised their 525 line System M, as used today. The bandwidth this would require, using the same type of calculation that we used earlier, is as follows:

Number of squares $= 525^2 \times 4/3$
30 pictures/s $= 525^2 \times 4/3 \times 30$ squares/s, and
Highest modulating frequency $= 525^2 \times 4/3 \times 30/2 = 5.5$ MHz

It was decided to limit the highest modulating frequency to about 4 MHz which, compared with System A, should result in a better picture quality. The resulting video double-sideband would be 8 MHz and, with a sound channel spaced +4.5 MHz and guard frequencies, results in a channel width of about 9 MHz. This wide channel requirement posed a serious problem. The US demands for channel space were considerable because, unlike the UK with its one channel serving the greater London area only, New York required about six channels, San Francisco about seven, with similar demands right across the country.

4.2 DSB, SSB, ASB AND VSB

The US engineers then modified their standard in a manner that resulted in a considerable saving in channel space. They developed what is known as a 'vestigial sideband' system (VSB), sometimes termed an asymmetric sideband (ASB) system.

4.2.1 AM distortions in sound and vision

In order to appreciate the operation of VSB it is necessary to digress into some important differences between sound and video amplitude modulation. It is well known that single-sideband AM systems have been in use for speech and music for at least 60 years. Why cannot SSB be used for video modulation and, thus, save half of the bandwidth?

We re-called the principles of DSB for sound in Fig. 1.4(a). Two important factors concern possible distortion in such a speech or music system. An amplitude constraint exists whereby any attempt at over-modulation results in severe waveform distortion, with very unacceptable audible quality. The second effect occurs in single-channel sound reproduction where the quality of the reproduced sound is totally unaffected by distortion of phase response. It must be pointed out that there are many who do not accept that our normal binaural experience does not distort. Nevertheless, simply turning one's head, for example, results in a severe phase change of say a 5 kHz sound, where about 4 cm movement represents about 180°. We live with this effect and do not notice it, even when we use the effect for directional information.

4.2.2 AM vector relationships

Figure 4.1(a) shows the vector relationship of the side frequencies with respect to the carrier during one cycle of a single modulating frequency of double-sideband AM at maximum modulation. The process was detailed in Chapter 2. Figure 4.1(b) shows the resultant vector addition of those in (a), from which is seen the well-known rise and fall of carrier amplitude and, notably, the phase of the resultant vector, which remains that of the carrier at all times during the modulation cycle.

Figure 4.1(c) shows the same as (a), but with the lower side-frequency suppressed. Figure 4.1(d) is the resultant of (c). We now see two new features: (1) the modulation, considered in terms of carrier amplitude, is halved, and (2) phase modulation, the angle of the resultant vector swinging with respect to the carrier.

It is worthwhile to digress and complete the story of SSB speech transmission. The commercial use of SSB sound transmits a single sideband, such as shown for a single tone in Fig. 4.1(e), with either no carrier or with a low-level carrier, often termed a 'pilot tone'. At the receiver, a carrier, either

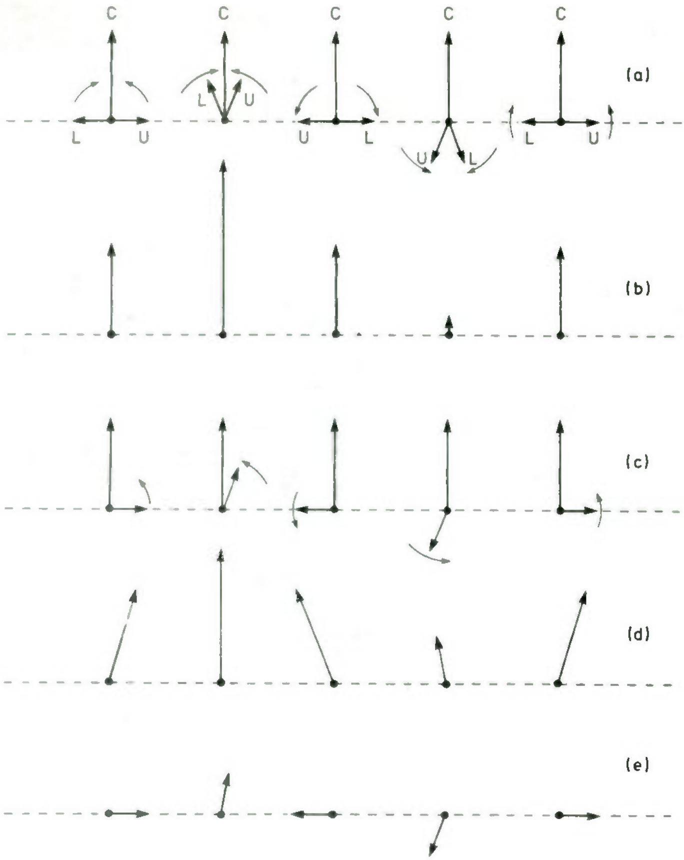


Figure 4.1

receiver-generated or consisting of the extracted pilot-tone which is amplified, is introduced for demodulation as shown in Fig. 4.1(c). The recovered modulating signal would be distorted severely if the relative levels shown in (c) were used, and phase distortion would be severe, as shown in (d). To reduce the distortion, the carrier is re-introduced for demodulation at a relatively high level, compared with the side-frequency amplitude. This reduces the amplitude and phase distortion.

4.2.3 Pulse and phase response

When compared with audio modulation, what is required of a video signal used for AM? Let us consider a simple video modulating signal, as shown in Fig. 4.2(a). It was shown in Chapter 2 that a pulse, such as shown in Fig. 4.2(a), could be analysed into a number of discrete harmonic frequency components, which extend out to infinity with a descending order of amplitude, as indicated in Fig. 4.2(b). Similarly, the square waveform of (a) could be synthesised by

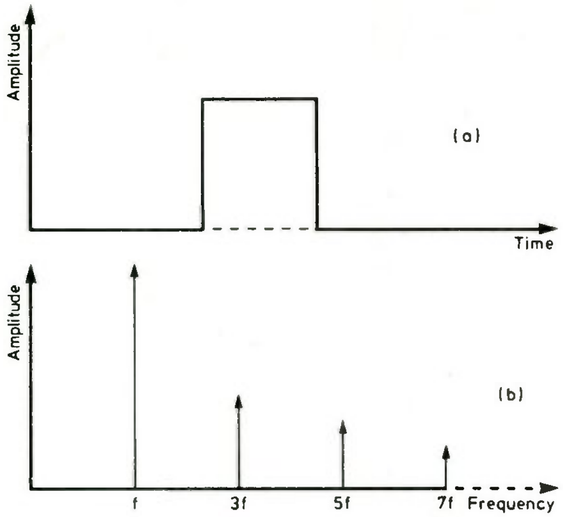


Figure 4.2

addition of the frequency components of (b), at the correct amplitudes and phases. We require the pulse signal to progress through the system in such a manner that the shape of the waveform that is finally modulating the display-tube beam current, is a faithful copy of the signal derived by the scanning process at the transmitter.

The pulse of Fig. 4.2 can experience degradation of its waveform in a number of ways as it goes through the system. For example

(1) An inadequate HF response at some stage in the system can round off the corners of the pulse, to a degree depending on the number of HF harmonic components affected. In the limit, when all of the harmonics are removed, we are left with the fundamental as shown in Fig. 1.5.
(2) The various discrete frequency components that constitute the signal waveform may experience differing transit times as they travel through the system. Some may go through faster or slower than others, resulting in wave-shape distortion to varying degrees at the point where they are supposed to arrive together, e.g. at the point where the display-tube beam current is modulated.
(3) Inadequate LF response could produce 'tilt' at the top of the pulse. The d.c. voltage may not hold up for the pulse duration.

Consideration of (1), (2) and (3) shows that (1) and (3) are concerned with straightforward circuit behaviour, and any deficiencies in this respect can be resolved, but the timing effect of (2) is much more important. The phase response of the system is, clearly, of the utmost importance for preservation of pulse shape. Each of the signal-frequency components must travel through the system in the same time. Within reason, the actual time of transit is of no importance.

Further examination of Fig. 4.2 will show this phase requirement more clearly. Assuming a time t taken by the fundamental frequency f to go through the system. This time can be interpreted as a phase change ϕ. The frequency $2f$ will go through twice this phase angle in the time t. $3f$ will shift three times the fundamental phase angle, and so on for each harmonic component. If, for example, one of the harmonic components goes through too great a phase shift, it means that it will travel too fast and will arrive at its final destination ahead of the correct time. A little thought shows that the phase response of the system should be such that the angle ϕ must be proportional to frequency, as shown in Fig. 4.3. Any one of the lines shown in Fig. 4.3 would indicate a satisfactory phase response. The linear relationship is essential, and the only significance of the differing slopes is that they relate to different transit times. (A horizontal line would indicate zero time and an instantaneous appearance of the input signal at the output which is, of course, not possible.)

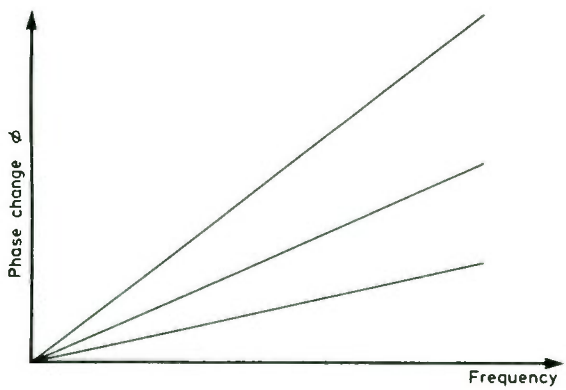


Figure 4.3

The main differences between audio and video signal processing are now more clearly seen. Audio signal processing can tolerate phase distortion to a high degree, but amplitude distortion must be kept to very low limits in order to be acceptable. Video signal processing must ensure that phase distortion is minimised to very low values, but amplitude overload effects are far less serious.

The peak value of Fig. 4.2 may represent peak white, for example, and any increase above this value would not therefore be of visual significance, although it may drive a transmitter to an overload condition. Alternatively, the peak value may correspond to a black signal value, in which case 'blacker-than-black' is of no visual consequence. Of course, no engineer would deliberately design a television system to operate under permanent overload conditions. (It is of note that, in techniques concerned with digital systems, 'slicing' the top of a pulse is often used to remove noise that has been picked up in transmission, to restore a 'clean' pulse.)

To return to the earlier question, 'Why not halve the bandwidth by the use of SSB?'. The answer is seen to be that the phase distortion produced by SSB (Fig. 4.1(c), (d) and (e)) may be acceptable for audio, but is totally unacceptable for video. However, it *is* possible to use partial suppression of one sideband (i.e.

vestigial sideband, VSB) for a video system and save considerable channel space by taking full advantage of the fact that (1) visual disturbance due to phase errors are severe and unacceptable where large picture areas are concerned, but (2) phase errors become difficult to see on small detail in the picture. Thus, low modulating frequencies must minimise phase distortion, whereas the high frequencies generated by fine detail are tolerant of phase distortion effects because, although they may be present, their effects are very difficult to see.

Figure 4.4 shows, at (a), the full channel bandwidth that would result from double-sideband video amplitude modulation, together with the sound channel, for the UK System I. The transmitted signals are shown in Fig. 4.4(b) where the lower sideband has had about 4 MHz filtered off, leaving a 'vestigial' 1.25 MHz. The 12 MHz channel has been reduced to 8 MHz. In order that VSB shall function correctly, it is necessary that the received signals be presented to the demodulator as shown in Fig. 4.4(c). Consider, firstly, a 50 Hz

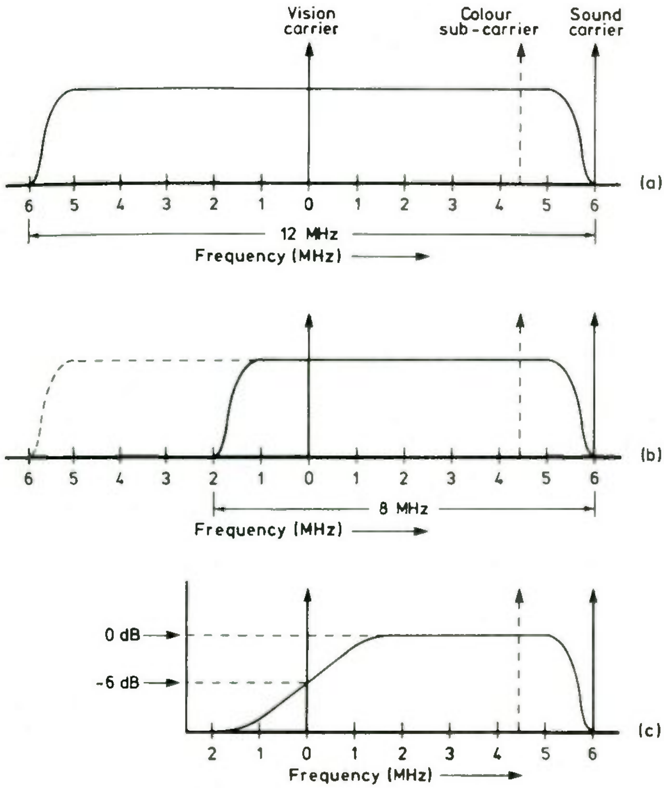


Figure 4.4 (a) The full bandwidth that would result from amplitude modulation. (b) Part of one sideband filtered off before transmission. (c) Receiver IF response presented to the demodulator

modulating frequency, for example, and its position in Fig. 4.4(b). This frequency would result from the scan of a large area, indeed a complete field, and the resulting side frequencies would be very close to the carrier. The signals presented to the demodulator by the receiver i.f. amplifier are shown as vectors in Fig. 4.5(a). The carrier and side frequencies are cut by 6 dB in the i.f. filter, but no phase distortion would result in the demodulation process because the signal is a full double-sideband AM signal at 100% modulation, as shown in Fig. 4.1(a). Consider now a modulating frequency resulting from a smaller detail than a full field size, producing a frequency of say 500 kHz.

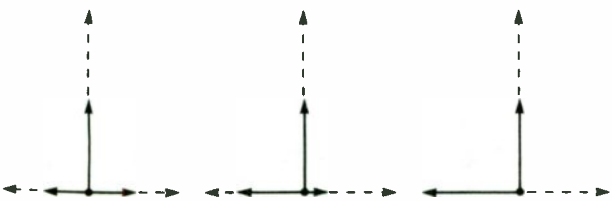


Figure 4.5

Figure 4.5(b) shows the vector relationship between side-frequencies and carrier. The 500 kHz signal might possibly produce the vector situation shown in Fig. 4.5(b), where the overall length of the upper plus lower side vectors add to the same as shown in (a), showing that the modulation factor has not changed. However, one vector is longer than the other, and some phase distortion is beginning to appear in the resultant addition of the three vectors. At frequencies of about 1.5 MHz and above, the vector situation is shown in Fig. 4.5(c). The single side-frequency resulting from the scan of a small detail has an amplitude equal to the carrier and, again, the level of modulation is unchanged. Phase distortion is present, but the detail is small and the errors are difficult to see.

The US problem of minimising channel space was considerably eased by the use of VSB and all standard systems now use VSB. It is of note that all UK System A stations that followed Channel 1 used VSB, and when the first station moved from its original site to Crystal Palace, the opportunity was taken to bring it into line and change it from DSB to VSB. The vestigial band was less than 1 MHz, and the channel width was 5 MHz.

4.3 NATIONAL STANDARDS

Table 4.1 gives details of most of the standard systems in use throughout the World. Firstly it will be noted that most of them use 625 lines. The exceptions are the UK System A, which is now phased out, the US System M on 525 lines, and the French System E on 819 lines. It is unlikely that the US will change their Standard M, because it is used by many countries other than the US, such as Japan. The French 819 line System E was developed before the last

Table 4.1 STANDARD TELEVISION SYSTEMS

	A	*B*	*C*	*D(K)*	*E*	*G*	*H*	*I*	*L*	*M*	*N*
Lines per picture	405	625	625	625	819	625	625	625	625	525	625
Field frequency (Hz)	50	50	50	50	50	50	50	50	50	60	50
Line frequency (kHz)	10.125	15.625	15.625	15.625	20.475	15.625	15.625	15.625	15.625	15.734	15.625
Video bandwidth (MHz)	3	5	5	6	10	5	5	5.5	6	4.2	4.2
Channel bandwidth (MHz)	5	7	7	8	14	8	8	8	8	6	6
Sound/vision carrier spacing (MHz)	3.5	5.5	5.5	6.5	11.15	5.5	5.5	6	6.5	4.5	4.5
Vestigial sideband width (MHz)	0.75	0.75	0.75	0.75	2	0.75	1.25	1.25	1.25	0.75	0.75
Vision modulation polarity	+ve	−ve	+ve	−ve	+ve	−ve	−ve	−ve	+ve	−ve	−ve
Sound modulation	AM	FM	AM	FM	AM	FM	FM	FM	AM	FM	FM
Deviation (kHz)		50		50		50	50	50		25	25
Pre-emphasis (μs)		50		50		50	50	50		75	75

War as an attempt to standardise a system that will provide a better picture quality than the 405 line standard then prevailing. This virtual doubling of the line structure does, indeed, provide an excellent picture quality, at the expense of channel space, of course.

The almost universal adoption of a 625 line standard resulted from an international meeting on standards that followed the end of the last War. At that time, all countries required to start a television service, and possible standards were discussed at length. It was hoped that such standards could be established to permit international links between European countries and 625 lines was considered to be a better compromise between picture quality and channel width. The CCIR 625 line standard has now been adopted by many countries but, unfortunately, did not result in a universal standard. Examination of Table 1 will show that, in Europe, there are many 625 line systems—and none compatible. Some have positive modulation, others negative. Some have AM sound, others FM sound and vision carrier spacing varies.

The addition of colour to broadcast television has added further differences. Chapter 8 deals with the first broadcast colour standard, the US-NTSC system, which was adopted for System M. A later European conference attempted to standardise a European colour system. It was considered that the NTSC system could be improved, and several proposals were made for systems based on the NTSC system, which uses a sub-carrier to convey colour information to the receiver. Germany proposed the NTSC modification known as PAL (Phase Alternating Line). The French proposal was for a frequency-modulated sub-carrier to be used for the colour information, the system being termed SECAM (Séquence et Mémoire). Other variants were considered, but the final outcome was that PAL was preferred and adopted by most countries, whilst France decided to use SECAM.

There is one common feature of all colour systems, they all use a sub-carrier for the colour information in the form of a 'burst' of a few cycles for use by the receiver as a reference. The burst usually 'sits' on the back porch of each line-sync pulse, as shown in Fig. 1.9.

4.4 BANDS AND CHANNELS

All users of the radio-frequency spectrum occupy some part of the spectrum which is classified in Table 4.2. The band classifications are international and in some of these bands broadcasting takes place. Medium wave broadcasting, for example, takes place in the MF band.

Table 4.2 RADIO-FREQUENCY SPECTRUM

Band	*Frequency*	*Band*	*Frequency*
VLF	<30 kHz	VHF	30–300 MHz
LF	30–300 kHz	UHF	300–3,000 MHz
MF	300–3,000 kHz	SHF	3–30 GHz
HF	3–30 MHz	EHF	30–300 GHz

Table 4.3 EUROPEAN BANDS

Band	*European band*	*Frequency range*	*Service*
VHF	I	41 – 68 MHz	Television
VHF	II	88 –108 MHz	FM sound
VHF	III	174 –216 MHz	Television
UHF	IV	470 –582 MHz	625 line, System I, colour
UHF	V	614 –854 MHz	625 line, System I, colour
SHF		11.7– 12.2 GHz	Satellite services

Table 4.4

European band	*UK Channel numbers*
I	—
III	—
IV	21–34
V	39–68

Table 4.5

Channel	*Frequency* (GHz)	
4	11.78502	All left-hand circular polarisation
8	11.86174	
12	11.93846	
16	12.01518	
20	12.09190	

The sections of the spectrum which are allocated to television are, of necessity, in the higher-frequency parts of the spectrum, due to the need to provide a wide frequency band for each channel. There are three World zones with differing frequency allocations for broadcasting, and Table 4.3 shows the European bands and their European designations. A recent international conference on frequency allocation introduced modifications to some of the bands shown in Table 4.3, the most important being that, in some bands, exclusive use for broadcasting will no longer be enjoyed, and the band will be shared with other services. The UK 405 line, Standard A has been closed down. However, the UHF Bands IV and V will be virtually unchanged, except for the possible addition of further channels.

In Europe, the division of bands into channels is a domestic decision for each country. The differing standards mean that channel widths are not the same for each country. The UK television channels for UHF are shown in Table 4.4.

Figure 4.6 shows the essential features of Tables 4.2 to 4.4 collected into a single presentation. All television and sound allocations are shown, together with the satellite band.

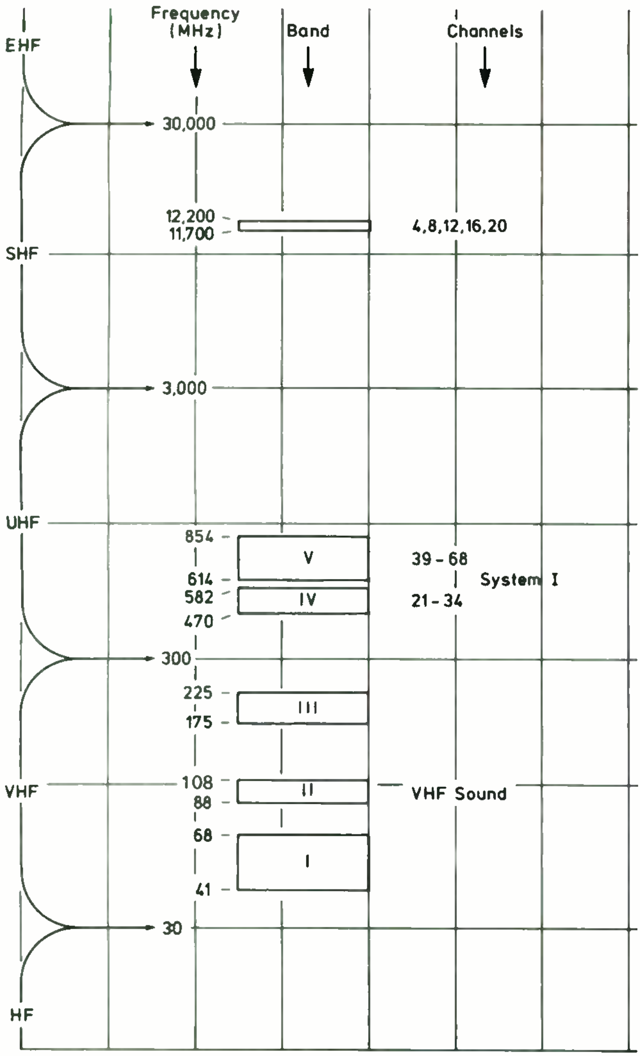


Figure 4.6

The satellite band covers 11.7 to 12.5 GHz and within this band there are 40 channels, each 20 MHz wide. Five of these channels have been allocated to the UK, as shown in Table 4.5. The form of the transmission has not been finalised, but the downcoming signals will be frequency modulated to ensure the highest-possible signal/noise ratio.

Chapter 5

Colour

R. S. Roberts

However good we consider the engineering of a television system to be, the final arbiter in this respect is the human eye. Visual satisfaction and acceptance by the eye is essential in the final display of the system*.

5.1 THE EYE

'Seeing' is a highly complicated process and our knowledge of the process is not complete, but we know sufficient to enable the engineering of a system to take full advantage of the possibilities and limitations of the eye behaviour.

Figure 5.1 shows a cross-section of a human eye. It has a lens that focusses the external scene on to the retina at the back of the eye. An iris controls the amount of light passing into the eye. The retina has, on its surface, a large number of light-sensitive cells which are stimulated to varying degrees by the focussed scene. The action is a chemico-electric one in which the incidence of light on a cell produces a change in its chemical constitution and this, in turn, generates a voltage which is transmitted as an electrical impulse. Removal of the light stimulus produces a reversal of the chemical change, with a time-lag, as mentioned in Chapter 1, which is an important part of the illusion of movement that television and film provides.

The cells on the inside wall of the retina are of two kinds, rods and cones. Most of the cells are rods, and they number about 120 million, occupying most of the inner surface of the retina, providing a visual experience over more than 180° for each eye. The rods have a maximum density near the centre of the retina, over an angle of about 10°–20°. The rods have a very high sensitivity, and have no colour response. Variations in brightness of the visual scene are interpreted very accurately by the rods which are also very sensitive to periodic variations in light intensity (i.e. 'flicker').

* This chapter will be using many terms concerning light, and these are defined and explained in the Appendix at the end of the chapter.

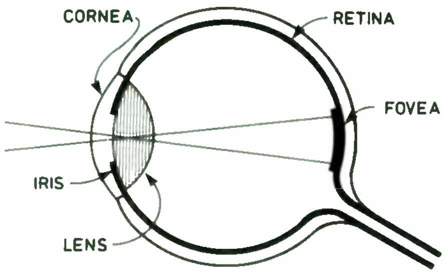


Figure 5.1

The simultaneous excitation of the millions of cells produces an electrical voltage from each cell, and these are conveyed to the optic nerve for processing in the brain. However, a recent count of the number of nerve fibres on the 'bundle' shows that the number is far less than 120 million and is of the order of 50,000 to 100,000. This suggests that a large amount of the processing of the visual information takes place in the retina and it is the pre-processed information that is passed to the brain.

The cones are cells of a different kind. There are about 6 million of these, centred over a narrow angle of about 5° in the foveal region of the retina. The cones are the interpreters of any fine detail in the image and they are the only cells with a sensitivity to colour. (Some night-predator animals, such as the tiger or the domestic cat, have no colour sensitivity.) In order that the cones can perceive colour, a minimum brightness level is required (about 0.5 ft-lambert).

A further part of the eye/brain processing is that somewhere along the signal path there must be a memory bank to which the visual signals are referred. Figure 5.2 presents some visual information which the brain will interpret in either of two ways, clearly using a memory bank.

5.2 LUMINANCE

In general terms, the 'brightness' of a scene is termed its luminance, of which there are many units (see Appendix). The range of luminance over which the eye can function is astonishingly large. The upper limit is about 10^4–10^5 ft-lambert, which is a value for bright sunlight on fresh snow. The cinema is a visual environment which is dark, and film is displayed at about 10 ft-lambert. Black and white television is, of course, just a luminance system, and is displayed at about 15–50 in the domestic environment. At the bottom of the scale is the light from faint stars which may be 4,000–5,000 light-years away, which is of the order of 1×10^{-6} ft-lambert. This enormous range of luminous intensity that the eye can accept and interpret is, thus, at least 10^{10} but it has, in addition to the automatic iris operation, a very efficient AGC system of some kind. Figure 5.3 shows the range of luminance values in more familiar engineering terms, the decibel. The 0 dB reference is the faint star value, which the physicist will appreciate as about two photons.

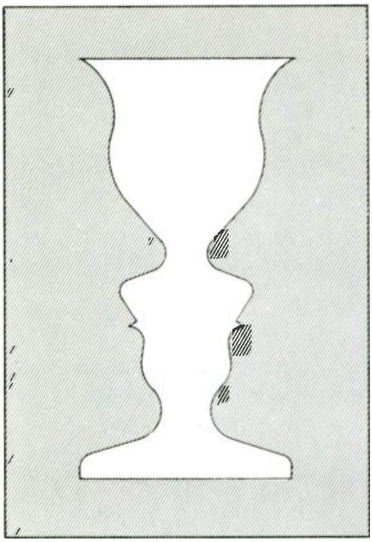


Figure 5.2

Although the eye can cope with this considerable range of brightness, the process has its limitations. During a bright sunny day, we know that there are stars in the sky, but we cannot see them. The reason is that the eye has another little understood process that comes into play. In any visual scene, the eye 'adapts' to the prevailing range of luminance values, establishing for itself peak white and black values and all the intermediate shades of gray. In fact, the term used for this eye adjustment is 'adaptation'.

5.3 CONTRAST

Another effect associated with luminance is 'contrast', or the difference in luminance values between white and black. The average brightness in a scene can be anywhere on the scale shown in Fig. 5.3. The eye, adapted to a particular value of average luminance, will establish its 'contrast' range between the black and white values.

It is of particular interest to note that neither black or white have absolute values. An absolute black would be a total absence of light, which would be very unusual in any television situation. A graduated range of gray could be regarded as a range of low-luminance white values, having a peak value at an arbitrary value which is termed 'white'. We shall consider this in more detail later in this chapter.

At low light levels, a contrast range of 10:1 will be fully acceptable, with peak white well defined, and a range of grays down to black. The summer sun at noon can provide a contrast range of 1,000:1. Domestic television receivers can be set up to a range of 100:1, but the effect of ambient lighting is to dilute the contrast range to a possible 30:1. In the case of black and white television, the effect of ambient lighting is just to reduce the range of contrast but, in the

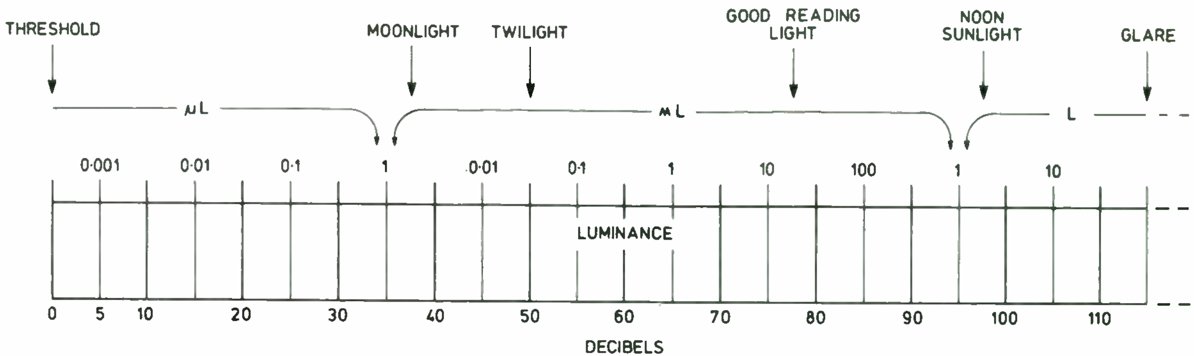

THRESHOLD
MOONLIGHT
TWILIGHT
GOOD READING LIGHT
NOON SUNLIGHT
GLARE
μL
mL
L
0·001
0·01
0·1
1
0·01
0·1
1
10
100
1
10
LUMINANCE
0
5
10
20
30
40
50
60
70
80
90
100
110
DECIBELS

Figure 5.3

case of colour television, the effect will be to reduce contrast and change perception of the displayed colour, to a degree depending on the colour of the ambient illuminant.

5.4 COLOUR

So far, we have only considered luminance. As distinct from luminance, the term for the colour component of a scene is 'chrominance'.

What is colour? A study of colour probably started with Newton in 1704. It was considered at that time that 'white' was a colour, in the same general terms as red, blue or green are colours. Newton demonstrated that the colour components of white light could be displayed by passing a beam of white light through a glass prism. The path of coloured rays will be bent in their passage through the prism, to a degree depending on their wavelength. Red is the least affected, and the short-wavelength blue experiences the most change in direction. The colours emerging from the prism can be displayed as the familiar rainbow-hued spectrum, ranging from red, through orange, yellow and green, to blue.

At this stage, it is appropriate to make the following four points:

(1) Radio waves and light waves are identical in their physical behaviour. They travel with the speed of light, and obey the same physical laws. Their only difference is that light is of very high frequency and consequently very short wavelength although, with the rapid improvement in the development of ever shorter radio waves, the gap between light and radio is closing fast.
(2) The difference between one colour and another is one of frequency or wavelength.
(3) The entire human visible world of colour is contained in about one octave of frequency range. The lowest visible red has a frequency of about 385 mega-mega Hz (or about 385 THz), and the highest visible blue-violet has a frequency of about 790 THz.

Considering light in terms of its frequency is somewhat unwieldy, and it is customary to consider wavelength in milimicrons, or nanometers (nm). The long-wave red is about 780 nm, and the short-wave blue is about 380 nm. (Another unit of wavelength is sometimes used, termed the 'Angstrom', which is 0.1 of a nm.)
(4) Colour cannot be observed without some luminance.

Any object is 'coloured' in either of two ways. It is either a radiator of energy at a wavelength corresponding to a particular colour or 'hue', or it reflects some particular colour when light from another source is impressed upon it. For example, a cathode-ray tube generates radiation from the phosphor screen at a particular colour, the hue depending on the type of phosphor, but a red snooker ball looks red because, under a white light, the blue and green are absorbed and the red is reflected. (It is of interest to note that the temperature of the snooker ball is increased in the process of absorbing the green and blue energy!)

'Reflectance' is an important factor in television, in that it is the usual source of visual energy that excites the camera. The amount of energy reflected from a surface will, obviously, vary with the nature of the surface. A totally absorbant surface will appear black but, as an example, white light impressed on a human forehead, produces a reflected 'colour' in the well-known flesh-tones, the reflectance of energy being about 30%.

Newton followed his original experiment with an obvious further experiment. By adding the colours displayed by the prism through a suitable optical system, he was able to re-constitute the original white light. He also found that if part of the spectrum was removed or reduced in the re-constitution process, the resulting colour would not be white. For example, if blue were to be removed, the resulting addition of red and green only, produced yellow. The most interesting of these experiments was to remove all colours except for the red and blue at each end of the waveband, and a section of green in the middle, these three being termed 'primary colours'. Adding the three primary colours produced white and, by adding the primaries in varying amounts, an infinite range of colours could be produced.

Very significant colour combinations are those produced by the addition of any two of the primary colours. Red plus green produces a range of colour (red–orange–yellow–green) centred on yellow. Red and blue mixtures generate magenta, and a green/blue combination produces a range of cyan hues.

In practice, as we shall see, the addition of colours is not so simple as '2+2=4', mainly due to another property possessed by the eye. Figure 5.4 shows that the sensitivity of the eye varies with different colours, being least sensitive at the red and blue ends of the visible range and most sensitive to the green/yellow region at the centre of the range. We would expect, therefore, that

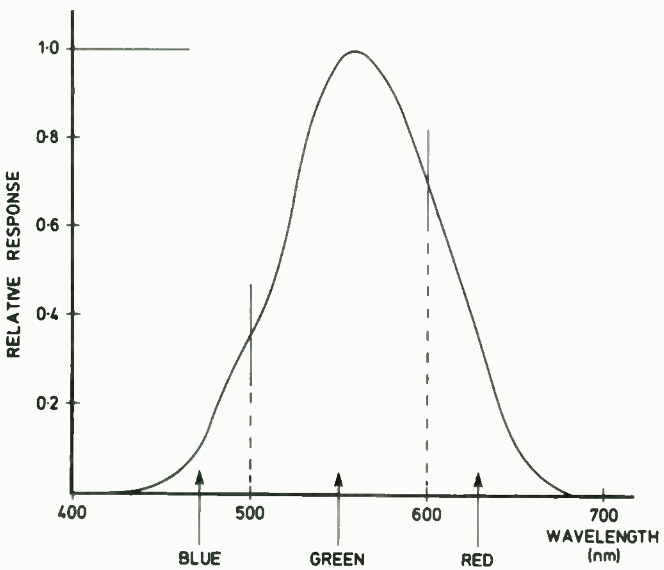


Figure 5.4

equal quantities of red, green and blue would not produce white, and this is indeed the case.

It was not clear at the time of Newton how an addition of two or three colours could produce another colour. It was assumed that the eye had a vague three-colour sensitivity in some manner, that it processed to provide the resultant colour information to the brain. This assumption has been the basis of all colour theory and, fortunately, has been recently verified experimentally. The overall average colour response of the eye, as shown in Fig. 5.4, is the total response of the three separate colour sensitivities possessed by the cones, as shown in Fig. 5.5. The monochromatic sodium yellow is included in this diagram. It shows how the sensation of the colour yellow is experienced, by the simultaneous excitation of the green and red sensors.

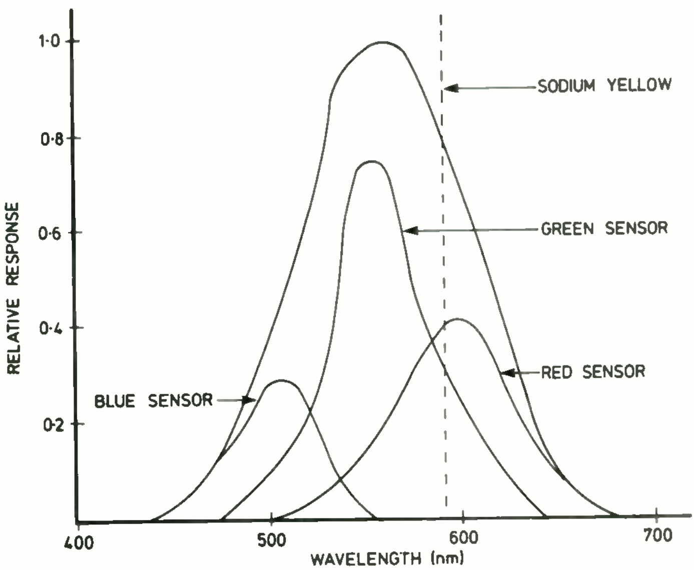


Figure 5.5

To summarise this section, white light can be analysed into appropriate quantities of red, green and blue light and white can be synthesised by the addition of red, green and blue in the correct proportions.

White light can produce any colour by subtracting a colour or colours from white, for example, white can appear yellow by the subtraction of blue. This subtraction process is used in colour photography, but addition processes are used in colour printing and television.

5.5 VISUAL ACUITY

The eye can resolve small variations of luminance over an astonishingly small angle of 0.5 to 1.0 minutes of arc (there are 60 minutes to one degree) and thus the eye is capable of resolving very small detail in the visual scene. In fact, as was discussed in Chapter 1, television system design is based on considerations of fine detail resolution and resulting channel bandwidth is decided by these considerations.

Visual acuity for colour is far less sensitive than for luminance and it depends to some degree on the range of colours involved. Over an orange/cyan range of colours, the resolution angle becomes 1.5 to 3.0 minutes of arc, and over a green/magenta line, detail resolution requires about 4 to 8 minutes of arc. (Note: these two lines are indicated in Fig. 5.10.) We thus arrive at our first engineering consideration for colour. A black and white system is only concerned with luminance, but we do not need the same bandwidth for transmitter and receiver processing of colour information. The bandwidth required for colour information can be reduced to something between 1/3 and 1/8 of the band required for luminance. It is interesting to note that an artist such as Constable knew these facts by instinct. He would paint a tree, for example, in some detail with respect to the trunk, the boughs and some of the smaller limbs, but the individual leaves were not detailed; a large area of green sufficed. In order to add colour to an existing black and white system, it is only necessary to find about 1 MHz of space for the low-detail colour information. The fine detail information is present in the normal wide-band luminance system and there is no purpose in colouring the fine detail if the eye cannot see the colour in detail.

5.6 SATURATION

Few colours in nature are monochromatic, i.e. exist as single frequencies. Most occupy some bandwidth, but there are exceptions such as 'sodium yellow', used in the lighting of traffic highways, and various lasers. Figure 5.6 shows what is meant by a red source, in terms of its energy and wavelength. The solid line represents a peak at the long-wave red and no output at the green or blue. If some white or a cyan mixture is added to the red source, we might obtain the curve shown dotted, which we would describe as 'pink'. The solid curve is termed 'saturated', and the dotted curve is a de-saturated red. Pastel shades of the other primary colours can be similarly represented. It is seen from Fig. 5.6 that an increasing degree of de-saturation would finish as white and this would

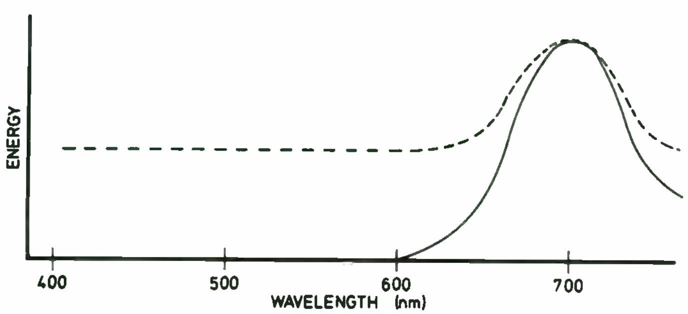


Figure 5.6

apply to any colour. It is very rare to find saturated colours in nature, most are de-saturated to a high degree.

It is now seen that colour is a three-dimensional quantity. In order to see colour we require some luminance, it will then have some colour or 'hue' and it will have some degree of saturation. Representation of the three interrelated quantities can be unwieldy. Two-dimensional representation is very much easier to manipulate and Maxwell showed how this can be achieved in representation of colour.

5.7 THE MAXWELL CUBE

The quantities involved in the representation of colour can be depicted as shown in Fig. 5.7. The two-dimensional green-red plane shows how suitable quantities of a green and red mixture can produce a range of colour, such as yellow. The green-blue plane shows mixtures that produce a cyan range and

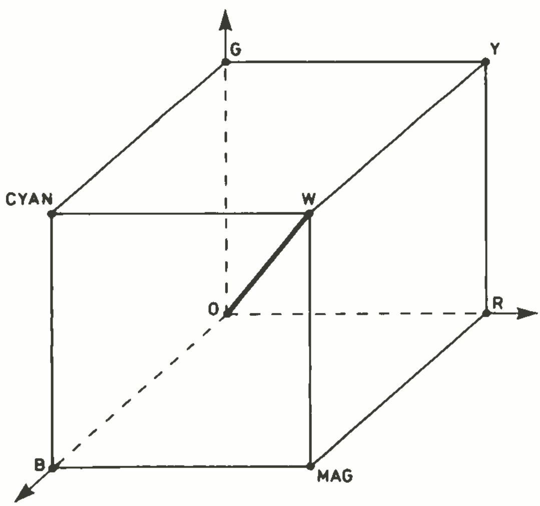


Figure 5.7

the red-blue plane identifies the magenta range. Completion of the cube by linking the cyan, yellow and magenta values produces a corner of the cube that is the mixture of all three primary colours and complimentary primaries, and it can only represent white. The corner marked *O* represents the absence of light, i.e. black, and the diagonal line *O–W* through the cube represents the gray scale from black to white.

Let us now consider the gray scale more closely. No dimensions have been given in Fig. 5.7, but it is clear that if we doubled the dimensions of the cube there would be more luminance, the white point would be brighter, but the proportions of the cube would be the same. Similarly, a reduction in the linear dimensions by some factor would reduce the luminance of the white point, and we can envisage a very small 'mini-cube' around the point O, with the luminance almost extinguished. Thus, the gray scale, represented by the line O–W, is of no significance when considering a colour resulting from a mixture of primaries; it would only indicate how bright the colour would be. (Note: in this argument it has been presumed that adding equal amounts of two primary colours will produce a complementary colour. This is not true, for many reasons, and, in fact, the figure would be matchbox-shaped instead of being a cube.)

5.8 THE COLOUR TRIANGLE

Now consider a slice through the cube of Fig. 5.7, cut through the plane represented by the corners G, R and B, as shown in Fig. 5.8. The revealed surface would appear as the equilateral triangle of Fig. 5.9, with the gray-scale

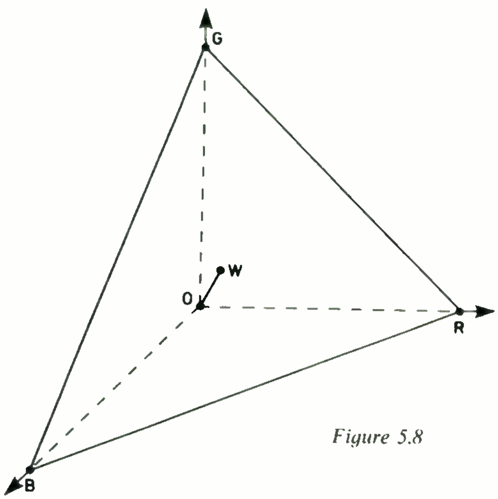


Figure 5.8

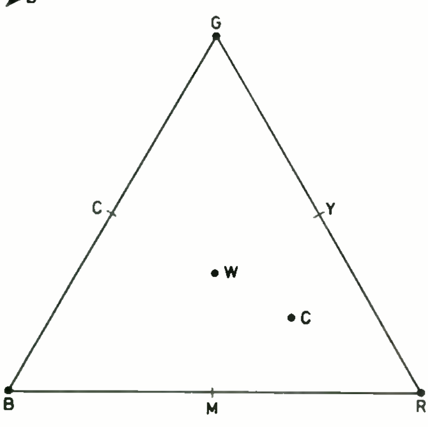


Figure 5.9

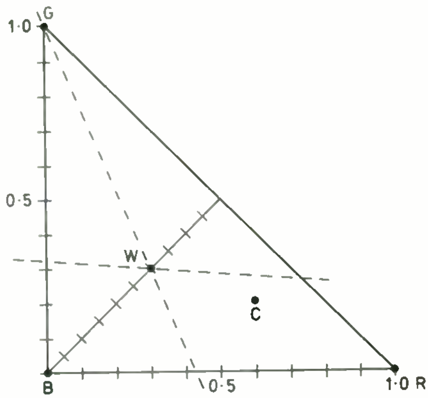


Figure 5.10

line passing through the triangle at the centre point marked *W*. We now have a two-dimensional representation which can identify any colour such as *C*, in terms of the relative proportions of *R*, *G* and *B*, the saturated primary values. We can simplify the triangle still further by converting it to the more convenient right-angled triangle, as shown in Fig. 5.10, where the white point *W* is shown at the centre of the triangle. Any colour, *C* for example, can now be specified by the amount of red (represented by the scale of the *B–R* line), the amount of green (shown by the scale of the *B–G* line), and the amount of *B* (by the scale of the normal to the *G–R* line, drawn through *B*). In other words, any 'hue' can now be represented by the proportions of the primary colour components. A very important fact emerges which results in further simplification of the engineering problems. If a colour has a hue resulting from a mixture from the three primaries of, say, 20% *R* and 50% *G*, it is not necessary to point out that *B* must be 30%. It is only necessary to specify the proportions of any two of the three values. Fortunately, colour mixing is a linear process, and the effective luminance of a mixture is the sum of the individual luminances of each colour component.

It is worth noting from Fig. 5.9 that white can be produced by the correct proportions of *R*, *G* and *B* or, by a two-colour mixture of a primary and its complementary colour, such as cyan and red, blue and yellow or green and magenta.

The concept of the two-dimensional triangle representation of a colour has been adopted internationally by the CIE (Committee International d'Eclarage), and Fig. 5.11 shows a CIE diagram. Fictitious ordinates are shown as *X* and *Y* axes, and the horse-shoe shaped locus of the saturated colours encloses the range of visible colours, which can be specified internationally in *X* and *Y* values. The locus specifies the position of saturated colours, and their wavelengths. The approximate position of the NTSC primaries and white are shown, the resulting triangle enclosing the range of colours that the NTSC system (and its variants, such as PAL and SECAM) can resolve. Any points within the locus will be de-saturated to a degree depending on their proximity to white, showing again that the limit to de-saturation is white.

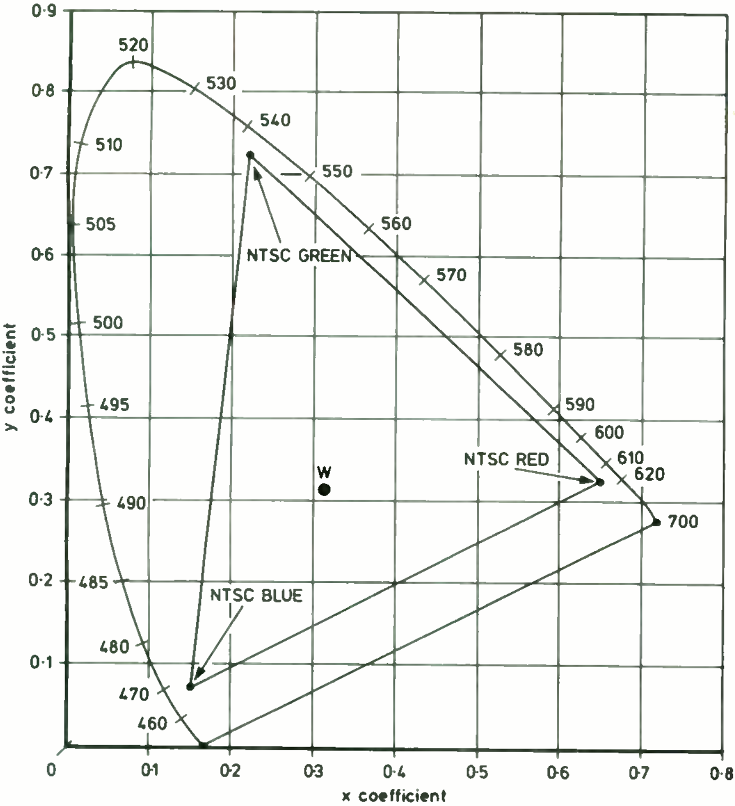


Figure 5.11

5.9 COLOUR TEMPERATURE AND ILLUMINANTS

It will be recalled that a coloured surface can only appear coloured if the source of the light that is being reflected from the surface includes the reflected colour. For example, a light source containing no red, such as a blue, cyan or green source, would show a red surface as black; no light would be reflected.

Bearing in mind that television depends almost entirely on reflected light, an immediate question springs to mind; what illuminant is to be used for television? Figure 5.12 shows a number of light sources and their behaviour over the visible range. The equal-energy source is a useful theoretical concept. The noonday summer sun in a northern sky peaks in the green-yellow, and this peak is an undoubted reason for our very acute eye sensitivity to this part of the spectrum. Fluorescent sources consist of the excitation of various phosphor materials which are luminous at various wavelengths over the band. Sodium is shown as an example of a monochromatic source in Fig. 5.5. The tungsten filament lamp is well known for its large output in the very long-wave infra-red, i.e. heat, and its low output at the short-wave blue.

Excluding monochromatic and fluorescent sources, it is possible to specify

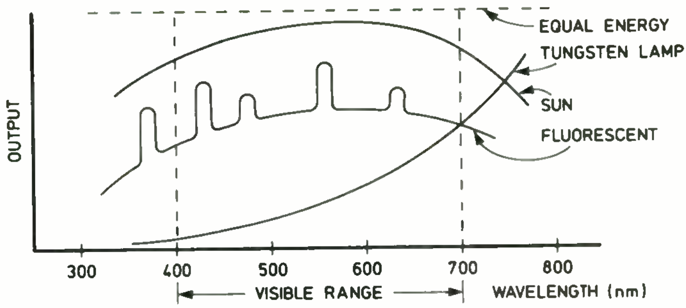


Figure 5.12

sources in what is termed 'colour temperature', using a theoretical concept of a 'black body'.

Consider a body, wholly black, on to which electromagnetic energy such as light or radio waves is being radiated. The energy will be totally absorbed, none will be reflected, and the temperature of the 'black body' will be raised by the energy absorbed. Assuming the 'black body' to be indestructible, consider the reverse process in which the 'black body' temperature is raised. It will now radiate in the manner shown in Fig. 5.13, for various temperatures. Lamps for studios and projectors operate at around 3,000°K, and the radiation from the noon sun approaches the theoretical equal-energy curve. It is clear that using tungsten-filament lamps and suitable light filters, the overall output from a tungsten light source can be changed. For example, the curve representing a temperature of 3,000°K can, by using a high-pass filter that heavily attenuates the long-wave red, provide a curve that is equivalent to a much higher temperature. By this means a number of 'standard illuminants' have been derived over the years. Standard illuminant *B* was about 4,900°K. *C* approximates to the colour temperature of 6,800°K and is near the 'blue white' of a display tube for a black and white television receiver. The standard finally adopted for colour television in the UK is the 6,500°K illuminant *D*.

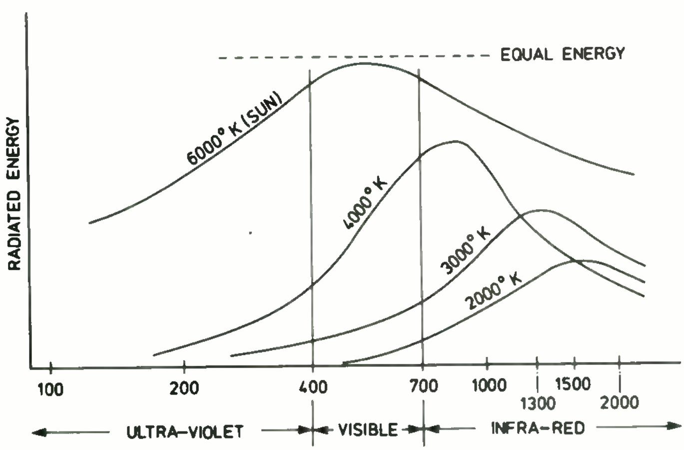


Figure 5.13

Studio illuminants are, conveniently, tungsten-filament lamps and it is not possible to filter each of them to be equivalent to Standard D for a number of reasons, the most important being the considerable loss of light that would be involved and the increase in heat that would be generated in the need to make up the loss. The filtering therefore is carried out at the camera where colour temperature correction filters are embodied in the optical system.

5.10 WHAT IS WHITE?

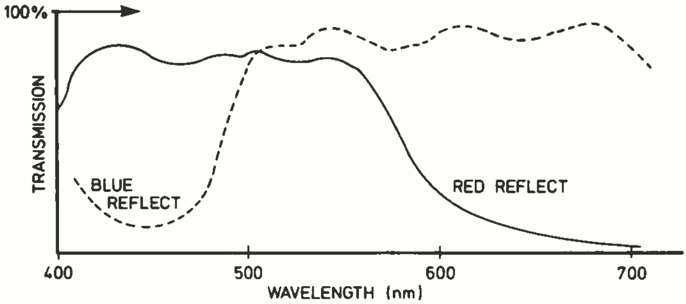
White has been mentioned many times, but what is white? Newton showed that it consists of an infinite range of hues, but examination of Fig. 5.11 shows that there is not a clearly-defined spot with precise $X-Y$ values that can be called 'white'. It is a large area of the CIE diagram, any part of which may provide an acceptable white. The actual primary colours used as the basis for any three-colour system are, clearly, of the utmost importance and those used by CIE are red (650 nm), green (530 nm) and blue (460 nm).

For a colour television system, another factor comes into play; what display system can be used? If a cathode-ray tube type of display is to be used, are phosphors available that will display the chosen three primary colours?

The early NTSC system used a three-gun display tube, and the primary colours chosen for the system were very near those used by CIE, being red (600 nm), green (540 nm) and blue (450 nm). Those primaries in conjunction with the availability of display-tube phosphors, provided a system where 1 lumen of near equal-energy white was provided by $0.3R+0.59G+0.11B$, as explained in Chapter 8. This forms the basis for design of colour television systems and it is of interest to note that, with the passage of time and the development of more efficient phosphors, a recent study has shown that it has not been found necessary to change this relationship.

5.11 DICHROIC FILTERS

In the transmission system, it is necessary for the camera or similar scanning equipment, to analyse the scene presented to it, into its red, green and blue components, for separate processing. This is carried out in the optical system, generally by means of dichroic filters. These filters consist of glass, on the surface of which are thin deposits of various materials. The deposits vary in thickness between $\frac{1}{8}$ and $\frac{1}{2}$ of a wavelength, and the combined effect of thickness and materials used for the coating will produce filters that will reflect or pass high and low frequencies, the characteristics being similar to high- and low-pass electrical filters as shown in Fig. 5.14(a). The filters are included in the optical path as shown in Fig. 5.14(b), where three separate camera tubes can provide the separate outputs for processing. Modern camera systems embody the dichroic assembly as a compact single unit which is part of the camera optical system, as will be seen in Chapter 11.



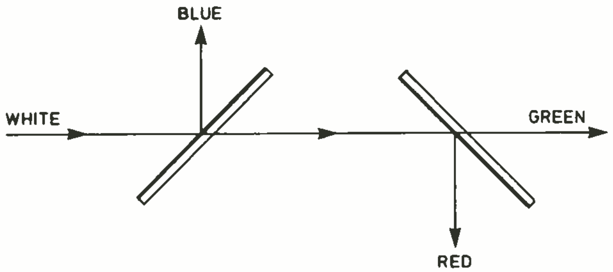


Figure 5.14

5.12 POSSIBLE COLOUR SYSTEMS

The addition of colour information to existing black and white systems has, over the years, produced many interesting possibilities and suggestions for methods by which it can be done. Compatibility has been an essential requirement for a colour system. Existing black and white receivers must be able to receive colour transmissions as black and white pictures; black and white transmissions must be displayed on colour receivers as black and white pictures.

We have seen that colour information requires only about 1 MHz of bandwidth, and a dedicated out-of-band channel has been considered. (At one time, the 2 MHz or so between Channels 1 and 2 in Band I was considered as a possible nationwide source of colour information for networked programmes).

A sequential colour system, in which sequential fields of red, green and blue are transmitted at a high field rate, was demonstrated by Baird in 1928 and was the principle adopted by CBS for a standard colour broadcast system in the USA, with regular broadcasts commencing in 1951. Electromechanical systems were devised, but broadcasts ceased shortly after the start when it was found that the system had bad flicker, colour 'break up' and mechanical failure; the system proved to be impractical for domestic broadcast entertainment.

The work of Merz and Gray showed, in 1934, how a sub-carrier, suitably modulated, can be used to carry colour information 'in band', with a minimum of interference between the black and white luminance information and any additional colour information on the sub-carrier. The NTSC system uses this method, as shown in Chapter 8, the first US broadcast being in 1955. The important later variants of the NTSC system are PAL (Phase Alternating Line), devised by Dr Bruch of Telefunken, and the French SECAM system, in which the sub-carrier is frequency modulated.

The use of satellites for broadcasting has caused much re-consideration of the present systems. Many consider that modern techniques have reached the point where a change in standards can now receive serious attention. The start of a satellite system could provide the opportunity, possibly the only one for many decades to come, for a change in standards towards the goal of full international exchange of programmes. Despite Merz and Gray, there *are* mutual interference effects between luminance and the 'in band' colour sub-carrier. Proposals for satellite systems offer wide-band channels which are generous with channel space and five such channels are already allocated to the UK. The systems will use FM in order to obtain all the possible advantages in signal/noise ratio that FM can provide and a dedicated receiver will process the satellite signals as received. However, there are a large number of very reliable receivers which are receiving several channels of terrestrial broadcasting and, for these to enjoy reception of satellite channels, some form of 'standards conversion' will be required. The use of a converter thus provides for the use of 'out of band' colour information.

Several proposals exist for a higher field rate (to reduce flicker and produce brighter pictures), and to raise the number of lines. With compatibility in mind for conversion to existing receivers, a field rate of 100/s and 1250 lines have been considered. A number of proposals have been made for new standards which will separate luminance and chrominance information. The BBC has proposed an Extended PAL (EPAL) system in which the luminance channel is split into two bands, one below 3.5 MHz and the other from 3.5 to 5.5 MHz which is frequency-shifted to a higher frequency outside the band. The colour sub-carrier thus exists in its present position in an isolated channel. Existing receivers will not be seriously affected by the loss of luminance signals above 3.5 MHz and dedicated receivers can re-introduce the high-frequency luminance signals from 3.5 to 5.5 MHz.

The IBA has proposed a system called MAC (Multiplexed Analogue Component), in which existing standards of interleaved luminance and chrominance information are abandoned. Each line scan time is devoted to a 20 μs colour period, and a 40 μs luminance period. Luminance and colour information is thus received without any possible interference between the two by a dedicated receiver or converter which can process the received information in any manner required. It can, for example, be stored and read out in compatible form for existing receivers.

APPENDIX: LIGHT UNITS IN SI

Units

There are four basic SI units used in studies relating to colour and light (Table A.1).

Table A.1

Quantity	*Unit*	*Abbreviation*
Luminous intensity	Candela	cd
Luminous flux	Lumen	lm
Illumination	Lux (lm/m^2)	lx
Luminance	cd/m^2	—

These four photometric units relate to four main quantities and each other, as shown below.

Luminous intensity

This term replaces earlier terms relating to brightness, such as 'candle power' and 'candles'. The unit is the candela (cd), and is the luminous intensity of a 'black body' at the temperature where molten platinum would just solidify.

Luminous flux

The flow of light from a source. A point source of light, with a luminous intensity measured in candelas, will emit light in all directions. A point source at the centre of a sphere will have a flux that can be specified in terms of a specific section of the sphere.

The unit is the lumen (lm) which is the flux from a point source of 1 candela, within a solid angle of 1 steradian. (Note on the steradian: a sphere of radius R has, on its inside surface, an area of $4\pi R^2$, the solid angle subtended by an area of R^2 is 1 steradian. The radius can be of any value and in any units, but the solid angle will always be the same.) The complete sphere comprises 4π steradians, i.e. 12.57, thus the total luminous flux from a point source of 1 candela is 12.57 lumens.

Illumination

The luminous flux from a point source at the centre of a sphere will illuminate the inner surface of the sphere. The unit of illumination is the lux (lx), which is produced when 1 lumen falls on an area of 1 square metre. Therefore, 1 lux is produced on an area of 1 m^2 at a distance of 1 m from a point source of 1 candela.

An alternative (but not preferred) unit is the phot, which is 1 lumen/cm^2.

Illumination can be defined in terms of any unit of area, and the English foot-candle (or lumen/ft^2) is widely used, being approximately equal to lux.

Table A.2 ILLUMINATION

Unit	*Lux*	*Phot*	*Foot-candle*
1 lux (lm/m^2)	1	10^{-4}	9.29×10^{-2}
1 phot (lm/cm^2)	10^4	1	929
1 foot-candle (lm/ft^2)	10.76	10.76×10^4	1

Table A.3 LUMINANCE

Unit	*cd/m² (nit)*	*cd/ft²*	*Lambert*	*Foot-lambert*
1 cd/m^2 (nit)	1	9.29×10^{-2}	$\pi \times 10^{-4}$	0.292
1 cd/ft^2	10.76	1	3.38×10^{-3}	π
1 lambert (lm/cm^2)	$1/(\pi \times 10^{-4})$	296	1	929
1 foot-lambert	3.43	$1/\pi$	1.076×10^{-3}	1

Luminance

A surface can be self-luminous, or illuminated from another source, and the luminance is expressed in terms of luminous intensity per unit area. There are several terms, the metric unit being candela per square metre (cd/m^2), sometimes termed the nit.

There are other units, and their relationship is shown in Tables A.2 and A.3.

Chapter 6

Video amplifiers, clamping and d.c. restoration

A. Fremont

6.1 SPECIAL REQUIREMENTS

A video amplifier must handle, without distortion, anything from a 50 Hz square wave to a sinusoid of, say, 8 MHz. Depending upon the particular application, we may require the -3 dB points of the response to be at 0.5 Hz and as high as 10 MHz. Within a range of about 50 Hz to 6 MHz we could be looking for a frequency response which is flat within ± 0.1 dB. Furthermore, the phase response must be held to about 1° and, additionally, the phase and gain performance over the amplifiers dynamic range must be rigidly controlled. (Typically we may require 0.2% for linearity, 0.2% for differential gain and 0.5° or better for differential phase.)

The exact specification for the performance of an amplifier will depend very much upon the use to which it is to be put. A device such as a video distribution amplifier (VDA) will have a very tight specification. Because any chain may use a number of such units, it follows that the unit should be virtually transparent. On the other hand, if we consider a unit that is to appear once only in the total video chain, we can relax the specification, especially if the unit is not required to handle colour encoded signals. This is because differential phase and gain are not so important. Indeed, we would find it hard to realise a VDA performance for a high gain camera head amplifier. To achieve the sort of performance required one should look for a straightforward design. It is not common to see very complex stages, more a succession of simple stages.

Having stated the outline requirements for a video amplifier, the implications of these will now be examined more closely. To do this a single transistor stage will be used as an example.

6.2 THE EMITTER FOLLOWER

Before even trying to get a stage with a voltage gain it is worthwhile examining an emitter follower (Fig. 6.1).

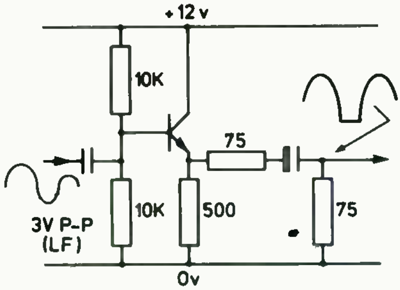


Figure 6.1 Emitter follower d.c. conditions

6.2.1 Emitter follower, d.c. conditions

To supply the load effectively it is required to swing a current of ± 10 mA in it. The d.c. conditions set give a quiescent emitter current of about 10.8 mA and will allow only 1.24 V negative swing. Hence serious clipping of the negative going portions of the signal will occur. To achieve our required linearity performance it is necessary to be well clear of this condition. Although this arrangement would not be used for a high quality 75 Ω output stage, the performance would be greatly improved by increasing the standing current and returning the emitter load to a negative rail.

6.2.2 Emitter follower, HF conditions

The foregoing considerations also apply at high frequency. In the positive direction there should be no problem, but when the drive is negative the output will have a limited slew rate. Considering the example in Fig. 6.2, it can be seen that the point slew rate will be $1/CR \times$ the standing voltage, or 25 V/μs maximum. Remembering that the maximum rate of change for a sinewave is $\omega \hat{V}$ volts, it can be noted that a 5 MHz sinewave of 1.6 V peak to peak has a zero crossing slew rate equal to that of the example. Problems introduced by the limited slew rate will show up a clipping effect on fast negative-going edges. Once again we must not allow ourselves to approach this kind of condition where the emitter load is unable to discharge the stray capacity.

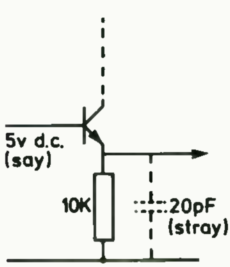


Figure 6.2 Emitter follower HF conditions

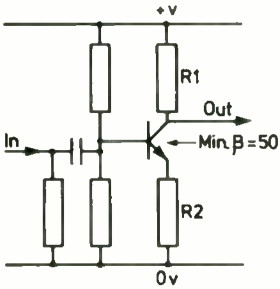


Figure 6.3 Simple amplifier stage

6.3 SIMPLE AMPLIFIER

Consider the design of a simple inverting amplifier with a gain of 2 (see Fig. 6.3). The input has been set at 0.7 V peak to peak. In the case of the emitter follower a sinewave was considered and hence the positive and negative excursions of the signal relative to the average level were equal. This is far from the case with a normal video signal (Fig. 6.4).

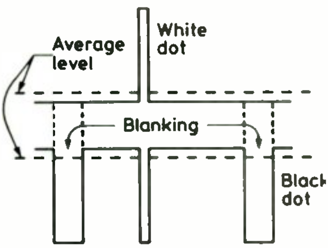


Figure 6.4 Average signal level variations

If we considered a small white dot on a black background the average signal would be near zero with a positive swing of 100%. Conversely, a black dot on a white background would yield an average level approaching 100% (the exact figure depends upon the width of line/field blanking): here the signal would swing nearly 100% down. Hence, for a start, we should be considering a signal of plus and minus the peak to peak input level when we have an a.c. input circuit. So, in the example, with ± 0.7 V at the base and with a gain of 2, we have ± 1.4 V at the collector. Additionally, one would normally allow for overload conditions but we will consider this later.

6.3.1 Standing current

The first thing is to choose the standing current such that the linearity will be good. In general, the less the current swing, the better the linearity. The point gain of the amplifier is given approximately by R_1/R_2, but the emitter impedance adds to R_2 becoming a modifying factor. This impedance is given by: $25/I_e$ where $25 \approx kT/e$ (a physical characteristic of semiconductors, having the dimension of voltage). This impedance adds to R_2 giving

$$\text{Gain} = \frac{R_1}{R_2 + \dfrac{25}{I_e}} \tag{6.1}$$

Further, if $I_e R_2 \gg 25$, we can write

$$\text{Gain} = \frac{R_1}{R_2}\left(1 - \frac{25}{R_2 I_e}\right) \tag{6.2}$$

It is not easy to calculate the required conditions and one must rely on experience to a large extent in order to get within the right area. For reasons to be discussed later, the conditions set are $R_1 = 1\ \text{k}\Omega$, $R_2 = 500\ \Omega$ and a standing current of 6 mA. Then the gain of the simple amplifier is

$$G = 2\left(1 - \frac{50}{I_e} \times 10^{-3}\right)$$

The gain at the quiescent point is

$$G = 2\left(1 - \frac{50}{6} \times 10^{-3}\right)$$

$$= -0.833\% \text{ w.r.t. } 2$$

With the extremes of the range (at 7.4 mA and 4.6 mA), -0.675% (w.r.t. 2) and -1.087% (w.r.t. 2) respectively, the linearity is about $\pm 0.2\%$ at low frequency. This is already approaching the limits for a complete amplifier system. It may therefore be concluded that either higher standing currents must be used or some form or feedback applied. The other alternative of higher emitter loads is not attractive for reasons that will be discussed shortly.

If the feed impedance had not been low, it might also have effected the linearity due to variations of the β with current.

6.3.2 Bias conditions

The bias for the stage can be set with a current significantly higher than the base current. For our example let us set 24 K and 4.7 K and assume an 18 V rail.

6.3.3 LF response (Fig. 6.5)

Having settled the bias conditions we can estimate the input impedance at LF to be 4 kΩ. The input coupling capacitor is the only component likely to effect the LF performance which is typically required to be 2.5% tilt on a 50 Hz square wave. The calculation in estimating the performance is as shown in Fig. 6.6.

Although our example does not include them, there are other places where poor LF response can originate, these are detailed in Fig. 6.7.

It is worth noting that tilt with a long time constant can be removed later whereas tilt with say, a few milliseconds time constant can only be removed with a circuit having a complementary tilt.

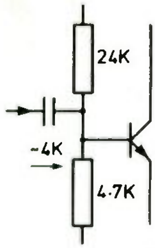


Figure 6.5 Effective circuit for input calculations

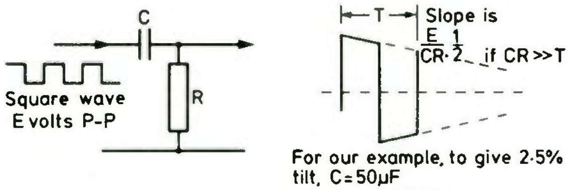


Figure 6.6 Calculation of LF tilt $\left(\% \text{ tilt} = \dfrac{25T}{CR}\right)$

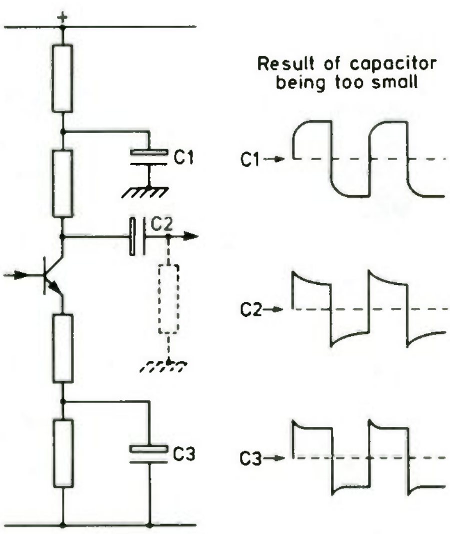


Figure 6.7 Results of poor LF coupling and decoupling

6.3.4 HF response (Fig. 6.8)

Looking now at the HF response we can say that stray capacitance across the load resistor of our original example is going to be the main problem. If we had 20 pF strays across our 1 kΩ collector load we would have an 8 MHz, 3 dB bandwidth. (It is worth remembering that 20 pF at 8 MHz has a reactance of 1 kΩ.) Now, fortunately, 20 pF is rather higher than the sort of strays

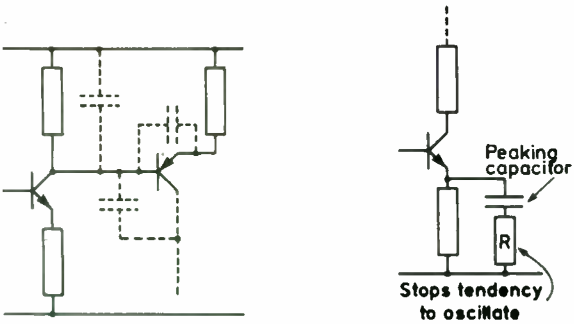


Figure 6.8 Stray capacitance in a simple amplifier; simple HF lift circuit

encountered in practice, 5 to 10 pF is more normal—4 pF would give us about −0.1 dB at about 5 MHz. While this might be acceptable for a normal video amplifier in a camera, the point here is that it is dangerous to make the collector load more than 2 kΩ or so. Incidentally the phase shift at sub-carrier frequency would have been about 7°, not nearly acceptable for a distribution amplifier.

If the loss of HF gain is not too serious there are ways of lifting the response to achieve reasonable results. A simple way is to add some capacitance across the emitter load, but we must take care not to cause oscillation. For this reason a resistor is usually added in series with the capacitor. Another method of improving response would be to add an inductor in series with the collector lead (see Fig. 6.9). Again, this cannot be done to excess without seriously affecting the phase response.

6.4 HIGHER-GAIN AMPLIFIERS

Well if we look back at our efforts so far, we have produced a marginal design with a gain of only 2. In practice we need gains considerably in excess of 2. Say we wanted a gain of 15. If we chose a 1.5 kΩ collector load the emitter resistor would be 100 Ω. To get ±0.25% linearity for 0.7 V we would need about 25 mA standing current and a 50 V rail. Well, this is not very attractive and not what one would normally want to do. So unless we want small gains (or we

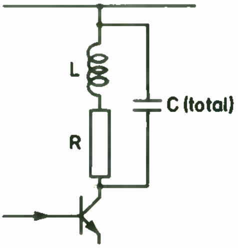


Figure 6.9 HF peaking circuit

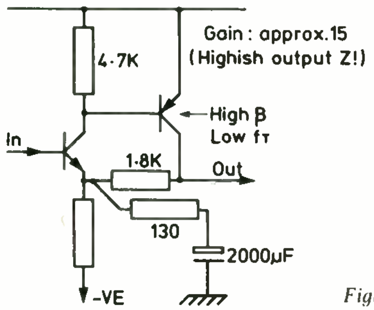


Figure 6.10 Simple feedback voltage amplifier

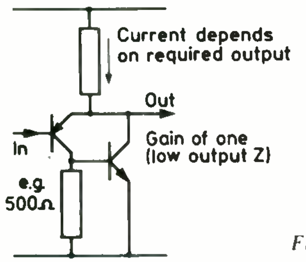


Figure 6.11 Simple low output impedance stage

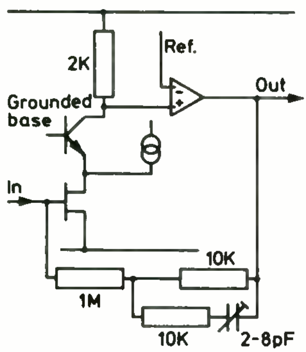


Figure 6.12 Typical camera head amplifier front end

have a small current swing) we should avoid the types of stage we have been discussing—however, the purpose was to demonstrate possible problems.

In practice then one uses feedback amplifiers whenever we require best performance. Some examples are shown in Figs. 6.10 to 6.13.

6.5 DIFFERENTIAL PHASE AND GAIN

Differential gain is a linearity measurement made at HF rather than LF, as we have been considering earlier. Obviously, the differential gain has a tendency to follow the LF linearity but it is also effected by factors such as variations of collector capacitance with voltage and stray capacitance at the emitter.

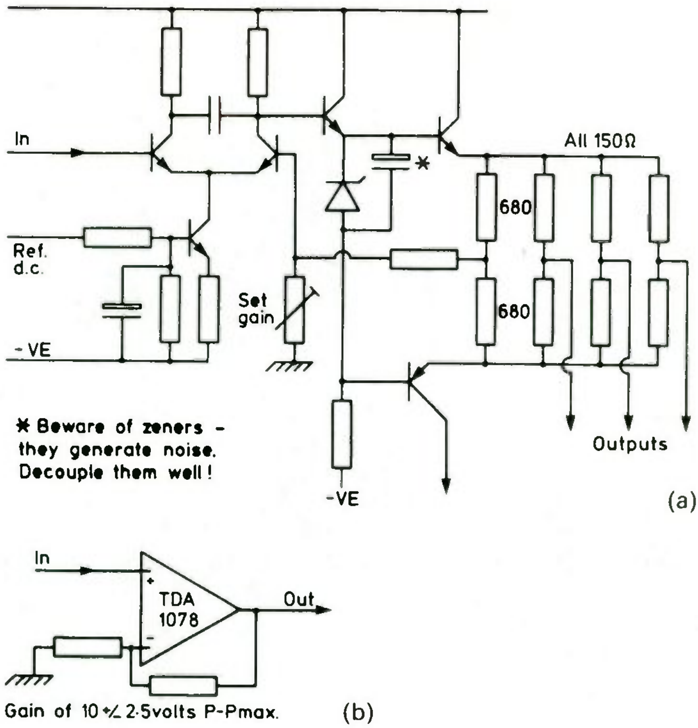


Figure 6.13 (a) Outline of a typical Vision Distribution amplifier circuit
(b) Simplicity with an integrated circuit

Differential phase is the term describing the phase shift accompanying the differential gain. Both these measurements are especially important when designing circuits handling full colour signals.

6.6 GAIN STABILITY

In addition to the foregoing considerations the video engineer must design his circuits to operate within the performance specification, over typically a 20°C range within an overall range of say −15°C to +45°C ambient. This brings in another set of variables which must be considered—change of R_e, β, etc.—and the resistors and capacitors themselves.

6.7 DC STABILITY

In a number of cases good d.c. stability is important. Apart from consideration of resistor stability we must look at other sources of d.c. shift. Obviously a designer should make his circuit largely independent of supply rail fluctuations, but he must be aware of other sources of drift. One common source of drift is the change of base-emitter voltage with temperature. This is typically 2 to 3 mV/°C and must be compensated in critical situations. A useful method is to use emitter followers in *npn/pnp* pairs and if they are on the same header, so much the better. However, this method of compensation cannot be precise and, if better performance is required, a feedback clamp should be used. This technique will be outlined later.

6.8 VIDEO SWEEP TESTING

A video sweep test is usually used to check an amplifier's frequency response and this is commonly done by looking at the peak detected signal. During design it is also worth looking at the actual envelope of the sweep waveform to detect possible problems that may not be obvious by the normal method. Figure 6.14 shows some typical waveforms. Figure 6.14(c) shows the waveform that can result when a gamma corrector (i.e. a circuit with approximately a square-root transfer characteristic) is swept at full level. The effect is due to the fact that most of the harmonics generated by the non-linear function are outside the bandwidth of the circuitry.

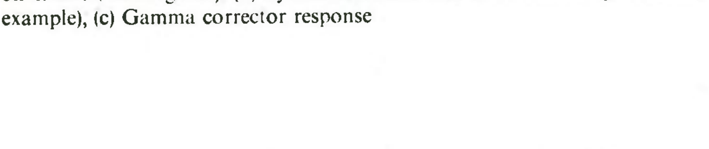


Figure 6.14 Idealised sweep performance: (a) Running out of current at HF (e.g. due to capacitance on emitter, see Fig. 6.2), (b) Symmetrical loss due to collector capacitance (as in Fig. 6.8, for example), (c) Gamma corrector response

6.9 GOOD DESIGN GENERALLY

A number of important considerations have been listed and one final point should be made here. The importance of rail voltage independence has already been made but we must also be careful not to amplify any 'rubbish' on the rails. In Fig. 6.15(a) it will be seen that the input signal is amplified relative to earth and the signal at the first collector is referenced to the positive rail and includes any disturbances on that rail. It is therefore important to do any further amplification with reference to the positive rail. In a similar way the circuit shown in Fig. 6.16(a) can be a problem, with the solutions shown in Fig. 6.16(b).

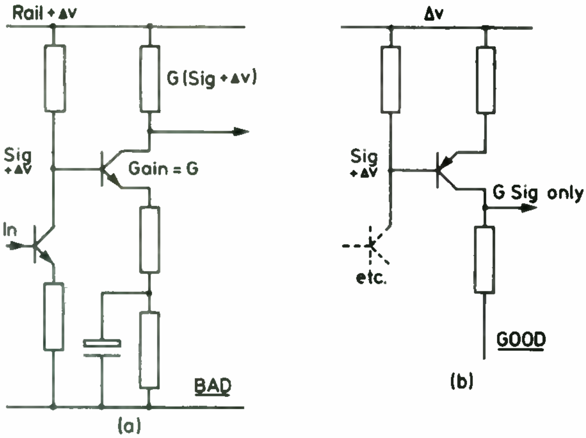


Figure 6.15 Using the correct reference for amplification

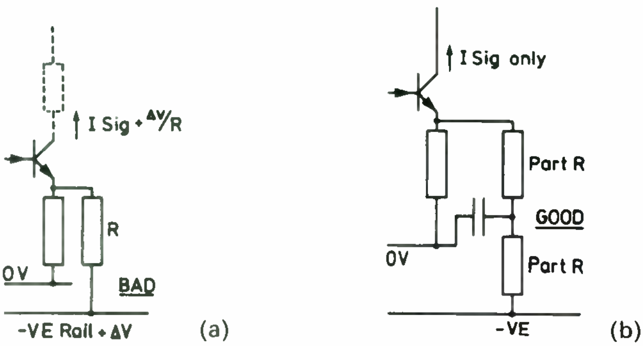


Figure 6.16 Decoupling emitter circuit (a + b)

6.10 IMPORTANCE OF DC COMPONENTS

In the days of black and white television, receiver manufacturers a.c.-coupled the video signal drive to the display tube and the picture black level wandered up and down, depending upon the average level of the signal. With colour, this sort of behaviour is not permissible. To put the problem into perspective, a black level shift of 1/4% or less is visible.

There are two main reasons for 'clamping' or defining the d.c. level of a signal. These are:

(1) to remove unwanted components such as hum or tilt from the signal and
(2) setting the signal to a precise d.c. level prior to a non-linear operation (such as gamma correction, etc.) or modifying the signal in some way such as blanking, clipping, mixing, etc.

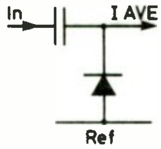


Figure 6.17 D.C. restorer

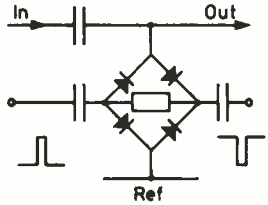


Figure 6.18 Four-diode clamp

6.11 METHODS USED FOR CLAMPING

6.11.1 The d.c. restorer (Fig. 6.17)

First it must be said that this is not a very elegant method, but it is suitable for some purposes. The diode will prevent the signal output from going more than a diode voltage drop negative of the d.c. reference. For the system to work, the average current flow must be in the direction shown otherwise the diode cannot do any work. Not a very precise method of setting a level, but acceptable if the signal is large and there is little or no tilt or hum. Some distortion may occur to the black part of the signal.

6.11.2 Four-diode clamp (Fig. 6.18)

The diodes are normally switched on during the line blanking period of the input video. The drive pulses need to be considerably larger in amplitude than the video signal (say $\times 2$), otherwise the diodes will switch on peak video; the top right diode on positive peaks and the top left on any negative spikes. The 'clamp off/clamp on' ratio is normally fairly large, say 30 (e.g. 62 μs off, 2 μs on). It follows that the current flowing through the capacitor during clamping is on average 30 times that during non-clamping. To achieve good clamping the diode current needs to be considerably higher than the capacitor current during the clamp period. This system gives good high level handling and is not direction sensitive. It was superseded by the next two methods for general use in the early 1960s but it has a place where high precision is required.

6.11.3 Transistor switch (Fig. 6.19(a))

This is a fairly stable and simple method but it can handle no more than 3 or 4 V negative swing due to the limitations of reverse V_{be} and the collector base diode becoming forward biassed. The clamp driving current must not be too large or else the offset voltage will increase, 1 to 2 mA is normal.

6.11.4 FET transistor switch (Fig. 6.19(b))

A FET could be used instead of a bipolar transistor, giving less offset. However, the pulse drive will be high and possibility of pick-up is increased. With careful PWB layout this may be overcome and we have the simplicity of a bipolar transistor clamp and the handling of a four-diode clamp.

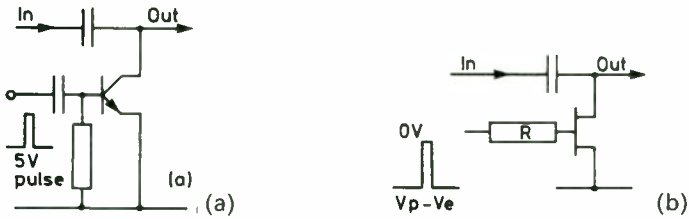


Figure 6.19 (a) Bi-polar transistor clamp. (b) Field Effect transistor clamp

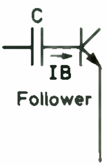


Figure 6.20 Clamp follower

6.12 CLAMP CALCULATIONS

6.12.1 Clamp follower (Fig. 6.20)

During the non-clamping period we have a simple calculation as we used for low frequency tilt. The minimum size of the capacitor is limited by the permissible line tilt. Obviously the higher the β of the follower, the smaller the magnitude of the tilt. In this case the aiming voltage for the tilt calculation is the signal plus the standing d.c. In calculating the input current of the follower we must take care not to use the small signal characteristics, since it is the total base current which must flow in the capacitor. An alternative way to work out tilt is as follows:

$$\text{Tilt} = \frac{I_b \Delta t}{C} \text{ volts} \tag{6.4}$$

For example, if we had 10 μA base current and a 0.1 μF capacitor we get:

$$\text{Tilt} = \frac{10 \times 10^{-6} \times 64 \times 10^{-6}}{0.1 \times 10^{-6}} = 0.64\%$$

on a 1 V signal, which is very poor if we are following it with gamma correction, especially if it is correction for negative film with a gain of 20 or 30.

6.12.2 Source/clamp resistance (Fig. 6.21)

The effectiveness of the clamp in removing tilt or hum from the signal will depend upon the time constant of the circuit with the clamp in the 'on' state.

$$\text{Time constant} = (R_e + R_c) \times C$$

What is happening is that the voltage across the capacitor has changed since the last clamp period or the d.c. level of the input signal has shifted. We want to

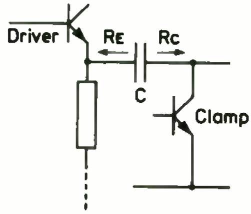


Figure 6.21 Effective impedance at clamp capacitor

reduce this shift so that the black level is set to be equal to the reference. For really hard clamping, the clamp period must be 3 or 4 times the time constant just mentioned. If the total circuit resistance were 10 Ω with the 0.1 μF capacitor it would just fall into this condition for a 3 μs clamp period. If we are clamping this hard we cannot improve upon a hum reduction of about 50:1 on peak values. This is decided as follows. The maximum rate of change of a sinewave is $\omega\hat{V}$ at zero voltage, so the voltage shift between clamp periods is $\omega\hat{V}t_L$, where t_L is the line period. For a 50 Hz hum and a 64 μs line period we get:

$$\begin{aligned}\text{Shift} &= 2\pi f t\hat{V}\\ &= 100\pi \times 64 \times 10^{-6}\hat{V} = 0.02\hat{V}\end{aligned}$$

6.12.3 'Soft' clamping

Now although hard clamping is ideal for removing hum or tilt, it can be a problem if there is any disturbance during the clamp period. For example, noise, spikes, etc. will cause a shift in black level until the next clamp pulse. This effect shows up as black streaks across the picture.

This problem is usually overcome by adding resistance in series with the clamp transistor collector/drain to soften its effect (actually increasing the time constant). This method works well if the follower current is low, otherwise clamping may be inaccurate and may be dependent upon video level and hence insert streaking.

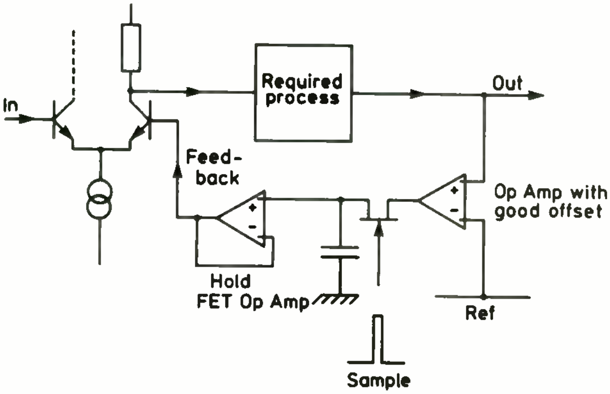


Figure 6.22 Outline of a typical feedback clamp circuit

6.12.4 Feedback clamps (Fig. 6.22)

In this method stability is improved by using a keyed d.c. feedback system. It is particularly useful where the input signal is low level and hence considerably affected by any normal clamp errors. Great care should be taken to ensure that the comparator system accurately measures the conditions that must be stabilised.

Chapter 7

The display system

R. S. Roberts

The display of a television image, first generated by Baird, was simple and uncomplicated. The picture, as transmitted, was generated by means of a rotating disc that had on its surface a spiral of holes for the passage of light used in the scanning process. The scan was produced by the spiral as a vertical pattern of 30 lines, with an aspect ratio of 1:2. For reception, a synchronised disk, having a similar pattern of holes, was interposed between a modulated light source and the viewer, the image being directly observed through a mask having a 1:2 aperture for the display.

7.1 THE CATHODE-RAY TUBE

Development of the cathode-ray tube overcame the problems associated with mechanical scanning, and modern display methods use such tubes.

The principles of CRT operation are well known, the essentials being shown in Fig. 7.1(a) and (b). A cathode emits electrons which pass through a hole in the centre of a disk, termed a 'grid' under the influence of a high positive potential on the final cylindrical anode. The general heater/cathode/grid structure is termed the 'gun'. Various cylindrical sections are used for acceleration and focussing of the electron beam into a small area where the electrons finally strike the screen. The relatively weightless electrons acquire considerable kinetic energy in their transit through the vacuum of the tube and, when they strike the screen, the kinetic energy is transformed mostly into light, with some heat and a very small amount of X-rays. It is of interest to note that the phosphor bright-up and decay times under the impact of electrons resembles the operation of the eye sensors and thus contributes to the illusion of movement.

Many materials are available for the fluorescent screen, and many colours can be produced with the right mixture of phosphor powders. The materials comprising the screen of a black-and-white tube have been developed over the years to produce a high light output, and the colour temperature of a modern

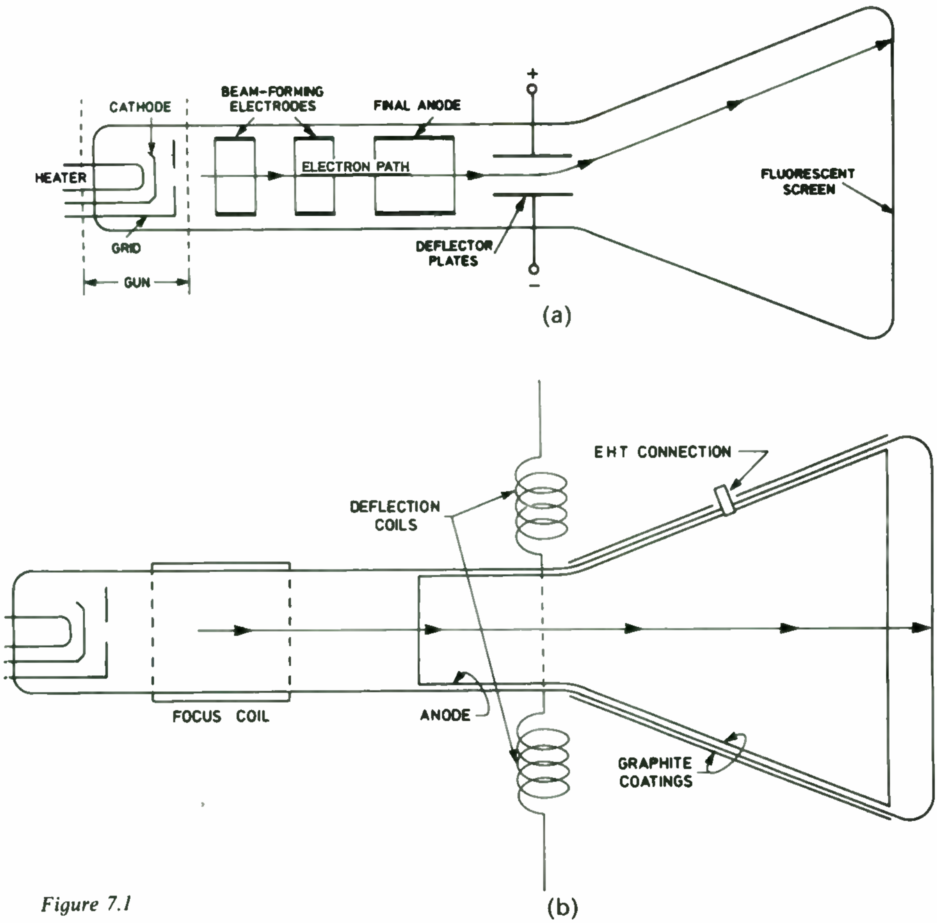


Figure 7.1

white tube is in the region of 10,000°K. The powders are deposited on the inside of the tube face, and then coated with a thin aluminium film which is transparent to electrons. The aluminium film serves two purposes: (1) it reflects light that would otherwise be passed down the tube towards the cathode and thus increases the light output from the screen and (2) the vacuum is never complete and some traces of gas exist in the tube; heavy negative and positive ions are formed, and the aluminium film prevents the ions bombarding the screen and damaging the phosphors.

Another feature of the tube for television receivers is shown in Fig. 7.1(b). The tube flare has an external graphite coating which, when installed in the receiver, is earthed. The glass envelope of the tube, with its internal and external graphite coatings, thus forms a capacitor. The EHT supply is generated by high-voltage pulses at line-scan frequency, and the tube provides the smoothing capacitor.

7.2 BEAM DEFLECTION

The beam can be deflected in two possible ways, one of which is shown in Fig. 7.1(a). A pair of plates are shown with electrodes marked +ve and −ve. As the electron beam passes between the plates, to which d.c. voltages are applied, the beam will be attracted to the positive plate and repelled from the negative plate during its passage through the field between the plates, thus being deflected from its original path. After leaving the deflection field, the beam continues in a straight line path to the screen. In the case of a cathode-ray tube intended for display of a television picture, two pairs of plates would be required, each pair being supplied with a 'saw-tooth' potential variation to produce a deflection of the spot on the screen which is linear with time. One pair would effect the horizontal line sweep across the screen, and the other pair of plates, disposed through 90° with respect to the horizontal deflectors, can affect the much slower vertical sweep.

The second method uses magnetic fields for deflection. Figure 7.1(b) shows another form of tube in which the beam is focussed by means of an external coil which establishes a co-axial field for beam focus. The final anode is, conveniently, a graphite coating inside the tube, to which EHT is applied. Deflection of the beam is carried out with external coils, as shown. The principle is shown in Fig. 7.2, where the external coils establish a uniform field across the beam path. Depending on the direction of current flow through the

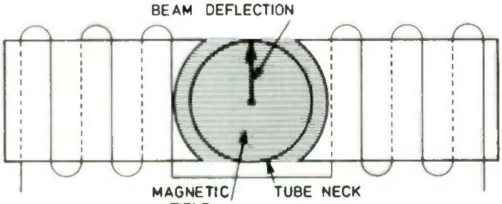


Figure 7.2

external coils, the beam path will be deflected either upwards or downwards as shown. Again, two sets of coils, with their axes at right angles, can be used to trace out a rectangular area on the phosphor screen with the required aspect ratio when the coil currents are of the right peak-to-peak values, and are varying in a 'saw-tooth' manner at line and field rates. Cathode-ray tubes for domestic television receivers generally use magnetic deflection and electric or magnetic focussing of the beam, but camera tubes use combinations of both deflection systems. Chapter 11 deals in detail with the principles of beam formation, focussing and deflection.

7.3 TUBE CHARACTERISTICS

The characteristic curve of any cathode-ray tube has the general shape shown in Fig. 7.3. The following features of this characteristic curve are important and they are common to all cathode-ray tubes:

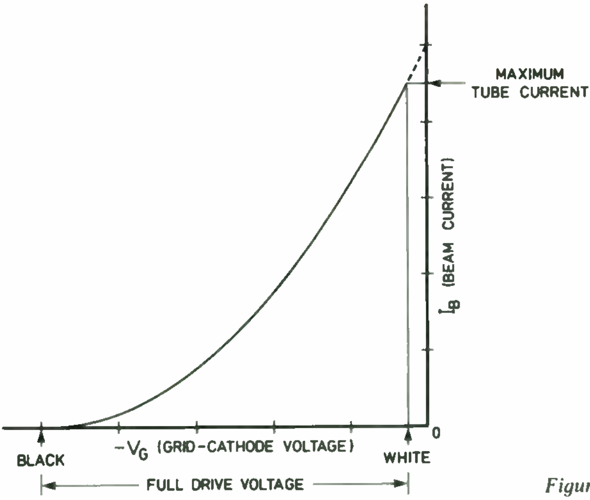


Figure 7.3

(1) The relationship between the control voltage and the beam current is not linear. The light output is proportional to the beam current, and this follows the law

$$I_b = K V_g^{\gamma}$$

Gamma (γ) is about 2.2 for black-and-white tubes and about 2.8 for colour tubes.

For a linear relationship between the luminance in the details of the visual scene being transmitted and the display tube output luminance, some gamma correction is required at the transmitter. The amount of this correction will depend on the type of camera tube used and, if film is being transmitted, the gamma of the positive or negative film being used must be taken into account.

(2) The cut-off value of V_g is quite sharply defined and, of course, is the black value.

(3) Maximum beam current is a value that is decided by the tube manufacturer. It is the value that provides maximum light output and must not be exceeded.

(4) In order to fully modulate the beam current, a peak-to-peak grid-cathode voltage swing is required over the range between the value that will provide maximum luminance, and the cut-off value that extinguishes the beam.

7.4 LARGE-SCREEN DISPLAY

Display systems for domestic use are for direct viewing of the phosphor screen. Optical projection of phosphor images has limitations for large-screen presentation due to the fact that the luminance of the projected picture relies on the basic luminance of the phosphor screen as the light source. When this basic source of flux is spread over a larger area than that of the phosphor screen, the ratio of the projected area of the image to that on the tube screen determines the reduction in available luminance on the screen. A ratio of 2 reduces the luminance to $1/2^2$ or 1/4 of the value at the cathode-ray tube screen. A ratio of 10 reduces the available light to 1/100. In order to generate a high-flux light source, the tube is run at EHT values which are very high and this introduces further limitations. One is that X-ray radiation, normally very low and negligible with domestic tubes, now imposes an upper limit which must not be exceeded, and the other is the need for circuitry to safeguard the tube from destruction in the event of scan failure.

Other systems do not use cathode-ray tubes, but use external light sources which are modulated in intensity by the video signals. The advantages of such systems are that, within reason, the light source can be as bright as may be desired and there are therefore no limitations on screen size. One such system is the Eidophor which uses a separate light source of high intensity. The light beam is directed on to an oil film, from which it is reflected through a grill and projected on to a screen. A scanning electron beam, modulated with a video signal, deforms the oil film surface to varying degrees in accordance with the video signal. The reflection angle of the light beam is varied by the oil surface deformation and, with the aid of the grill which cuts off light as the deflection angle is changed, the projected light is modulated.

7.5 COLOUR DISPLAY TUBES

There are some obvious methods for producing a coloured display from a received colour transmission. The demodulated colour signals can be separated and used to drive three cathode-ray tubes, as shown in Fig. 7.4(a) for projection on to a screen, through suitable optical systems. An alternative system is to use dichroic filters, as shown in Fig. 7.4(b). Each of these systems suffer from some fundamental problems, although the basic system of Fig. 7.4(a) is used for some projection installations. Accurate registration of the three images can be very difficult to engineer. Simultaneous scan drive for the three tubes can present many difficulties.

7.5.1 Laurence tube or Chromatron

Several tubes have been developed for colour television, designed to use only a single deflection system for a three-colour display. Figure 7.5(a) shows the principle of operation for the Laurence tube or Chromatron. A single beam is used, and the screen is striped horizontally with colour phosphors, as shown. A network of 'grid' wires is set up in the path of the electron beam, their function being to deflect the beam vertically by means of a 'switching' signal applied to the grid, so that the beam alights on the correct phosphor at the time when the correct colour signal is modulating the beam, as shown in Fig. 7.5(b). At the centre is shown how the beam strikes the green stripes when the switching signal on the grid is at zero volts. On the right, the switching signal on the grid

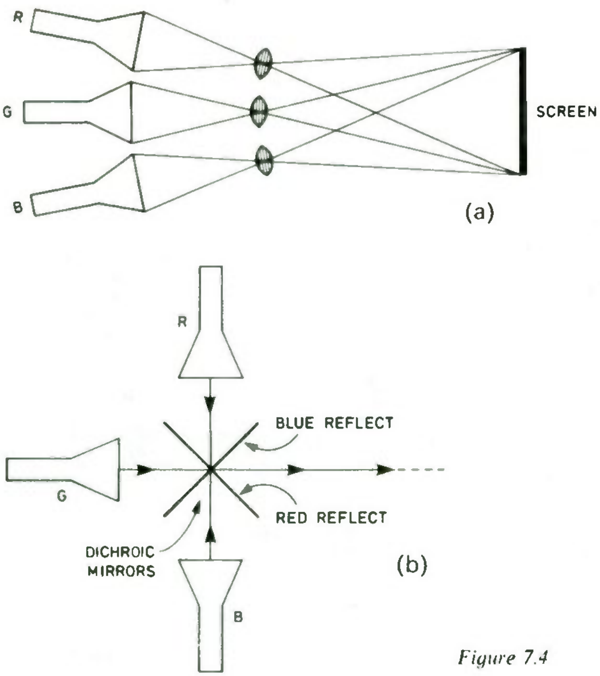


Figure 7.4

wires is such as to deflect the beam to the red phosphors only, whatever position the beam may be in. A reversal of switching polarity will result in the blue phosphors being excited, as shown on the left.

The Laurence tube proved to be difficult to construct on a production-line basis and there are operational difficulties. For example, the switching cycle can be operated at sub-carrier frequency, a complete cycle of colour strips requiring one sub-carrier period. However, the grid structure has a capacitance of about 1600 pF and, tuned to resonate at the sub-carrier frequency, requires about 50–60 W of r.f. power.

7.5.2 Index tube

Another example of a tube similar in principle to the Chromatron is the index tube, under various names such as 'Apple' and 'Zebra'. This type uses stripes of colour phosphor, backed by an aluminium layer which, in turn, is backed by secondary-emitting stripes. The gun generates two beams, one for the display and the other unmodulated, termed the 'guide' or 'pilot' beam, which scan together. The secondary emission is collected and the signals due to the pilot beam are extracted and used to time the switching of the colour circuitry. A variant uses stripes that generate ultra-violet when hit by the pilot beam, collection of the ultra-violet signal by means of a photocell being relatively easy.

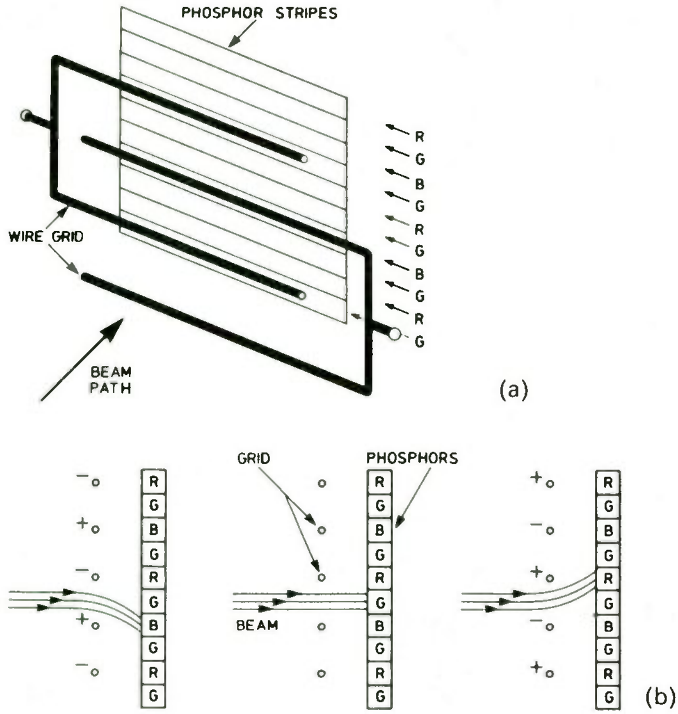


Figure 7.5

7.5.3 The shadow-mask tube

The breakthrough in colour display tubes came with the development and production by RCA of the shadow-mask tube. In this tube there are three guns, separately modulated with red, green and blue signals, situated in a 'delta' formation at the end of the tube, as shown in Fig. 7.6. All three beams are deflected by a common external magnetic deflection coil system. The screen is covered with groups of phosphor dots which, when excited by an electron beam, will display red, green or blue. The dots are grouped in threes, as shown in Fig. 7.6, each group of three being termed a 'triad'. Close to the phosphor screen is a metal screen, the shadow-mask, and this is perforated with holes through which the electron beams can pass. The holes and triads are carefully aligned so that the green beam, for example, whatever its position, will only fire at a green phosphor dot. Similarly the blue gun, firing through the same hole, as shown in Fig. 7.6, will only excite a blue phosphor. The red phosphors are similarly excited by the red gun. Figure 7.6 shows, in a very over-simplified manner, the principle of this operation. Only two holes of the 300,000 or so are shown and, in fact, each focussed beam diameter covers several mask holes and several triads, but the required selective excitation of the colour dots takes place.

The early RCA tubes operated with a final anode voltage of 25 kV and a

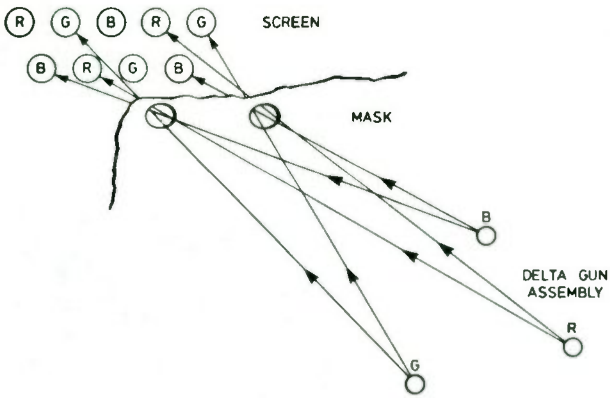


Figure 7.6

mean current of about 1 mA (i.e. 25 W). A large proportion of the heat was collected by the mask, with a risk of permanent damage due to over-heating and distortion.

The geometry of the deflecting operation over the full screen has posed formidable problems. Figure 7.7(a) shows in exaggerated manner how the focal point of a single-gun tube beam requires the screen to be part of a sphere. With two guns the effect is shown in Fig. 7.7(b), but now with an additional error known as 'mis-convergence'. This occurs because the two guns are not occupying the same position. It is possible for the two beams to coincide at the centre of the screen, but they diverge to an increasing extent as they are deflected.

By the use of carefully-disposed static magnetic fields and the injection of dynamic magnetic correction, whereby field deflection signals can be introduced into the line scan and line signals into the field scan, a high degree of correction for convergence errors is obtained.

It will be appreciated that the beam landing on the triad screen is a precision achievement. It is very susceptible to errors that can be produced by external magnetic fields. The steel parts involved in the construction of the receiver can become magnetised and influence the beam path. The tube mounting is surrounded by a coil assembly, through which a 'de-gaussing' current is passed. This current is drawn from the a.c. mains power supply each time the receiver is switched on. Suitable circuitry allows the de-gaussing current to decay to a very small value after a few seconds, the whole process de-magnetising all of the metal parts associated with the tube.

7.5.4 Modern colour tubes

Over the years, considerable improvements in display-tube techniques have taken place. The deflection angle has increased from 90° to 110°, and this has resulted in the overall length of the tube being reduced, thus enabling a more compact receiver design.

Three later developments have resulted in the modern colour display tube. The first was the use of an 'in-line' gun assembly instead of the 'delta'

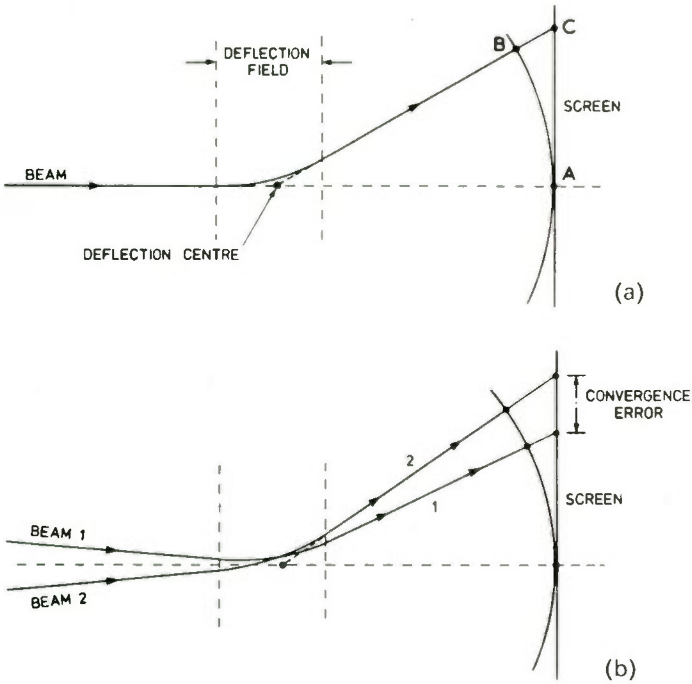


Figure 7.7

construction. Mis-convergence errors are produced in the delta gun system because the guns do not occupy the same position. When the three guns are in a line side by side, mis-convergence is substantially reduced in one direction, resulting in a considerable reduction in the need for convergence error correction. An associated development has abandoned the use of the colour dot triads. The mask now has slots instead of holes and the screen has stripes instead of triads.

The cost of additional components for convergence-error correction is substantial and any development towards their elimination is a very useful improvement. The latest tubes use a 'self-convergence' technique. The deflection coils can produce well-known distortions of a rectangular display, which are known as 'pin-cushion' or 'barrel' distortion, as shown in Fig. 7.8. These distortions can be corrected by the use of external magnetic fields, the inclusion of correcting currents during deflection or by designing the deflection coils so that the deflection field through which the beam passes is not uniform (see Fig. 7.2). Later types of colour tube use the last method, in which the tube and its deflection-coil system become a complete self-converging unit. It will be appreciated that quantity production for domestic television receivers requires very close tolerances in the production of tubes and in the winding of the coil assemblies, for any coil unit to be fitted to any tube.

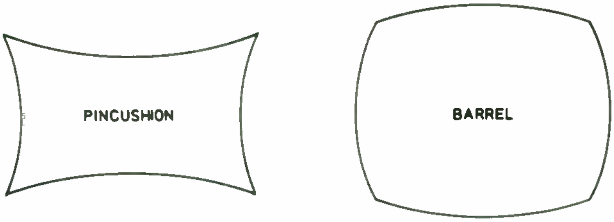


Figure 7.8

7.6 FUTURE DISPLAY SYSTEMS

For very many years the prospect of a 'flat' display unit that could be hung on a wall or stood on a shelf, has been investigated at length. Many attempts have been made to construct and use variants of the cathode-ray type of display, whereby a large area of striped colour phosphors constitute the viewed picture area, the phosphors being excited by a beam (generated from a gun along the bottom or side of the picture area) in a selective manner for total scanning and phosphor excitation over the picture area.

Another interesting possibility lies in the use of small semi-conductor elements such as LEDs, or liquid crystal units. A matrix of such elements can constitute the viewing area, and a digital system can be used to excite any unit in the matrix by means of a suitable code for X–Y positioning. This type of display is being used in some types of test equipment where an a.c. waveform display, for example, can be obtained by selection of a sequence of matrix elements, using an appropriate code for X–Y positioning in the matrix. At the present state of the art the light output from semi-conductor LEDs or LCD elements cannot compare with that obtained from the cathode-ray tube and the 'flat plate' display would seem to be far from being generally available in the near future.

REFERENCES

1. Morrell, Law, Ramberg and Herold, 'Shadow-mask technology', in *Advances in Image Pick-up and Display*, supplement 1, *Colour Television Picture Tubes*, Academic Press (1974)
2. Haantjes and Lubben, 'Errors of magnetic deflection I and II', *Philips Research Reports*, Vol. 12, No. 1 (1957) and Vol. 14, No. 1 (1959)
3. Heijmans, Nieuwendijk and Vink, 'The deflection coils of the 30AX colour TV system', *Philips Technical Review*, Vol. 39, No. 6/7 (1980)

Chapter 8

NTSC and PAL systems

R. D. A. Maurice

PART A: THE NTSC SYSTEM

8.1 INTRODUCTION

This system will be explained using the 525-line/60-field parameters since all countries that use the NTSC system apply it to those standards. When we come to the PAL system it will be based on the parameters adopted for UK System I, which is a 625-line/50-field system.

It is useful at the start, therefore, to have in mind the 525/60 parameters listed in Table 8.1.

8.2 COMPATIBILITY AND THE CONSTANT LUMINANCE PRINCIPLE

A simultaneous three-wire colour television system is shown in Fig. 8.1 in which one wire carries the red signal, another the green and a third the blue. On American standards this could be frequency division multiplexed on a single wire using a theoretical minimum bandwidth of 12.6 MHz. Obviously this is not compatible with the monochrome standard in which the full channel width is only 6 MHz. We have therefore to think of a means of fitting our complete colour television signal into a 4.2 MHz bandwidth if it is to be compatible with the monochrome video bandwidth of the same value.

The first thing therefore is to constitute a black-and-white signal that will transmit correctly the luminance values of the various colours in the scene to be televised. This luminance signal must therefore be panchromatic, that is to say it must be capable of the correct reproduction of the luminances of all the colours in the scene, so that in a monochrom receiver a proper panchromatic image is presented in a suitable colour, usually grey. The composition of the luminance signal will be dealt with presently. Its bandwidth will be just like that of a monochrome video signal, that is zero to 4.2 MHz. This is the first

Table 8.1*

Parameter	*Magnitude or polarity*
Field frequency	59.94 Hz
Line frequency	15,734.264 Hz
Video band	0 to 4.2 MHz
Chrominance sub-carrier frequency	3.579545 MHz
Frequency of inter-carrier sound	4.5 MHz
Maximum deviation of frequency modulated sound carrier	± 25 kHz
Audio pre-emphasis	75 μs
Power ratio of sound carrier to that of vision carrier	-7 to -10 dB
Unattenuated width of attenuated sideband of vision signal	0.75 MHz
Width of radio frequency channel	6 MHz
Polarity of vision modulation	Negative
Ratio of peak picture signal magnitude to that of synchronising pulses	10/4

* CCIR XIIth Plenary Assembly, New Delhi, 1970, Vol. V, Pt 2, Report 308-2.

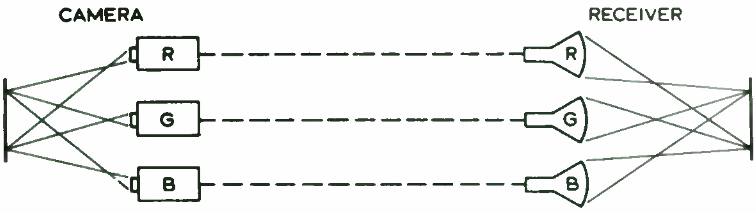


Figure 8.1 Simultaneous 3-wire colour television

part of the composite colour television signal and it is so arranged that monochrome receivers will be able to display monochrome pictures from a colour television transmission. The other property of a beam of light, namely chromaticity or saturation and hue, will require our close attention.

The first fundamental principle of modern colour television is the constant luminance principle and stated very simply it is 'that the luminance signal shall contain no information other than that referring to luminance and that the chromaticity signals shall make no contribution to luminance'. A corollary is that interfering signals appearing in the chromaticity channel do not appear as luminance effects on the receiver. One could invent a constant chromaticity principle that would read 'that chromaticity signals contain no information other than that referring to chromaticity and that the luminance signal shall make no contribution to chromaticity'. The equivalent corollary would be that interfering signals in the luminance channel do not appear as chromatic effects on the receiver screen. This latter constant chromaticity principle is not used.

8.3 LUMINANCE SIGNAL AND COLOUR DIFFERENCE SIGNALS

The luminance signal that will satisfy monochrome receivers and thus ensure a panchromatic image is

$$Y = 0.3R + 0.59G + 0.11B \tag{8.1}$$

in which the coefficients of R, G and B represent the contributions to luminance of each of those particular primary colours that have been so chosen that equal amounts of each are required to produce a given specified white called CIE Illuminant C. Thus, for one lumen of Illuminant C white, 0.33 lumens of the red primary added to 0.59 lumens of the green primary added to 0.11 lumens of the blue primary will be needed. The luminance of any other given colour, for which we might want a proportion (a) of red, (b) of green and (c) of blue, would then become

$$0.3a + 0.59b + 0.11c$$

The particular primaries R, G and B were chosen by the NTSC as reasonably saturated practical colours obtainable from real phosphors available at the time. R, G, B, a, b, c are called tristimulus values.

Colorimetry being based on properties of the eye, which, by and large, have been found to be linear, is a technology which makes great use of linear transformations. Since almost all perceptible colours can be reproduced with the aid of three primaries, rather than a greater number, we now observe that having set up the luminance signal we require two further signals in order to be able to transmit sufficient information for the reconstitution of a full-colour image. The other two signals could be chosen from a vast number of combinations of the original three primary signals, R, G and B. A convenient pair of such combinations are

$$C_{dR} = R - Y \tag{8.2}$$

$$C_{dB} = B - Y \tag{8.3}$$

We could have chosen $R-B$ and $R-G$ or any other similar pair, but we shall see later on that $R-Y$ and $B-Y$ are particularly convenient in that they contain explicitly the luminance signal, Y.

If we substitute for Y its value in terms of R, G and B we get

$$C_{dR} = 0.7R - 0.59G - 0.11B \tag{8.4}$$

$$C_{dB} = -0.3R - 0.59G + 0.89B \tag{8.5}$$

8.4 COLORIMETRY

A little elementary colorimetry is now in order. Figure 8.2 shows a right-angled colour triangle in normalised units. First, note carefully that the three tristimulus values or 'amounts' R, G and B of primary lights have been changed into trichromatic units or chromaticity co-ordinates as follows:

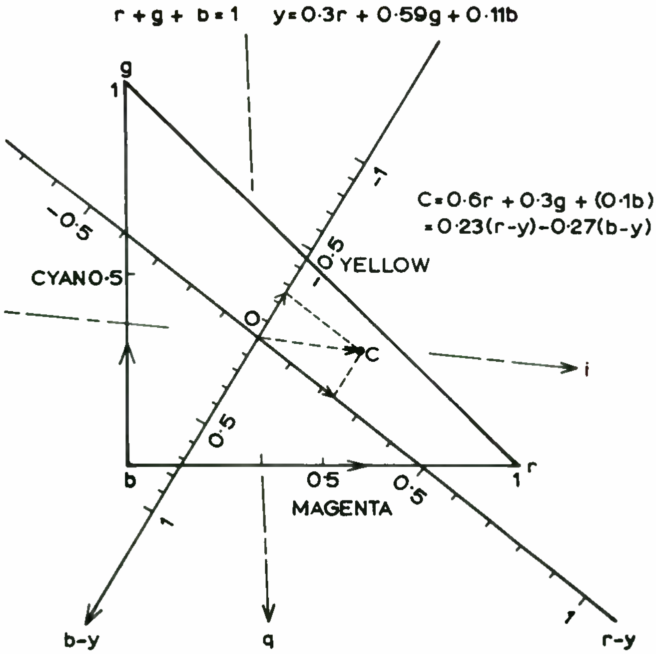


Figure 8.2 Colour triangle

$$\begin{aligned} r &= R/(R+G+B) \\ g &= G/(R+G+B) \\ b &= B/(R+G+B) \\ y &= 0.3r+0.59g+0.11b \end{aligned} \tag{8.6}$$

thus ensuring that

$$r+g+b=1 \tag{8.7}$$

What colours, or rather chromaticities, are represented by

$$C_{dr}=r-y \tag{8.8}$$

$$C_{db}=b-y \tag{8.9}$$

First, observe that two axes of co-ordinates are required, rather than three, since

$$b=1-(r+g) \tag{8.7}$$

and so if we know r and g we can calculate b.

If we can establish the locus of points lying on the line $C_{db}=0$ we shall have found the line C_{dr} and vice versa. The two lines will cross wnen we have, simultaneously

$$C_{dr}=0 \tag{8.10}$$

and

$$C_{db}=0 \tag{8.11}$$

that is, both colour difference signals simultaneously equal zero. This condition will leave only a luminance signal Y or in normalised units, y. From the choice of primaries, we should hope to find that this condition leads to equality between the three chromaticity co-ordinates and thus to the defined white or grey of Illuminant C. Putting the equivalent of Equations 8.4 and 8.5 in normalised units into 8.8 and 8.9 and then equating to zero as indicated in 8.10 and 8.11 and then using 8.7 we get

$$r=0.3\dot{3} \qquad g=0.3\dot{3} \qquad b=0.3\dot{3} \tag{8.12}$$

The r and g values, above, define the point 0 in Fig. 8.2. This is the chromaticity of Illuminant C in terms of the chromaticity co-ordinates r and g that we have adopted.

Now to find one other point on each axis, C_{dr} and C_{db}, so as to be able to draw them in Fig. 8.2. First, for a point on C_{dr}, we let $C_{db}=0$, and secondly we can, for convenience, choose a point that is the junction between the C_{dr} axis and the axis of abscissae, namely the r axis, for which $g=0$. Thus, explicitly, $C_{db}=0$ gives, from Equation 8.5 with normalised values,

$$-0.3r-0.59g+0.89b=0$$

and with $g=0$ and $b=1-(r+0)$ from Equation 8.7 we have

$$-0.3r-0.59\times 0+0.89(1-r)=0$$

whence

$$r=0.75$$

The C_{dr} axis is, therefore, the straight line passing through the points $r=g=0.33$ and $r=0.75, g=0$. We can put a scale on this axis which will enable us to mark it off in units of $C_{dr}=r-y$. These units will differ from the numerics r and g. We know that when $r=0.75$ and $g=0$, that is, at the intersection of C_{dr} with the r axis, we have

$$C_{dr}=0.7\times 0.75-0.59\times 0-0.11(1-0.75)$$

from Equation 8.4, in normalised units, of course. From the above equation we get

$$C_{dr}=0.49$$

Thus, the C_{dr} axis is now completely defined. In similar manner we can define the C_{db} axis and this, too, is shown in Fig. 8.2.

8.5 BAND SAVING

After the constant luminance principle, the second fundamental principle used for modern colour television is that of band saving. This allows the colour-difference signals to be transmitted in a much smaller bandwidth than that allocated to the luminance signal. This results from a careful study of the properties of vision. The acuity of the eye to colour (at constant luminance) in fine detail is about five times lower than it is to luminance contrast in fine detail. That is, to observe a given detail in black and white (or in any two colours whose luminances are very different), the spectator can move five times further away from it than for the same geometric detail in two colours each having the same luminance. There is one condition underlying the above experiment and that is that the common luminance of the two colours shall be significantly lower than that of the brighter colour in the detail involving luminance contrast.

Figure 8.3 shows the Bailey experiment which, following the RCA 'mixed-highs' experiment, demonstrated the relative insensitivity of the eye to the presence of small-area flickering chromaticity fluctuations as compared with similar fluctuations of luminance.

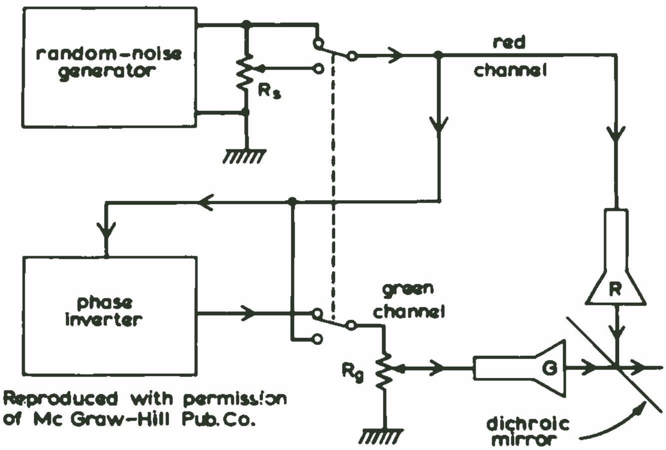


Figure 8.3 The Bailey experiment

In this experiment the observer looks at two display tubes by means of a dichroic or a semi-reflecting mirror. On one tube there is a plain red raster and on the other a plain green one. With the switches in the positions shown in Fig. 8.3, the observer adjusts R_g for minimum random noise on the yellow raster. When the switches are changed to the alternative positions, the observer decreases R_s until the noise is subjectively the same as before. It is found that the polarity-inverted noise (first switch positions) is 6 to 8 dB less obtrusive than the in-phase noise which retians the original yellow colour, but manifests itself as a brightness or luminance change. The polarity-inverted noise when adjusted by means of R_g to maximum visibility becomes a chromaticity change at virtually constant luminance.

Figure 8.4 shows the photographed results of a repeat of the experiment carried out at the BBC's research department. If the constant luminance condition, mentioned at the start, is maintained, then Bailey helps to show that we could send the colour-difference signals in a relatively narrow band and

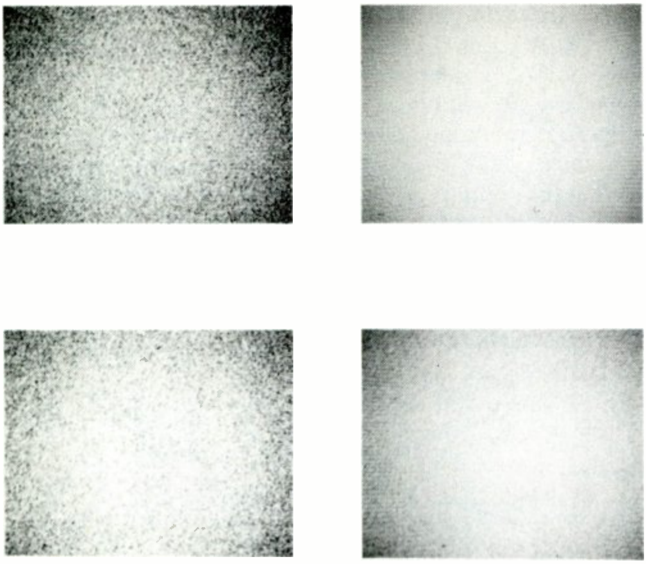


Figure 8.4 Fluctuations: chromaticity versus luminance (top, red/cyan; bottom, green/magenta)

thus transmit no colour information at all about coloured detail that is finer than about one-fifth of the maximum fineness of the luminance detail.

If, for example, we consider the finest luminance detail that the 4.2 MHz video band will pass, to be that of an abrupt transition from black to white lasting for

$$1/(2 \times 4.2\ \text{MHz}) = 120\ \text{ns}$$

or $0.12/52.7^* = 0.0023$ of the active picture width then the most abrupt transition from one colour to another at constant luminance may occupy a time interval of about

$$5 \times 120\ \text{ns} = 0.6\ \mu\text{s}$$

or $5 \times 0.0023 = 0.01$ of the active picture width. The video bandwidth required is therefore about

$$1/(2f_C) = 0.6$$

whence

$$f_C = 0.8\ \text{MHz}$$

In fact, the US figure for colour-difference signal bandwidth is about 0.5 MHz, but there is a little more to say about this.

The orange–cyan transformation, although nowadays a somewhat academic matter, is a most ingenious way of ensuring the maximum attainable information about colour transitions, without occupying extra bandwidth in a way that would cause deleterious effects. Experiments have shown that the

* Duration of active line-scan interval is $(63.5 - 10.8)\ \mu\text{s}$.

colour of a patch of small, but not very small, size can be matched by two rather than three primaries, but the two primaries must not be chosen at random; they are orange and cyan. It is believed that since the orange/cyan axis in a colour triangle lies close to the Planckian locus of chromaticities of a 'black-body' radiator at various temperatures, it is not surprising that the eye† has slightly greater acuity for these colours than for others.

The NTSC system transmits the colour information in such a way that a pair of colour-difference signals like C_{dR} and C_{dB} carry the necessary three-colour information for colour areas of horizontal dimensions that would be passed by a 500 kHz low-pass filter, whilst for horizontal colour detail that requires a wider pass-band, one of the pair of colour-difference signals with a 1.5 MHz pass-band adds the necessary information between 500 kHz and 1.5 MHz. This colour-difference signal lies along the orange–cyan axis. Figure 8.2 shows the two new colour-difference signals, i and q, in terms of our chromaticity co-ordinates chosen for that diagram. In order to ensure that one of the axes lies on the orange–cyan line, all we have to do is to rotate $C_{dr}=r-y$ and $C_{db}=b-y$ by about 33° counterclockwise. We then transmit C_{dQ} in a 500 kHz video band and C_{dI} in a 1.5 MHz video band.

Since C_{dq} and C_{di} are merely the axes C_{db} and C_{dr} rotated by 33°, there is a linear relationship between them as follows:

$$\begin{aligned} C_{di} &= 0.765C_{dr} - 0.55C_{db} \\ C_{dq} &= 0.55C_{dr} + 0.917C_{db} \end{aligned} \tag{8.13}$$

or, in terms of tristimulus (non-normalised) signal values

$$\begin{aligned} I_1 &= 0.765(R-Y) - 0.55(B-Y) \\ Q_1 &= 0.55(R-Y) + 0.917(B-Y) \end{aligned} \tag{8.14}$$

The subscript, 1, has been added to I and Q, because these particular values are not those actually transmitted. There are coefficients put in front of the colour-difference signals $R-Y$ and $B-Y$ before transmission, for reasons that we shall consider in some detail.

If, for the moment, we consider the transmission of Y, $R-Y$ and $B-Y$ rather than Y, I and Q, we see (Fig. 8.5) that if we were to combine the three-wire signals shown, in a frequency-division-multiplex system, the total bandwidth

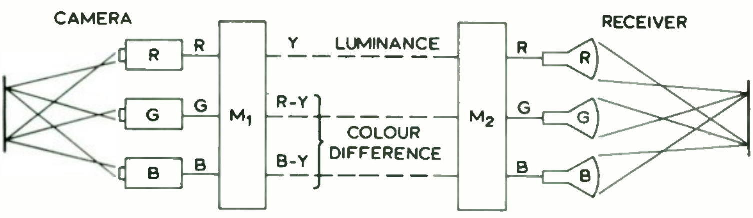


Figure 8.5 Separate Y, R-Y, B-Y system

† 'Naked apes' running about in a sunlit environment with little or no pollution.

required would be $4.2+0.5+0.5=5.2$ MHz, which is a very considerable saving over the figure of 12.6 MHz required for the frequency-division multiplexing of R, G and B. It is still much greater than the 4.2 MHz of video bandwidth that is all that is available, however. We now consider the third fundamental principle of modern colour television.

8.6 BAND SHARING

The trick adopted, still considering Y, $R-Y$ and $B-Y$ rather than Y, I and Q, is to modulate a sub-carrier simultaneously with $R-Y$ and $B-Y$ and let the sub-carrier have a frequency near the high end of the video band; in fact, the sub-carrier frequency is such that the cut-off frequency of the video band at 4.2 MHz is just high enough to allow the upper $B-Y$ and $R-Y$ sidebands of the sub-carrier to be passed without attenuation. But will not the sub-carrier and its sidebands cause interference with the Y or luminance signal? The answer is 'yes', but means are adopted to abate this interference somewhat and the amplitude response of most monochrome receivers at the sub-carrier frequency is far from 100%, whilst, for another reason, the colour sub-carrier is adequately attenuated in the video luminance channel of NTSC and PAL colour receivers. Interference from high-frequency luminance components that appear in the chrominance band is called 'cross-colour'.

How do we modulate a carrier simultaneously with two independent signals, $R-Y$ and $B-Y$? We use double sideband quadrature amplitude modulation of a suppressed carrier. The fact that the carrier is suppressed need not surprise us, because all television systems use this method since they are nominally d.c. transmission systems. Amplitude modulation sound broadcasting does not use a suppressed carrier simply because the audio band does not allow for signals having very low frequencies including d.c.

Figure 8.6 shows a vector or phasor picture of this type of modulation. Diagram (C) shows two suppressed carriers in phase quadrature R' and R'' being modulated by two sidebands, each, S'_-, S'_+ and S''_-, S''_+ respectively. The two modulated carriers (suppressed) form a resultant R whose amplitude is the root of the sum of the squares of R' and R'' and whose phase angle is the arc whose tangent is the ratio of R' to R''.

We may now write as a first attempt at a chrominance signal

$$C_1 = (B-Y)\sin\omega_0 t + (R-Y)\cos\omega_0 t \qquad (8.15)$$

where ω_0 is the angular frequency of the sub-carrier. The words 'colour difference signals' apply, as we have seen, to $B-Y$, $R-Y$, I and Q whereas when a pair of these modulate a sub-carrier we call the result a 'chrominance' signal. The word is highly significant. It is a combination of 'chromaticity' and 'luminance' and it signifies that the colouring information being transmitted is not chromaticity, but the product of chromaticity and luminance, thus, for example,

$$R-Y = Y(R/Y-1)$$

$R/Y-1$ is a chromaticity, although not expressed in our normalised units of r and y.

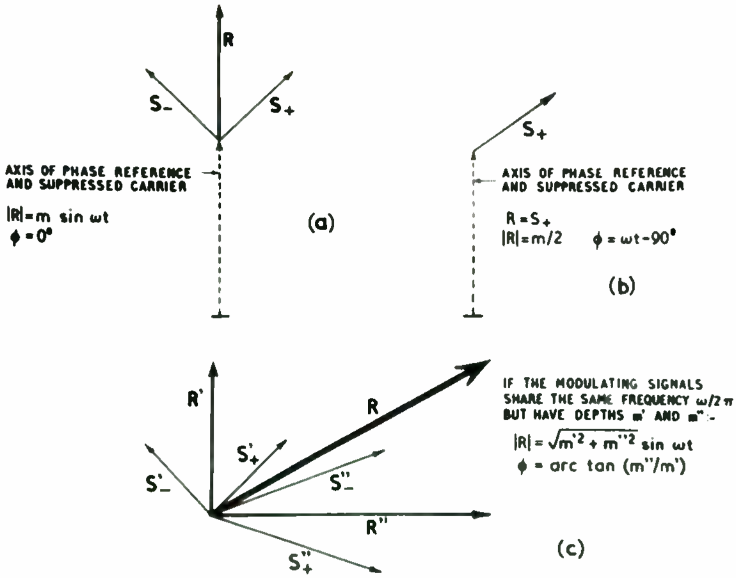


Figure 8.6 Methods of modulation. (a) Double sideband suppressed-carrier modulation. (b) Single-sideband suppressed carrier modulation. (c) Double-sideband suppressed carrier quadrature modulation. (Note: S_- and S_+ are sidebands resulting from the modulation of a carrier wave sin ω_0t by a modulating signal m sin ωt. R is the resultant wave of the entire process and has an amplitude |R|, a frequency $\omega_0/2\pi$ and a phase angle ϕ)

If we transmitted pure chromaticity we could fulfil my 'constant chromaticity' principle, but we should encounter the grave disadvantage that since 'black' need not necessarily mean the absence of 'white', but might be the absence of some colour or other, the chromaticity signal would not necessarily go to zero for black parts of the scene. It would be an absurdity to be transmitting colouring information when 'black' was being televised and the interference on monochrome receivers caused by the sub-carrier would be greater than need be.

We now return to Equation 8.15, which allows us to write

$$|C_1| = \{(B-Y)^2 + (R-Y)^2\}^{1/2} \tag{8.16}$$

$$\phi_1 = \arctan(R-Y)/(B-Y) \tag{8.17}$$

If we choose to write

$$C_1 = \{(B-Y)^2 + (R-Y)^2\}^{1/2} \sin(\omega_0 t + \phi_1) \tag{8.18}$$

the composite picture signal, excluding synchronising pulses is then

$$M_1 = Y + C_1 \tag{8.19}$$

The excursions of the chrominance signal amplitude must now be considered (Table 8.2). $|C_1|$ would be the chrominance signal amplitude obtained directly by modulation of a sine and cosine wave whilst $|C|$ would be

Table 8.2 CHROMINANCE AND LUMINANCE AMPLITUDES

	Blue	*Green*	*Red*	*Cyan*	*Magenta*	*Yellow*
R	0	0	1	0	1	1
G	0	1	0	1	0	1
B	1	0	0	1	1	0
$R-Y$	−0.11	−0.59	0.7	−0.7	0.59	0.11
$B-Y$	0.89	−0.59	−0.3	0.3	0.59	−0.89
$\|C_1\|=\{(R-Y)^2+(B-Y)^2\}^{1/2}$	0.89	0.834	0.76	0.76	0.834	0.89
$0.56(B-Y)$	0.5	−0.33	−0.168*	0.168	0.33	−0.498
$\|C\|=\{(R-Y)^2+[0.56(B-Y)]^2\}^{1/2}$	0.51	0.676	0.72	0.72	0.676	0.51
Y	0.11	0.59	0.3	0.7	0.41	0.89
$Y\pm\|C_1\|$	1.0	1.424	1.06	1.46	1.244	1.78
	−0.78	−0.244	−0.46	−0.06	−0.424	0
$Y\pm\|C\|$	0.62	1.266	1.02	1.42	1.086	1.4
	−0.4	−0.086	−0.42	−0.02	−0.266	0.38
$Y\pm 0.88\|C\|$	0.56	1.184	0.934	1.334	1.005	1.339
	−0.34	−0.005	−0.334	0.0664	−0.184	0.441

the chrominance signal amplitude after multiplication of $B-Y$ by the number 0.56. This factor reduces the discrepancy in the relative magnitudes of the two colour-difference signals $R-Y$ and $B-Y$. For example, when blue is transmitted $B-Y$ is $0.89/0.11=8$ times greater than $R-Y$. Reduction of the $B-Y$ signal by multiplication by 0.56 reduces this discrepancy to $0.5/0.11=4\frac{1}{2}$ times. This process is limited, however, by the lowest value attained by $0.56(B-Y)$; namely, that for red or cyan, which now gives the ratio

$$(R-Y)/0.56(B-Y)=0.7/0.168=4.15$$

The factor 0.56 may also be explained by saying that it reduces the rate of change of chromaticity with respect to chrominance signal phase in the yellows and blues where the eye is most sensitive to colour changes. The penultimate line of the table shows that the composite (luminance magnitude plus chrominance amplitude) colour video signal would exceed peak-white and would go below black level by about 40%. Experiments concerned mainly with compatibility have shown that such excursions are excessive and in fact the composite colour video signal magnitude has been made equal to

$$Y\pm 0.88|C|$$

which restricts the excursions above peak-white and below black level to about 33%. This is a compromise.

So the final expression for the composite colour picture signal expressed in terms of $R-Y$ and $B-Y$ (not I and Q, yet) is

$$M=Y+0.88|C|\sin(\omega_0 t+\phi) \tag{8.20}$$

$$=Y+0.49(B-Y)\sin\omega_0 t+0.88(R-Y)\cos\omega_0 t \tag{8.21}$$

and

$$|C|=\{(R-Y)^2+[0.56(B-Y)]^2\}^{1/2} \tag{8.22}$$

$$\phi=\arctan(R-Y)/[0.56(B-Y)] \tag{8.23}$$

$|C|$ is a measure of saturation and ϕ is a measure of hue.

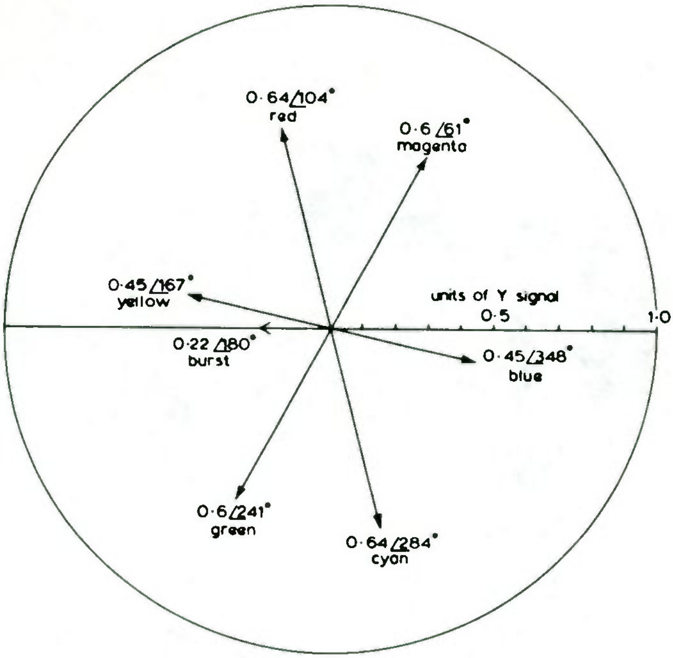


Figure 8.7 Chrominance signal phases

Figure 8.7 shows the vectors or phasors for fully saturated colours at maximum brightness, that is R, G and B each equal to one as in Table 8.2. Figure 8.8 shows the same signal, but as a time function or waveform.

We must now discuss the orange–cyan transformation, again. We have siad we want to transmit I in 1.5 MHz and Q in 0.5 MHz. Assume, for the moment, that $R-Y$ and $B-Y$ (voltages) are restricted to a 0 to 1.5 MHz bandwidth. Then we can effect the orange–cyan transformation, which amounted to a 33° rotation of the $R-Y$ and $B-Y$ axes on Fig. 8.2, by rotating by an angle 33° the two quadrature quantities $0.88(R-Y)$ and $0.49(B-Y)$ in Equation 8.21.

In place of Equation 8.20 we may now write

$$M = Y + \sqrt{I^2+Q^2}\,\sin(\omega_0 t+\psi+33^\circ) \tag{8.24}$$

where

$$\psi = \arctan I/Q \tag{8.25}$$

or

$$M = Y + Q\sin(\omega_0 t+33^\circ) + I\cos(\omega_0 t+33^\circ) \tag{8.26}$$

Identification of Equation 8.26 with Equation 8.21 yields

$$\begin{aligned} I &= -0.27(B-Y)+0.74(R-Y) \\ Q &= 0.41(B-Y)+0.48(R-Y) \end{aligned} \tag{8.27}$$

Equation 8.26 does not show that the signal $I(t)$ is restricted to a 0 to 1.5 MHz bandwidth whilst $Q(t)$ is restricted to a 0 to 0.5 MHz bandwidth. This

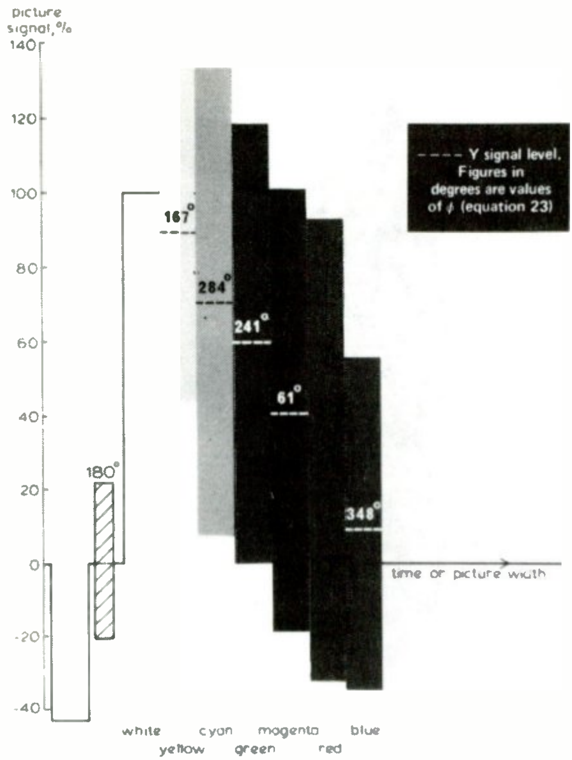


Figure 8.8 Colour bars – waveform

restriction is imposed by the presence of appropriate low-pass filters in the I and Q channels after formation of I and Q signals by matrices which solve, electrically, the equations

$$\begin{aligned} I &= 0.60R - 0.28G - 0.32B \\ Q &= 0.21R - 0.52G + 0.31B \end{aligned} \tag{8.28}$$

which derive from Equation 8.27.

Figure 8.9 shows the spectrum of $M(t)$ and the fact that whilst Q is transmitted double sideband suppressed-carrier AM, I is transmitted as a vestigial sideband suppressed-carrier AM. This causes the phase of the chrominance vector $\sqrt{(I^2+Q^2)}\angle\psi+33°$ to vary independently of hue at frequencies between 0.5 and 1.5 MHz, because only one I sideband will exist at these frequencies and 'balance' will no longer be maintained. Such phase variations, due to phase variation of the I vector, mean that between 0.5 and 1.5 MHz the I and Q vectors are no longer at right angles. We shall see later how the resulting 'cross-talk' is eliminated.

Table 8.3 summarises the formation of the composite colour video signal.

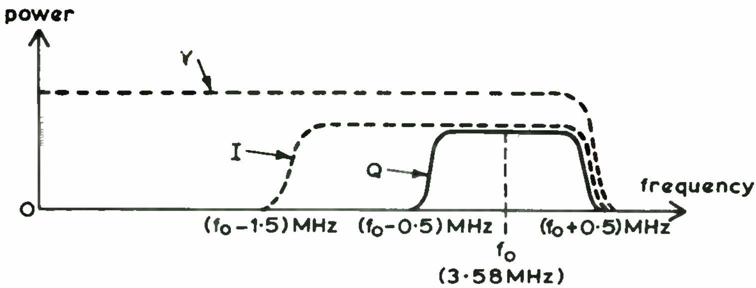


Figure 8.9 Spectrum of the composite video signal

Table 8.3 SPECIFICATION OF NTSC SIGNAL

Composite signal	*Description*
$E'_Y = 0.30E'_R + 0.59E'_G + 0.11E'_B$	Luminance voltage
$E'_Q = 0.21E'_R - 0.53E'_G + 0.31E'_B$	Colour-difference voltage, quadrature (narrow band)
$E'_I = 0.60E'_R - 0.28E'_G - 0.32E'_B$	Colour-difference voltage, in phase (wider band)
$E'_Q = 0.41(E'_B - E'_Y) + 0.48(E'_R - E'_Y)$	
$E'_I = -0.27(E'_B - E'_Y) + 0.74(E'_R - E'_Y)$	
$E_M = E'_Y + 0.88[0.56(E'_B - E'_Y)\sin\omega_0 t + (E'_R - E'_Y)\cos\omega_0 t]$	For narrow-band colour receivers
$E_M = E'_Y + E_Q\sin(\omega_0 t + 33°) + E'_I\cos(\omega_0 t + 33°)$	For wider-band colour receivers
$E_S = 0.23\sin(\omega_0 t + 180°)$	Colour sub-carrier phase-reference signal or 'burst' of at least 8 cycles at a frequency, $\omega_0/2\pi$, of $(3{,}579{,}545 \pm 10)$ Hz; $\omega_0/2\pi = 455f_L/2$ where f_L = line-scan frequency

We write E'_R, etc. because we are referring to electrical voltages. The primes refer to the fact that gamma-corrected signals are used.

To demodulate the colour information from the chrominance signal requires synchronous detection and this requires a signal at the receiver which is locked in frequency and phase (synchronism) to the colour sub-carrier oscillation used at the picture source and on to which are modulated the two signals I and Q. The colour synchronising burst is E_S. The phase of the burst gives it a yellowish hue should it ever appear in a colour receiver during line flyback. Since it lies on the E'_B–E'_Y axis it is immediately available, along with its quadrature, for synchronous detection in a receiver which detects on the E'_B–E'_Y and E'_R–E'_Y axes.

8.7 SPECTRAL INTERLEAVING

Figure 8.10 shows a seven-line television system, each field having $3\frac{1}{2}$ lines. The sub-carrier frequency is an odd multiple of half the line-can frequency (in fact, one times), in order to minimise interference in monochrome receivers due to

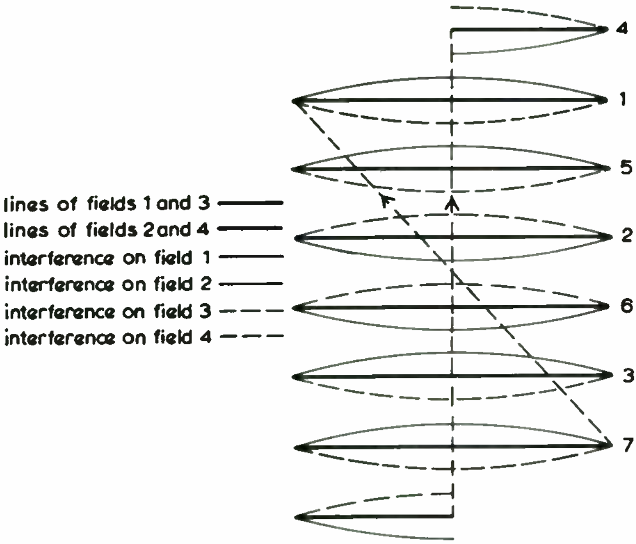


Figure 8.10 A seven-line television system. Spectral interleaving displayed in time. Sub-carrier frequency = one times $f_L/2$

band sharing. This ensures (on a linear display tube) that the bright and dark half-cycles of the interfering sub-carrier are cancelled on alternate fields; thus, four fields or two pictures are required before complete cancellations occurs. The USA is better off than we are due to their 60 Hz field frequency. This is sometimes called 'spectral interleaving'. The precise relation between sub-carrier and line-scan frequencies results in the necessity for asynchronous* working because the sub-carrier frequency cannot be allowed to vary in view of the use of crystal oscillators as sub-carrier regenerators in colour receivers.

8.8 SOUND/CHROMINANCE BEAT

Intermodulation in the r.f. or i.f. stages of receivers can cause a beat-frequency pattern between the sub-carrier and the sound signal to appear at a frequency of 4.5 MHz − 3.58 MHz = 0.92 MHz. To minimise this, the sound carrier, in the absence of (frequency) modulation, is arranged to have an odd-multiple-of-half-line-frequency relationship to the sub-carrier, and so an integer multiple relationship to the vision carrier:

$$f_{\text{sound}} = f_{\text{vision}} + 286 f_L$$

In the USA it was decided that it would be unwise to shift the inter-carrier sound frequency from its 'monochrome' value of 4.5 MHz and so the rather odd values for f_L (line-scan frequency) and f_F (field frequency) came about as

* That is, field frequency not synchronised with that of the electricity supply mains.

follows:

$$f_L = \frac{4500\ \text{kHz}}{286} = 15{,}734.26\ \text{Hz}$$

$$f_F = 2f_L/525 = 59.94\ \text{Hz}$$

and, as indicated in Table 8.3, the colour sub-carrier lies halfway between the 227th and 228th harmonics of the line-scan frequency.

Figure 8.11 shows a block schematic of the transmission equipment.

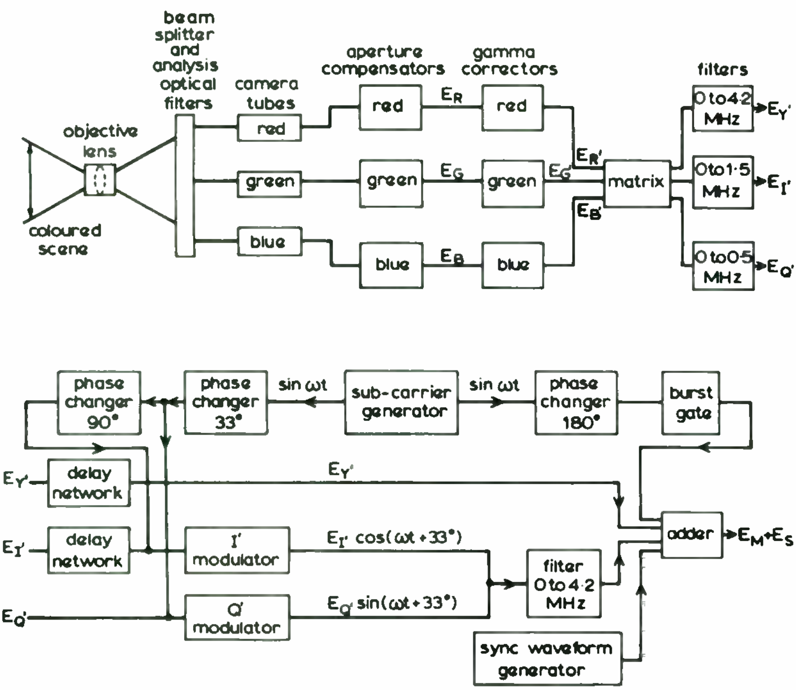


Figure 8.11 Transmission equipment

8.9 SYNCHRONOUS OR PRODUCT DEMODULATION

8.9.1 Narrow-band colour-difference signal reception

We now consider the reception of an NTSC colour television signal. First we shall consider a receiver designed to detect on the $R-Y$ and $B-Y$ axes rather than on the I and Q axes. In this case there will be only one video bandwidth required, namely 0 to 0.5 MHz; and there will be no $(R-Y)$-to-$(B-Y)$ 'cross-talk' to worry about. Consider Equation 8.21

$$M(t) = Y + 0.49(B-Y)\sin\omega_0 t + 0.88(R-Y)\cos\omega_0 t$$

For colour-difference signal frequencies between 0 and 0.5 MHz, the above equation is the same as Equation 8.26, which is written in terms of I and Q, because I and Q (Equations 8.27) were calculated by identifying Equation 8.26 with Equation 8.21. By the way, we should really use the notation E'_B, E'_R and E'_Y instead of B, R and Y, but dropping the Es saves effort and makes the processes we are about to examine somewhat clearer to visualise. The receiver first demodulates the entire $M(t)$ signal, including the colour 'burst' on the back porch of the line blanking interval, from the radio-frequency vision carrier. A 4.2 MHz low-pass filter then allows $M(t)$ to pass into the luminance or Y channel that contains a fairly narrow band-elimination filter centred on the colour sub-carrier frequency. In effect, therefore, the luminance signal loses components much above a frequency of about $3.58-0.5 \approx 3$ MHz. This loss of resolution in colour pictures is acceptable. The reason for the band-elimination or 'notch' filter will be given later; suffice it to say that all traces of the chrominance signal are thus adequately removed. Before reaching the 'notch' filter, some of the signal, $M(t)$, is tapped off and passes through a band-pass filter centred on sub-carrier frequency and so arranged to pass the narrow-band chrominance signal, but to reject all components of the luminance signal save those that happen to lie in the chrominance band ($f_0 \pm 0.5$ MHz). It then reaches the chrominance channel where synchronous demodulation takes place. The signal emerging from the chrominance band-pass filter consists, of course, of the second and third terms of the right-hand side of Equation 8.21, thus

$$0.88C = 0.49(B-Y)\sin\omega_0 t + 0.88(R-Y)\cos\omega_0 t \tag{8.29}$$

or using Equation 8.20, the equivalent of Equation 8.29,

$$0.88C = 0.88|C|\sin(\omega_0 t + \phi) \tag{8.30}$$

Now assume that in the receiver there exists an oscillator locked in frequency and phase, i.e. synchronous with the colour burst $0.23\sin(\omega_0 t + 180°)$. After pole-changing the oscillator output, which can be written $-0.23\sin\omega_0 t$, to $+0.23\sin\omega_0 t$, we feed it into two parallel channels. In one of these there is just an amplification factor of 17.9 which produces a signal $4.1\sin\omega_0 t$ and in the other an amplification of 10 times followed by a 90° phase shift thus producing a signal $2.3\cos\omega_0 t$. The signal 8.29 or 8.30 is then multiplied by $4.1\sin\omega_0 t$ in one channel to obtain $B-Y$ and by $2.3\cos\omega_0 t$ in another channel to obtain $R-Y$. Thus using, first, Equation 8.29, we have

$$\begin{aligned}(0.88C)(4.1\sin\omega_0 t) &= 4.1\sin\omega_0 t[0.49(B-Y)\sin\omega_0 t] \\ &\quad + 4.1\sin\omega_0 t[0.88(R-Y)\cos\omega_0 t] \\ &= 2(B-Y)\sin^2\omega_0 t + 3.6(R-Y)\sin\omega_0 t\cos\omega_0 t \\ &= (B-Y)(1-\cos 2\omega_0 t) + 1.9(R-Y)\sin 2\omega_0 t\end{aligned}$$

If we now pass this signal through a low-pass filter with a cut-off frequency anywhere between 0.5 MHz and a value less than twice sub-carrier frequency, 7.16 MHz, we have

$$(0.88C)(4.1\sin\omega_0 t) = B-Y \tag{8.31}$$

Similarly, the product $(0.88C)2.3\cos\omega_0 t$ will, after low-pass filtration, give us $R-Y$.

If we use Equation 8.30 instead of 8.29 we have, for $B-Y$,

$$(0.88C)(4.1\sin\omega_0 t) = 0.88|C|\sin(\omega_0 t+\phi)(4.1\sin\omega_0 t)$$

Now

$$\sin(\omega_0 t+\phi)\sin\omega_0 t = \tfrac{1}{2}[\cos\phi-\cos(2\omega_0 t+\phi)]$$

Again, low-pass filtration will remove the second term of the right-hand side of the above equation, leaving

$$(0.88C)(4.1\sin\omega_0 t) = 0.88\{(R-Y)^2+[0.56(B-Y)]^2\}^{1/2}(4.1/2)\cos\phi$$

But since

$$\cos\phi = \frac{0.56(B-Y)}{\{(R-Y)^2+[0.56(B-Y)]^2\}^{1/2}}$$

we have, finally

$$(0.88C)(4.1\sin\omega_0 t) = B-Y \tag{8.31}$$

as before and similarly for $R-Y$.

It is, as we see, a relatively simple matter to demodulate the chrominance signal and obtain the two narrow-band colour-difference signals $B-Y$ and $R-Y$. It is usual practice in NTSC colour receivers that demodulate the chrominance signal on the narrow-band axes $R-Y$ and $B-Y$ to use a very simple method of exciting the three electron guns of the shadow-mask display tube. Thus, having obtained $R-Y$ and $B-Y$, a matrix gives $G-Y$ from the following formula

$$G-Y = -0.51(R-Y)-0.19(B-Y) \tag{8.32}$$

Armed with the three narrow-band colour-difference signals, we put $-Y$ on to the three paralleled cathodes and the three colour-difference signals on to the three grids of the electron guns as shown in Fig. 8.12.

8.9.2 Reception based on detection of *I* and *Q* signals*

Consider the chrominance signal expressed in terms of the I and Q colour-difference signals. The second and third terms of the right-hand side of Equation 8.26 give

$$C = I(t)\cos(\omega t+33^\circ)+Q(t)\sin(\omega t+33^\circ) \tag{8.33}$$

Now assume we have a pair of sinusoidal colour-difference signals

$$I(t) = I\sin\beta t$$

$$Q(t) = Q\sin(\delta t+\phi)$$

$$C = I\sin\delta t\cos(\omega t+33^\circ)+Q\sin(\delta t+\phi)\sin(\omega t+33^\circ)$$

* No domestic receivers have taken advantage of reception on the I and Q axes. The reader may omit this section at a first reading.

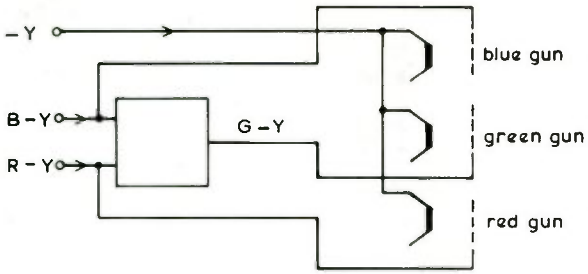


Figure 8.12 Electron gun excitation for narrow-band colour-difference receiver

Let $2\pi \times 0.5 < \beta < 2\pi \times 1.3$ and $2\pi \times 0.5 > \delta$, β and δ in Mrad/s.

$$C = \tfrac{1}{2}I \sin[(\omega+\beta)t + 33^\circ] + \tfrac{1}{2}I \sin[(\omega-\beta)t + 33^\circ] + \tfrac{1}{2}Q \cos[(\omega-\delta)t + 33^\circ - \phi] - \tfrac{1}{2}Q \cos[(\omega+\delta)t + 33^\circ + \phi]$$

Since $(\delta/2\pi) < 0.5$ MHz, $\omega \pm \delta$ is not filtered out, but $\omega+\beta$ is. Thus

$$C = \tfrac{1}{2}I \sin[(\omega-\beta)t + 33^\circ] + \tfrac{1}{2}Q \cos[(\omega-\delta)t + 33^\circ - \phi] - \tfrac{1}{2}Q \cos[(\omega+\delta)t + 33^\circ + \phi]$$

Now in a receiver designed to take advantage of I and Q rather than $R-Y$ and $B-Y$, the local synchronised colour-reference oscillator which is locked in phase to that of the burst also undergoes a 33° phase shift and then, like an $(R-Y)/(B-Y)$ receiver, two separate versions of the oscillation are obtained thus,

$$2 \cos(\omega t + 33^\circ) \quad \text{for the } I \text{ channel}$$

and

$$2 \sin(\omega t + 33^\circ) \quad \text{for the } Q \text{ channel}$$

The Q channel demodulator will produce $2 \sin(\omega t + 33^\circ)C$ whilst the I channel demodulator will produce $2 \cos(\omega t + 33^\circ)C$. *Thus for the Q channel*

$$I \sin(\omega t + 33^\circ) \sin[(\omega-\beta)t + 33^\circ] + Q \sin(\omega t + 33^\circ) \cos[(\omega-\delta)t + 33^\circ - \phi] - Q \sin(\omega t + 33^\circ) \cos[(\omega+\delta)t + 33^\circ + \phi]$$

$$= \tfrac{1}{2}I \cos \beta t - \tfrac{1}{2}I \cos[(2\omega-\beta)t + 66^\circ] + \tfrac{1}{2}Q \sin[(2\omega-\delta)t + 66^\circ - \phi] + \tfrac{1}{2}Q \sin(\delta t + \phi)$$

The first term will be eliminated by the 0.5 MHz Q channel low-pass filter because $\beta/2\pi > 0.5$ MHz. The second and third terms will likewise be eliminated by the same filter since $2\omega/2\pi$ is 7.16 MHz. Finally, we are left with

$$\tfrac{1}{2}Q \sin(\delta t + \phi)$$

which, if the Q channel gain is 2, becomes

$$Q \sin(\delta t + \phi) = Q(t)$$

For the I channel, we have

$$I\cos(\omega t+33°)\sin[(\omega-\beta)t+33°]+Q\cos(\omega t+33°)\cos[(\omega-\delta)t+33°-\phi]$$
$$-Q\cos(\omega t+33°)\cos[(\omega+\delta)t+33°+\phi]$$
$$=\tfrac{1}{2}I\sin[(2\omega-\beta)t+66°]+\tfrac{1}{2}I\sin(-\beta t)+\tfrac{1}{2}Q\cos[(2\omega-\delta)t+66°-\phi]$$
$$+\tfrac{1}{2}Q\cos(\delta t+\phi)-\tfrac{1}{2}Q\cos[(2\omega+\delta)t+66°+\phi]$$
$$-\tfrac{1}{2}Q\cos[-(\delta t+\phi)]$$

All the terms in $2\omega t$ will be filtered out by the 1.3 MHz I channel low-pass filter. The two terms in $\delta t+\phi$ are equal in magnitude and opposite in sign and therefore add to equal zero.

Finally, we are left with

$$-\tfrac{1}{2}I\sin\beta t$$

which, if the I channel gain is -2 (a polarity reversal), becomes

$$I\sin\beta t=I(t)$$

Thus, we see that the 'cross-talk' term in the Q channel due to the I signal (the term $\frac{1}{2}I\cos\beta t$ is filtered out), whilst the two 'cross-talk' terms $\pm\frac{1}{2}Q\cos(\delta t+\phi)$ due to the Q signal in the I channel cancel one another.

A receiver designed to demodulate the chrominance on the I and Q axes does not use the method shown in Fig. 8.12, but rather uses a matrix to calculate R, G and B from I, Q and Y. The three primary signals may then be fed to the grids of the three electron guns. There are many variations that may be used, but so far as is known, there are no receivers that use the I and Q axes, because it is slightly more expensive to use them than to detect on $R-Y$ and $B-Y$ axes. Figure 8.13 shows a block schematic of a possible I/Q receiver. Omit, in Fig. 8.13, the 33° phase changer and apart from filters, not shown, we have an $(R-Y)/(B-Y)$ narrow chrominance band receiver.

The purpose of the 'notch' filter in the luminance channel of the colour receiver is to attenuate the chrominance signal, because if it appeared as an interfering carrier (with its sidebands) on the colour display screen, it would not only cause a dot pattern or 'knitting' pattern as it does in the monochrome receiver, but also it would desaturate the colours by an amount that would depend upon its amplitude at any given moment. Such desaturation occurs because the non-linearity of the beam current as a function of grid-cathode excitation voltage of the electron guns causes partial rectification, making the positive half-cycles of sub-carrier greater in amplitude than the negative ones and thus adding some 'd.c.' to each gun. Since equal amounts of d.c. would be added to R, G and B, the resulting added colour would be white (or grey). If the amount of d.c. is Δ, the added white lumens would be $\Delta=0.3\Delta+0.59\Delta+0.11\Delta$. Figure 8.14 shows the effect of a notch filter in a colour receiver and Fig. 8.15 the effect in a monochrome receiver.

The delay circuits shown in the Y and I channels of the receiver in Fig. 8.13 are required to slow down the wider bandwidth I and Y signals so that they coincide, in time, with the narrow-band Q signal. Remember that the delay

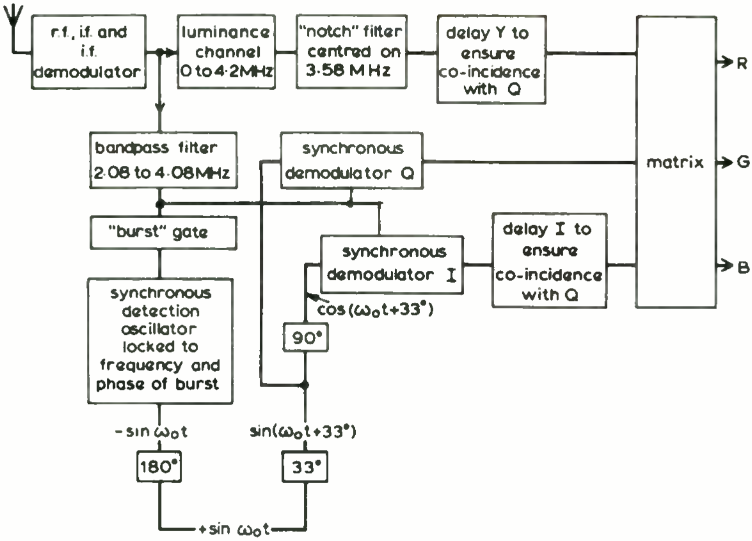


Figure 8.13 An I/Q receiver

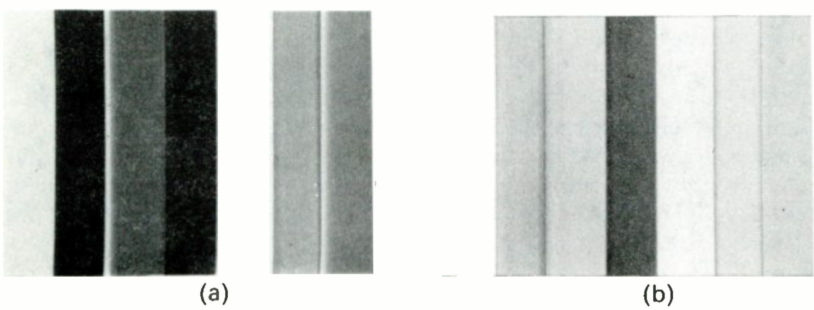


Figure 8.14 Colour bars (a) with notch, (b) without notch

through a low-pass filter is inversely proportional to the cut-off frequency. Consider an ideal low-pass filter having a uniform pass-band from 0 to ω_c and a linear phase characteristic from 0 to ω_c at which it reaches π*. The phase angle can be written

$$\phi = \pi\omega/\omega_c$$

Now the characteristic delay τ of a signal can be expressed as

$$\cos \omega(t-\tau) + j \sin \omega(t-\tau) = e^{-j\omega\tau} e^{j\omega t}$$

where the phase is $-\omega\tau$. If we equate the two expressions for phase we get

$$\tau = -\pi/\omega_c = -1/2f_c$$

* Practical low-pass filters try to approach π, but do not quite reach it.

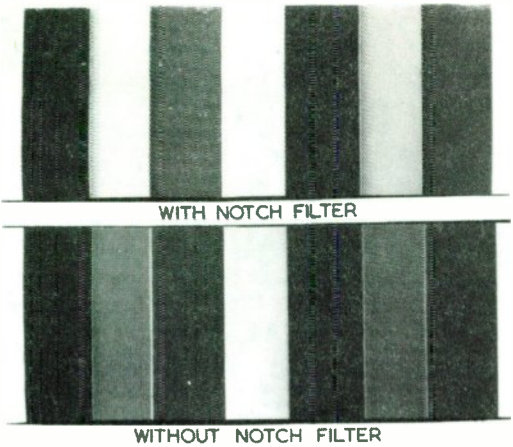


Figure 8.15 Colour bars on black and white receiver

Thus, the Y signal, in its 4.2 MHz band, will suffer a delay of about $1/8.4 = 0.12$ μs; the I signal, in its 1.5 MHz band, will suffer a delay of about $1/3 = 0.33$ μs; whilst the Q signal, in its 0.5 MHz band, will suffer a delay of $1/1 = 1$ μs In order to achieve coincidence, therefore, Y must be delayed by $1 - 0.12 = 0.88$ μs, whilst I must be delayed by $1 - 0.33 = 0.67$ μs.

8.10 THE COLOUR-REFERENCE OSCILLATOR

Figure 8.16 shows a very simple automatic phase-control circuit for synchronising the local sub-carrier oscillator to the 'burst'. This circuit relies on equality of frequency between the stable oscillator and the incoming 'burst'. The phase

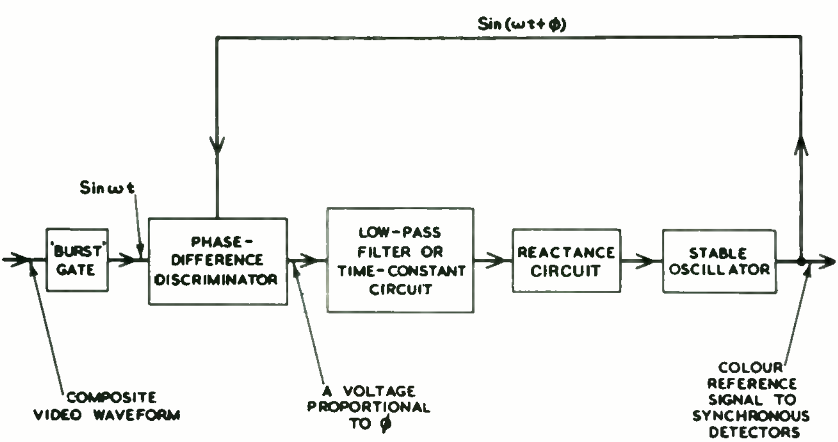


Figure 8.16 Simple phase-control circuit

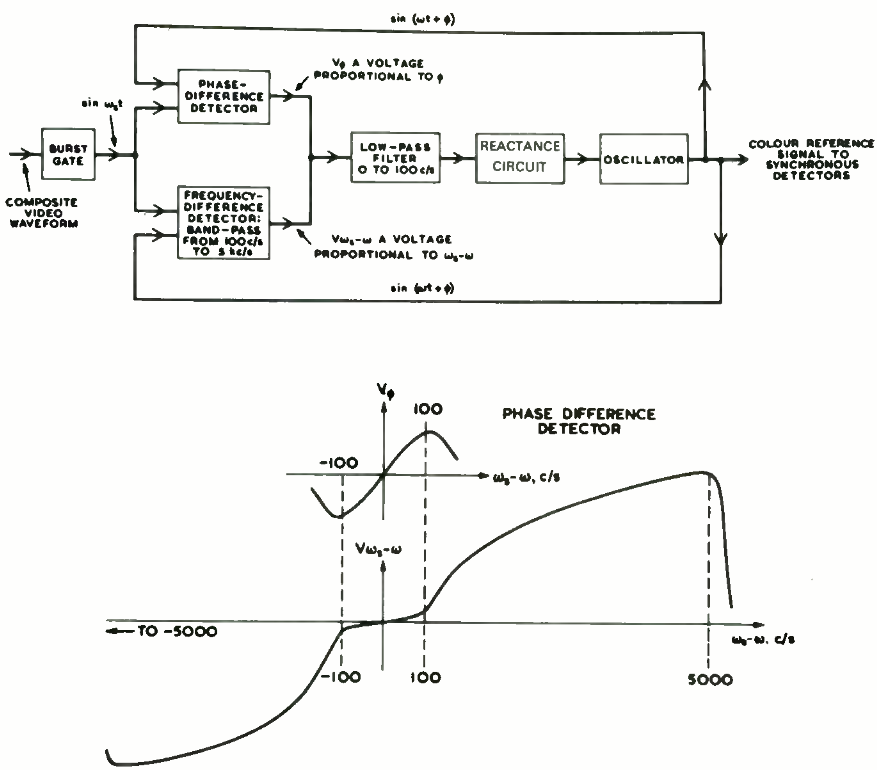


Figure 8.17 A frequency and phase-control circuit

discriminator is then able to ensure phase equality between local oscillation and 'burst'. The low-pass filter should have a low cut-off frequency (about 100 Hz) to combat random noise, but a high enough one to ensure that the local oscillator, on switching on, has a frequency within the filter's pass-band.

Figure 8.17 shows the diagram of a circuit that allows more latitude for the frequency stability of the local oscillator. It is a two-mode circuit whose operation is self-evident.

8.11 FAILURE OF CONSTANT LUMINANCE

All the foregoing work has been based on the assumption that the television signal is entirely linear, although mention has been made of the non-linear beam current-to-excitation-voltage characteristic of the electron guns in the display tube. This non-linearity requires that the primary signals R, G and B be gamma corrected just as is the case for the brightness signal for monochrome television.

The luminance signal, for colour television, is

$$E'_Y = 0.3E'_R + 0.59E'_G + 0.11E'_B \qquad (8.34)$$

where

$$E'_R = E_R^{1/\gamma} \qquad E'_G = E_G^{1/\gamma} \qquad E'_B = E_B^{1/\gamma}$$

and where γ is the exponent of the power law expressing beam current as a function of excitation voltage of the colour display tube. But a true luminance signal, gamma corrected to compensate for the display-tube characteristic would be

$$E_Y^{1/\gamma} = (0.3E_R + 0.59E_G + 0.11E_B)^{1/\gamma} \qquad (8.35)$$

Thus, the ratio of NTSC transmitted luminance to true luminance as they would be displayed on, for example, a monochrome receiver is

$$r = (E'_Y/E_Y^{1/\gamma})^{\gamma} = E_Y'^{\gamma}/E_Y \qquad (8.36)$$

If we let $Y = 2.5$—the real value is about 2.8 and the official US specification value is 2.2—we obtain the values shown in Table 8.4.

Table 8.4

*Colour**	*r*
Grey	1
Red	0.16
Green	0.45
Blue	0.037
Yellow	0.84
Magenta	0.29
Cyan	0.58

* The colours in the table are assumed to have maximum luminance and maximum saturation: E_R, E_G and E_B each equal to one.

First, we note that when $E_R = E_G = E_B$, the NTSC luminance is correct. Secondly, to take the worst case, we find that for saturated blue, the NTSC luminance signal carries only 4% of the required amount. Nevertheless, the luminance of a saturated blue patch as displayed on the colour receiver screen is correct, because the blue gun will be excited by the voltage E'_B or $E_B^{1/\gamma}$ and the other two guns will be left un-excited. Where does the 96% of blue luminance come from? It must come from the chrominance signal which will, therefore, have a greater amplitude than it would have had if the constant-luminance principle had applied fully. Furthermore, the vertical edges of fully saturated blue patches will be unsharp, since 96% of the luminance required by them will have come through the relatively narrow-band chrominance channel—at best only 1.5 MHz wide if an I/Q receiver were used and only 0.5 MHz wide in the case of the popular $(R-Y)/(B-Y)$ receiver.

On a monochrome receiver, the result of this partial failure of constant luminance working is to make red make-up on lips look too dark and blue skies look thundery. In the unlikely event of the monochrome receiver having a uniform response over the whole video band, including the upper region covering the chrominance signal frequencies, the chrominance signal will appear on the screen as a 'knitting' pattern. Due to non-linearity of the electron gun's transfer characteristic the resulting rectification will add a good deal of the lost luminance, thereby restoring red lips and blue skies to a value of brightness equal to about one-half of the correct value. The lack of horizontal resolution in patches of saturated colour is, however, permanently lost and is not, of course, restituted by rectification of the low-resolution chrominance signal.

8.12 COMPENSATION AT TRANSMITTER FOR NON-UNIFORMITY OF GROUP-DELAY AT RECEIVER

Figure 8.18 shows typical receiver i.f. amplitude and group-delay characteristics. If the overall group-delay of the transmission system is not the same at sub-carrier frequency as it is at low video frequencies the coloured patches of a scene as displayed on a colour receiver will not coincide with their corresponding luminance patches. On a monochrome receiver the patches of dot-pattern interference and the consequent increased luminance due to rectification will, again, not coincide with the corresponding luminance patches. Group-delay compensation at the transmitter can largely overcome

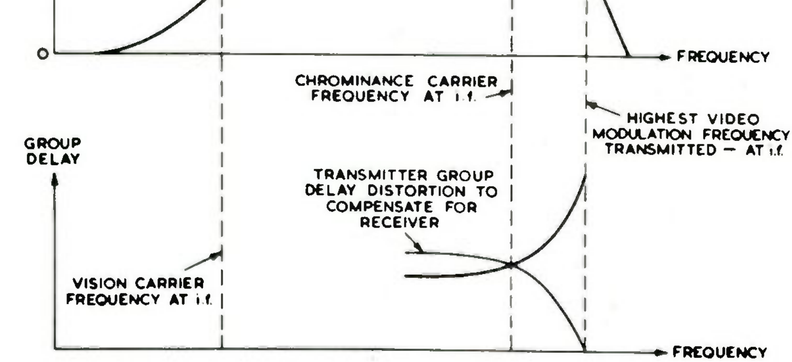


Figure 8.18 Phase and amplitude characteristics of typical receiver IF circuits

these errors. Group-delay errors across the chrominance signal bandwidth can cause chromatic errors, but these are, as a rule, negligible.

PART B: THE PAL SYSTEM

8.13 SIMPLE PAL

The NTSC chrominance signal was given by Equation 8.29

$$0.49(B-Y)\sin\omega_0 t + 0.88(R-Y)\cos\omega_0 t$$

Synchronous detection of this signal enables us to extract the two colour-difference signals $R-Y$ and $B-Y$. Let us revert to vector or phasor notation.

Figure 8.19 shows a chrominance signal represented by a colour phasor C_ϕ having a hue represented by the phase angle ϕ and a saturation represented by the length S. Now whereas in the NTSC system the phasor C_ϕ is sent

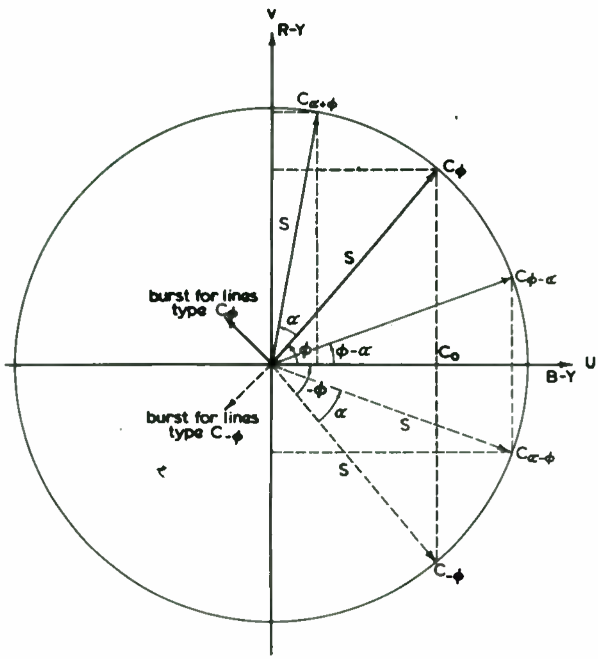


Figure 8.19 Chrominance signal phasors: PAL

unchanged, scanning-line by scanning-line, as long as the colour C remains unchanged (say a whole scene consisting of a plain coloured background with no variations of either chromaticity or luminance) in the PAL system the phasor that is transmitted is alternately, line by line, C_ϕ, $C_{-\phi}$, C_ϕ, $C_{-\phi}$, etc. The only difference between C_ϕ and $C_{-\phi}$ is that $C_{-\phi}$ is the mirror image of C_ϕ with respect to the $B-Y$ axis. Another way of saying this is that $C_{-\phi}$ is the complex conjugate of C_ϕ or that the $B-Y$ projections of C_ϕ and $C_{-\phi}$ are equal but the $R-Y$ projections are each the negative of the other.

The equation for the chrominance signal for the PAL system is

$$C = E'_U \sin \omega_0 t \pm E'_V \cos \omega_0 t \tag{8.37}$$

where

$$\left.\begin{aligned} E'_U &= 0.493(E'_B - E'_Y) \\ E'_V &= 0.877(E'_R - E'_Y) \end{aligned}\right\} \tag{8.38}$$

It will be noticed that Equation 8.37 seems to result from a geometrical figure derived from Fig. 8.19 by an exchange of U $(B-Y)$ and V $(R-Y)$ axes. It means that in Fig. 8.19 cos $\omega_0 t$ is an ordinate and sin $\omega_0 t$ is an abscissa, contrary to normal academic practice; but this will not worry us in what follows. This is an unfortunate circumstance due to the desire to keep the similarity between NTSC and PAL to the fore. Correctness is maintained, however, if the burst, whose phase with respect to the $+U$ axis is alternately $\pm 135°$, is written as

$$E_S = 0.23 \sin(\omega_0 t \pm 135°) \tag{8.39}$$

Now, why use such a transmission method? First, let us make the assumption that at a normal viewing distance from the receiver screen the eye cannot distinguish any chromaticity difference between two adjacent lines of a television field. Thus, if one line were, say, red and the following one were, say, green, the two colours would merge and, in the eye, would add to give the impression of a horizontal yellow band. If an NTSC receiver were used to decode a PAL signal, all colours would appear to lie along the $B-Y$ axis, because the addition in the eye of the colour C_ϕ to the colour $C_{-\phi}$ would give the colour C_0. A PAL receiver must therefore perform one extra task over and above those required of an NTSC receiver. It must, either during or after synchronous demodulation, reverse the phase or polarity of the $R-Y$ component of the phasor $C_{-\phi}$ during every other scanning-line—in other words, the receiver must undo the mirror imaging which the coder did around the $B-Y$ axis. Now the eye will receive the impression of the colour C on every line and all will be well. We see that if the PAL decoder is equipped with a polarity reversing switch operating at line-scan frequency the colour reproduction is correct. But, why do this? Well, let us now assume that during its passage from studio to receiver, the chrominance signal has suffered an unwanted phase change (due maybe to equipment which imposes a change of phase on the signal passing through it that is a function of the magnitude of the luminance signal upon which the chrominance signal 'rides'. Such a phase error is called differential phase or level-dependent phase.) Consider such a phase error, α. In the NTSC system this error will cause the reproduced hue to correspond to the angle $\phi + \alpha$ instead of to the desired angle ϕ. In the PAL system, on the other hand, the chrominance phasor being alternately, line by line, $C_{\alpha+\phi}$ and $C_{\alpha-\phi}$, we see that after polarity reversal of the $R-Y$ component of $C_{\alpha-\phi}$ we convert it to $C_{\phi-\alpha}$ and the vector addition performed by the eye:

$$\vec{C}_{\alpha+\phi} + \vec{C}_{-\alpha+\phi} \tag{8.40}$$

yields a phasor whose angle with respect to the $B-Y$ axis is ϕ, independent of α. We have thus eliminated the hue error due to the phase error α. This is the basic improvement of PAL over NTSC. There is left a saturation error that can be calculated by performing the vector addition 8.40.

Let the $B-Y$ or U component of the resultant of $C_{-(\alpha-\phi)}=C_{\phi-\alpha}$ and $C_{\alpha+\phi}$ be X and the $R-Y$ or V component of the same resultant be Y; then

$$X = S\cos(\phi-\alpha) + S\cos(\phi+\alpha) = 2S\cos\alpha\cos\phi \tag{8.41}$$

$$Y = S\sin(\phi-\alpha) + S\sin(\phi+\alpha) = 2S\cos\alpha\sin\phi \tag{8.42}$$

The factors $\cos\alpha$ in each component is an amplitude or saturation multiplication factor which is near to unit if α is small. The factor 2 is a constant independent of α and represents the fact that the eye is integrating the colour in two scanning lines covering twice the area that would be covered by only one scanning line.

The switch in the PAL receiver that reverses either the phase of the colour reference oscillation used for the $R-Y$ demodulator or the polarity of the already demodulated $R-Y$ colour-difference signal requires an identification signal sent with the composite colour signal.

This identification signal must tell the polarity reversing switch in the receiver whether the scanning line being received is one in which the colour phasor is of type C_ϕ or of type $C_{-\phi}$. This is accomplished by ensuring that the phase of the colour burst that occurs in the line-blanking interval has a value of $+135°$ with respect to the plus $B-Y$ axis during lines of C_ϕ type and $-135°$ during lines of $C_{-\phi}$ type. Pulses to operate the identification switch may be obtained by passing the gated burst through the $R-Y$ synchronous demodulator thus obtaining positive pulses during C_ϕ lines and negative pulses during $C_{-\phi}$ lines.

The process of polarity inverting one of the components of the chrominance phasor produces double the number of chrominance signal sidebands that occur with NTSC. Furthermore, the sidebands due to the $R-Y$ component would coincide in frequency with the harmonic components of the line-scan frequency, because the NTSC half line-frequency offset of the chrominance sub-carrier will have been nullified by the polarity reversals. Compatibility of PAL with NTSC half line-frequency offset is not acceptable. The PAL sub-carrier frequency should therefore be given a quarter line-frequency offset from the nearest appropriate line-scan harmonic. This is near to, but not exactly a precision-best offset. An exact precision-best offset would not actually be the optimum because it would not suit simultaneously both the $R-Y$ and $B-Y$ components. The final frequency chosen has the following relationship with the line-scan frequency:

$$f_0 = (283\tfrac{3}{4} + \tfrac{1}{625})f_L = 283\tfrac{3}{4}f_L + 25\text{ Hz} \tag{8.43}$$

This is $6\frac{1}{4}$ Hz below a precision-best offset. With

$$f_L = 15{,}625\text{ Hz} \qquad \text{we have} \qquad f_0 = 4{,}433{,}618.75\text{ Hz} \tag{8.44}$$

You will see from the above figures that we are now considering a 625-line, 50 field system, since PAL colour has only been used with 525/60 systems in

Brazil. The complete cycle of sub-carrier interference pattern has a duration of eight fields as compared with NTSC's four-field cycle.

8.14 STANDARD OR DELAY-LINE PAL

Now the assumption that the eye will integrate the chromaticities of two adjacent lines of a television field, giving to the brain the impression of a single chromaticity is, in fact, adequately valid for repetition frequencies at which flicker does not occur. (The Bailey experiment applied to very small area coloured dots, noise; not whole television lines.) The frequency at which the colour of a given television line will oscillate in a simple PAL receiver is only $12\frac{1}{2}$ Hz, assuming a given phase error in the chrominance vector. W. N. Sproson, formerly of BBC Engineering Research Dept., has pointed out in a written communication that the integrating property of the eye at $12\frac{1}{2}$ Hz is probably very weak indeed and that, given a chrominance signal phase error of sufficient magnitude, the eye will see sets of horizontal bars of flickering luminance and chromaticity. These are known as Hanover Bars or Venetian Blinds. 'Hanover' Bars because the inventor of the PAL system was Walter Bruch of Telefunken Company in Hannover. Simple PAL decoding is scarcely more tolerant of phase errors than is NTSC; not because the phase error is translated into a simple hue error as in NTSC, but because it is transformed into a set of horizontal bars of alternating chromaticity and luminance. The Appendix deals with this phenomenon in more detail.

If, however, the vector addition of $C_{\alpha+\phi}$ and $C_{-\alpha+\phi}$ (α being the phase error and ϕ the hue-bearing chrominance phase) can be done electrically, instead of in the eye, then the two colour errors can be averaged and the average value can be applied to the display equally in each scanning line and the objectionable Venetian blind structure eliminated.

Electrical addition or averaging requires the simultaneous availability of the chrominance signal from two adjacent lines of a field. This can be achieved by passing each scan line's chrominance signal through a delay-line having a delay equal to the duration of a scanning line. At any given instant, therefore, the input to and the output from the delay-line may be added, thus forming the required sum. One way of achieving this would be to synchronously detect $C_{\alpha+\phi}$ and $C_{-\alpha+\phi}$, thus obtaining, after appropriate polarity reversing of $R-Y$, the two colour-difference signals $R-Y$ and $B-Y$. This would require two delay-lines. The method usually adopted is to operate upon the modulated chrominance signal itself. In this case, the delay-line, which can be a bar of steel or preferably quartz or glass, is used as an ultrasonic or acoustic transmission line with electro-mechanical (piezo-electric) transducers at input and output. Acoustic velocities of propagation are much lower than electrical ones, hence a glass bar is much smaller and cheaper than an electrical delay-line would be. Since we shall operate upon the modulated chrominance signal, the delay-line must have an integer number of half-cycles of sub-carrier signal in it if we are to be able to add the delayed to the un-delayed signal. The actual number of sub-carrier cycles in one scanning line is given by $n = f_0/f_L$ which, from Equation

8.43, is seen to yield

$$n = 283\tfrac{3}{4} + 1/625$$

We cannot, therefore, make the delay of the delay-line exactly equal to the duration of a scanning line. The nearest integer value of n would be $n = 284$ (although $n = 283\tfrac{1}{2}$ would be acceptable). This will mean that in order to achieve perfect addition at radio frequency (4.43 MHz) we must sacrifice perfect timing between ultimately demodulated colour-difference signals and their appropriate patches of luminance. The timing error is 56 ns, the delay in the delay-line being 64.056 μs.

Figure 8.20 shows a possible circuit for a standard PAL demodulator. If C_+ (Equation 8.37) is emerging from the delay-line, the (upper) adder output will be the sum of

$$C_+ = U \sin(\omega_0 t + \theta) + V \cos(\omega_0 t + \theta) \tag{8.45}$$

where

$$\theta = 2\pi \times 284 \text{ rad (the delay measured in radians)}$$

and

$$C_- = U \sin \omega_0 t - V \cos \omega_0 t \tag{8.46}$$

But since θ is an integer number of complete cycles

$$C_+ = U \sin \omega_0 t + V \cos \omega_0 t \tag{8.47}$$

Therefore

$$C_+ + C_- = 2U \sin \omega_0 t \tag{8.48}$$

Similarly, the subtractor output will be

$$C_+ - C_- = +2V \cos \omega_0 t \tag{8.49}$$

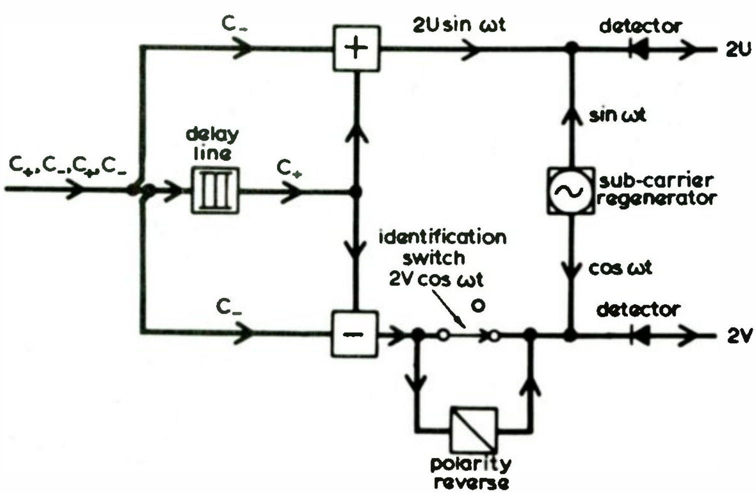


Figure 8.20 A PAL demodulator

when C_- is emerging from the delay-line we have from the adder

$$C_+ + C_- = 2U \sin \vartheta t \tag{8.50}$$

and from the subtractor

$$C_- - C_+ = -2V \cos \omega_0 t \tag{8.51}$$

For this situation, the identification switch must allow the polarity reverser to come into circuit so as to change $-2V \cos \omega_0 t$ into $+2V \cos \omega_0 t$. Note that the delay-line and adder/subtractor combination has separated the chrominance vector (phasor) C into its two radio-frequency components. Each of these can be detected with an ordinary diode detector as shown in Fig. 8.20. Re-inserted carrier must be used, however, because each radio-frequency component is suppressed-carrier modulated. The diode detectors may be replaced by synchronous demodulators if it is so desired.

The PAL system has been described as if it is based on $R-Y$ and $B-Y$ colour-difference signals rather than on I and Q signals, and so, in fact, it is. For the UK System I with the sound carrier at 6 MHz above vision carrier and a video bandwidth of 5.5 MHz, the two PAL colour-difference signals U and V require a video bandwidth of about 1.3 MHz each.

8.15 DISTORTIONS AND DISTURBANCES

First, between coder and decoder.

8.15.1 Differential or level-dependent phase

We have already examined the effect of an error of phase in the transmitted chrominance phasors C_ϕ, $C_{-\phi}$. The effect of this error α, that converts the chrominance phasors into $C_{\phi+\alpha}$ and $C_{-\phi+\alpha}$, is to reduce by a factor $\cos \alpha$ the lengths of the U and V projections, X and Y, of the phasors resulting from the vector addition of $C_{\phi+\alpha}$ and $C_{-\phi+\alpha}$. Thus, a loss of saturation is experienced, but no change of hue.

Normally, phase errors such as α are not constant, but vary with luminance level. Such a condition may arise from the non-linearity of resistive circuit components like valves, diodes and transistors in the presence of reactance. A simple example is shown in Fig. 8.21. Here, the resistance r_2 depends upon the voltage across it. r_2 could represent the grid-cathode resistance of a transmitter modulator or amplifier valve, or the input resistance of a

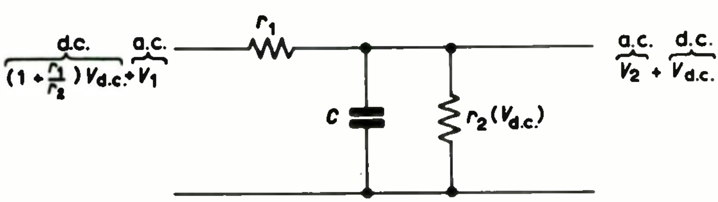


Figure 8.21 Non-linearity causes differential phase

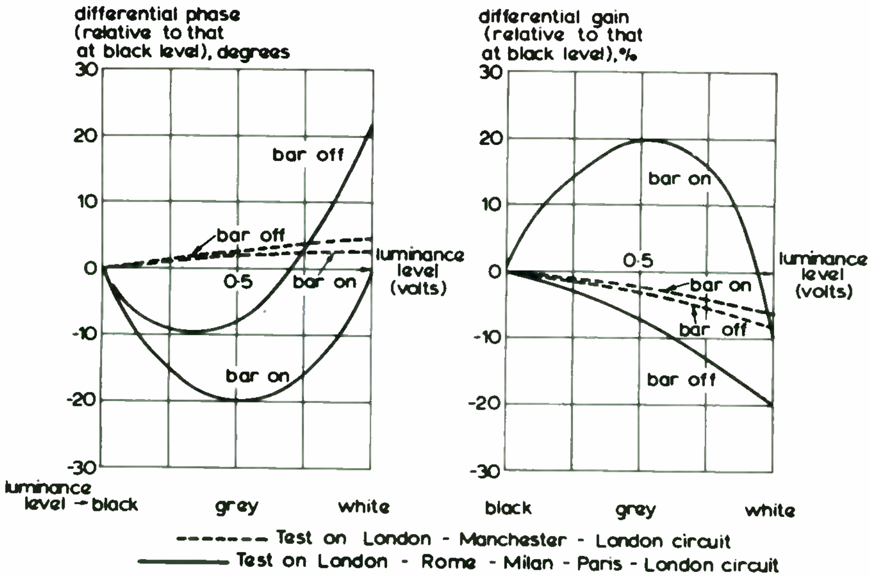


Figure 8.22 Examples of differential phase and of differential gain

transistor, or a diode clamp. If we assume that $V_2 \ll V_{\text{d.c.}}$, we can write

$$\frac{V_2}{V_1}=\frac{1}{1+r_1/r_2(V_{\text{d.c.}})+j\omega r_1 C} \tag{8.52}$$

where $\omega/2\pi$ is the sub-carrier frequency. That is, we have not accounted for the d.c. component (a low-frequency luminance component) except for its effect upon the value of r_2, whence the phase shift through the network is

$$\alpha = -\arctan\frac{\omega r_1 C}{1+r_1/r_2(V_{\text{d.c.}})} \tag{8.53}$$

which is a function of the d.c. output voltage, $V_{\text{d.c.}}$.

If $r_2 \rightarrow \infty$, α becomes constant or if r_2 ceases to be a function of $V_{\text{d.c.}}$, α becomes constant. It is easy to see from Equation 8.53 that the phase of a chrominance signal superimposed upon a low-frequency luminance signal will vary with the amplitude of this luminance component. This is an example of differential phase. Figure 8.22* shows some typical shapes. Differential phase is often measured by comparing the phases of bursts of chrominance sub-carrier that are superimposed upon a luminance staircase waveform. As a matter of fact, V_2 is often greater than $V_{\text{d.c.}}$ and the mean phase, α, can vary with chrominance signal amplitude, but I am not aware that this effect is significant, in practice.

* The test signal used was CCIR Test Signal No. 3 of Recommendation 421 with the intermediate three lines at peak-white (bar on) and at black level (bar off).

8.15.2 Differential or level-dependent gain

An unwanted change of gain suffered by the chrominance signal unaccompanied by the same change of gain imposed upon the luminance signal will cause an unwanted change of saturation. In this case, the PAL signal behaves in exactly the same way as NTSC. Differential gain results from non-linear resistance and does not need the presence of reactance as did the production of differential phase. Usually, however, non-linear resistance is accompanied, intentionally or otherwise, by reactance and differential gain and differential phase co-exist. For example, Equation 8.52 has a modulus or gain

$$\left|\frac{V_2}{V_1}\right| = \frac{1}{[\{1 + r_1/r_2(V_{\mathrm{d.c.}})\}^2 + \omega^2 r_1^2 C^2]^{1/2}} \tag{8.54}$$

which is clearly dependent upon $V_{\mathrm{d.c.}}$, the luminance component of the output voltage.

Differential gain can be measured using the same test signal as for differential phase. Examples are shown in Fig. 8.22 where the percentage of differential gain is shown as ordinates.

8.15.3 Mis-timing of chrominance signal with respect to luminance signal

This is also referred to as 'chrominance-to-luminance delay inequality'. If a patch of colour is displayed in such a way that it is misplaced horizontally with respect to its associated patch of luminance there results what is termed in the USA the 'funny-paper' effect. Such a condition arises if, for example, the luminance delay-line in the colour receiver is not of the correct electrical length. A severely non-uniform curve of group-delay $[d\alpha/d\omega = f(\omega)]$ as a function of frequency will give rise to the same effect. This was discussed earlier and Fig. 8.18 gave an example.

8.15.4 Non-uniform group-delay across the chrominance pass-band

This manifests itself as an unwanted phase shift between the upper and lower sidebands of each of the two components E'_U and E'_V of the chrominance signal. The burst phase, whatever it may be, is always taken as reference. Reference to Fig. 8.6 shows that, for example, S'_+ and S''_+ may be rotated more clockwise than S'_- and S''_- anti-clockwise. Thus R' and R'' will be rotated clockwise from their correct angles and consequently so will be R. The PAL system, as we have already seen (Fig. 8.19), will convert this phase or hue error into one of saturation, which is far less visible. This is such an effective advantage of the PAL system that BREMA* have stated that they do not require pre-correction at the transmitter for group-delay errors that occur in the receiver in the range of frequencies occupied by the chrominance signal.

8.15.5 Non-uniform amplitude response across the chrominance pass-band

Again referring to Fig. 8.6, if S'_+ and S''_+ are more attenuated than S'_- and S''_-, there will be errors of phase in R' and R'' and there will also be reductions in the

* British Radio Equipment Manufacturers Association.

amplitudes of these two phasors. The amplitude errors will result in a decrease of saturation and, after delay-line decoding, so will the phase errors, in the same way as we have already seen when discussing the U and V components of the vector sum of the switched and unswitched chrominance phasors, Equations 8.41 and 8.42. The fact that the phases of the quadrature components R' and R'', and hence R, will oscillate at the frequency of the sidebands under consideration, does not invalidate the foregoing argument.

We now turn to distortions that occur upstream of the coder and downstream of the decoder.

8.15.6 Gamma tracking errors

These can occur wherever gamma-correcting circuits are used in an R, G, B transmission. Examples are gamma correctors in cameras and telecine equipment, unequal transfer characteristics in the three (or four) pick-up tubes of cameras and unequal beam-current against excitation-voltage characteristics of three-gun display tubes. Apart from the errors that can arise if the gamma correction law is not the inverse of the display-tube gamma law, there are the much more visible effects resulting from unequal laws applying to each of the three primary signals. For example, it may arise in a display tube that although equal excitation voltages on the three guns may produce the desired colour of peak-white and that three equal black-level voltages may give rise to an 'achromatic' dark grey of the same 'colour' as the peak-white, an intermediate set of three equal voltages may give rise to a mid-grey that is not achromatic. This not uncommon fault can be found in all equipments, both of the professional broadcasting type and in domestic receivers.

8.15.7 Colour balance

Even when all non-linear circuits of the gamma-law and gamma correcting type are satisfactory, the background or 'sit' levels must be sufficiently stable to be able to maintain their alignment values. The eye is perfectly capable of seeing the colour appearing in a 'black' part of a scene if the misalignment of the three black levels is of the order of 10% of the black level.

8.15.8 Misregistration

In a receiver, the static and dynamic convergence circuits are relied upon to enable adequate registration to be achieved and maintained. For satisfactory performance, the three electron beams in the display tube must register to well within one picture element—say, a 600th of the picture height. The requirements for the camera are no less stringent; although it is claimed that some latitude is permissible for four-tube cameras.

8.15.9 Random noise

As far as the luminance channel is concerned, there is no difference between the behaviours of monochrome and colour, but, due to the partial application of the 'constant-luminance' principle, noise appearing in the chrominance channel contributes very little to the displayed luminance and fine detail chromaticity noise is, as we have seen, Fig. 8.14, somewhat less visible. Perhaps it should be added that noise in that part of the luminance channel in the colour receiver that is occupied by the chrominance signal never directly reaches the display tube (other than via the chrominance decoder), because of the presence of the anti-chrominance notch filter centred around 4.4 MHz in the luminance channel at a point downstream of the chrominance 'take-off'. Thus, the service area of a colour transmission is, effectively, the same as that of a monochrome transmission—transmitters, aerials, and all other things being equal, of course. Some radio links employ frequency modulation and the spectrum of the noise contributed by these may have a triangular form; that is, the noise voltage per unit bandwidth may be approximately proportional to video frequency. For these equipments, colour is more susceptible to random-noise interference than is monochrome, because the noise per unit bandwidth in the region of the video-frequency pass-band occupied by the chrominance signal can exceed that in the part occupied by the luminance signal by a significant amount.

8.15.10 Co-channel interference

If one takes an interfering c.w. signal—a sinewave—and slowly varies its frequency from below the vestigial sideband of a Standard I colour emission right through the range of frequencies up to the limit of the vision pass-band, and one adjusts the amplitude so that, at all frequencies, the interference could be described as lying between the subjective criteria of 'definitely perceptible but not disturbing' and 'somewhat objectionable', one arrives at the curve

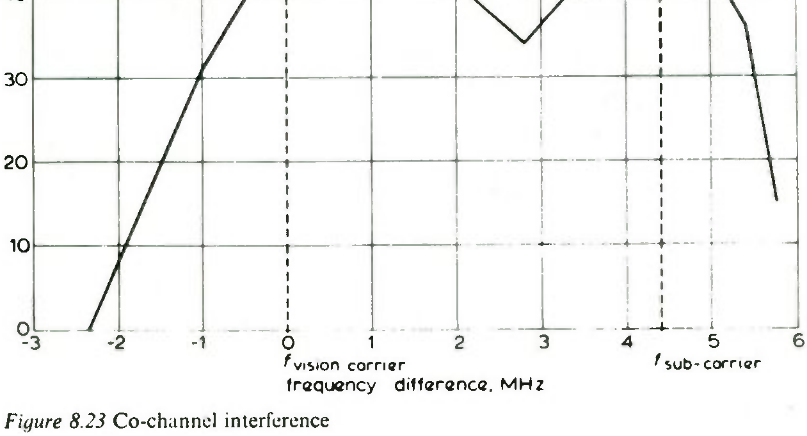


Figure 8.23 Co-channel interference

shown in Fig. 8.23. The effect producec by the receiver circuits designed to accept the chrominance signal is shown by the second peak situated around the sub-carrier frequency, f_0. The answer to the question: 'Why does chrominance require almost as much protection from interference as luminance?' (near-equality in heights of the two peaks), is that an interfering sinewave close in frequency to that of the sub-carrier produces coloured interference patterns whose detail is no longer small.

Now, if instead of an interfering sinewave we consider another, but unwanted, colour emission in co-channel relationship with a wanted colour emission, we find that the protection ratio, that is, the ratio of wanted-to-unwanted vision carrier amplitudes for a given subjective criterion, is the same as if either or both emissions were monochrome ones. In other words, the interference between the two chrominance signals does not predominate over that between the two luminance signals. This is because, on the one hand, the wanted chrominance signal is no more susceptible to sinewave interference than is the wanted luminance signal (refer to the two near-equal peaks in Fig. 8.23*) and, on the other hand, the unwanted chrominance signal is always at least 14 dB weaker than the unwanted luminance signal.

8.15.11 Cross-colour

See Sections 8.6 and 8.7.

APPENDIX: LINE-TO-LINE VARIATION OF COLOUR DUE TO A PHASE ERROR IN THE CHROMINANCE SIGNAL—'HANNOVER BARS' IN A SIMPLE PAL RECEIVER

As stated in Section 8.14, the standard or delay-line PAL receiver will average the chrominance signal or the demodulated colour difference signals over each pair of scan lines in each field so as to eliminate the colour changes that occur line-by-line if there is an error of phase in the chrominance signal. Whilst the receiver decoder switch cancels the transmitted switching of the chrominance phase angle, thus converting every ϕ_- into a ϕ_+; if there is an error angle α this will be switched to $-\alpha$ on all ϕ_--type lines. In the simple receiver with no delay line there will be luminance and chromaticity changes from line to line that we shall investigate.

First we consider a chrominance vector or phasor

$$\psi = 0.88C = 0.88[(R'-Y')^2 + \{0.56(B'-Y')\}^2]^{1/2}\sin(\omega_0 t + \phi)$$

similar to the chrominance part of Equation 8.20. If we assume an error angle α we can write for lines of the type ϕ_+

$$\psi_+ = 0.88[(R'-Y')^2 + \{0.56(B'-Y')\}^2]^{1/2}\sin(\omega_0 t + \phi + \alpha) \qquad \text{(A.1)}$$

and for lines of the type ϕ_-, after the receiver switch has cancelled the action of

* CCIR Report No. 306, Oslo.

the transmitter switch,

$$\psi_- = 0.88[(R'-Y')^2 + \{0.56(B'-Y')\}^2]^{1/2} \sin(\omega_0 t + \phi - \alpha) \tag{A.2}$$

The receiver switch has changed $-\phi+\alpha$ into $\phi-\alpha$. Note that, for convenience, I have assumed that we have switched the chrominance-carrier signal rather than the demodulated colour-difference signals. The switching may be performed either before demodulation or after it.

If we now synchronously detect the signals ψ_+ and ψ_- and filter out all terms having the frequency $2\omega_0/2\pi$ we have

$$\psi_+ \times 4.1 \sin \omega_0 t = 1.786[(R'-Y')^2 + \{0.56(B'-Y')\}^2]^{1/2} \cos(\phi + \alpha) \tag{A.3}$$

$$\psi_- \times 4.1 \sin \omega_0 t = 1.786[(R'-Y')^2 + \{0.56(B'-Y')\}^2]^{1/2} \cos(\phi - \alpha) \tag{A.4}$$

From the paragraph following Equation 8.31 on page 161, we see that Equations A.3 and A.3 represent the $B'-Y'$ signal for lines ϕ_+ and ϕ_- respectively. For $R'-Y'$ we have

$$\psi_+ \times 2.3 \cos \omega_0 t = [(R'-Y')^2 + \{0.56(B'-Y')\}^2]^{1/2} \sin(\phi + \alpha) \tag{A.5}$$

$$\psi_- \times 2.3 \cos \omega_0 t = [(R'-Y')^2 + \{0.56(B'-Y')\}^2]^{1/2} \sin(\phi - \alpha) \tag{A.6}$$

Now expand the cosine and sine factors in Equations A.3 to A.6 and replace $\cos \phi$ by

$$\frac{0.56(B'-Y')}{[(R'-Y')^2 + \{0.56(B'-Y')\}^2]^{1/2}}$$

and $\sin \phi$ by

$$\frac{R'-Y'}{[(R'-Y')^2 + \{0.56(B'-Y')\}^2]^{1/2}}$$

and we obtain for the distorted (due to the phase error α) $B'-Y'$ signal on lines ϕ_+

$$\phi_+ : (B'-Y')^{*}_{\text{dis}} = (B'-Y') \cos \alpha - 1.786(R'-Y') \sin \alpha \tag{A.7}$$

and on lines ϕ_-

$$\phi_- : (B'-Y')_{\text{dis}} = (B'-Y') \cos \alpha + 1.786(R'-Y') \sin \alpha \tag{A.8}$$

For the distorted $R'-Y'$ signal, we have

$$\phi_+ : (R'-Y')_{\text{dis}} = (R'-Y') \cos \alpha + 0.56(B'-Y') \sin \alpha \tag{A.9}$$

$$\phi_- : (R'-Y')_{\text{dis}} = (R'-Y') \cos \alpha - 0.56(B'-Y') \sin \alpha \tag{A.10}$$

From Equation 8.32 we calculate the distorted $G'-Y'$ signal

$$\begin{aligned}\phi_+ : (G'-Y')_{\text{dis}} = &-0.51\{(R'-Y') \cos \alpha + 0.56(B'-Y') \sin \alpha\} \\ &-0.19\{(B'-Y') \cos \alpha - 1.786(R'-Y') \sin \alpha\}\end{aligned} \tag{A.11}$$

* The subscript 'dis' is intended to distinguish distorted signals from undistorted signals.

$$\phi_-: (G'-Y')_{dis} = -0.51\{(R'-Y')\cos\alpha - 0.56(B'-Y')\sin\alpha\}$$
$$-0.19\{(B'-Y')\cos\alpha + 1.786(R'-Y')\sin\alpha\} \qquad (A.12)$$

Figure 8.12 shows that to obtain the nominal (distorted) values of the primary signals, we must add Y' to each of the three colour-difference signals; thus, for example:

$$(B'-Y')_{dis} + Y' = B'_{dis}$$

Adding Y' to both sides of equations A.7 to A.12 we obtain for lines ϕ_+

$$B'_{dis} = B'\cos\alpha - 1.786R'\sin\alpha + Y'(1-\cos\alpha + 1.786\sin\alpha) \qquad (A.13)$$

$$\phi_+ \quad R'_{dis} = R'\cos\alpha + 0.56B'\sin\alpha + Y'(1-\cos\alpha - 0.56\sin\alpha) \qquad (A.14)$$

$$G'_{dis} = R'(-0.51\cos\alpha + 0.339\sin\alpha) - B'(0.19\cos\alpha + 0.286\sin\alpha)$$
$$+ Y'(1 + 0.7\cos\alpha - 0.0537\sin\alpha) \qquad (A.15)$$

And for lines ϕ_-

$$B'_{dis} = B'\cos\alpha + 1.786R'\sin\alpha + Y'(1-\cos\alpha - 1.786\sin\alpha) \qquad (A.16)$$

$$\phi_- \quad R'_{dis} = R'\cos\alpha - 0.56B'\sin\alpha + Y'(1-\cos\alpha + 0.56\sin\alpha) \qquad (A.17)$$

$$G'_{dis} = R'(-0.51\cos\alpha - 0.339\sin\alpha) + B'(-0.19\cos\alpha + 0.286\sin\alpha)$$
$$+ Y'(1 + 0.7\cos\alpha + 0.0537\sin\alpha) \qquad (A.18)$$

As a check, we note that if $\alpha = 0$ in Equations A.13 to A.18, we obtain for both ϕ_+ and ϕ_- lines

$$B'_{dis} = B' \qquad R'_{dis} = R' \qquad G'_{dis} = G'$$

When the three primary signals B'_{dis}, G'_{dis} and R'_{dis} appear as grid-to-cathode voltages of the colour display tube, they are raised to the power gamma, say 2.5, as they become the three beam currents and the luminance displayed on each line is

$$Y_+ = 0.299(R'_{dis})^{2.5} + 0.587(G'_{dis})^{2.5} + 0.114(B'_{dis})^{2.5}$$

for ϕ_+ lines and similarly for Y_- on ϕ_- lines.

Equations A.13 to A.18 clearly show that the colour on ϕ_+ lines is different from that on ϕ_- lines; the word colour meaning chromaticity and luminance. W. N. Sproson[1] had kindly calculated the quantity ΔE^*_{uv} occurring for a phase error of 10°. ΔE^*_{uv} is the most accurate quantitative measure of the total difference between two colours. It is the internationally agreed measure of subjective perceptibility. The calculations of ΔE^*_{uv} from the distorted red, green and blue signals of Equations A.13 to A.18 require the raising of their values to the power gamma = 2.5; the calculation of the values of the internationally agreed 'unreal' primaries (XYZ) of the CIE† system of colorimetry; the calculation of the CIE 1930 chromaticity co-ordinates (xyz); the calculation of the $u'v'$ CIE 1976 'uniform' chromaticity co-ordinates; the calculation of the differences in 'uniform' co-ordinates between the ϕ_+ and ϕ_- colours and,

† Commission Internationale d'Eclairage.

Table A.1 THE VISIBILITY OF HANOVER BARS

Rank order of impairment in the subjective experiment	*Theoretical studies*			
	W. N. Sproson		*R. D. A. Maurice*	
	NTSC + Illuminant C	*System I*	ΔE^*_{uv}	*NTSC + Illuminant D65*
Cyan	Cyan	Cyan	116	Magenta
Magenta	Magenta	Magenta	77	Yellow
Yellow	Yellow	Yellow	58	Cyan
Red	Red	Blue	46	Blue
Blue	Blue	Red	10	Red
Green	Green	Green	5	Green

finally, the square root of the sum of the squares of the luminance difference and the two 'uniform' chromaticity differences. The above is no simple task. For the details of the method the interested reader should refer to Sproson's book, Sections 2.1.2, 2.1.3, 2.1.4, 1.9 and 1.10.

Following Sproson's method, I have calculated the quantity ΔE^*_{uv} for NTSC phosphors and Illuminant D65 whereas Sproson's calculations assumed NTSC phosphors with Illuminant C on the one hand and System I phosphors and Illuminant D65 on the other hand. Some of the results are given in the table below.

M. B. Tancock, BBC Engineering Training School, arranged a subjective assessment of the visibility of Hanover Bars using eight test volunteers. The colour bar signal used was the same as that assumed for the theoretical calculations mentioned above, namely 100% amplitude and 100% saturation. Various values of phase error were introduced and the test subjects were asked to record the colours in rank order of impairment due to the Hanover Bar phenomenon. The experimental results are summarised in Table A.1.

The agreement between Sproson's NTSC + Illuminant C calculations and the experimental results is excellent, although it is likely that the phosphors used in the experiment were of the System I variety (PAL with Illuminant D65). My own results seem to be rather erratic!

REFERENCE

1. Sproson, W. N., *Colour Science in Television and Display Systems*, Adam Hilger

Chapter 9

Pulse and carrier generation

G. B. M. Claydon

9.1 INTRODUCTION

Within a television centre it is necessary to provide a central source of pulses to trigger the line (horizontal) and field (vertical) scanning circuits of the various picture-originating equipments, cameras, telecines, video-recorders, so that they run in synchronism, thereby enabling the producer to switch or mix between these sources without causing a disturbance at the receiver. The accuracy of the scanning frequencies is important, especially in colour television, and must meet the standards laid down by the broadcasting authorities.

In addition, in colour television systems, the colouring information is conveyed on a modulated sub-carrier. This is superimposed on the black and white, or luminance part of the signal, and in most systems, in order to minimise the visibility of this sub-carrier on monochrome receivers, the sub-carrier and line frequencies must be related in a special way.

It is the function of the sync pulse generator (SPG) to supply the various correctly-related line, field and sub-carrier synchronising signals for distribution throughout the television studio centre. Generally, two or more sync pulse generators are installed, to permit independent operation of certain studios, and to provide a standby.

Figure 9.1 shows a typical arrangement of two SPGs, a changeover (CO) unit feeding pulse and video (for the sub-carrier) distribution amplifiers (PDA/VDA). The output of these is routed to the various equipments, video tape recorders (VTR), studio, telecine (TC), etc.

9.2 PULSES

In a monochrome signal there are two types of pulse signal, generated by the SPG, which are transmitted and used by the receiver for blanking and sync signals. The frequency and duration of these pulses are accurately specified by

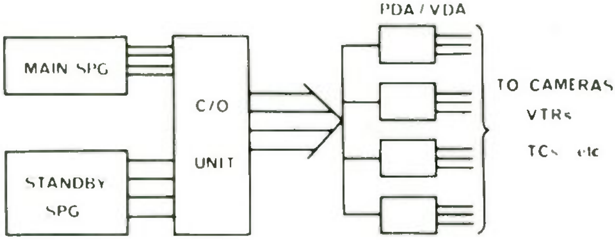


Figure 9.1 Typical arrangement of SPGs

the broadcasting authorities in order to ensure that pictures from different sources may be displayed and combined easily.

In addition to the transmitted blanking and sync pulses, it was useful for the SPG to generate other signals that were used only within the studios, and not transmitted. These were line drive and field drive. They are not used in modern equipment.

The line drive pulse is mainly obsolete now, but some older systems, still in service, may use it. It occurs line by line right through the field blanking period. Its main use was to trigger the camera line timebase. Historically, some cameras used directly-driven time bases, and it was common practice to advance the position of the line drive pulse, by up to 8 μs, relative to sync, to enable the cameras to be used with long camera cables, up to 3,000 feet.

Modern cameras use phase-locked oscillators in the line timebase circuits, and this has removed the need for advancing the line drive pulse.

There was no need to advance the position of the field drive pulse, since the camera cable delay is only a small fraction of the field period.

9.2.1 Types of pulse

Blanking is used, as the name implies, to blank or suppress the video signal from the camera output during the flyback period. It is also used to generate pedestal or 'set-up' signals at black level. Line blanking proceeds the line sync pulse by the width of the 'front porch'. The purpose of the front porch is to ensure that the sync pulse always starts at black level. If this is not the case, the effective position of the sync pulse would vary with the picture content, with the result that the start of the line scan would vary and give a ragged edge or streaky picture.

The duraction of field blanking has tended to increase in recent years to allow for the insertion of vertical interval test signals (VITS) and Teletext. If these signals are not to be visible, they must be positioned late enough to allow the receiver field flyback to be completed, and yet not too close to the start of the active picture. Lines 16 to 20, and 329 to 333 have been allocated for VITS in the 625-line System I, line 1 being at the start of the first broad pulse on the even field. For Teletext, lines 17 to 18, and 330 to 331 have been allocated.

Equalising pulses are at twice line frequencies. This is necessary because of interlacing, which requires that the start of every second field should occur half way between two line pulses. Interlacing two fields to produce one complete picture reduces the effect of flicker. It is worth mentioning here that the 525-line systems where the field frequency is 60 Hz can be brighter than the 50 Hz field rate of the 625-line systems. For the same flicker visibility, the 525-line picture can be six times brighter.

Separation of the field sync pulses from the line pulses at the receiver or monitor is most easily done by integrating the signal or, in other words, removing the line pulses by means of a low-pass filter. The purposes of the equalising pulses, before and after the broad pulses, is to improve this process. The equalising pulses ensure that the integrator sees the same signal before and after the broad pulses, in both odd and even fields. Without equalising pulses, the broad pulses would be preceded by a whole line period in one field, and by a half line period in the next. This could give different outputs from the integrator on the two fields, and so cause poor interlacing.

However, by using slightly more advanced field pulse separators, the effect of the odd and even fields can be eliminated, even without equalising pulses, and there have been various proposals for reducing the number of equalising pulses. The latest suggestion is for one equalising pulse to precede one field only.

Figure 9.2 shows the timing for the 625-line System I line pulses. The timing reference point is the leading edge of the sync pulse. The line period is 64 μs, line blanking 12.05 μs, line synchronising pulse width 4.7 μs, and front porch 1.5 μs. Where a line-drive pulse is used, it will have a width of 6.5 μs.

The field pulses shown in Fig. 9.3 consist of field sync, field blanking and field drive. The field frequency is 50 Hz, the field blanking interval is 21 to 25 lines, and the field sync pulse consists of five broad pulses, each of 27.3 μs duration. The pulse train includes two groups of five equalising pulses, each

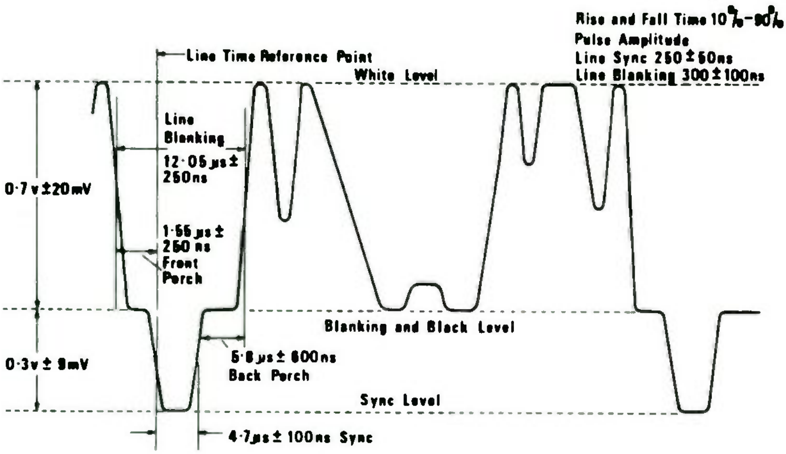


Figure 9.2 625 line system I line pulses. The waveform of a typical line showing synchronizing signals. Pulse duration is measured at half amplitude points. Blanking duration is measured at half amplitude points with a white level signal of line duration and for this reason the picture signal has been shown starting and finishing at the white level

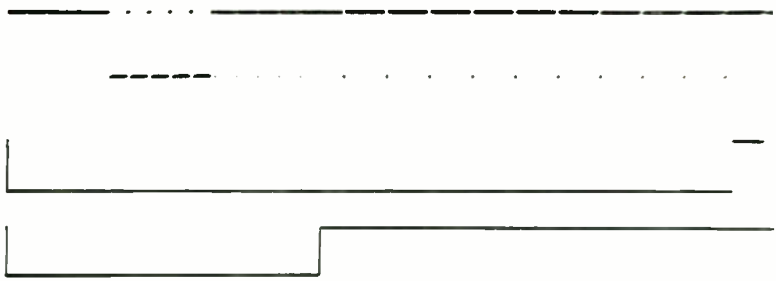


Figure 9.3 625 line system I field pulses, field sync, field blanking and field drive

pulse with a duration of 2.3 μs. The field-drive pulse starts coincident with field blanking, and has a duration equal to the total equalising and broad-pulse period of $7\frac{1}{2}$ lines.

In a colour SPG additional pulses are required. These are

(1) Burst gate pulse.
(2) PAL identification (ident).
(3) Sub-carrier.

These signals are used in the colour coders. Figure 9.4 shows the burst gate pulse which determines the position and width of the reference sub-carrier burst. About 10 cycles of sub-carrier are gated in on the back porch, starting 0.9 μs after the trailing edge of the sync pulse. The PAL ident signal ensures that all coders produce outputs whose carrier phase is switched line by line in the correct sequence. In modern coders, the PAL ident square-wave is generated from the burst gate.

In addition to these extra outputs, the SPG must be driven from clocks that are precisely related to the sub-carrier frequency. These clocks are usually run at 2 × line frequency, f_{2h}. The sub-carrier oscillator counting circuits, used to

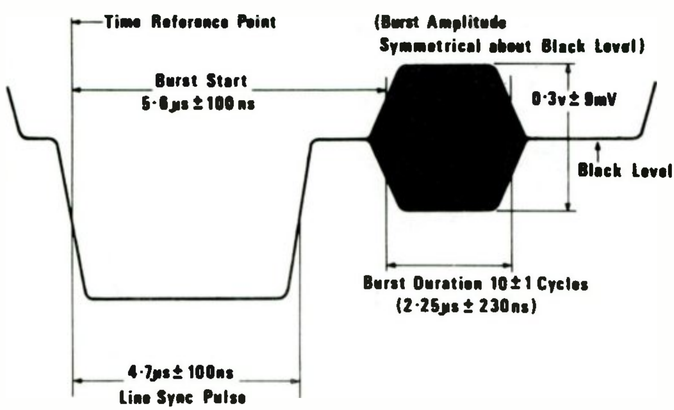


Figure 9.4 Position and width of colour burst on a standard level signal (700 m white level)

generate the twice line frequency, make up the colour sync unit. In older SPGs this was a separate unit that could be added to monochrome SPGs.

9.3 GENERATION OF THE COLOUR SUB-CARRIER

The frequency tolerance required for the PAL system is ± 1 Hz or 2 parts/10^7, and this can be met by a quartz crystal oscillator in a temperature-controlled oven. A rubidium frequency standard is a more expensive alternative.

The quartz oscillator requires periodic checks of its frequency due to changes with age. A typical unit has an ageing rate of 5 parts/10^8/month, producing a rate-of-change of sub-carrier frequency of $(5/10^8)\ 4.43 \times 10^6 = 0.22$ Hz/month, giving a life between frequency adjustments of about four months.

A rubidium standard has a much lower ageing rate of 2 in 10^{11}/month. In this case, the rate of frequency change is $(2/10^{11})\ 4.43 \times 10^6 = 8.8/10^{-5}$ Hz/month.

Checks or adjustment of oscillator frequency must be made against a standard which has a high order of frequency precision. A digital frequency counter would not be good enough, being possibly of the order of 1 part in 10^6. An 'off-air' standard is used, this is derived from the BBC Droitwich transmitter, which transmits on 200 kHz to an accuracy of 1 part in 10^{11}.

9.3.1 Factors that determine the choice of sub-carrier frequency

The sub-carrier frequency is 4.43361875 MHz, and it was shown in Chapter 8 that many factors were involved in the determination of this value.

The sub-carrier is situated at the top end of the video band, where the luminance energy is low, and interference effects between luminance and colour sub-carrier are minimised. It has been shown that representation of

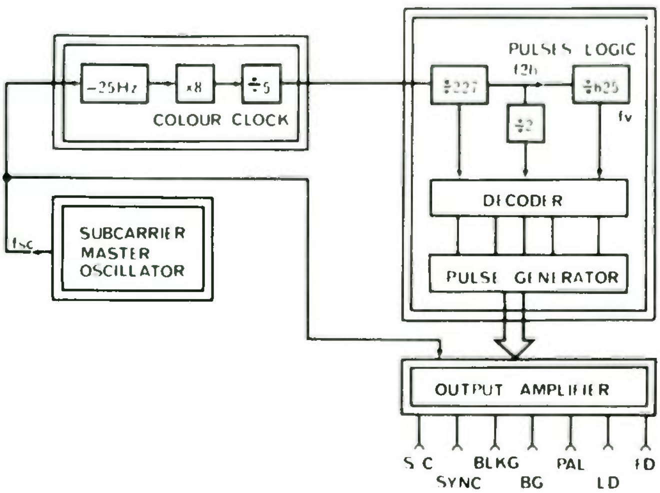


Figure 9.5 SPG block diagram (Marconi)

colour does not need a wide bandwidth and the colour sidebands are, consequently, restricted in width. The sound carrier is spaced 6 MHz from the vision carrier and is thus 1.57 MHz from the sub-carrier, but beat effects are negligible due to the sound carrier being frequency-modulated.

The lowest visibility of interference effects between the colour sub-carrier and the vision (luminance) carrier is produced when a quarter-line offset is used, i.e.

$$f_{sc} = (n - \tfrac{1}{4})f_h + f_v/2$$

where f_{sc} = sub-carrier frequency, f_h = line frequency, f_v = field frequency and n = an integer.

An important feature of this relationship is that f_h and f_v must be generated from the high-precision sub-carrier, as explained in Chapter 8.

For the PAL System I, $n = 284$ is selected, and the sub-carrier frequency becomes

$$f_{sc} = (284 - \tfrac{1}{4})f_h + f_v/2$$

If this expression is re-arranged to give a value for twice line frequency, $2f_h$, it becomes

$$\begin{aligned} 2f_h &= (f_{sc} - f_v/2)\left(\frac{8}{5 \times 227}\right) \\ &= (f_{sc} - f_v/2)\left(\frac{8}{1{,}135}\right) \end{aligned}$$

An early Marconi SPG used this relationship, as follows:

$$\begin{aligned} f_{sc} &= \frac{1{,}135}{4} f_h + \frac{f_v}{2} \qquad \left(\text{but } f_v = 625 f_{2h} \text{ or } \frac{625}{2} f_h\right) \\ &= \left(\frac{1{,}135}{4} + \frac{1}{625}\right) f_h \\ &= \frac{709{,}379}{2{,}500} f_h = \frac{11 \times 64{,}489}{5{,}000} \times 2f_h \end{aligned}$$

Unfortunately, 64,489 is a prime; however,

$$64{,}489 = 5{,}000 + 59{,}489 = 5{,}000 + (19 \times 31 \times 101)$$

Now

$$f_{sc} = 11 f_{2h} \left\{ \frac{5{,}000 + (19 \times 31 \times 101)}{5{,}000} \right\}$$

and

$$5{,}000 = 5^4 \times 2^3$$

Let $11 f_{2h} = f_0$, then

$$f_{sc} = f_0 \left(1 + \frac{19 \times 31 \times 101}{5^4 \times 2^3}\right)$$

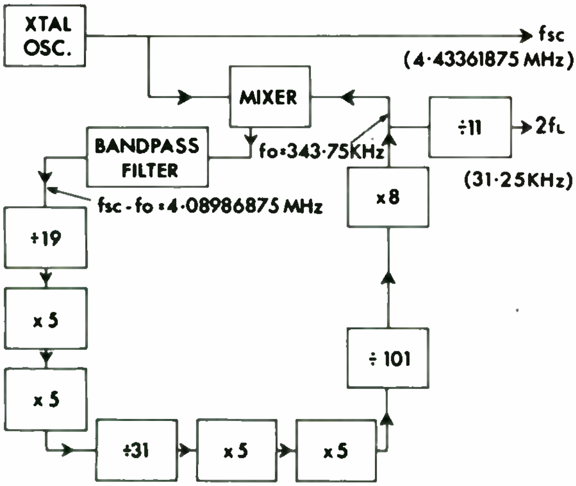


Figure 9.6 Generation of twice line frequency from subcarrier using a mixer

or

$$f_0 = (f_{sc} - f_0)\left(\frac{5^4 \times 2^3}{19 \times 31 \times 101}\right)$$

Figure 9.6 shows the block diagram of this colour sync unit. The sub-carrier is fed to one input of the mixer and the frequency f_0, derived from the counter chain, to the other. From the mixer output, the difference component $(f_{sc} - f_0)$ is selected and fed to the divider chain. The twice line frequency is obtained by dividing f_0 by 11.

The wanted frequency at the mixer output is spaced from the sub-carrier by 340 kHz and by twice that amount from the unwanted sideband $(f_{sc} + f_0)$ so that the required selectivity of the band-pass filter is not critical.

The mixer is actually a balanced modulator which gives considerable suppression of the sub-carrier. The $\times 8$ stage uses a phase-locked oscillator, which provides the means for starting the closed-loop system.

A later SPG uses a different method. Again, starting from

$$f_{sc} = (284 - \tfrac{1}{4})f_h + f_v/2$$

this can be re-arranged as

$$2f_h = (f_{sc} - f_v/2)\left(\frac{8}{5 \times 227}\right)$$

In this case we have to use a dividing and multiplying technique. One method uses a balanced modulator, as shown in Fig. 9.7. The vector diagrams of Fig. 9.8 show the principle. The carrier is suppressed by the balanced modulator, and the 90° phase shift of the modulating signals produces cancellation of one pair of sidebands, and an addition of the other pair. The remainder of the system is shown in Fig. 9.9. The multiplication may be carried out by tuned harmonic amplifiers, but circuits of very high Q are required, and

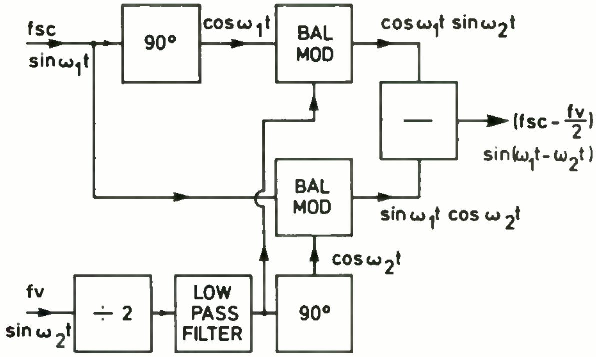


Figure 9.7 Generation of the 25 Hz offset ($\sin \omega_1 t \cos \omega_2 t - \cos \omega_1 t \sin \omega_2 t = \sin(\omega_1 - \omega_2)t$)

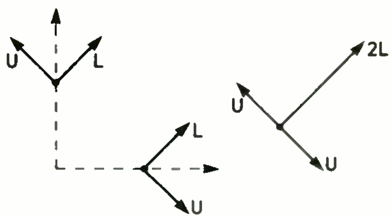


Figure 9.8 Vector diagram of 25 Hz offset

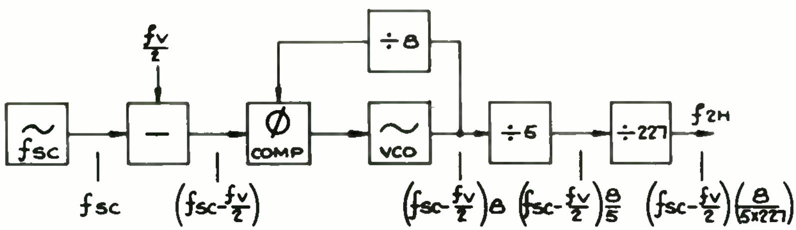


Figure 9.9 Generating twice line frequency from subcarrier

a clipper stage is needed to produce the harmonics for selection by the tuned circuit (Fig. 9.10(a)).

A better method is shown in Fig. 9.10(b), using a locked oscillator. The voltage-controlled oscillator can be an L/C system or a quartz crystal controlled system. The low-pass filter will have a long time-constant. A disadvantage is that, if the input fails, an unlocked output is provided.

The operation of the loop can be understood if, initially, the assumption is made that the loop is locked, i.e. the phase comparator is receiving inputs from the reference input and the voltage-controlled oscillator (VCO) which are equal in frequency. Under these conditions the phase detector produces an output voltage proportional to the phase difference of its inputs. This output is

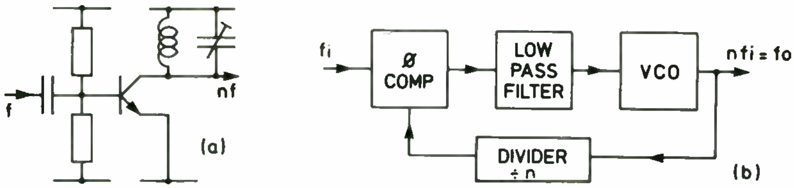


Figure 9.10

filtered to remove transients, and is fed back to control the frequency of the VCO. Hence the loop acts as a proportional-controller with respect to phase and, in the locked condition, the phase error will be proportional to the reciprocal of the open-loop gain. However, since phase angle is the time integral of frequency, the loop may also be considered as an integral controller with respect to frequency and, as it is characteristic of this form of control, no average frequency error exists in the locked condition.

9.4 PULSE GENERATION

The widths and timings of the line pulses from an SPG must be accurately controlled, and must be stable once set up to the required values. The permitted tolerances are sometimes very small, e.g. the width of the line sync pulse on the 625-line System I is 4.7 ± 0.1 μs. Furthermore, this tolerance should be held over the specified temperature range of the equipment.

There are a number of ways in which pulse timings can be obtained:

(1) delay lines
(2) monostables
(3) clock pulses and counting circuits.

Figure 9.11(a) shows a simplified delay-line arrangement for, say, sync and blanking, and (b) shows the relative timing in various parts of the circuit. The input at *A* is a line-frequency pulse of short duration, which is fed into the line. The outputs *B*, *C* and *D* are taken from the line and used to trigger two bistable flip-flops. The blanking flip-flop is triggered on by pulse *A* and off by pulse *D*. Similarly, the sync bistable is triggered on and off by pulses *B* and *C*. The widths of the output pulses can be chosen by adjusting the taps on the delay line. The main disadvantage of this scheme is the size and cost of suitable tapped delay lines. An advantage is that the accuracy and stability can be very good since they depend on passive components only.

Monostables may be conventional RC types using discrete components or integrated circuits, or they may use LC timing. With the availability of cheap integrated circuits, digital systems for the determination of pulse widths are very attractive.

One method uses a high-frequency crystal oscillator and, by means of counters and gates, selects certain of the oscillator pulses. These pulses then fix the position of the edges of the SPG output waveforms. Figure 9.12 shows this method. The start pulse, which is a line pulse, triggers the bistable and, at the

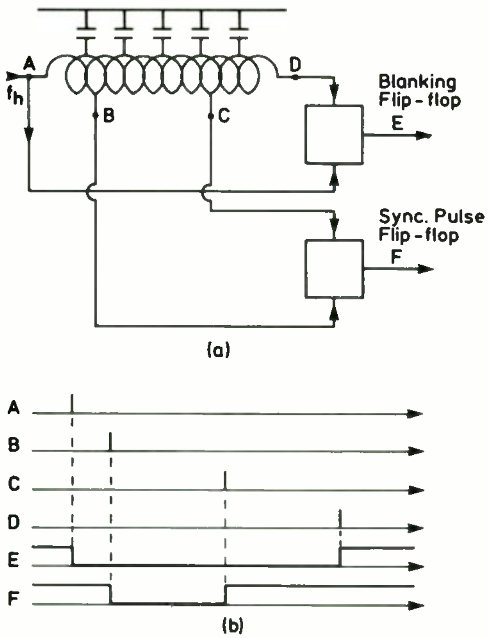


Figure 9.11 Generating pulses using a delay line

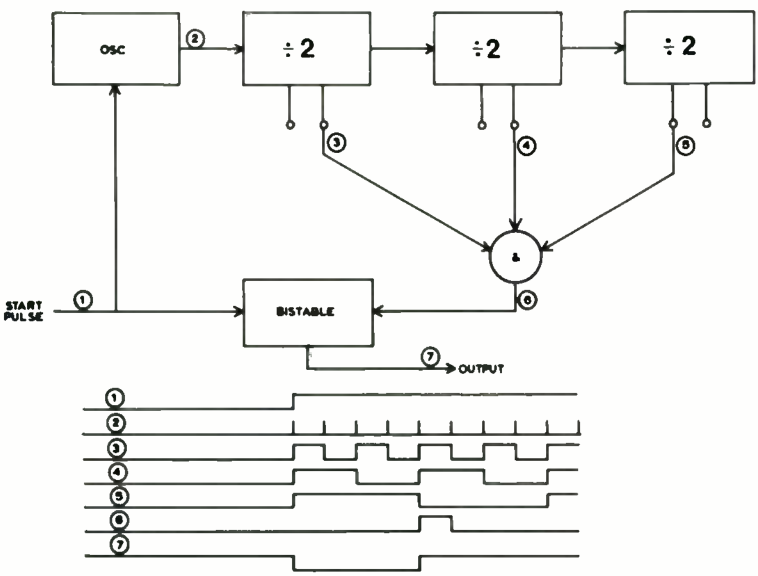


Figure 9.12 Generating pulses using a high-frequency crystal oscillator

same time, starts the oscillator. An AND gate is used to decode the binary counter waveforms. Depending on the connections to this AND gate, an output is obtained after a fixed number of oscillator pulses, in this case, five. The output from the decoding AND gate resets the output flip-flop and fixes the back edge of the pulse.

The chief merit of this scheme is the high order of stability of pulse duration that can be achieved, depending only on the oscillator stability. It has the slight disadvantage that the output pulse duration can only be changed by fixed amounts equal to the oscillator period.

In Section 9.3.1 on colour sync units we saw that part of the divider to f_{2h} was a divide by 227 circuit. This can be made to perform a second function in addition to that of dividing down to twice line frequency, and that is fixing the widths and relative timings of all the output pulses. Figure 9.13 shows the principle of operation. The divide by 227 is carried out by a programmable

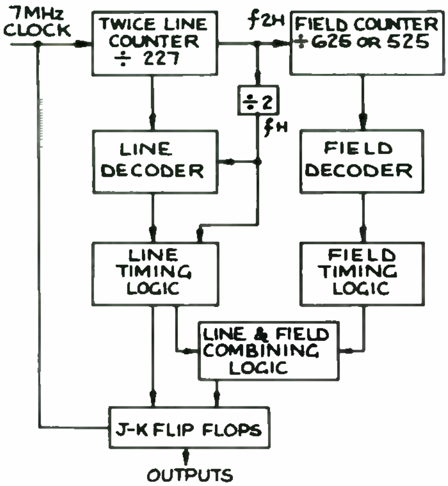


Figure 9.13 Generating pulses using the colour sync unit dividers

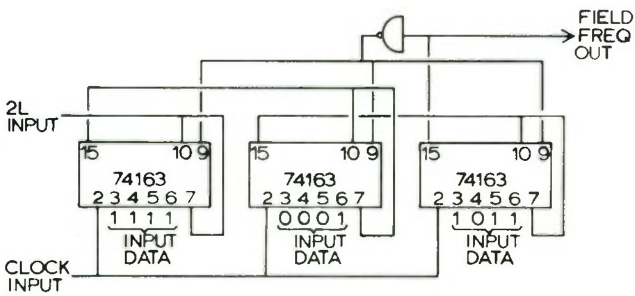
divider consisting of two 4-bit counter packages, which are automatically loaded with the number 29, that is 256 − 227, and which then count in 227 steps to 256. The eight outputs of the counter chain are connected to a pair of 4-line to 16-line decoders and, by gating together any pair of decoder outputs, pulses are obtained which are timed according to the connections to the decoder output and by the clock period of the counter input. Since the clock pulse frequency is approximately 7 MHz, pulse timings can be selected at intervals of 0.14 μs.

In order to generate an actual pulse, the decoder outputs are gated by a line-frequency signal to eliminate every second pulse, and then applied to the *J* and *K* inputs of a flip-flop. This flip-flop is clocked by the original 7 MHz input pulse train, which virtually eliminates any timing drift or jitter inherent in the counting and decoding process.

In the case of output pulses which contain field as well as line-frequency components, decoded pulses obtained from the 625 counter are combined with those from the twice line counter before being applied to the J–K inputs of the output flip-flop.

The field period of the SPG outputs is determined by a counter. The counter input is twice line frequency, and the division ratio is 625, giving $312\frac{1}{2}$ lines/field. A number of different circuits have been used for field dividers, e.g. synchronised monostable multivibrators, step counters, magnetic ferrite cores, etc.

Most field counters use a binary counter. Each stage of the binary counter divides by two so, for a chain of 'n' cascaded stages the division ratio is 2^n. The TTL MSI counters may be pre-set, so that the count starts not from zero, but from any number required. Figure 9.14 shows the implementation of a 625 counter.



$$1024 - 625 = 399$$
$$399 = 256 + 128 + 8 + 4 + 2 + 1$$
$$= 2^8 + 2^7 + 2^3 + 2^2 + 2^1 + 2^0$$
$$\therefore \text{DATA INPUT} = 110110001111$$

Figure 9.14 625 field counter using TTL MSI

The output of the divide by 625 may be decoded, similarly to the divide by 227, and used to fix the number of equalising and broad pulses, the duration of field blanking, field drive and burst blanking.

It is not practicable to transmit the burst during the field sync pulse as there is no back porch. However, if a burst blanking that is fixed with respect to the field pulses is used, the first burst to appear after the field blanking will alternate in sign from field to field. This alternation gives a small phase jitter to the reference oscillators which are separating the sub-carrier reference from the burst. In principle, this would give rise to a saturation flicker at the top of the picture. The effect can be avoided by shifting the burst-blanking waveform by one half-line each field, for four fields. In this way, the first burst to appear after the field period is always a $+V$ burst, as shown in Fig. 9.15. Similarly the last burst is always a $+V$ burst. This form of burst field blanking also provides a means of identifying the field number, which is useful when two signal sources are to be brought into synchronisation. The field which has its first burst closest to the end of the field synchronisation is field 3.

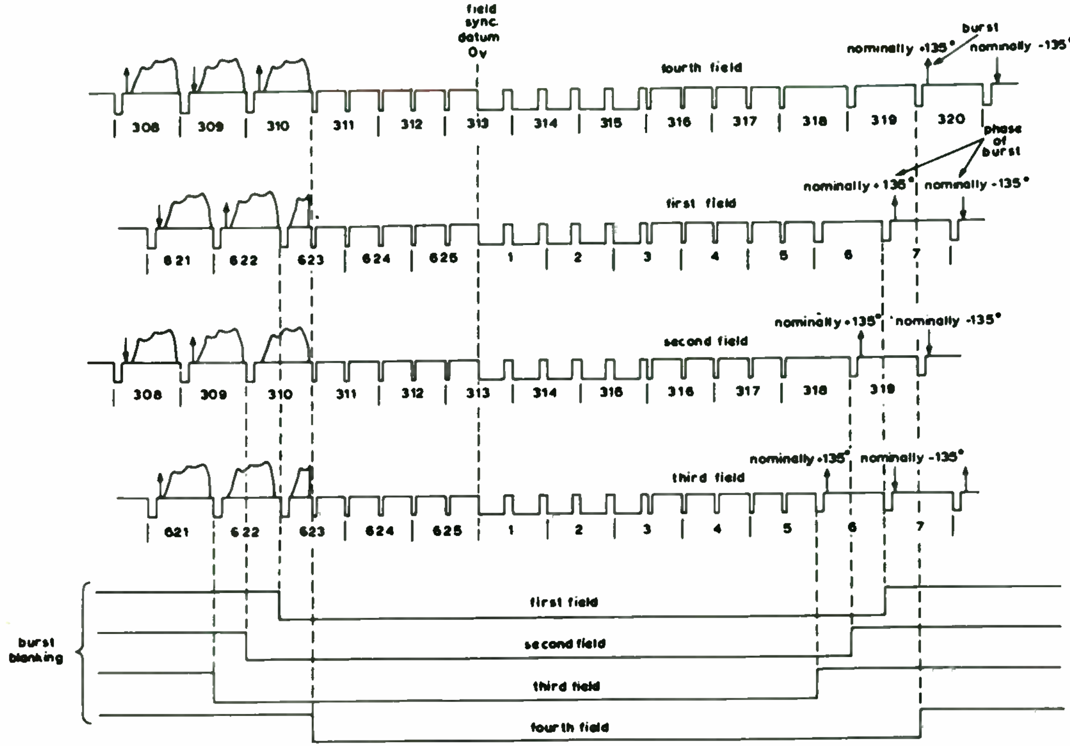


Figure 9.15 PAL: field blanking interval

9.5 MINIATURE SPG

With the advent of ENG, the demand for small, self-contained cameras has grown. It meant that the SPG had to be part of the camera chain. TTL large-scale integration has made possible the miniature SPG. The whole of the pulse-generating circuitry can be contained on one chip.

The Ferranti ZNA 134 IC utilises a 2.5 MHz crystal to generate all the line and field pulses, sync and blanking waveforms that are required for a monochrome SPG. It also operates on both 625 and 525 standards by having a mode-select input. However, for colour operation, a burst gate pulse still has to be generated and, as we have seen, the line frequency pulses have to be derived from the sub-carrier.

The mini SPG uses a technique known as 'fractional dividers' to count directly from sub-carrier to f_h. Fractional dividers use programmable counters to achieve, on average, a fractional division by dividing by two or more numbers in an appropriate ratio. For example, a division of 4.5 could be performed by dividing alternately by 4 and 5. A division of 4.333 could be performed by dividing by 4 and 5 in the ratio of 2:1 respectively. The problem is that the output exhibits a large amount of jitter, but this can be eliminated with a phase-locked loop (Fig. 9.16).

The sub-carrier can be broken into lower multiples:

$$f_{sc} = 4.433618.75 = 11 \times 25 \times 25 \times 644.89 \text{ Hz}$$

$$2f_h = 31{,}250 = 2 \times 25 \times 25 \times 25 \text{ Hz}$$

$$\frac{f_{sc}}{11 \times 25 \times 25 \times 644.89} = \frac{2f_h}{2 \times 25 \times 25 \times 25} = 1$$

i.e.

$$2f_h = \frac{f_{sc} \times 25 \times 2}{11 \times 644.89} = f_{sc} \times \frac{5}{11} \times \frac{1{,}000}{64{,}489}$$

$$2f_h = \frac{f_{sc}}{2.2 \times 64.489}$$

The first divider is pre-set to divide by 11, and produces five pulses at the QA output, thus dividing by 2.2 (Fig. 9.17). The second stage divides by 64.489.

Dividing by 64 and 65 alternately would produce 645. Dividing by 64 a few

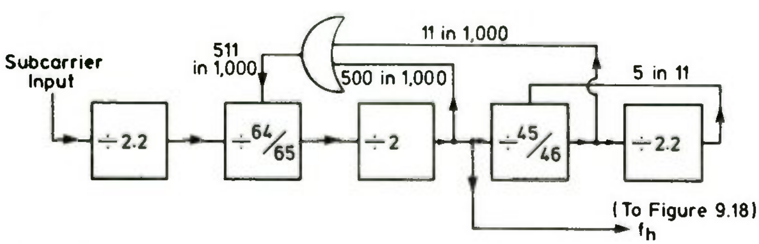


Figure 9.16

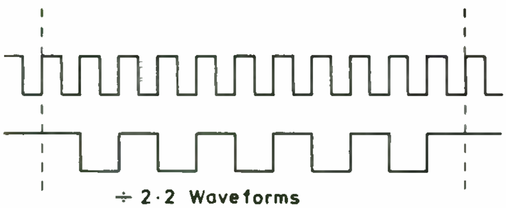


Figure 9.17 92.2 waveforms

more times will reduce the average count to the required value. In fact, the ratio of '64' counts to '65' counts is 511 : 489 = (500 + 11): (500 − 11); thus it requires 11 in 1,000 of the '65' counts to be changed to 64. The output of the divide by 64/65 is divided by 2 and fed back via the OR gate, to change the count alternately. The output of this divide by 2 also goes to the further stages to produce the 11 pulses in 1,000 necessary to modify the count.

To minimise jitter on the output, it is necessary to have the 11 pulses as evenly spaced as possible. Eleven in 1,000 requires a division ratio of 90.909. Since the first stage is a divide by 2, the counter following it is used to divide by 45.45. This is achieved by dividing by 45 for 6 counts in 11, and 46 for 5 counts in 11. As we have seen earlier, a divide-by-2.2 stage produces 5 pulses in 11 and, therefore, it is used to pre-set the divide by 45/46. The 11 pulses thus produced are gated with the output from the divide-by-2 stage to modify the pre-set inputs of the divide by 64/65.

The line frequency output of the fractional divider is used to control the frequency and phase of a 2.5 MHz crystal oscillator that, in turn, drives the SPG chip, as shown in Fig. 9.18. The loop also serves to eliminate the jitter due to the fractional divider.

9.6 SPG MEASUREMENTS

A simple way of checking for the correct operation of the SPG's dividing and multiplying circuits is to make use of the A/B facility of some digital counters. Figure 9.19 shows the sub-carrier connected to the A input, and field divided by 8, to the B input, therefore

$$\frac{A}{B} = \frac{4{,}433{,}618.75}{50} \times 8 = 709{,}379$$

In other words, the number of sub-carrier cycles in eight fields is displayed.

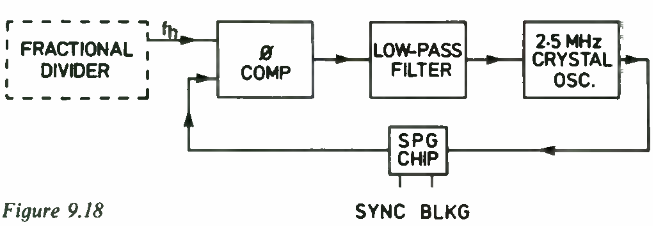


Figure 9.18

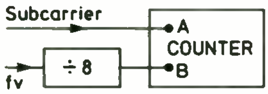


Figure 9.19 Checking the number of subcarrier cycles in 8 fields

With the advent of VTRs and the need for editing tape, more and more attention is being paid to the line–sub-carrier relationship. Because of the 25 Hz offset, the line–sub-carrier relationship is changing through an 8-field pattern. To make a clear edit, the field number has to be known, consequently the line–sub-carrier relationship has to be known. How is it measured? Clearly, strobing on an oscilloscope is out of the question for one sub-carrier pulse in eight fields. The method used by Marconi's is shown in Fig. 9.20. Field

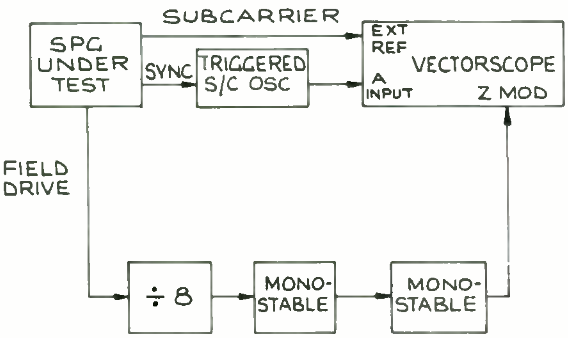


Figure 9.20 Measuring subcarrier-to-line phase relationship

drive is divided by 8, two monostables gate out one line in eight fields which is used for *Z* modulation of a vectorscope beam, thereby enabling the vectorscope display for that period of time. The sync output of the SPG is used to trigger an oscillator that runs at sub-carrier frequency, giving a burst of sub-carrier. This is connected to the *A* input of the vectorscope, where the external reference is connected to sub-carrier from the SPG. The display shows the sub-carrier–line relationship, and the size of the spot is an indication of line sync jitter. When using this test arrangement to test an SPG that already has the line–sub-carrier phase fixed, care must be taken to phase-lock the two divide-by-8 circuits.

9.7 GENLOCK

Turning now to genlock. It is frequently necessary to lock the picture from the studio cameras to one coming from a remote source, for example, on outside broadcast. Locking the two sources together enables mixes and special effects to be carried out in the same way as between the studio cameras themselves.

Genlock is the most commonly used method of locking two sources and is usually in the form of an optional extra attachment to an SPG. The composite video signal from the remote source is fed into the unit and the line and field sync pulses on this video are used to control the SPG in such a way that the local pulses are locked in both frequency and phase to the remote signal.

The locking process can be considered in two parts, line locking and field phasing. When genlock is fitted, the 31 kc/s twice line signal is obtained, not from the SPG master oscillator, but from a locked oscillator in the genlock unit. This oscillator forms part of a phase locked control loop and this can be locked to either the line sync pulses on the remote video or to a local reference line signal. The locking range of the oscillator is usually a few hundred cycles either side of the normal line frequency. A line phasing control is usually provided to enable the leading edges of the local and remote syncs to be accurately aligned.

Once line locking is complete, the local and remote field pulses will also be locked in frequency, but it is still necessary to adjust the phase of the local field pulses. This is done by making the field counter miscount. If the division ratio of the counter is changed slightly from 625, or if extra pulses are added or taken away at the counter input, which amounts to the same thing, then the local field frequency will be slightly different from the remote and the field pulses will drift into phase. When they come into phase, the field counter reverts to normal operation.

In order to preserve an interleaved picture during field phasing, the field counter must be modified by an even number, and to keep the field disturbance to a minimum, the count is usually modified by 2, i.e. 625 is changed to 623 or 627. This gives a field phasing rate of one line per field.

Field phasing takes place via the shortest route. That is to say the count is increased or decreased depending on the relative phase of the local and remote signals. Half a field is, therefore, the greatest error that has to be corrected. The time taken for this in the worst case is

$$\frac{312\frac{1}{2}}{2} \times 20\ \text{mS} = 3.125\ \text{s}$$

In normal operation f_{2h} passes through the inhibit gate only and the count is divide-by-625 (Fig. 9.21).

During phasing, the local field gating pulse falls within the +ve or −ve part of the remote field square wave and causes pulses to be removed from or added to the f_{2h} feed to the divide-by-625. During the final stage of phasing, the count may need changing by 1, therefore the field gating pulse is foreshortened.

One of the most important features of genlock design is that of minimising the picture disturbance during the genlocking process. It is important to remember that some disturbance is unavoidable since the local and remote signals can only be brought into synchronism by changing the frequency of the local SPG. The line locking disturbance can be minimised by restricting the locking range of the genlock and by taking a longer time to complete the locking.

With XTAL controlled SPGs, the line frequency is very accurate and the genlock locking range can be narrow. This contrasts with mains locked SPGs where the variations in line frequency can be a few percent.

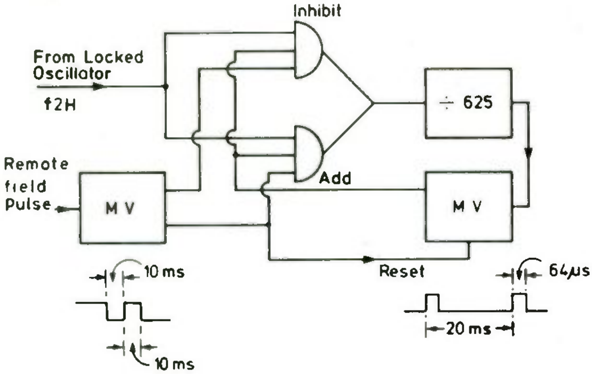


Figure 9.21 Field phasing

The picture disturbance during genlocking is very dependent upon the viewing monitor or receiver. Direct lock valve monitors were very insensitive to line locking and the disturbance only occupied a few lines of the picture. Flywheel sync receivers, however, cause the disturbance to be drawn out over a much longer period, a few fields in some cases.

One system which is most effective in reducing line disturbance is to make the line lock cuts in the genlock only operate when the local and remote pulses are in phase. This technique is used in the colour genlock unit described in the next section.

9.8 COLOUR GENLOCK

In addition to the requirements already mentioned for monochrome, i.e. line locking and field phasing, two further processes are necessary for colour. Firstly, the local sub-carrier must be locked to the reference burst of the remote signal, and secondly, the PAL phase alternation of the local and remote signals must correspond.

Figure 9.22 shows a block diagram of a sub-carrier locking unit.

The local sub-carrier from the colour sync unit, instead of being distributed directly, is taken through the colour lock unit. It passes through the electronic switch, shown in the local or non-genlocked position, to a phase locked XTAL oscillator. This oscillator provides sub-carrier for distribution to the local colour coders. The advantage of using this oscillator is that it ensures a smooth and controlled phase transition when switching between local and remote sub-carrier sources. This is because of the characteristics of the phase locked loop and the high Q of the XTAL oscillator.

The remote phase reference is provided by the burst, separated from the remote video input. In PAL the burst phase swings $\pm 45°$ on alternate lines and a filter is included in the XTAL oscillator control circuit to ensure that the oscillator locks to the average phase of the remote bursts. When the genlock button is pressed on the control panel, a signal is passed to the AND gate. A

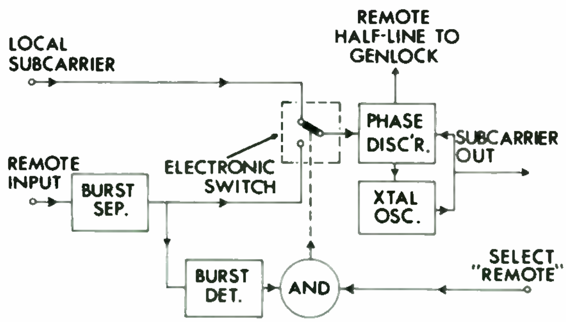


Figure 9.22 Colour genlock system, subcarrier locking

burst detector also feeds the AND gate, ensuring that the electronic switch will only operate if burst is present, and will at once return the oscillator to local lock should the remote signals or burst fail. With both signals applied to the AND gate, the electronic switch operates and the sub-carrier lock is transferred to the remote burst.

As is the case with PAL receivers, the phase discriminator gives out another signal. This is a 7.8 kc/s half-line frequency which is passed to the genlock unit where it is used to fix the correct PAL line alternation.

In the genlock unit the twice line-frequency locked-oscillator is controlled, not at line frequency, as in monochrome, but at half-line frequency. Figure 9.23 shows the half-line frequency locking circuits. The twice line frequency from the SPG master oscillator is fed into the genlock unit where it is divided by 4 to produce the local reference signal. For colour operation this master oscillator twice line signal is derived from the colour sync unit.

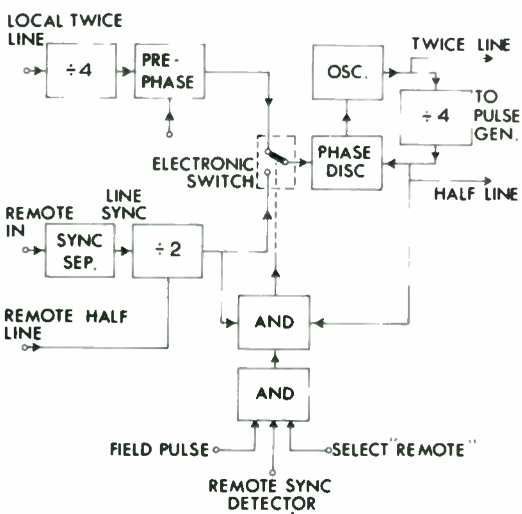


Figure 9.23 Colour genlock, line locking

The local half-line frequency signal passes through a pre-phasing stage and then to an electronic changeover switch, shown in the local lock position. The purpose of the pre-phasing stage is to bring the local half-line frequency signal into phase with the remote. This pre-phasing stage automatically introduces an increasing phase change into the path which reaches 360° in approximately 3 s, and starts to operate when the genlock or local lock buttons on the remote control panel are operated.

From the electronic switch the half-line frequency signal passes to a phase discriminator which locks an LC oscillator. The LC oscillator drives the SPG and, like the locked oscillator used for sub-carrier, allows a clean transition to be made between local and remote locking. Abrupt changes in phase are prevented by the properties of the phase locked loop.

With the electronic switch in the remote position, the half-line frequency signal is obtained from the remote input. Line sync pulses are separated from the remote input and drive a divide-by-2 stage to give a half-line frequency square wave. The correct phase of this square wave is determined, in PAL, by forcing it into step with the half-line frequency signal from the colour lock unit.

The operation of the electronic switch is controlled by a logic system represented here by two AND gates. By this means the electronic switch is prevented from operating until the phase error between local and remote signals has been reduced to a very small amount by the action of the pre-phasing circuit.

To run through the locking sequence: when the genlock button on the control panel is pressed first, sub-carrier locking takes place, and for PAL the half-line frequency derived from the remote syncs is brought into phase, as defined by the swinging burst. Next, the remote AND gate is activated and the pre-phasing sweep starts. When the remote and local half-line frequency square wave have been brought into phase, the second AND gate produces an output, and changes the electronic switch to the REMOTE position.

In order to further reduce any picture disturbance, the changeover of the electronic switch is made to take place during field blanking by means of a field pulse fed into the AND gates. Changeover from genlock to local lock is carried out in a similar way.

The purpose of the remote sync pulse detector input to the lower AND gate is to return the system automatically to local lock if the remote signal fails.

Field phasing is carried out in the same way as for monochrome by miscounting.

The total time taken to complete the genlocking operation depends on the relative phasing of the local and remote signals. In the worst case it will take 7 s, with sub-carrier locking, 1 s, line locking, 3 s and field phasing, 3 s.

9.9 GENLOCK: PAL AND NTSC

As already mentioned, genlocking is accomplished by three separate processes. First, the sync generator sub-carrier is locked to the reference burst of the incoming video signal. Second, the line phase of the local pulses is

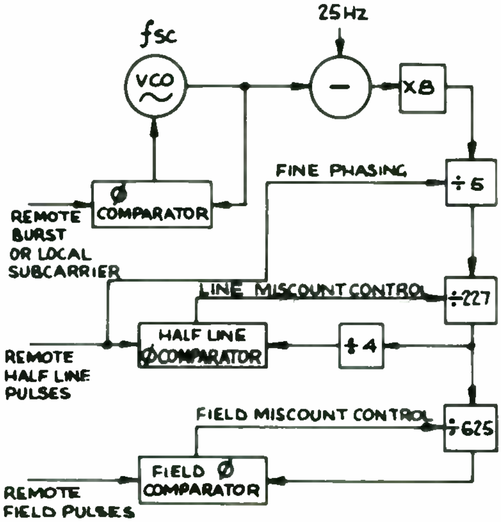


Figure 9.24 Colour genlock system

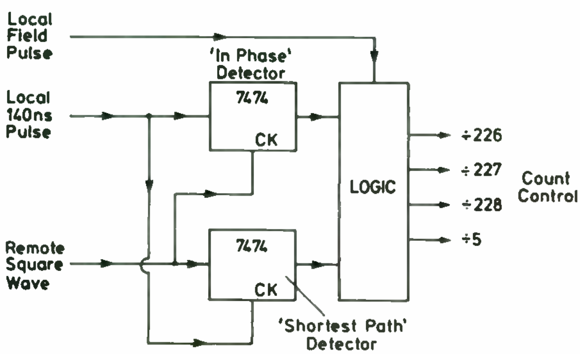


Figure 9.25 Line phasing, colour genlocking

adjusted to align them with the remote line syncs, and third the phase of the field pulses is adjusted in a similar manner.

The line-phasing operation is effected by changing the counting ratio in the divide-by-227 and divide-by-5 stages until the local sync pulses are aligned with the remote signal. When this is achieved, the dividers automatically revert to their normal operation. Figures 9.24 and 9.25 explain the operation of the line-phasing circuit. Two separate detectors compare the phase of the two half-line-frequency signals, one a 140 ns local pulse and the other a square-wave derived from the remote video signal and properly related in the case of PAL to the line alternation sequence. The square wave is correctly phased by reference to the PAL alternating burst. Comparison at half-line frequency in this way ensures that the sequence is correctly established at the local SPG.

The outputs from the two detectors give the three possible states of the phase of the local and remote pulses: (1) in-phase; (2) out-of-phase 'early'; and

(3) out-of-phase 'late'. The two different 'out-of-phase' signals indicate whether the frequency of the local pulses should be decreased or increased in order to achieve the 'in-phase' state in the shortest time.

Logic circuitry combines the outputs of the two detectors to provide control signals for the divide-by-227 counter to give division ratios of 226, 227 or 228, corresponding to the three possible states of phase. A field pulse is also fed to this logic circuitry to ensure that the count only changes during three lines of the field blanking interval, at all other times division is maintained at its correct value of 227.

Since the frequency of the clock pulses feeding the divide-by-227 counter is 7 MHz, variation of the division ratio to 226 ot 228 changes the phase of the local pulses in steps of 140 ns, one clock pulse period. In order to achieve a closer phasing of the local and remote pulses, the line phase comparator also controls the divide-by-5 counter from which the 7 MHz pulses are obtained, thereby improving the phasing accuracy to 28 ns.

9.10 GENLOCK: MONOCHROME AND SECAM

For operation on these standards the colour block pulse unit is replaced by a mono clock pulse version, in which the 7 MHz clock pulses are derived from an oscillator locked to the remote line sync pulses. Having locked the line and field frequencies, the local pulses are phased to the remote signal the same way as has been described for PAL and NTSC operation.

When genlocking to a SECAM signal one additional operation is necessary to ensure synchronisation of the local and remote pulses—the SECAM colour sequence identification pulse must be correctly related to the incoming D_R and D_B modulated pattern. This operation is performed in a SECAM colour lock board, replacing the colour lock used in PAL and NTSC operation. This ensures that the remote half-line frequency square wave for the phase comparator is correctly phased for the identification lines in the field blanking interval of the incoming SECAM signal.

REFERENCES

1. Iles, G. D., 'A digital sync pulse generator', *Communications and Broadcasting*, p. 38, Autumn (1976)
2. Kitchen, H. D. and Tucker, A., 'Monosync pulse coding', *Sound and Vision Broadcasting*, pp. 23–32, Vol. 9, Winter (1968)

Chapter 10

Timing a colour studio system

R. M. Stevens

10.1 INTRODUCTION

A television studio system should be designed so that all the vision signals arriving at each mixing point (such as a studio, presentation or outside broadcast vision mixer, a video tape editing mixer or a network switcher) are synchronous in line, field and sub-carrier phase and frequency. If synchronism of any of these parameters is not achieved between two or more sources, then techniques of mixing, wiping or special effects between them cannot satisfactorily be achieved. For example, cutting between sources that are not in field synchronism will produce a frame roll on the home viewer's television screen.

Further, if the line sync pulses of two vision signals are out of synchronism by only 1 μs (equivalent to about 20% of the sync pulse width) and those two signals are mixed together, the picture will jump sideways by 8 mm during the mixing operation on a 510 mm diagonal television screen. If the sub-carrier of the two signals is out of phase, then, during the mixing operation, the combined signals will be adversely affected in hue and saturation—in the extreme case, when the two signals are 180° out of phase, the resulting picture could lose colour altogether, depending on the operation of the mixer. Of course, two signals may be in line, field and sub-carrier phase synchronism at any given time but, if they are not in frequency synchronism, they will gradually drift out of phase. The speed of drift will depend on the difference in frequencies.

In addition to these immediately visible effects of mistimed signals, a videotape recorder will be unable to satisfactorily record and therefore replay a programme if the act of switching or mixing between the various vision sources during recording produces discontinuities in the final signal.

10.2 TOLERANCES

To overcome the problems outlined above some general tolerances are accepted for the phase or timing errors between vision signals at a mixing point. Signals within these tolerances can be manipulated in a vision mixer or recorded on a VTR without any deleterious effects. Line sync is measured at the half amplitude point of the leading edge of line sync—this is the point which triggers the flyback in a television receiver and which is, therefore, used as the timing reference. For eight signals arriving at a mixing point the tolerances are:

Leading edge of line sync pulses—within ± 25 ns
Sub-carrier phase—$\pm 1.5°$ (equivalent to about ± 1 ns).

The field synchronism must be absolute, i.e. field 1 of each signal must be in phase with each other.

10.3 ACHIEVING SYNCHRONISM

Essentially, frequency synchronism of vision sources in a television system is achieved by triggering them with the synchronising pulses and sub-carrier from a central sync pulse generator (SPG), and phase synchronism is achieved by ensuring that the various path lengths from the central SPG, via the different picture sources (cameras, telecines, VTRs, etc.) to the mixing points, are equal.

The best way of achieving these criteria in a television system depends on many factors and it will not be possible in this chapter to give details of how any particular station can be timed. The optimum timing solution for a small provincial station or a facilities house, with perhaps only one studio, will be different from that employed in a large regional television complex, with many studios, telecines, VTRs and presentation facilities. Similarly the geographical layout of the technical areas (perhaps new facilities have been built in new premises some distance from the main complex), the production uses of the facilities (news and sport requirements will differ from drama—particularly in the number of inputs from remote sources) and the age of the complex will influence the timing technique to be employed. For example, some regional stations in the UK, built towards the end of the 1960s, still use techniques that would not be considered today.

10.3.1 A typical studio centre

Figure 10.1 shows a typical studio complex in block diagram form. The videotape and telecine machines are generally located in a central area and their functions assigned to the various studios or presentation as programme schedules demand. Cameras tend to be associated with one particular studio. It can be seen from this diagram that there needs to be two 'timing levels'—one at the inputs to the studios and one at the input to the presentation mixer. All equipment and cable introduces a delay into the vision and pulse paths (from

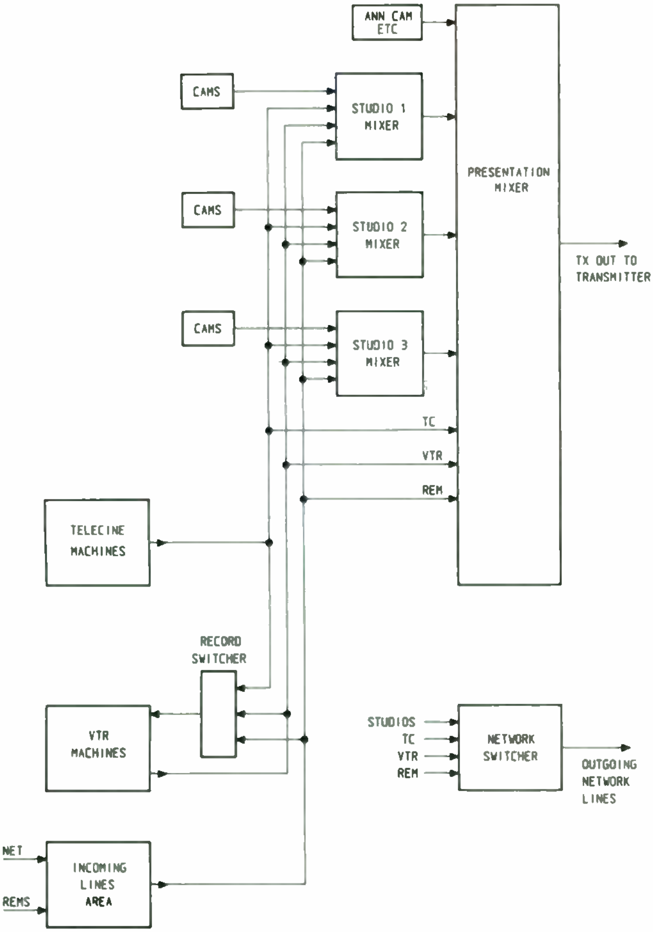


Figure 10.1 Block diagram of a typical studio centre

about 25 ns for a distribution amplifier to perhaps 600 ns or more for a vision mixer) and so a telecine, for instance, that passes through a studio mixer will be delayed at the input to the presentation mixer with respect to a telecine that has been assigned directly to presentation. If these two telecines are fed with the same pulses down identically timed paths they will not be in synchronism when they reach presentation because of their differing vision path lengths. Hence signals assigned to presentation need to be delayed with respect to signals feeding studios.

The question arises of where in the signal path should the delay be put in order to achieve the two timing levels. The choice is between inserting it in the vision path feeding the presentation mixer or inserting it in the pulse feeds to the telecine, where it would be switched out of circuit when the source is feeding the studio. As far as possible the latter course should be adopted as it is a good policy to interfere with the vision signal as little as possible. Any form of vision delay requires equalisation and amplification to restore the original signal level and, whilst this may not in itself contribute in a major way to a deterioration in the stability of the signal or its parameters, the effects will be added to the distortions occurring elsewhere throughout the system. It is better to include the delay in the pulse and sub-carrier feeds, where it cannot influence the vision parameters.

10.3.2 Timing codes

Sync pulse generators provide six different pulse waveforms and sub-carrier. Some old equipment still in use will require, perhaps, all these pulses but equipment being designed in more recent years may only need sub-carrier, mixed line and field syncs, mixed line and field blanking, and burst gate pulse. Even so, with only these four signals, it would not be easy to distribute them to all picture generating equipments, delay them as necessary, switch the delays in and out and switch the pulses from various SPGs to the machines as the assignments vary. To ease this problem various coding systems were designed to code information about the start and finish times of the various pulses onto a single cable.

The waveform of the system most widely used in the Independent Television network is shown in Fig. 10.2. It has no l.f. content and thus hum and l.f. distortions can be removed by simple filtering. The sine squared pulses, which time the start and finish of the television pulses, have a bandwidth of approximately 3.3 MHz and thus do not interact with the two 10 μs bursts of sub-carrier. The coder generates a delayed and an undelayed set of pulses for driving studio and presentation timing level sources. In the decoder the sub-carrier bursts are removed by filtering and continuous sub-carrier is regenerated in a phase locked oscillator with high inertia. A sub-carrier phase shifter can be included in the output for adjusting the phase of the picture

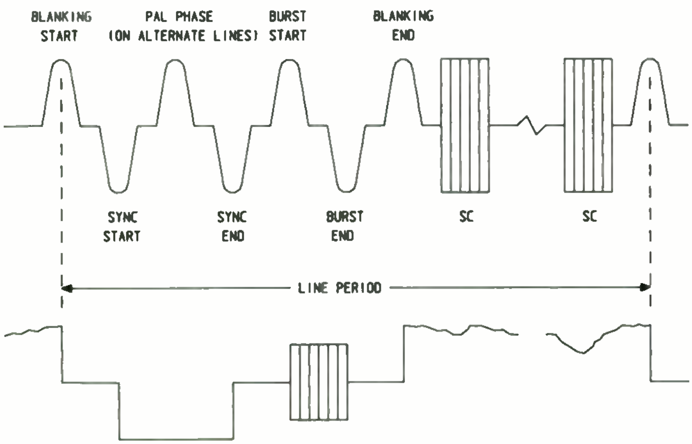


Figure 10.2 Coded pulse line period waveform

source being fed. The pulses pass through a tapped lumped delay line for system timing adjustments, and are decoded in logic circuitry into the constituent pulses.

10.3.3 Cable

The items of vision equipment in a television system are connected together using coaxial cable. The delay and attenuation of the cable are the two parameters which affect the design and timing of a studio system. The cable most widely used has the following parameters:

(1) The delay caused by 1 m of cable = 5 ns = 8° of sub-carrier.
(2) In addition, a change in temperature of the cable will alter its electrical length. If a 100 m length of cable has its temperature altered by 10°C, the subsequent phase change of the sub-carrier will be about 5°.
(3) The attenuation of 100 m of cable at sub-carrier frequency is about 2 dB. Hence signal paths over 10 m in length should be equalised.

10.3.4 SPG assignments

In the typical studio system diagram (Fig. 10.1) it can be seen that, in addition to the local picture sources, which are fed with pulses from the station SPG, there may be remote inputs to the system, such as outside broadcasts or a feed from network. Using the SPG in the 'genlock' mode, where it takes its timing and frequency reference from the remote input and locks its generation of pulses to it, the local sources can be synchronised to the remote, but the act of genlocking—the changing from the local to the remote reference—disturbs local station pulses, which in turn upsets any VTRs connected to them. It is therefore necessary to have multiple SPG systems, so that one SPG can genlock while others feed other areas of the system.

One system provides an SPG per area so that each studio has its own SPG, and the operation of genlock only affects the picture sources assigned to that studio. Studios can be timed into presentation by presentation genlocking to the studio. It is only with the advent of integrated circuits that stable SPGs can be made relatively inexpensively. Over a decade ago SPGs were less reliable and in some cases an SPG assignment system was found to be more appropriate as it gave greater flexibility and reduced the number of SPGs required. A block diagram of such a system is shown in Fig. 10.3. The pulse coding and phasing details have been omitted for clarity and the block marked with a 'D' not only refers to a delay in line pulse timing but also a change in sub-carrier phase. Each area can be fed from any SPG through the 'SPG assignment' matrix and therefore there is no need to have a reserve standby SPG. In addition to the source machines receiving the pulses from the studio or presentation area with which they are working through the 'pulse matrix', a feed of the stable master pulses is also made available to each machine for use during post production activities such as film to tape transfers, editing, etc.

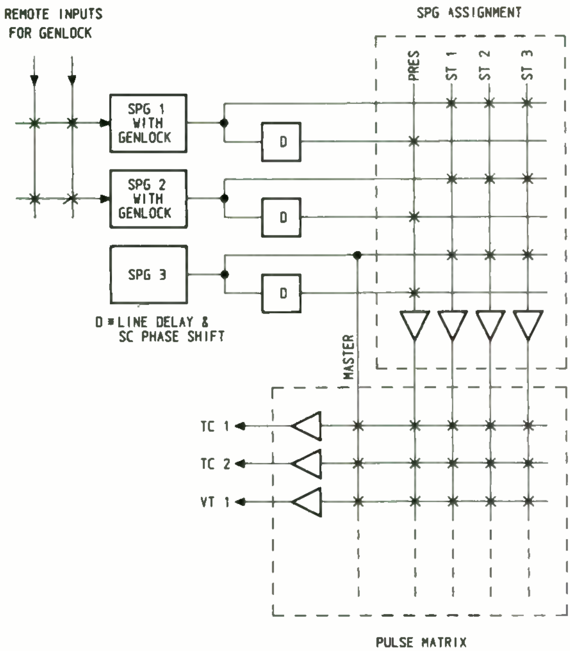


Figure 10.3 SPG assignment system

10.3.5 The timing chart

Armed with the foregoing knowledge of techniques, the studio system can now be timed. Decisions will have already been made on the number of studios, telecines, VTRs, etc. and the layout of the equipment with cable runs, etc. It is appropriate, however, at this stage to make allowance for future expansion of the system. A system block diagram can then be prepared, as shown in very simplified form in Fig. 10.4. The main pulse paths—in this case using a pulse coding system—and vision paths can be shown, together with the positions of pulse delays and sub-carrier phase adjustments, so that the sources can be timed at the various time coincidence levels. Here is shown an assignable telecine and a dedicated camera working into a studio and an assignable VTR and a dedicated announcer's camera working into presentation.

From this block diagram a timing bar chart can be prepared. A simplified chart, following the system block diagram of Fig. 10.4 is shown in Fig. 10.5. The delays are shown in nanoseconds and are examples to give some idea of the practical delays in a system. The chart is best drawn to scale with a suitable scale representing time. The delay through an item of equipment is usually obtained from manufacturers' data and is measured between the half amplitude point of the leading edge of line syncs into and out of the equipment. A detailed bar chart would show the delay down every piece of cable in the system but, for simplicity, a lumped delay for the total of all the lengths of cable has been included in Fig. 10.5.

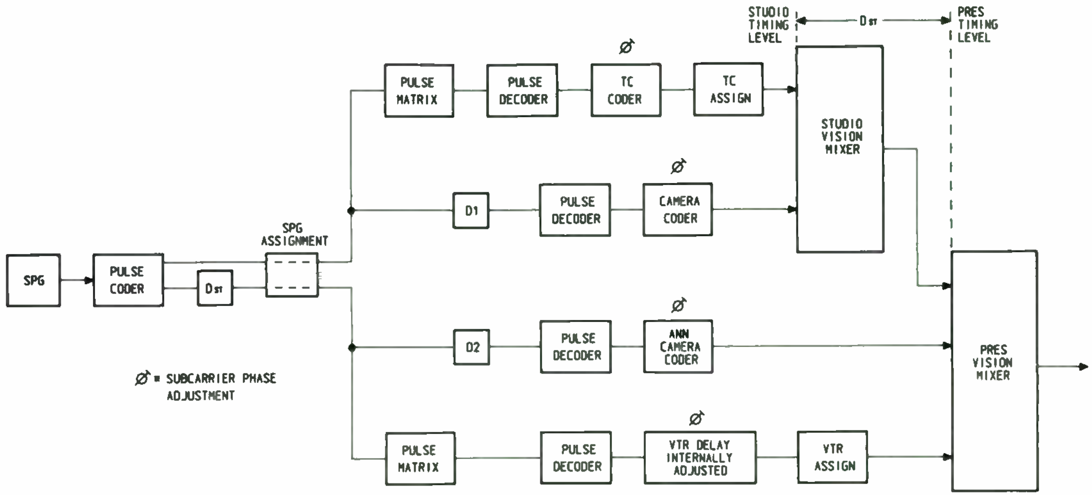


Figure 10.4 System block diagram with pulse coding

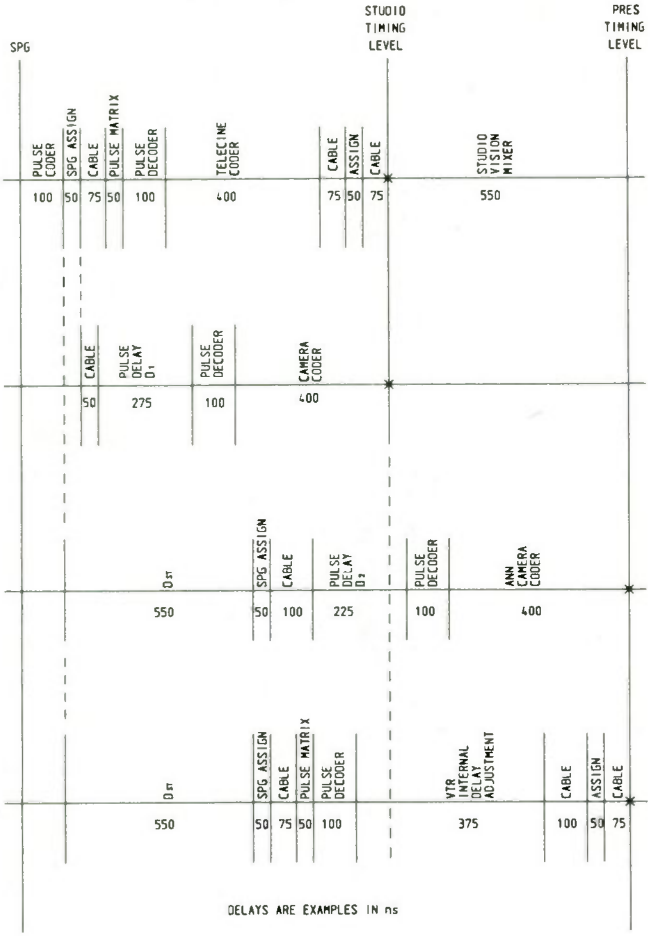


Figure 10.5 Timing bar chart

In order that the leading edge of line syncs of all sources can be within ± 25 ns of each other at the input to each mixing point, all paths from the SPG to these mixing points must be of equal length with a tolerance of ± 25 ns, which is equivalent to a tolerance of ± 5 m of coaxial vision cable. Hence the longest natural path in the system must be found and the other paths delayed to the same length. In the bar chart in Fig. 10.5 the telecine has the longest path. It may, however, be advantageous to put some delay into even the longest path to provide flexibility if, in the event of the later installation of new equipment, an even longer path is created. The new source could then be timed without any need to alter the existing installation.

Having prepared the bar chart, the equipment installation can be carried out and final timing adjustments made. The timing bar chart will narrow down delays to perhaps the nearest 50 ns, but some changes will have to be made on the practical system. The equipment should be allowed to reach its operating temperature and all pulse and vision levels and frequency responses should be set before detailed timing is started. Synchronism should be checked at the actual input to the mixer and not just at the nearest patch panel, as cable lengths between patch panel and mixer may not be identical.

10.3.6 Sub-carrier phase

After the line timing has been completed the sub-carrier phasing has to be adjusted. From Fig. 10.4 an arbitrary phase reference is taken at the first mixer—perhaps camera 1 or the longest telecine path—and the phases of all the other contributing sources are adjusted to be the same ($\pm 1.5°$) into the mixer. The vision path from each assignable machine (in this case the telecine) should be made the same length through the assignment and distribution system to each studio, in which case its phase will be the same into each studio. The other dedicated sources can then be adjusted to the same phase. The telecine should then be fed through the studio system and into presentation. Its phase on arrival at presentation will be the reference phase of presentation and the local sources can be adjusted to be the same. If the telecine is now fed directly into presentation it will be out of phase (unless the delay through the studio was exactly a whole number of sub-carrier cycles). However, the delay D_{st}, which is included in the pulses for the telecine when it is assigned to presentation, should also include a sub-carrier phase shifter, which can be adjusted purely for sources feeding presentation and thus bring them back into phase at the presentation mixer.

10.3.7 Measuring errors

The timing of line sync edges can easily be measured on a good oscilloscope or television waveform monitor, triggered from external syncs. The timing of sub-carrier can be measured on a vectorscope display or a phase comparator, which drives a moving coil meter with an error voltage. If the line sync delay between pulses from the SPG and the input to the mixer is, for example, 1,175 ns, then, at first sight, there would seem to be a need to delay the sub-carrier by the same amount. This is not the case, however. If the sub-carrier is delayed by 1,175 ns this is equivalent to about 5.2 cycles of sub-carrier but, as sub-carrier is a sinusoidal repetitive signal, it need only be delayed by 0.2 cycles, which is about 45 ns. It is usually not necessary to delay it as most equipments, such as PAL encoders, VTRs, etc. include some form of phase shift circuitry.

10.4 GENLOCK AND SPG STABILITY

Earlier mention was made of the technique of 'genlock' to time a remote signal into a system. The SPG can only be genlocked to a single source at a time and this restricts the action of the studio for major events—such as sports programmes—where with genlock alone it is not possible to wipe between two remote sources. Various techniques have been devised to allow the simultaneous synchronisation of more than one remote source.

If it were possible to make SPGs that are so stable that they run at exactly the same frequency, then genlock would not be necessary. Provided phase synchronism could be achieved before the start of the broadcast the various remote sources would remain synchronous. Unfortunately this is not the case. The CCIR PAL specification requires a sub-carrier frequency of 4,433,618.75 ± 1 Hz. Hence two SPGs on either limit of this tolerance will drift apart by 1 television line (64 μs) in just over 2 min. Average SPGs have a stability of a few parts in 10^{-8} and, from Fig. 10.6, it can be seen that the line sync timing of equipments on the edge of this tolerance would drift out in only a few seconds. Using a rubidium frequency standard with a stability of a few

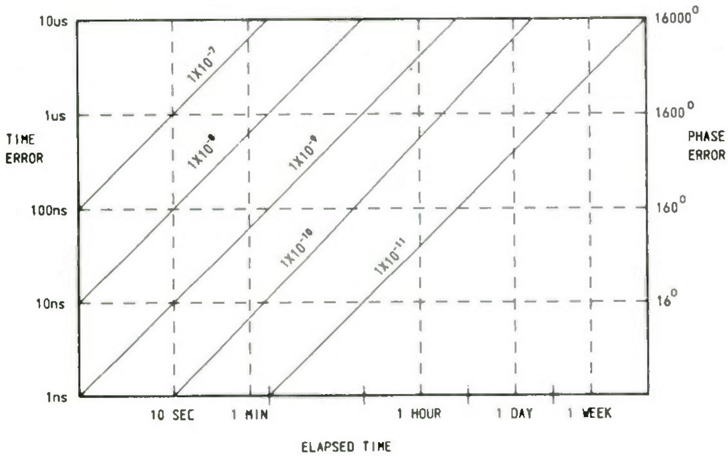


Figure 10.6 SPG stabilities

parts in 10^{-11} (equivalent to a clock error of 1 s in about 600 years) it would take about 1 h for the line sync phase to drift out of the ±25 ns tolerance but the sub-carrier tolerance of ±1.5° would be breached in less than 2 min. Hence the stability of SPGs alone cannot be relied upon. Variable delay lines or modified time base correctors can be used in conjunction with high stability SPGs to remove their drift, however.

Figure 10.7 shows the difference between genlocking and reverse locking techniques. In reverse locking the local station analyses the drift between the local and remote pulses in comparator circuits and sends an error signal to the remote SPG which controls the sub-carrier frequency and phase and the line and field phases. Firstly the remote source is moved into synchronism by a series of coarse corrections and then synchronism is maintained by occasional fine corrections. The error signal may be fed back to the remote source either by land line or by radio techniques. The advantages of this type of system are twofold: firstly more than one source can be synchronised at any one time and secondly the synchronising process can be made sufficiently slow so that it does not affect VTR recordings. The disadvantage is that a return path has to

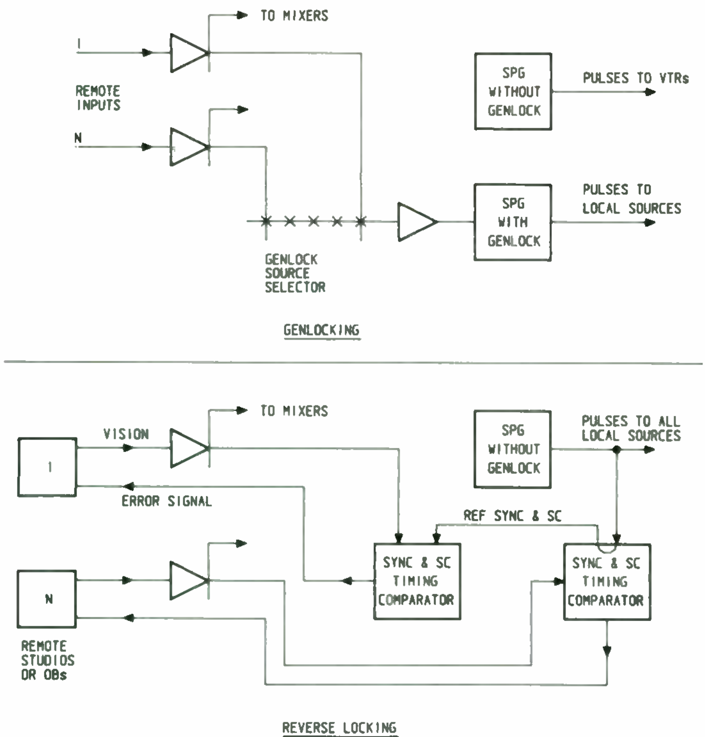


Figure 10.7 Principles of genlocking and reverse locking

be provided for the error signal and, if this is broken for any reason, synchronism is lost. BBC Natlock is an example of this technique.

10.5 MODERN DEVELOPMENTS

Consideration has been given thus far to techniques that have been in use in colour television studios for over a decade. In recent years further techniques have become available and, whilst the older techniques are still in use in a number of UK television stations, the new techniques are being incorporated into new stations and considered by those involved in the updating of the older systems.

10.5.1 Sub-carrier-to-line phase relationship

Ideally a pulse system should have the following characteristics:

(1) If it fails it should not cause a major system failure involving the loss of all picture sources. The SPG and coded pulse systems mentioned above partially conform to this. If a pulse assignment path or pulse decoder fails

then only one source is lost, but if an SPG, its assignment system, or a pulse encoder fails, there is a major system failure until standby equipment is switched in.

(2) It should be operationally flexible in that it should be possible to correct easily for drifts in the system, due perhaps to temperature or equipment age. This is possible for sub-carrier phase as it can be adjusted in the phase shifters, but adjustment of line timing involves the alteration of tappings on lumped delay lines or, perhaps, the inclusion of a suitable length of cable in the pulse path.

(3) It should be easily modifiable to accommodate the ever changing requirements for equipment facilities. Again the systems already described can accommodate sub-carrier changes easily, but it would be convenient to be able to change the line timing of each source by operating a single control.

In addition, modern VTR editing techniques have shown up another problem with the timing of pulses. Whilst in the 625-line PAL specification there is a fixed relationship between the *frequency* of line syncs and sub-carrier, there is no such relationship between the *phase* of line syncs and sub-carrier. Up to now it has been possible to adjust the sub-carrier phase independently of the line sync phase—in fact these adjustments have been purposely separated.

However, with the increased sophistication in VTR editing, it has been found that if two programme sequences are edited together each with a different phase relationship between line sync and sub-carrier, there will be a disturbance at the edit point when the programme is played back. Differing phase relationships can be caused by a number of factors:

(1) The originating SPG can either drift in phase between line sync and sub-carrier or can take up a new phase after being switched off, so that a recording made one day may have a different phase to that recorded on a subsequent day.

(2) A programme may comprise edited items recorded in the studio and on location by an outside broadcast vehicle. This will involve editing together material from two SPGs—each of which will have a different phase relationship.

(3) Programme content may even be recorded in two studios and, whilst the same SPG may provide the pulses, a different phase relationship at the mixer inputs may exist for each studio.

What is required, therefore, is a further specification parameter for the 625-line PAL system to define the phase relationship between sub-carrier and line sync at mixing points. The EBU has produced the following recommendation:

'The sub-carrier-to-line sync phase is defined as the phase of the E'_u component of the colour burst extrapolated to the half-amplitude point of the leading edge of the synchronising pulse of line 1 of field 1. In the definition of field 1 of the eight PAL fields, the EBU has adopted the value of 0° for the central value of the sub-carrier-to-line phase.

For maximum protection against picture disturbances on edit points, the EBU recommends that the video signals on tapes to be used in sophisticated

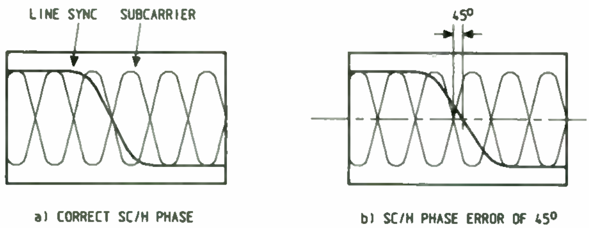


Figure 10.8 Sub-carrier-to-line phase errors

editing procedures are recorded with a value of 0° ±20° of the sub-carrier-to-line phase.'

This is shown diagrammatically in Fig. 10.8.

However, it is not sufficient for the SPG to be designed to this standard. The SPG needs to have a stable sub-carrier-to-line phase but the EBU recommendation needs to be met at the mixing point. If it is met at the SPG and the sub-carrier phase of a PAL encoder is altered it could take the sub-carrier-to-line phase outside the ±20° window. The recommendation must, therefore, be monitored at the mixing point and systems adjusted in line sync to sub-carrier phase to meet it. Suitable measuring and monitoring devices can now be obtained.

In the coded pulse system previously described, the sub-carrier was separately timed from the line syncs. It was thus not possible to maintain a fixed phase relationship between the two. In recent years, however, the coded pulse system has been replaced by a colour black (mixed syncs and colour burst) distribution system. With the advances in integrated circuitry, SPGs have become less expensive and a colour black system, with an SPG per source genlocked to a distributed reference colour black signal, is a viable proposition.

10.5.2 The 'colour black' reference

The use of colour black as the reference with an SPG, when required, at the picture source as a pulse regenerator has the following advantages:

(1) Many items of equipment are now being manufactured which take in a colour black as the timing reference instead of discrete pulses. This eliminates the need for a separate regenerator in many cases.
(2) If the master SPG, which is generating the reference colour black, fails then each source continues to receive pulses from its local SPG and, although it loses frequency synchronism, this is adequate for emergency purposes.
(3) If the regenerating SPG at the source fails then only one source is lost.
(4) Retiming a system is eased because the line and sub-carrier timing can be adjusted at the source SPG without the use of tapped delay lines.

An SPG system using colour black as the reference has a master SPG providing the station frequency and further SPGs per studio or presentation

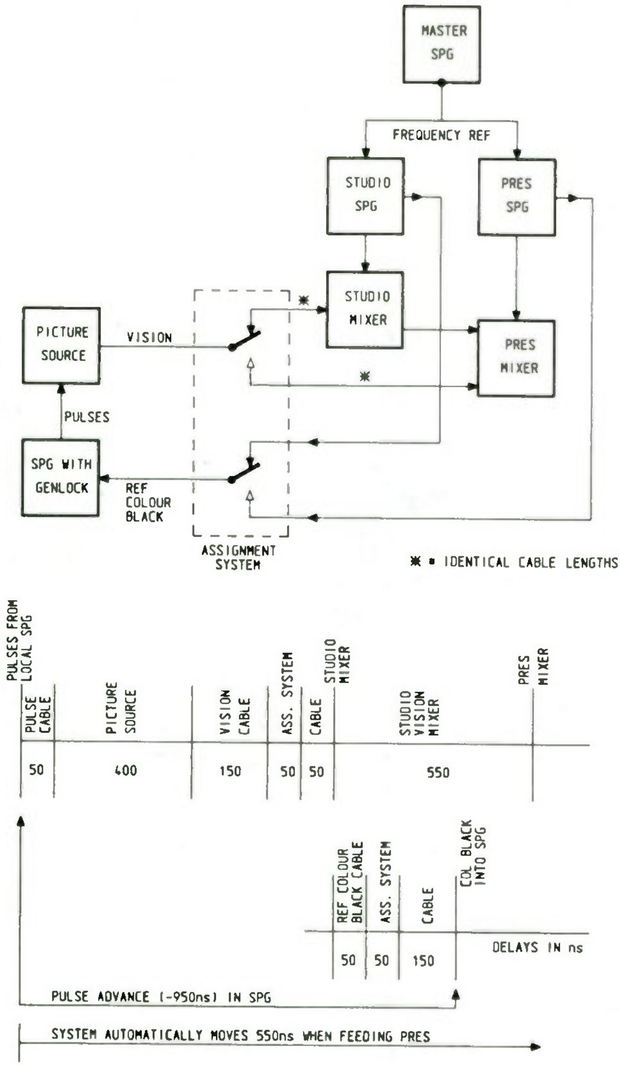


Figure 10.9 Colour black distribution with zero loop delay

to derive the relevant timing for each area. This is shown in Fig. 10.9. The reference colour black, which drives all the local studio sources, is also assigned back to the source regenerator SPG, which can be adjusted to advance its output pulses so that the pulse and video delay through the system equals the SPG pulse advance, thus giving an overall zero loop delay. By making the vision cables from the assignment system to the various mixers equal in length, the delay will be the same into studios or presentation. The different timing levels are achieved by adjusting the line timings of the studio and presentation reference SPGs.

Discipline is required if signals are to arrive at the mixing point correctly timed. However, equipment using techniques similar to the reverse locking previously mentioned is available, which will automatically time the system using a feedback correction system. Systems vary in their sophistication but ultimately the line, field and sub-carrier phase can be inspected at the input to the mixer and error correction information fed back to the source. The advantage of these systems is that any equipment can be installed without worrying about the timing as the system will automatically carry out the corrections. However, they tend to be expensive as each source needs the correction circuitry and a complex feedback path is needed through the assignment system for the correction signal.

10.6 THE DIGITAL SYNCHRONISER

On paper the most elegant solution to the timing of sources is the digital synchroniser, which will accept a vision signal, convert it to an 8-bit digital data stream, store it in memory, convert it back to an analogue signal and deliver it synchronised to a local supply of discrete pulses or colour black. However, synchronisers are expensive (typically from four to eight times the price of an SPG), they introduce a delay between the vision signal and the accompanying sound signal (which can be quite noticeable if three or more synchronisers are in the signal path) and it is not desirable to subject the signal to analogue/digital/analogue signal processing unless it is absolutely necessary. It is possible to use synchronisers within a television complex on each input to each vision mixer to synchronise both internal sources as well as remote inputs. Alternatively the synchronisers can be put inside the vision mixers between the input switching buses, where non-synchronous cuts will be carried out, and the effects units, where the signals need to be synchronous. If this solution is followed there is no need to time the station picture sources as the synchroniser deals with all mistiming. However, it requires many synchronisers. The best solution for optimum signal quality is to time the local sources and the assignable sources into the studio system and use synchronisers purely for remote inputs to the station.

Synchronisers that are available store the incoming signal in a variety of forms:

(1) Composite storage of a single field (store size about 2 Mbit).
(2) Composite storage of a single frame (store size about 3.2 Mbit).
(3) Storage of a frame decoded into the luminance Y and colour difference $R-Y$ and $B-Y$ components (store size about 4.7 Mbit).

Each of these systems has advantages. Storage of a single field tends to be relatively inexpensive because of the smaller memory size but odd and even fields can be exchanged on the output and the device is sensitive to disturbances, such as genlocking, on the input signal. The other systems increase in cost with, in some cases, more analogue circuitry to adjust, but they may be more tolerant of non-standard PAL input signals.

A general block diagram of a synchroniser is shown in Fig. 10.10. The input

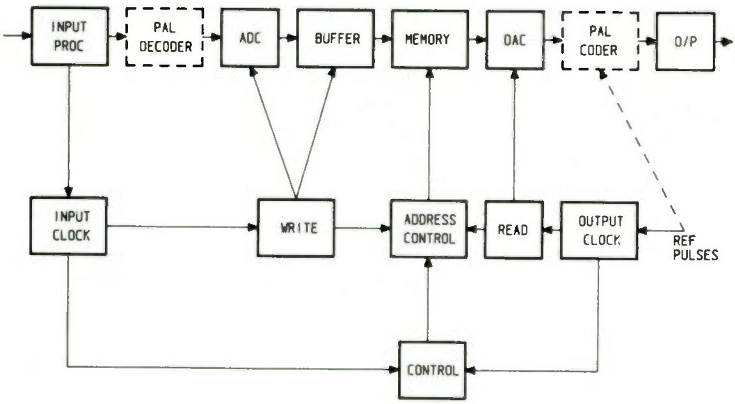


Figure 10.10 Block diagram of digital synchroniser

processor provides adjustment of the gain, etc. of the input signal and the PAL decoder (in the synchronisers using this technique) generates the luminance and colour difference components. The analogue signal is then converted to a digital signal at three times sub-carrier rate for coded signals or a multiple of line rate for decoded signals. The buffer is used to prevent the writing of data into the memory and reading occurring at the same time. The output circuits essentially provide the reverse process, with suitable processing to reinstate sync and burst. Input timing information is derived from the incoming signal and used to generate the write addresses. Similarly the read timing and addressing is controlled by the input reference pulses, which are either discrete pulses or colour black driven from the local SPG.

Circuitry is also provided to stop writing into the store in the event of a non-synchronous picture change at the input. Reading of the previous information in the store continues while the addressing and control circuitry lock up to the new input signal—the whole process taking no more than a field and maintaining a synchronous output. Some synchronisers also generate certain picture manipulation functions and special effects, such as multiple images, picture reversions, compressions and freezes.

10.7 FUTURE DEVELOPMENTS

It is doubtful whether pulse systems will change much in their concept in the immediate future. Obviously integrated circuitry will continue to improve and units will get smaller but it will not be until vision signals are distributed digitally around the station that systems will change. Then the signals can be easily synchronised at any point by clocking the digital stream through a suitable storage process.

Chapter 11

The camera tube and the camera

I. E. Gibson

11.1 IMAGING DEVICES

The camera tube is a specific type of imaging device. Imaging devices can be divided into two basic categories, non-storage and storage. This arises from the fact that the television system is arranged so that the image is scanned element by element, and the information obtained is transmitted serially. In the UK the upper frequency limit of this information is approximately 5 MHz, and 25 complete pictures or frames are transmitted each second. If we consider that each cycle of the 5 MHz contains two picture elements, one black, one white, then the rate of transmission is 400,000 elements per frame. In a non-storage imaging device the light from each element is gathered only while that particular element is scanned. All the light energy that contributes to the information contained in that element is gathered in 1/400,000 of the frame period. This is obviously not a very sensitive way of gathering light from a normally illuminated scene. In a storage device on the other hand the light forming each element of the image is continuously gathered between successive scans and the integrated information is read off by the scanning mechanism. Each element, therefore, receives 400,000 times more light energy than in a non-storage device if the scene is continuously illuminated; so a storage device is far more sensitive and is the obvious choice for a television camera.

If the scene is not continuously illuminated, but we are able to illuminate only that part of the scene that is forming the image element being scanned, then all the light illuminating the scene and being gathered by the lens could be fully used by a non-storage device. In this case the illumination itself becomes the scanning mechanism, the principle of the original Nipkow disk used by Baird. This is obviously not very practical for a camera, but it is widely used in telecine machines in which the flying spot is produced by the raster on a cathode-ray tube.

Imaging devices of the storage type, which are invariably used in television cameras, can be further subdivided by the mechanism whereby the light energy is converted into electrical information. Early camera tubes, including the Iconoscope, the Orthicon and the Image Orthicon, used photoemissive materials. Modern cameras use photoconductive tubes, including vidicon, lead oxide and Saticon types. Recent developments in solid state imaging devices, primarily charge coupled devices (CCDs), have been used in monochrome and experimental colour cameras, but they have some way to go before they can be used in broadcast quality cameras. However, they are already being used in telecine machines.

11.2 ELECTRON OPTICS

All camera tubes are vacuum tube devices and the methods of scanning and focussing are governed by the laws of electron optics. A brief description of electron optics will be an aid to understanding the operation of camera tubes.

11.2.1 Uniform electrostatic field

Consider the uniform electric field produced between two parallel plates (Fig. 11.1). If the potential difference is V and the separation is d then the field E is given by

$$E = -V/d$$

The force on an electron of charge $-e$ is $-eE$, so

$$m\frac{dv_x}{dt} = -eE$$

Integrating gives

$$v_x = -\frac{eE}{m}t + v_0 = \frac{dx}{dt}$$

Integrating again gives

$$x = -\frac{eE}{2m}t^2 + v_0 t + x_0$$

If we assume the electron starts from rest at $x=0$ then

$$t = -\frac{mv_x}{eE}$$

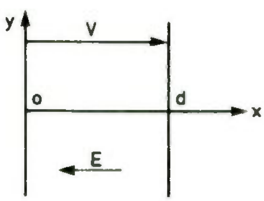


Figure 11.1 Uniform electric field

and

$$x=-\frac{eE}{2m}t^2=-\frac{mv_x^2}{2eE}$$

$$\therefore\ \tfrac{1}{2}mv_x^2=-eEx$$

At $x=d$ the kinetic energy is therefore

$$\tfrac{1}{2}mv^2=-eEd=eV$$

$$\therefore\ v=\sqrt{\left(\frac{2eV}{m}\right)}$$

Now consider what happens if the electron passes through a small hole in the plate at $x=d$ and enters an opposite field produced by a plate at $x=3d$. If the potential difference is $2V$ then the electron will decelerate until $v=0$ at $x=2d$, and will then accelerate back towards the hole. In other words a reflection occurs at $x=2d$.

If the electron has an initial velocity v_y perpendicular to the field, then v_y is not affected by the field. The trajectory is, therefore,

$$y=v_yt \qquad \text{and} \qquad x=-\frac{eE}{2m}t^2$$

$$\therefore\ x=-\frac{eE}{2m}\frac{y^2}{v_y^2}$$

which is a parabola. This is what happens when an electron beam passes between two electrostatic deflection plates. While the beam is in the field between the plates it follows a parabolic path.

Using the above results it is possible to demonstrate that an electron beam obeys laws of reflection and refraction. Consider a region of uniform potential separated from a region of uniform electric field by a fine wire mesh. If an electron beam passes through the mesh at an angle to the mesh and the field decelerates the beam, then the beam will follow a parabolic path and pass back through the mesh, and the angle of incidence will equal the angle of reflection.

When an electron beam passes from one region of uniform potential to another at a different potential, we can show that the beam will be refracted. An analogy can be drawn with optical refraction, Fig. 11.2(a), where

$$\frac{\sin\theta_1}{\sin\theta_2}=\frac{\mu_2}{\mu_1}$$

$\mu=$ refractive index. In the case of electron optical refraction, Fig. 11.2(b),

$$\frac{\sin\theta_1}{\sin\theta_2}=\frac{v_2}{v_1} \qquad (v=\text{velocity})$$

$$\therefore\ \frac{\sin\theta_1}{\sin\theta_2}=\sqrt{\left(\frac{V_2}{V_1}\right)}$$

$V=$ potential (relative to zero velocity potential).

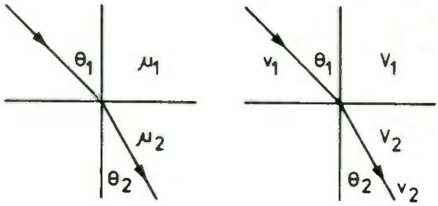


Figure 11.2 (a) Optical refraction, (b) electron optical refraction

11.2.2 Uniform magnetic field

If an electron is at rest or is moving along the field then the field has no effect on the electron. If the electron is moving at right angles to the field then a force is exerted that is perpendicular to both the field and the direction of motion of the electron. The direction of the force is given by the left-hand rule, where the forefinger represents the field, the thumb is the motion, and the second finger the current, remembering that the current direction is opposite to the electron motion. The magnitude of the force is Bev. This causes the electron to move in a circle where

$$\frac{mv^2}{r} = Bev$$

$$\therefore \quad r = \frac{mv}{eB}$$

If the electron has been accelerated from rest by a potential V, then

$$v = \sqrt{\left(\frac{2eV}{m}\right)}$$

$$\therefore \quad r = \frac{1}{B}\sqrt{\left(\frac{2Vm}{e}\right)}$$

The time taken to complete one revolution is

$$T = \frac{2\pi r}{v} = \frac{2\pi m}{eB}$$

which is independent of the velocity.

If an electron is travelling at any other angle to the magnetic field the components of motion along the field and perpendicular to the field can be treated separately. In the general case this will produce a path in the form of a helix, where the axis of the helix is parallel to the field, Fig. 11.3. If the component of velocity along the field is v_1 then the distance travelled during one revolution is given by

$$D = v_1 T = v_1 \frac{2\pi m}{eB}$$

Figure 11.3 Electron motion in a uniform magnetic field

If v_1 is the velocity attained by an electron accelerated from rest to a potential V_1 then

$$v_1 = \sqrt{\left(\frac{2eV_1}{m}\right)}$$

$$\therefore \quad D = \frac{2\pi}{B}\sqrt{\left(\frac{2mV_1}{e}\right)}$$

Consider a source which is emitting electrons into a small solid angle along the direction of the magnetic field. Looking along the field the electrons will depart from the axis of the solid angle, go round in a circle, and then come back to the axis at a distance D from the source, where D is a loop length given above. If the source is extended then each point on the source will produce a corresponding point at a distance D along the field, so an image of the source is produced at D, and again at $2D$, $3D$ and so on, Fig. 11.4. This magnetic focussing is commonly used for focussing camera tubes and is achieved by immersing the tube in a long focus coil.

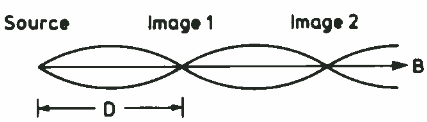


Figure 11.4 Magnetic focussing in a uniform magnetic field. Note that the illustration shows the envelope of all paths of electrons emitted within a solid angle, not the path of any one electron

11.2.3 Magnetic deflection

If an electron beam enters the magnetic field produced by a pair of coils situated either side of the beam then, in the absence of any other field, the beam will be deflected in a direction at right angles to the field. This is the type of deflection normally employed with a display CRT.

A different type of deflection is produced if the deflection coils are immersed in a long focus coil, Fig. 11.5. In this case the deflection and focus fields add vectorially and the electrons travel in a helix which has its axis parallel to the resultant field. This is the same as the magnetic focussing described above, and the image is produced along the direction of the resultant field. The beam is therefore deflected in the same direction as the deflection field (assuming both fields are uniform). This is the type of scanning normally employed with broadcast camera tubes.

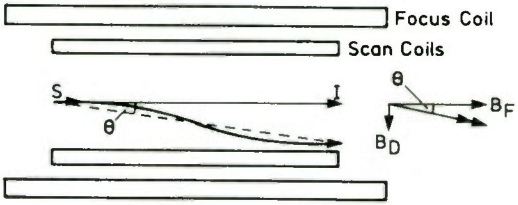


Figure 11.5 Magnetic deflection, showing the path of an electron with and without the scan field

11.3 PHOTOCONDUCTIVE TUBES

Many semiconductors exhibit an increase in conductivity when exposed to light. This mechanism is usually exploited in substances which are poor semiconductors in the absence of light. A semiconductor is a material which has a completely filled valence band and an empty conduction band at low temperatures. This is also true of insulators, but in a semiconductor the energy gap between the bands is low enough for some electrons to occupy the conduction band at normal temperatures. When a photoconductor is exposed to light the photons transfer sufficient energy to some of the electrons to push them up into the conduction band. The energy in a photon is $h\nu$, where h is Planck's constant and ν is the frequency of the light. So if $h\nu \geqslant E_g$ where E_g is the energy gap, then it is possible for an electron to escape from the valence band by absorbing a photon. Below a certain frequency the photon energy will be less than E_g and photoconduction will not occur.

Almost all camera tubes used in broadcast colour cameras are of the photoconductive type. They include the vidicon, the lead oxide types, including Plumbicons and Leddicons, and the Saticon type. They are all very similar in construction and the distinguishing feature is the material used to make the target. In most cases the name of the tube is derived from the target material. Although there are detailed differences in the construction, and several variants have been developed including anti-comet tail types and diode guns, none of these differences is restricted by the target material.

The first photoconductive tube to be developed was the vidicon. Early versions used selenium as the target material but this had stability problems and was replaced by antimony trisulphide. The term vidicon is sometimes confusingly applied to other types of photoconductive tube. The vidicon appeared at about the same time as the image orthicon, that is in the late 1940s, but it suffers from an effect called lag, which causes smearing of moving pictures, and so the image orthicon became the dominant camera tube in broadcasting during the 1950s and 1960s. Vidicons are relatively inexpensive so they are widely used in surveillance and domestic cameras. They are also used in some telecine machines. In a telecine there is plenty of light available and the lag can be kept to an acceptable level.

The next type of photoconductive tube to be developed was the lead oxide tube, and since it exhibits a much lower level of lag than the vidicon it was rapidly adopted for use in colour cameras. Before it was introduced the choice for a colour camera was between the vidicon which resulted in an undesirable degree of lag and the image orthicon which was much larger. Some early cameras actually used a mixture of both types. It is fortunate that the lead oxide tube arrived on the scene when it did, when colour television was in its infancy. The lead oxide tube was introduced by Philips under the trade name Plumbicon. The other major manufacturer of lead oxide tubes is EEV who use the trade name Leddicon, and the tube is usually referred to by one or other of these trade names.

Several further types of photoconductive tube have been developed since the lead oxide type, but the only one which is used to any extent in broadcasting is the Saticon. The name Saticon is an NHK trade name and is derived from the initials of the photoconductor, selenium-arsenic-tellurium. Its performance is comparable to the lead oxide tube.

Photoconductive tubes are available in a variety of shapes and sizes, but three diameters are most frequently used. They are 30 mm ($1\frac{1}{4}$ in), 25 mm (1 in) and 18 mm ($\frac{2}{3}$ in). The standard vidicon size is 1 in. Lead oxide tubes are available in all three sizes and the Saticon is available in $\frac{2}{3}$ in and 1 in sizes. The 30 mm size is the traditional size for the larger full facilities studio and outside broadcast cameras. However, there has been a steady improvement in the resolution of the smaller targets and many studio cameras are designed around the 1 in size. The $\frac{2}{3}$ in size is primarily used in electronic news gathering cameras.

11.4 CONSTRUCTION AND OPERATION OF THE LEAD OXIDE TUBE

The layout of a photoconductive camera tube is shown in Fig. 11.6. The tube can be divided into three main sections covering the functions of emission of an electron beam, focussing and deflection, and imaging.

11.4.1 The triode emission section

The emission section is normally a triode electron gun. The cathode is a nickel tube with one end closed and a non-inductive tungsten heater fitted in the open end. The closed end is coated with barium oxide which is a good thermionic emitter of electrons. The control grid, G1, and the anode, G2, of the triode section are flat metal discs, with central apertures. G2 is held at a potential of the order of 300 V to accelerate the electrons away from the cathode, and the potential on G1 is adjusted to control the beam current emitted by the gun. In

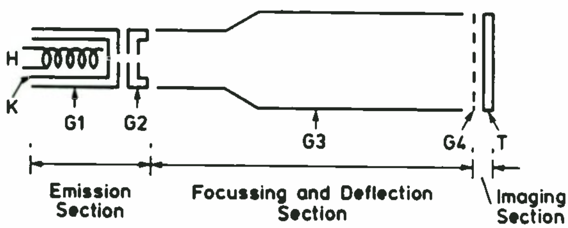


Figure 11.6 Structure of a photoconductive tube

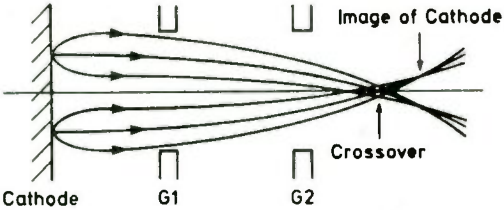


Figure 11.7 Triode electron gun (not to scale)

a standard triode gun the potential on G1 is typically −20 V to −50 V for normal beam currents. This configuration of G1 and G2 produces an electron optical lens which converges the emitted electrons into a beam. The lens is difficult to visualise because the refractive index changes rapidly as the electrons move away from the cathode, and the cathode is actually immersed in the lens. However, all the electrons emitted from one point on the cathode are focussed beyond G2, and at that point there is an image of the cathode, Fig. 11.7. Before reaching that point all the electrons pass through an image of the aperture in G1. This is the exit pupil of the system and is usually called the crossover. The crossover is several thousandths of an inch in diameter and in some tubes it is used as the object for the following electrostatic lens or magnetic focus coil. However, the position of the crossover changes as the potential on the control grid is varied. In a standard lead oxide tube the beam current is limited by placing a further disk in the path of the beam just beyond the crossover with a central aperture of the order of 0.001 to 0.002 in in diameter. This disc is part of the structure of G2 and the aperture provides a well-defined and stable object for the following lens.

It can be shown that the radius of the cathode emitting area decreases as the tube cutoff voltage is increased. Since a reduction in the size of the emitting area reduces the beam divergence, camera tubes are normally designed to have a large cutoff voltage. Beam control circuits typically have to be able to work down to −90 V. Because only a small part of the cathode surface is used, typically a spot about 0.010 in in diameter, the surface has to be particularly smooth.

11.4.2 Focussing and deflection section

Although vidicon tubes are available with electrostatic focussing, they suffer from a number of aberrations, the principal one being spherical distortion. Spherical distortion results in the peripheral electrons passing through a lens coming to focus at a different point from the paraxial electrons. This type of distortion is difficult to correct in an electrostatic lens. Spherical distortion is negligible in a magnetic focus arrangement using a long solenoid, so it is possible to obtain good resolution at high beam currents. Magnetic focussing is invariably used in broadcast cameras, together with magnetic deflection.

The focussing and scanning take place almost entirely inside the uniform electrostatic field inside G3, which is a long cylindrical electrode situated just beyond G2 and extending almost to the target. The electron beam emerging from the aperture in G2 is accelerated by the potential on G3. A high voltage on G3 produces a stiff beam, i.e. one that is less easily deflected by a magnetic field. The advantage of a stiff beam is that the mutual repulsion of the electrons has less effect so a higher resolution is possible. The disadvantage is that a greater focus field is necessary, since the field required is proportional to the square root of the accelerating voltage, and the deflection field also has to be correspondingly greater. Broadcast cameras are designed to achieve the best picture quality that is practical, so the G3 voltage is usually chosen to be near the manufacturer's recommended maximum, a typical figure being 600 V.

When magnetic focussing is used the electron trajectory is a series of loops and an image of the source is produced at the end of each loop. The spacing of the images, D, is given by

$$D=\frac{2\pi}{B}\sqrt{\left(\frac{2mV}{e}\right)}$$

Photoconductive tubes are normally operated with one loop between G2 and the target. This has the disadvantage that the deflection field cannot occupy a whole loop because it is undesirable for the scan field to enter the electron gun or target areas. However, using two loops to overcome this would require that the deflection takes place in half the tube length, so the deflection angle and associated aberrations would be larger. In addition the focus field would have to be twice as strong, so the yoke would be considerably heavier and would consume more power. Single loop operation is normally chosen together with as long a deflection coil as possible. Fortunately, it is possible to correct for the consequences of scanning less than one loop by controlling the flare of the focus field at the target end.

Electrical focus can be controlled either by varying the current in the focus coil, the focus current or by varying the potential on G3, the focus voltage. A common arrangement in a colour camera is to connect the focus coils in series and to provide adjustment of the common current. Individual adjustment is provided by a fine control of each focus voltage.

The final element in the focussing and deflection section of the tube is the mesh electrode. This is an extremely fine wire mesh which is situated just in front of the target. The spacing of the wires in the mesh is of the order of 1000 wires to the inch. The purpose of the mesh is to provide a uniform decelerating electric field as the beam approaches the target. In the simplest form of construction the mesh is placed across the end of G3; tubes of this type are called integral mesh. However, there are several reasons for operating the mesh at a slightly higher potential then G3, in which case the mesh becomes a separate electrode, G4, and a connection to it has to be provided. Separate mesh tubes are much more widely used than the integral mesh type.

One of the aims of the focussing and deflection system is to achieve orthogonal landing all over the target. If the path of the beam is perpendicular to the target the electrons will land on the target until the surface potential is the same as the cathode potential. At approximately cathode potential the kinetic energy of the electrons falls to zero, so no more electrons land and the beam is reflected back towards the mesh. However, if the beam is not normal to the target then the trajectory is parabolic so the beam grazes the surface at its nearest point. The beam still has some kinetic energy at this point so the target will be stabilised at a slightly higher potential. The difference in the target potential between the centre and the edge is called the beam landing error. The effect of this is particularly noticeable in a vidicon where a large beam landing error produces severe shading or portholing. The control of beam landing is mainly a function of the yoke, but in a separate mesh tube the difference in potential between G3 and the mesh produces an electrostatic lens. This lens can be used in conjunction with the flare of the focus field to improve the beam landing at the edge of the target area.

11.4.3 Imaging section—the target

The target is the element in the tube which converts the incident light energy of an optical image into a corresponding pattern of electrical charge, which can then be released by the electron beam. The target is produced by vacuum deposition onto the glass faceplate, Fig. 11.8.

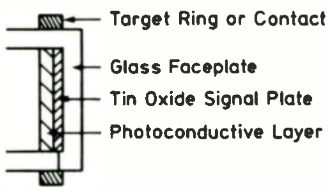


Figure 11.8 The target of a photoconductive tube

First the signal plate is deposited on the inside surface. This is a thin transparent layer of tin oxide, an *N*-type material with a high conductivity, which provides the electrical path for the output signal. Next the photoconductive material is deposited on top of the tin oxide. It is the type of material used, the thickness and the doping that largely determine the behaviour of the tube. In a standard lead oxide tube the layer built up to a thickness of about 20 microns using intrinsic lead oxide, and this is covered with a thin layer of *P*-type oxygen-doped lead oxide.

Originally the faceplate was processed separately and attached to the end of the glass envelope by a compression weld using a soft metallic plug of indium. This avoided having to use heat which would damage the target. The indium plug was surrounded by a stainless steel ring. This method of construction is still used for some tubes, but the majority of tubes now have the target deposited directly onto the inside of the closed-off glass envelope, with an electrical connection through the envelope to a contact on the outside surface. A spring contact inside the camera connects the target ring or contact to the first video amplifier, which is usually called the headamp or preamp.

A cross-section of the target is shown in Fig. 11.9. The tin oxide layer is connected to a positive potential relative to the cathode, usually about +45 V. The electron beam lands on the opposite surface of the target if a positive potential exists and lowers the potential to approximately 0 V. A uniform

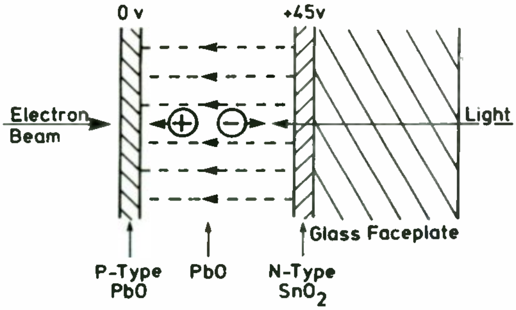


Figure 11.9 Cross-section of the target in a lead oxide tube, showing the generation of an electron-hole pair in the photoconductive target layer (not to scale)

electric field therefore exists across the target. Light from the camera lens passes through the faceplate and the transparent tin oxide layer and is focussed to form an optical image on the target. The photons produce electron-hole pairs in the photoconductive intrinsic lead oxide. In the presence of the electric field the liberated electrons move towards the tin oxide layer and the holes move to the surface of the lead oxide. Between scans a positive charge builds up on the surface, the magnitude of the charge depending on the intensity of the light and the time between scans. In the absence of light the lead oxide has a very high resistance, so the dark current is low and there is little tendency for the collected charge to move sideways. The thin, oxygen-doped layer is added to provide a blocking contact on the surface of the lead oxide to prevent the injection of electrons. Similarly the N-type tin oxide prevents the injection of holes at the positive surface, and together these two blocking contacts maintain the very low dark current in the lead oxide tube.

The target can be thought of as a capacitor with the photoconductor as the dielectric. The tin oxide layer acts as one plate of the capacitor while the other plate is the essentially insulated surface of the dielectric on which the positive charge image has been formed. The connection to this surface is the electron beam, and since this only connects to a small part of the surface at any one time, it is convenient to think of this surface as a large number of tiny isolated plates. In other words the target can be thought of as a large number of tiny capacitors on which the charge image is stored, all connected together by one signal plate. The signal plate is connected through a resistor to +45 V. To complete the equivalent circuit we should add a photoresistor in parallel with each capacitor. In the presence of light, the potential on the free ends of the capacitors rises by a few volts between scans, and is suddenly restored to 0 V by the beam. A momentary current flows in each capacitor in turn and the sum of all these currents flows in the signal plate. The instantaneous current depends on the charge on the target element being scanned. A typical signal current for peak white is 300 nA.

Lead oxide is fairly reflective, so a portion of the incident light is scattered back towards the lens. If no action is taken the scattered light is internally reflected at the surface of the faceplate and causes optical flare on the surrounding target. This problem is overcome by cementing a thick glass stud onto the front of the faceplate. Light scattered at an angle which causes total internal reflection is reflected onto the light absorbing walls of the stud.

11.5 CHARACTERISTICS OF THE LEAD OXIDE TUBE

11.5.1 Sensitivity

The signal current is reasonably independent of the target voltage above about 30 V, indicating that most of the charge carriers reach the surface of the target without recombining. Above this saturation point the sensitivity is typically 400 μA per lumen. The dark current also saturates, and is normally less than 1 nA at 45 V. It has been suggested that the observed dark current may be caused by stray light from the tube heater and that the true dark current is considerably less.

11.5.2 Transfer characteristic

The transfer characteristic or tube gamma is quoted as being between 0.9 and 1.0, where gamma is the constant in the equation

$$I_s = \phi^\gamma$$

relating the signal current, I_s, to the luminous flux, ϕ. For most purposes the characteristic can be assumed to be linear, i.e. the gamma is unity.

11.5.3 Spectral response

Up to about 500 nm wavelength the sensitivity approaches 100% quantum efficiency. This includes the blue and most of the green part of the spectrum. However, at longer wavelengths the sensitivity falls rapidly since the band gap of 2 eV of pure lead oxide gives a cutoff wavelength of 640 nm. The red sensitivity can be increased by doping the layer with a small quantity of sulphur. A complex lead oxide–lead sulphide molecule is formed which has an energy gap of about 0.4 eV. The response of the extended red tube extends into the infra-red region of the spectrum, and this can cause some unexpected colorimetry effects because some non-red objects reflect infra-red wavelengths strongly. Extended red tubes are therefore normally used in conjunction with an infra-red filter to remove the undesirable infra-red sensitivity. The filter can either be incorporated in the camera optics or on the front of the tube.

A disadvantage of the extended-red layer is the increased image retention and for this reason it is not always chosen instead of the standard layer.

11.5.4 Resolution

The resolution of a camera tube is normally quoted in terms of the modulation depth obtained using a square wave test chart. A typical figure for a standard tube at 400 lines/picture height (just over 5 MHz) is 40%. Square wave charts are used because of the extreme difficulty in accurately reproducing a sine wave chart. Although the square wave chart gives a different depth of modulation, because the amplitude of the fundamental is slightly more than the amplitude of the square wave, it is easily reproduced and easily verified.

A side effect of the doping with sulphur to extend the red end of the spectral response is an increase in the resolution of the tube. This is of limited value in the red channel; however, the resolution of a green tube can also be improved by a small degree of doping. These high resolution green tubes typically have a resolution of 60% at 400 lines, and this is approaching the limit set by the cross-sectional area of the scanning beam.

11.5.5 Lag

If the light intensity on a target element is suddenly changed, the signal current should change to the corresponding value after the next complete scanning period. However, in a real tube it can take many fields to reach the new value. This effect is called lag and it produces a smearing of movement. In colour cameras the smearing becomes coloured if the individual tubes exhibit different degrees of lag. Lag can be particularly troublesome in photoconductive tubes. It is caused by trapping of the charge carriers in the photoconductor and by the spread of energies in the electron beam.

Trapping of charge carriers causes two effects. It delays the charge carriers themselves and if it occurs in the region of a blocking contact it can temporarily affect the potential barrier and cause additional dark current. The poor lag performance of the vidicon is mainly caused by trapping effects in the antimony trisulphide.

In the lead oxide tube trapping effects are negligible and what lag there is is caused by the spread of beam energy. Consequently the lead oxide layer is much faster than the vidicon. However, it is not completely lag free, and if left uncorrected the lag can still be unacceptable in colour cameras.

To understand why the spread of electron energy in the beam causes lag it is necessary to examine the mechanism of target stabilisation. Electrons leaving a cathode by thermionic emission have a velocity distribution defined by the Maxwell–Boltzmann distribution function

$$f(E) = A \exp(-E/KT)$$

where A is a constant, T is the temperature and K is Boltzmann's constant. If an electron leaves the cathode with just enough energy to escape then it will be brought to rest at the target when the target is at cathode potential. An electron leaving the target with a higher energy will just come to rest at the target when the target is below cathode potential. The current that can land on the target at a target potential V_t is given by

$$\begin{aligned} I &= I_b && \text{for } V_t \geqslant 0 \\ I &= I_b \exp(eV_t/KT) && \text{for } V_t \leqslant 0 \end{aligned}$$

where I_b is the beam current and T is the effective cathode temperature. This beam acceptance curve can be measured experimentally using a tube with no photoconductor so that the electrons land directly on the tin oxide signal plate. The measured curves indicate a much greater spread of electron energy than the Maxwell–Boltzmann distribution predicts, Fig. 11.10. This

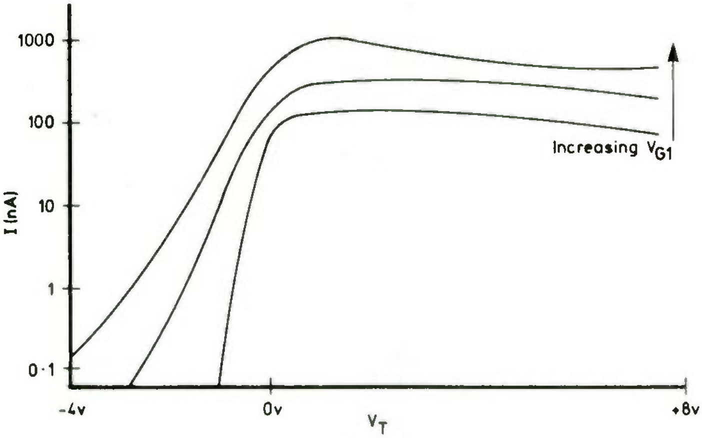


Figure 11.10 Beam acceptance curves

anomalous distribution is caused mainly by the mutual repulsion of electrons in the beam and this effect increases with beam current.

In a typical tube the total capacitance of the scanned target area might be 1000 pF and the peak white signal current might be 300 nA. The spot size is normally large enough to scan the entire area during one field, despite interlace, so the change in target potential between scans for this tube would be

$$\Delta V_T = \frac{300 \text{ nA} \times 20 \text{ ms}}{1000 \text{ pF}} = 6 \text{ V}$$

The beam current would normally be set to just discharge twice the peak signal current to allow some overload capability. Examination of the beam acceptance curves (Fig. 11.10) shows that the beam can supply 300 nA signal current provided the target potential remains above about 0 V. The target potential therefore oscillates between just below 0 V and nearly 6 V.

When the illumination is removed there is only dark current in the photoconductor. If this is 0.3 nA then the target potential changes by 6 mV between scans. The first scan after the illumination has been removed drives the target potential below 0 V, but this reduces the current that can land on the target. This in turn reduces the change in potential caused by the scanning beam. The target is driven more negative at a progressively slower rate until eventually only just enough current can land to balance the dark current. The signal current, which is the current landing on the target, therefore decays over a period of many fields, and this is called decay lag.

With no illumination, the target has stabilised at a potential which is several volts below cathode potential. If 10% of the peak white illumination is suddenly applied, the target potential will rise by 0.6 V between scans. Very little beam current will land at first, but after several fields the potential will be approaching 0 V and a significant current will be able to land. Eventually the target potential reaches equilibrium when the current landing on the target

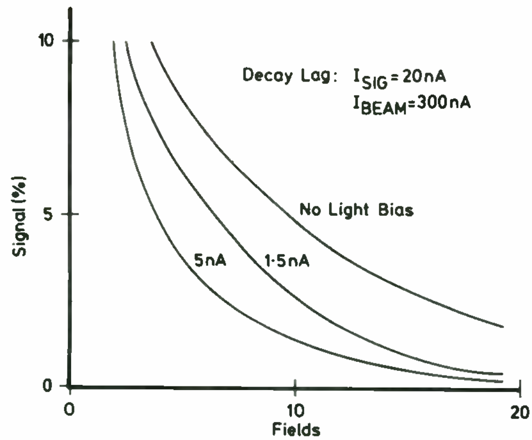


Figure 11.11 Improvement in decay lag brought about by using light bias

balances the number of positive charge carriers released in the target. The signal current therefore builds up over a number of fields, and this is called build-up lag.

It can be seen that the lag would be less severe if the dark current did not have such a low value, because then the target would be stabilised at a less negative potential in the dark. A high dark current in a semiconductor would be sensitive to temperature, but the dark current can be increased artificially by flooding the target with a low level of illumination. This is called light bias. The improvement in decay lag and build-up lag, produced by adding light bias, is indicated in Figs. 11.11 and 11.12. In addition to improving the absolute level of the lag, if the bias is adjustable it can be used to balance the lag

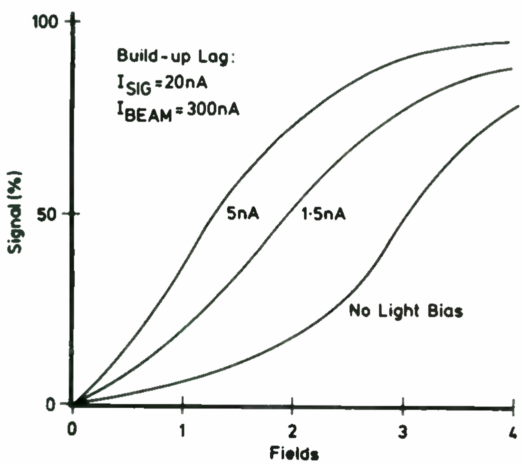


Figure 11.12 Improvement in build-up lag produced by using light bias

in the tubes in a colour camera. The chosen bias level is a compromise between the lag reduction and the black shading that any unevenness in the bias lighting might introduce. The light bias bulbs can be incorporated in the camera optical assembly, but the more usual arrangement is to incorporate the light bias inside the tube or in the tube socket. The bulbs are connected across the tube heater supply, with a series rheostat if the bias is adjustable. Provision is made in the tube structure for transmitting the light from the heater end of the tube to the target.

11.6 THE HOP OR ACT TUBE

The beam is capable of stabilising the target potential provided the equilibrium signal current remains less than the available beam current. When the illumination is increased beyond the point where the signal current is the same as the beam current, the target potential rises more between scans than the beam can pull it back at the next scan. The target potential therefore rises towards the potential on the signal plate. When the illumination is produced by focussing a small area highlight on the target, the target potential under the highlight rises while the surrounding target remains stabilised at cathode potential. The electric field in the target is no longer uniform and, as a result, the positive charge carriers released in the target layer at the edge of the highlight tend to move away from the centre of the highlight, Fig. 11.13. At the same time the scanning beam is pulled slightly towards the more positive target potential under the highlight. Both these effects cause the highlight to spread outwards on the displayed picture, producing a puddle of peak white signal around the highlight. This is called blooming. The puddle spreads until the beam is able to cope with the number of charge carriers being released.

When the highlight is removed it takes several fields for the beam to discharge the large charge image that has built up on the target. Taking the previous example, where the 600 μA available beam current was able to change the target potential by 12 V per field, it will take four fields before the target is discharged from near 45 V to 0 V. During these four fields a peak white signal will be seen on the displayed picture. If the beam current is set to a

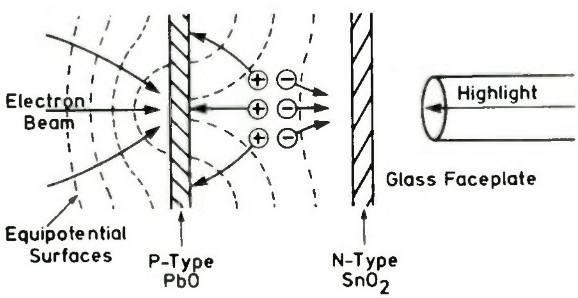


Figure 11.13 Blooming, caused by charge spreading and beam pulling in the vicinity of a highlight

lower figure, then it will take longer to discharge the overload. When a highlight moves across the target the delayed discharge causes a peak white tail following the highlight. This effect is called comet-tailing. In a colour camera the lengths of the tails will usually be different in the different colours, though they can be balanced to a certain extent by adjusting the beam currents.

Comet-tailing is not normally too much of a problem in the controlled lighting of a studio, though even there reflections can be troublesome on occasions. However, on outside broadcasts and, in particular, in the more adventurous use of portable cameras there is a real need to be able to eliminate or control comet-tailing. For this reason a new type of tube structure was developed. Philips call it the ACT or Anti-Comet-Tail tube, and English Electric call it the HOP or Highlight Overload Protection tube.

In the HOP or ACT tube an additional electrode, G5, is introduced inside G2, and the limiting aperture in G2 is increased in diameter. During the normal scan the G5 electrode is held at the same potential as G2, so the operation is similar to the standard tube, Fig. 11.14. However, it is the

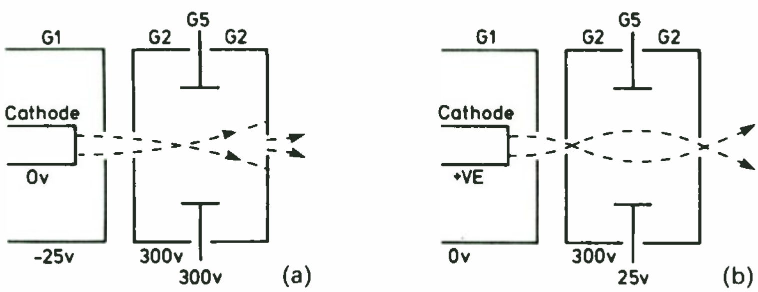


Figure 11.14 (a) Operation of the electron gun in an HOP or ACT tube during normal scan, (b) operation during flyback, showing the creation of a second crossover near the limiting aperture in G2

crossover which is focussed onto the rear of the target, not the limiting aperture in G2. During flyback G1 is pulsed to 0 V to turn the beam hard on. At the same time G5 is pulsed to about 25 V thereby producing a strong electrostatic lens inside G2, which produces a second crossover near the limiting aperture in G2. Consequently nearly all the electrons emitted by the cathode pass through the aperture, giving several hundred times the normal beam current. Since we do not want to discharge the normal image on the target during flyback the cathode potential is pulsed above the potential corresponding to peak white. If peak white illumination produces a potential of +6 V on the target then the cathode would be pulsed to about +9 V. Since the second crossover is further forward, the spot on the target is defocussed during flyback. The highlights are therefore discharged just ahead of the normal scan. This ensures that by the time the scan reaches the highlight it has already been discharged to about 9 V and since the beam is able to discharge up to 12 V (in our example) it can handle what is left of the highlight.

HOP/ACT tubes are able to handle highlight overloads up to five stops above peak white.

11.7 OTHER CAMERA TUBES

11.7.1 The Saticon

The Saticon was developed by the combined efforts of NHK, the Japanese Broadcasting Corporation, and Hitachi. It is based on selenium, which, on its own, lacks red sensitivity and has a tendency to crystallise at room temperature, which produces white spots on the picture. In the Saticon the selenium is stabilised by arsenic doping and a thin layer of cerium oxide on the surface of the tin oxide signal plate, Fig. 11.15. The red sensitivity is improved by adding tellurium to part of the selenium layer close to the signal plate. Too

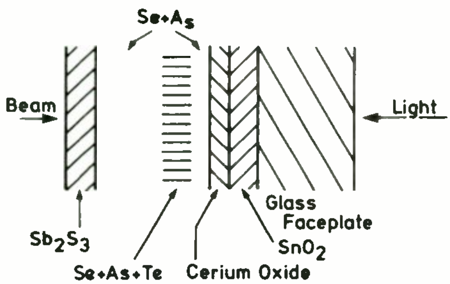


Figure 11.15 Cross-section of the target in a Saticon tube

much tellurium increases the tendency to recrystallise and also increases the photoconductive lag. A thin layer of antimony trisulphide is added to the outer surface of the selenium to act as a blocking contact. This outer layer also serves a second purpose. At the optimum target voltage of about 50 V, selenium is close to the potential at which it might switch over to the high velocity mode of operation. In the high velocity mode electrons in the beam cause secondary emission from the target resulting in a net loss of electrons and an increasing potential on the surface of the target. The antimony trisulphide layer reduces the secondary emission and prevents any tendency for this mode to occur. The name Saticon is derived from the initials of Se + As + Te.

The performance of the Saticon is very similar to the lead oxide tube. The resolution is slightly better, the lag slightly worse and the spectral response lies between the standard and extended red versions of the lead oxide tube at the red end of the spectrum. It is a popular tube, particularly among the Japanese camera manufacturers, in the $\frac{2}{3}$ in size. The gun structure has been modified to reduce the charge density in the crossover and thereby improve the lag performance. The modified gun requires a more negative G1 potential. A new gun structure has also been introduced in some types of lead oxide tube. This is the diode gun, so called because G1 is operated at a slightly positive potential, and is therefore conducting like an anode.

Both the modified gun of the Saticon and the diode gun can be operated at higher beam currents without degrading the picture. However, such currents would shorten the cathode life so circuits have been developed which automatically increase the beam when a highlight is detected. It is too early to tell whether the automatic beam control circuits are going to completely displace the HOP/ACT tube.

11.7.2 Other photoconductive tubes

The lag performance of the Vidicon, which uses an antimony trisulphide photoconductor, has been improved by using a multi-layer structure of alternate solid–porous–solid layers. It is the only photoconductive tube with a gamma significantly less than unity. It is also the only one in which the sensitivity can be controlled by changing the target voltage. This characteristic is widely used for automatic sensitivity control in surveillance cameras.

The Silicon Vidicon has a photoconductor consisting of an array of discrete silicon diodes, a typical array being 1000 × 750 diodes. Soft X-rays produced by electrons hitting the mesh can cause the dark current to increase gradually, so the tube has to be operated at a lower mesh voltage. They are widely used for surveillance because of the wide infra-red response.

The Chalnicon, announced by Toshiba in 1972, uses a cadmium selenide target with a coating of arsenic sulphide. The target is thin so the lag is worse than the lead oxide tube.

The Newvicon, announced by Matsushita in 1974, uses layers of zinc selenide and zinc cadmium telluride in the target, and has a high quantum efficiency. The lag performance is similar to the vidicon.

None of the above tubes is widely used in broadcast cameras. Figure 11.16 compares the spectral responses of several types of photoconductive tube.

11.7.3 Photoemissive tubes

The Iconoscope, patented in 1923 by Zworykin, was the first electronic means of producing a television picture. The tube consisted of a large glass envelope and an angled neck containing the electron gun, Fig. 11.17. The electron beam scanned the illuminated side of the target, hence the angled neck. The target was made from a sheet of mica coated with a metallic film signal plate on one side and a mosaic of isolated silver droplets coated with silver and cesium oxides on the other. Photo-electrons were released when light fell on the mosaic, forming a positive charge image. The Iconoscope operated in the high energy mode because the secondary gain was greater than unity and the target potential stabilised at about the potential of the anode. A high proportion of the secondary electrons fell back to the target, producing severe flare and shading effects. The shading had to be corrected almost continuously by a variety of controls. It was relatively insensitive and the angled scanning produced keystone distortion which had to be corrected in the scan circuits. The tube had to be handled carefully and the target was easily damaged by exposure to any bright light.

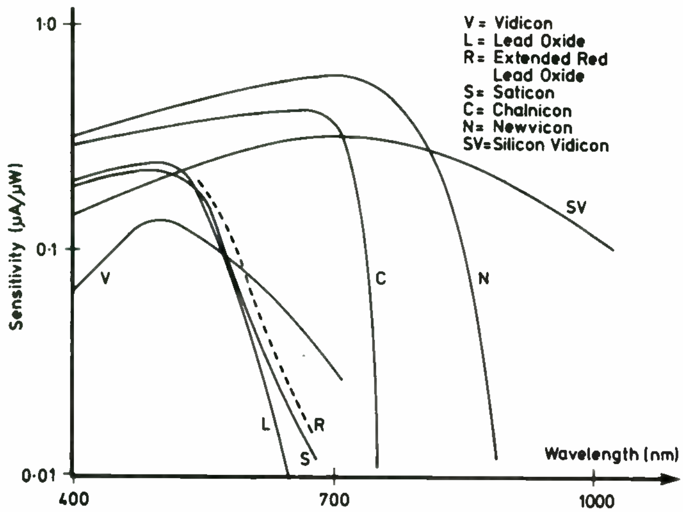


Figure 11.16 A comparison of the spectral responses of different types of photoconductive tubes

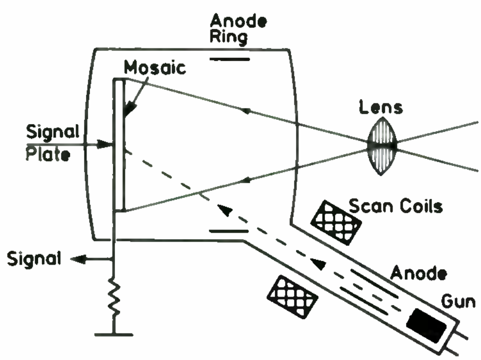


Figure 11.17 The Iconoscope camera tube

The Emitron was virtually identical to the Iconoscope. A development of the Emitron was the Super Emitron which had a photocathode positioned in front of the mosaic, and the mosaic was made electron sensitive rather than light sensitive. The optical image was formed on the photocathode and the electrons released by photoemission were accelerated towards the mosaic and focussed magnetically.

The Orthicon and CPS Emitron tubes both operated in the low velocity mode, CPS standing for Cathode Potential Stabilisation. They both used translucent signal plates which allowed the optical image to be projected onto the signal plate side of the target. The electron gun was located on the optical axis and the target was situated close to the end of the tube so short focal length lenses could be used.

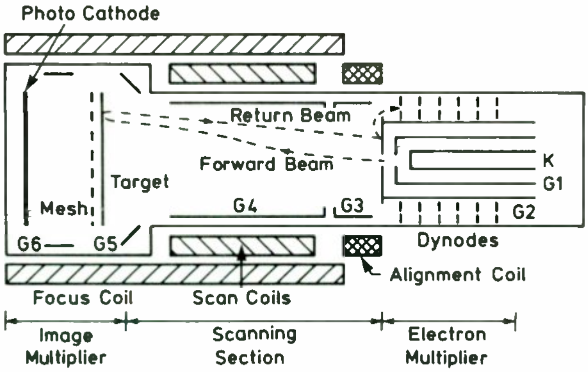


Figure 11.18 The Image Orthicon camera tube

The Image Orthicon is a large complex tube containing an image multiplier at the target end and an electron multiplier surrounding the electron gun. Fig. 11.18. The image multiplier consists of a photocathode situated in front of a thin glass target. Secondary electrons released at the target are collected by a mesh close to the target. When the target is scanned by the electron beam those electrons which do not land are accelerated back towards the gun. This return beam contains a signal current component since it is depleted when electrons land on the target. The return beam is multiplied by a five stage electron multiplier surrounding the gun. The Image Orthicon produces high quality monochrome pictures, but it is too large for use in colour cameras. Replacement tubes are still manufactured for cameras that use Image Orthicons.

11.7.4 Charge coupled devices

CCDs are not camera tubes, but they are grouped in this section for convenience. A CCD is an MOS integrated circuit. In its simplest form it consists of an array of closely spaced electrodes separated from an *n*-type or *p*-type silicon substrate by a thin insulating layer of silicon dioxide. In the case of a *p*-type substrate a positive potential is applied to the electrodes, forming a depletion region in the substrate immediately under the electrodes. When the substrate is exposed to light, electron-hole pairs are generated in the depletion region and the electrons are attracted towards the electrodes. The charge image can be shifted sideways to produce a signal current in an output diode by rippling the potentials on the electrodes. The rippling is achieved by connecting every third or fourth electrode together, and connecting them to suitably phased clock waveforms. Since it is not necessary to address individual elements in the array, a CCD is called a self-scanning device.

One-dimensional, or line array, CCDs are already being used in broadcast telecine machines and experimental colour cameras have been produced using two-dimensional (area) arrays. However, several problems must be overcome before CCDs can be used in broadcast cameras. It takes a finite time to clock the image out of the array as the usual arrangement includes a second array

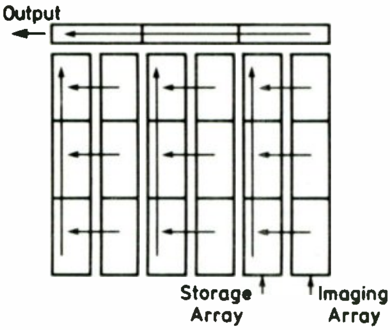


Figure 11.19 Interline transfer CCD imaging device

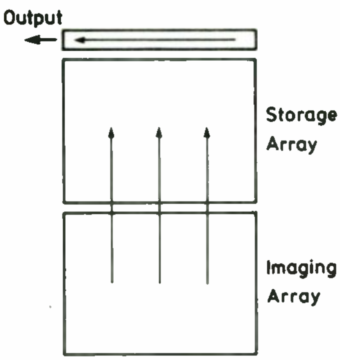


Figure 11.20 Frame transfer CCD imaging device

into which the image is clocked at the end of each integration period. The imaging array can then start to gather the next image while the image in the storage array is clocked out serially. The two arrays can either be interleaved (interline transfer) or butted together (frame transfer). The interline transfer type (Fig. 11.19) suffers from beat patterning or aliasing caused by the opaque stripes over the storage array. The frame transfer type (Fig. 11.20) suffers from the relatively long exposure to the light image while the charge image is being clocked into the storage array. Apart from the topological difficulties, the main problems that have to be overcome are the lack of blue sensitivity caused by absorption in the transparent polysilicon electrode structure and the difficulties involved in trying to produce large devices of satisfactory quality.

11.8 THE YOKE

The primary requirements of the yoke are first that it should sharply focus the electron beam onto the target, second that it should scan the image with a minimum of geometric distortion, and third that it should cause the electron

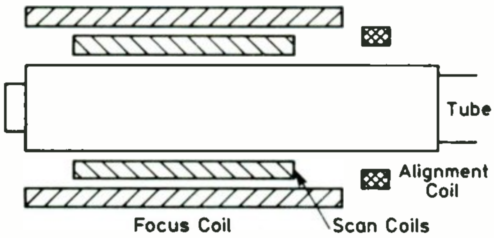


Figure 11.21 The yoke, showing the positions of the wound elements

beam to land orthogonally. The main elements in the yoke are shown in Fig. 11.21. The focussing is achieved by the focus coil. We have already seen that, in the absence of a scan field, the focus field causes the electron path to be a single loop of a helix. Optimum geometry is achieved if the scan coils produce a uniform magnetic field across the tube. The coils are wrapped around the tube to minimise the size of the yoke. A cosine distribution of the turns is required to obtain a uniform field, where each turn has a characteristic saddle shape. In a wound scan coil this type of distribution is normally simplified to a number of sections with more turns in the outer sections. Printed circuit scan coils are sometimes used since precise and consistent distribution patterns can be achieved, but they normally have fewer turns than a wound coil.

Given the length of the scan coil and the dimensions of the scanned area on the target it is possible to determine the deflection angle. Although this is typically only a few degrees there will nevertheless be a shorter focussing loop when the scan field is added to the focus field. The surface of beam focus will curve away from the target causing a loss of resolution in the corners. It is therefore desirable to make the deflection coils as long as possible, consistent with other requirements, to keep the deflection angle small. Corner defocussing can also be reduced by modulating the focus voltage to flatten the surface of focus.

The alignment coils are included in the yoke to compensate for non-orthogonal landing at the centre of the target. This occurs if the electron gun is not parallel to the axis of the tube. The alignment coils are short deflection coils which bend the electron beam as it leaves the gun. A short deflection coil can be produced by winding two coils on opposite sides of a toroid and connecting the windings so they oppose each other. This gives a reasonably uniform field across the centre of the toroid. Two pairs of coils are wound on the same toroid to provide horizontal and vertical alignment. If the alignment is incorrect the undeflected beam will approach the target at the same angle to the axis as it left the gun. Changing the focus voltage will cause the spot to move slightly in addition to going out of focus. Correct alignment can therefore be obtained by rocking the focus voltage and adjusting the currents in the alignment coils for no movement at the centre of the picture. A focus rock generator is normally included in the camera circuits.

With the tube correctly aligned the landing would be orthogonal all over the target if the focus field and deflection fields were uniform over the entire loop length, since after one complete loop the electron path is parallel to its initial path. However, it is undesirable for the scan fields to enter either the target or electron gun areas, so the length of the deflection coils has to be less than one complete loop. Consequently the electron beam will leave the scan field at an angle. In a uniform focus field this would result in non-orthogonal landing away from the centre of the target. However, the focus field is not uniform, it inevitably flares out towards the end of the focus coil and it is possible to use this flare to compensate for the shorter deflection coils. The flare can be controlled by adding a lip to the front end of the mumetal can or by adding extra turns (overwind) to the front of the focus coil. Other variables involved in optimising the beam landing include the positions of the tube and scan coils relative to the flared focus field, and the electrostatic lens between G3 and G4 in a separate mesh tube.

The yoke is shielded from external magnetic fields by enclosing it in a mumetal can. This also provides a return path for the focus field. A separate return path is provided for the scan fields by wrapping narrow mumetal tape outside the scan coils but inside the can. An electrostatic screen is fitted on the inside of the scan coils to prevent the scan voltages from being coupled to the tube electrodes. A further screen is fitted between the scan coils in the form of narrow strips of copper along the length of the yoke to prevent eddy current losses.

11.9 CAMERA OPTICS

The CIE chromaticity diagram is bounded by a triangle formed by three imaginary primary colours which can be mixed to produce any real colour, Fig. 11.22. The colours of the spectrum lie on a horseshoe-shaped curve. These are the colours that consist of monochromatic light. The light reflected by everyday objects usually contains a mixture of all wavelengths, and these colours lie well inside the horseshoe curve. In the receiver the colours are

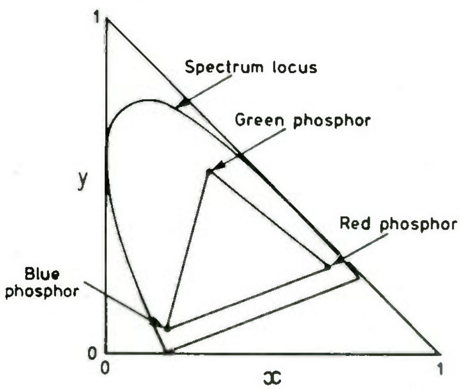


Figure 11.22 CIE x, y chromaticity diagram, showing the approximate coordinates of the display tube phosphors

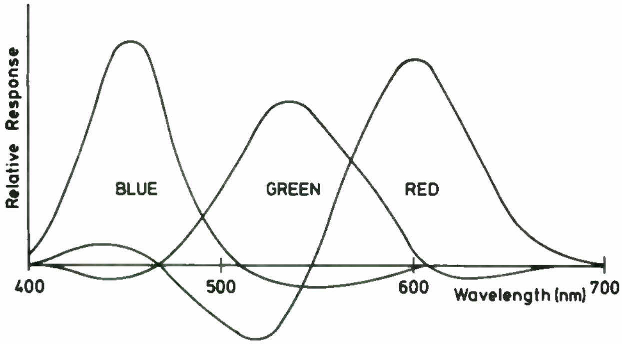


Figure 11.23 The ideal taking responses of a colour camera. The curves are dependent on the phosphors used in the display tube

reproduced on a display tube containing green, red and blue phosphors. It is only possible to reproduce colours that are contained within the triangle formed by the phosphor primaries. This is not a severe limitation because highly saturated colours do not occur very often. For any given set of phosphor primaries it is possible to derive the ideal taking responses for a colour camera, Fig. 11.23. The PAL phosphors have been chosen to be representative of the phosphors used in picture tubes today. Modern phosphors are more efficient than the original NTSC primaries but they do not include quite as many colours.

The ideal taking responses include some negative lobes, and in order to simulate these it is necessary to incorporate a linear matrix in the camera video circuits. For instance the negative lobe of the red response is obtained by subtracting a proportion of the green signal from the red signal, and increasing the amount of the red signal to compensate. This makes the signal to noise ratio slightly worse so it is best to keep the matrix factors as small as possible. It is not possible to transmit the negative lobes, so if a highly saturated colour is present the negative signals are clipped by the black clipper in the video processing. As noted above, this is an unusual situation since most colours in real life are very desaturated and give positive green, red and blue outputs.

When the ideal taking responses are normalised for equal areas under the curves the red and blue peaks are higher than the green, indicating that the red and blue tubes would have to be more sensitive than the green tube to obtain equal signal currents. In fact the spectral response of the tube falls rapidly towards the red and blue ends of the spectrum, the spectral response of the zoom lens normally falls at the red and blue ends, and the spectrum of the light used to illuminate the scene is not flat. The net result is that, in studio lighting, the red and blue signal currents are typically one-quarter to one-third of the green signal current. In daylight the spectrum of the scene illumination can vary considerably. In average daylight the signal current in blue is nearly twice the studio value, and the red signal current is nearly halved. It is possible to colour balance the camera using these signal currents, but the noise and lag in red are more pronounced than in studio lighting. Instead a colour conversion filter is usually used in daylight to convert the colour temperature to somewhere near the studio value of 3000°.

The lower signal currents in red and blue cause more noise and lag relative to green. In three tube cameras the high-frequency detail is often obtained from the green tube. The red and blue signals are bandwidth limited to reduce the noise, and then have the detail from green, or 'highs' from green, added in the video processing. This also tends to mask small registration errors. The higher lag in red and blue can be reduced by using variable light bias. Before the introduction of light bias it was necessary either to deliberately reduce the green sensitivity, or to use smaller optical images in red and blue to reduce the coloured lag.

The taking responses are achieved by using colour filters between the lens and the tubes. This could be achieved by using semi-silvered mirrors and absorption filters but this would be an inefficient arrangement since each absorption filter would remove about two-thirds of the light entering it. A better method is to use dichroic mirrors. A dichroic mirror or filter is a colour selective filter made by adding alternate quarter wave layers of high and low refractive index to the surface of a plate of glass, which gives a high value of reflectivity at the chosen wavelength. Early colour cameras used dichroic mirrors, but they invariably required the use of a relay optical system because there was not enough room to fit the mirrors between the lens and the tubes. A relay system contains a lens to relay the image formed by the zoom lens onto the tube. A further lens, called the field lens has to be inserted in the plane of the first image to concentrate the light onto the relay lens. A relay system is therefore quite complicated, and since it has a large number of glass surfaces the transmission losses are quite high. Also any dirt on the field lens is in focus.

A much more compact arrangement can be achieved by using a prism assembly with dichroic filters on the prism surfaces, Fig. 11.24. Not only is this smaller but also the effective light path is increased inside the glass. As a result it is possible to fit the prism assembly between the zoom lens and the tubes without using a relay lens. Modern broadcasting colour cameras almost invariably use a prism assembly.

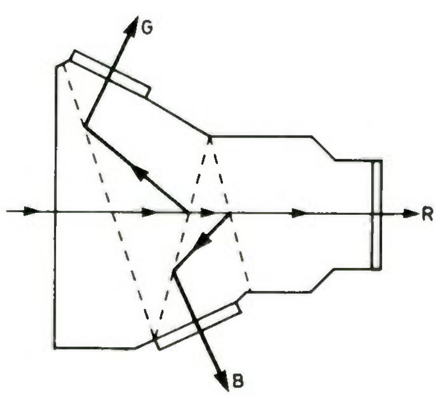


Figure 11.24 Prism optical splitter block

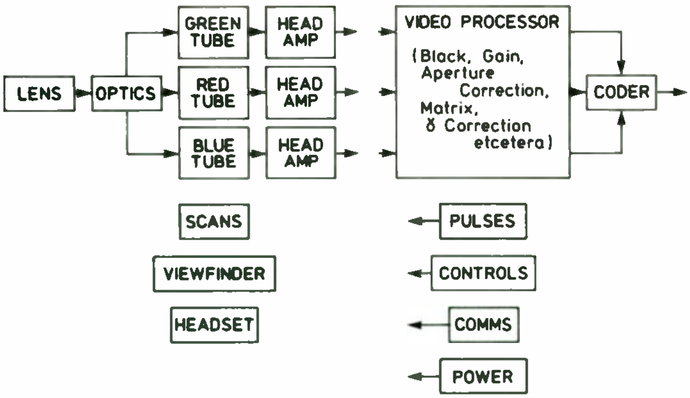


Figure 11.25 Camera block diagram

11.10 CAMERA CIRCUITS

Figure 11.25 shows a block diagram of the television camera. The camera tubes and the associated deflection yokes are the interface between the optical image and the electrical signal. There are one or two things that can be done to enhance the signal, the principle ones being aperture correction, noise reduction and shading correction, but with these reservations the output signal can only be as good as the performance of the tubes.

The block diagram is in two parts indicating the circuits which might be contained in the camera head and those which might be included in the camera control unit. Many cameras, especially the smaller types, have all the circuits inside the camera head. When two units are used they are connected together either by multicore camera cable, which contains several coaxial cables and several tens of single core conductors, or by a single coaxial cable onto which all the signals and controls are multiplexed. In the latter case the power is sent down the coax as well so there is a voltage drop along the outer of the coax. The coax is therefore surrounded by an earthed sheath and the composite cable is therefore a triaxial cable, or 'triax'.

11.10.1 Scan circuits

The scans provide the line and field sawtooth currents for the yokes. The main requirements are that they should meet the very demanding geometry and registration specifications. Registration specifications are typically better than 25 to 50 ns in the central area of the picture. Since the picture is just over 50 μs wide this represents 0.05 to 0.1% of the peak to peak scan amplitude. Normally a common scan generator provides most of the scan waveform to each yoke to reduce the demands on the generator linearity and stability. The output stages are normally feedback amplifiers, possibly combined with switching output stages in the case of the line scan. Various forms of individual geometry correction are frequently provided, including dee or bow, trapezium and barrel-pincushion, as well as the usual linearity and skew. Recently there has been a trend towards providing multi-point registration correction. Continuous horizontal and vertical scan correction waveforms are generated, normally automatically, by examining a large number of points on a special chart.

An indication of the demands on the stability of the circuits can be seen by considering what is required to maintain the registration within 0.05% over a temperature range of 20°C. This is equivalent to 25 parts per million per degree, so if the circuits can drift in opposite directions they have to be better than 12.5 ppm per degree. Even if the scan circuits could be made entirely drift free, there will almost inevitably be some drift in the optics, tubes and yokes, and this will be mainly a centring drift. Most cameras therefore include provision for automatic correction of the centring while the camera is in use. This is achieved by monitoring the picture being produced by the camera and searching for possible registration errors. When an error is detected an appropriate correction is applied to the centring circuit.

11.10.2 Head amplifier

A 30 mm tube produces a peak white signal current of 300 nA, in the case of the green tube. If this is fed into a 1 MΩ resistor it will give a signal of 300 mV. However, the capacitance between the tube target and ground will be of the order of 10 pF, and the following amplifier will also be connected across the resistor and might contribute a further 10 pF. The amplitude response will therefore roll off at 6 dB/octave above approximately 8 kHz. A multistage amplifier can be used to compensate for this with a rising response up to the highest video frequency. The noise produced by the first stage of the amplifier will therefore have a triangular spectrum after it has passed through the following stages. In practice a virtual earth amplifier is normally used with the 1 MΩ resistor connected as the feedback resistor, Fig. 11.26. A cascade first stage is used to reduce the effect of Miller capacitance on the gate of the input FET. With this arrangement, and with the capacitance values mentioned above, it is possible to achieve a signal to noise ratio of the order of 50 dB at black. This is mainly determined by the noise generated in the first FET. Care is needed to ensure that this is the dominant noise source, which is the reason for the high value resistance in the gate circuit. Above 8 kHz the open loop signal amplitude on the gate of the input FET depends on the value of the

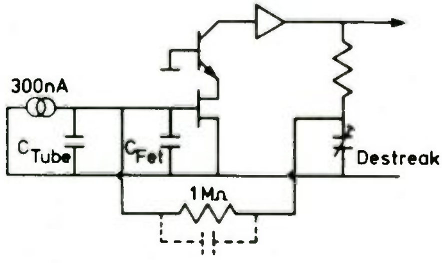


Figure 11.26 Camera head amplifier (or pre-amplifier)

capacitance between the gate and ground, so the capacitance is made as low as possible to obtain the best signal to noise ratio. However, the best choice of FET is one with an input capacitance equal to all the other contributions to the capacitance, and with a high ratio of g_m to input capacitance. If the input capacitance of the FET is reduced then the g_m is also reduced and the noise produced in the channel of the FET increases faster than the signal amplitude on the gate, since the total input capacitance is then dominated by the tube and wiring capacitance.

One circuit arrangement that can be used to improve the noise performance is to separate the tube capacitance and FET with an inductor called the Percival coil. The coil is tuned to give a peak in the open loop response near the top end of the video band. The resulting closed loop noise spectrum is no longer triangular, instead it falls at high frequencies. When optimised a Percival coil can give an improvement of over 2 dB in the signal to noise ratio. Many camera designs, particularly those using small and/or low capacitance tubes, opt for achieving the lowest possible stray capacitance by mounting the first FET inside the front of the yoke, close to the target, and dispense with the Percival coil.

11.10.3 Video circuits

The video signals pass through several stages of processing before the output. Most of these processes are described in detail elsewhere in the book, so only a brief summary is given here. Particular cameras differ in the degree of processing employed, and the sequence of processes, but the following is typical.

(1) DC restoration, e.g. line clamping.
(2) Cable correction, if required.
(3) Establish black level. This includes black shading correction and the removal of any set-up produced by dark current and light bias.
(4) Set gain. This includes white shading correction.
(5) Aperture correction. Horizontal correction is used to increase the resolution at 400 lines to 100%. Further horizontal and vertical correction is provided for subjective improvement of the picture sharpness.
(6) Noise reduction. Usually associated with the aperture corrector. Techniques include combing, coring and level dependent correction.
(7) Linear matrix.
(8) Gamma correction.
(9) Coding. Individual green, red and blue outputs are normally provided in addition to the coded output.

11.10.4 Monitoring and control

The viewfinder allows the cameraman to frame and focus the picture. The viewfinder signal is normally fully processed video. There is provision for selecting other sources for special purposes, including an external signal from the mixer which would be needed for framing chroma key shots or split screens. The cameraman invariably has control of the zoom and focus. Servo zoom is normally used, often with provision for presetting a number of shots by means of a shot box, but the focus control may be either servo or direct. The cameraman wears a headset to hear the production sound, and usually has provision for hearing programme sound and engineering sound. Talkback from the cameraman can be routed into the production or engineering circuits.

Operational controls, such as iris, black, trim controls for colour balance, etc. are available at the remote operational control panel. These are the controls which the operator might want to adjust while the camera is on air or between shots. Many more controls exist which require periodical adjustment. Monitoring signals are provided to aid the setting of these controls, and these signals are displayed on a picture monitor and a waveform monitor. Misregistration, for example, is much more visible when viewing colour difference signals on a picture monitor, so $R-G$ and $B-G$ signals are provided to aid or check registration adjustments. Varying degrees of automatic line-up are provided in modern cameras, making the cameras easier to use. Some automation of the operational controls may also be provided, auto iris being an example. With the introduction of microprocessor control and digital storage of control settings there is a trend towards providing a single central controller for lining up and/or controlling several cameras. Further progress in the introduction of digital techniques can be expected. However, the camera, and particularly the portable camera, is likely to be one of the last items of studio equipment to go fully digital.

Chapter 12

Propagation theory

J. L. Eaton

12.1 GENERAL PROPAGATION CONSIDERATIONS

This chapter deals with the principal concepts needed for some understanding of radio wave propagation effects with relation to television broadcasting in the United Kingdom. Certain effects, well known from the theory of optics, are mirrored in UHF radio propagation. This is a useful fact to remember, providing the analogy is not pressed too far.

12.1.1 Reflection

In optics, total reflection is often taken to be the case. In radio theory, it is more realistic to assume partial reflection in many circumstances.

12.1.2 Refraction

In the theory of optics, this is typified by the re-direction of light rays at glass/air interfaces or by ray bending in a medium whose refractive index varies along the path of the ray. The effect is important in UHF propagation where the troposphere* is the transmission medium.

12.1.3 Diffraction

Light casts sharply defined shadows because the wavelength of light is extremely small when compared with the size of most obstructions. However, fringing effects due to diffraction can be observed in light experiments. Diffraction can be important in radio propagation as it leads to reception beyond the line-of-sight.

The consequences of these effects for television broadcasting are seen in what follows but first it is necessary to consider *free-space* propagation which

* The troposphere is the lowest part of the atmosphere, extending from ground to a height of about 10–15 km, characterised by large air movements and 'weather' effects.

is used as a yardstick against which practical reception conditions can be gauged.

12.2 FREE-SPACE PROPAGATION

The concept of free-space propagation is theoretically simple but only arises in practice when all obstructions are far enough away from the transmission path to be ineffective, and the transmission medium is uniform.

An isotropic point source of light is fairly easy to imagine; that is a very small lamp emitting photons uniformly in all directions. If the lamp is situated at the centre of a hypothetical sphere of radius d and generates a constant power W, the power flux w at the surface area is

$$w = W/4\pi d^2 \quad (\mathrm{m}^2) \tag{12.1}$$

It is useful to have the concept of an isotropic source of radio waves even if it is not a very practical thing. Again, in broadcast studies, radio waves are thought of in terms of the electric and magnetic fields that they generate. In the model of a plane wave front, the electric and magnetic fields at any point are at right angles to one another and both at right angles to the direction of

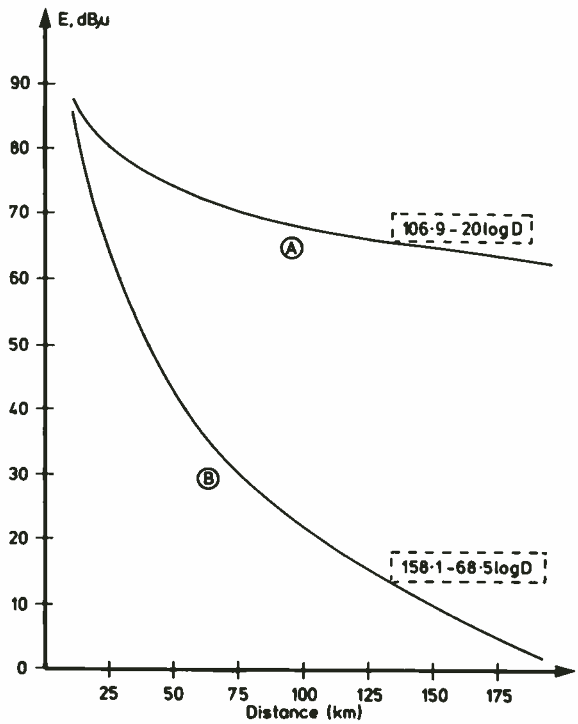


Figure 12.1 Field strength v. distance curves: (A) free-space curve, (B) median curve over land

propagation. At sufficiently large distances, the spherical surface is substantially plane.

Concentrating attention on the electric field component E, it can be shown that, in a plane wave, the power flux is

$$w = E^2/120\pi \tag{12.2}$$

(Compare Equation 12.2 with $w = E^2/R$.) Combining the two expressions for w gives

$$E = \frac{\sqrt{30W}}{d} \tag{12.3}$$

The field strength thus falls off as $1/d$, and not as $1/d^2$ shown in Equation 12.1. The fundamental dimensions of the electric field strength are volts/metre.

The isotropic source of radio waves is not a practical entity. The conventional reference source is the linear, half-wave, centre-fed dipole antenna. The radiation pattern of the dipole is not uniform (i.e. it is not isotropic). In a plane through the centre of the dipole and at right angles to it, the power flux is maximum and is 1.64 times as greast as that from an isotropic source radiating the same total power.

If W watts is fed to the reference dipole in free-space the maximum field, at distance d, is thus:

$$\begin{aligned} E &= \frac{\sqrt{(30 \times 1.64W)}}{d} \\ &= \frac{\sqrt{(49.2W)}}{d} \\ &= \frac{7\sqrt{W}}{d} \quad \text{(V/m)} \\ &= \frac{7000\sqrt{W}}{D} \quad (\mu\text{V/m}) \end{aligned} \tag{12.4}$$

when D is the distance expressed in km. E is, of course, the r.m.s. value of the field. The field value is often expressed in decibels relative to 1 μV/m (dBμ).

$$E = 76.9 + 10 \log W - 20 \log D \quad (\text{dB}\mu) \tag{12.5}$$

If the power W is standardised to 1 kW,

$$E = 106.9 - 20 \log D \quad (\text{dB}\mu) \tag{12.6}$$

This latter expression is seen as curve A in Fig. 12.1.

12.3 SHADOWING AND DIFFRACTION

As the wavelength associated with television colour broadcasting is short (about 0.5 m), obstructions in the path between the transmitting antenna and the viewer's antenna will appear relatively large and therefore shadowing effects will be much more prevalent than at the longer wavelengths associated with Band I (405-line television) or Band II (FM sound broadcasting) for example. In hilly or mountainous areas, the range of satisfactory service provided by a transmitter is mainly determined by the terrain due to the shadowing effects. The shadows are not sharply defined, however, and reception beyond the line-of-sight condition can be possible for limited distances.

Because of the shadowing effect, many more transmitters are needed to cover the United Kingdom than are needed at the lower frequencies. The transmitters encompass a wide range of sizes with output powers from 80 kW down to a few watts. In simple cases a theoretical prediction may be made of the attenuation incurred by the presence of an obstruction. Figure 12.2 illustrates this for the case of diffraction over a rounded hill at UHF (for either horizontal or vertical polarisation). Because the mathematical model is only a simplified representation of reality, such calculations can only be treated as approximate estimates.

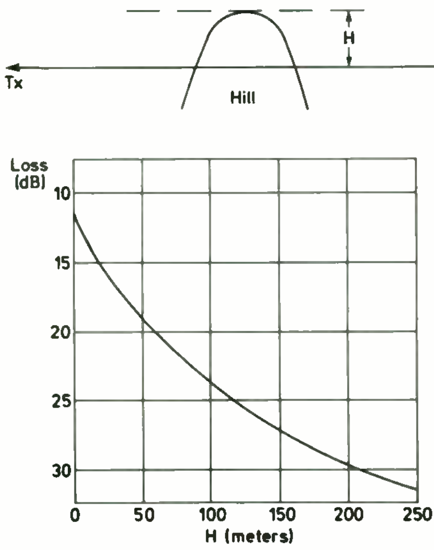


Figure 12.2 Example of theoretical diffraction loss over a rounded hill

The receiving antenna is rarely situated so high as to be completely 'clear' of obstructions and it is frequently surrounded by roofs, chimneys, trees and a variety of other 'clutter'. The result is a random resultant of diffraction and reflection effects that leads to additional signal attenuation of perhaps 8 to 10 dB in typical situations.

12.4 REFRACTION AND LONG-DISTANCE PROPAGATION

The effect of the standard or average atmosphere on radio waves is to make them travel along slightly curved paths. It is as if earth curvature was less than actual when the paths are drawn as straight lines. This fact is often allowed for by ascribing a hypothetical mean radius to the earth which is 4/3 times actual.

During certain anomalous meteorological conditions, often associated with the onset of anticyclonic still weather and high pressure, the refractive index profile in the troposphere may undergo a sharp increase in rate of change at some height above the ground. Then radio wave paths suffer increased bending and, in fact, signals may become trapped between the ground and the refractive layer or in an elevated duct. This so-called ducting mechanism can mean that strong signals are propagated over long distances and the received field strength can be in excess of that due to free-space propagation. In general, the effect is more prevalent over sea-paths or those in which the majority of the distance is over sea. Consequently, co-channel interference due to anomalous propagation over long distances for small percentages of the time must be taken into acount in planning. Long term measurement campaigns were conducted in the past, over suitable paths, to obtain an appreciation of the magnitude and time statistics of the anomalous fields. Existing planning curves for small time percentages and long distances were largely based on the results obtained in these campaigns.

12.5 REFLECTION AND MULTIPATH PROPAGATION

In the context of television reception at UHF, reflection of the transmitted waves can lead, most importantly, to multipath propagation. This is the case where the receive antennas in built-up areas may pick up more than one incoming signal due to reflections from buildings, etc. The case where two principal signals are picked up is probably most frequently encountered. When the relative time delay between the two signals is long enough (more than 10 μs, say) double images may be observed.

Short time delays can cause other undesirable effects. When there are just two similar signals with relative time delay T seconds, a radio-frequency sweep exhibits a 'standing wave' pattern of amplitude variation. The frequency difference between maxima and minima in this pattern is equal to $1/2T$ Hz. If the frequency difference is comparable, say, with the channel spacing, difficulties may be experienced in equalising the performance on different channels. But suppose, for instance, that the vision carrier frequency occurs at a maximum and the nearest upper minimum is 4.43 MHz away. In this case, colour performance may be seriously degraded. (The total path difference for the two signals in this case is about 34 m: it is a local effect).

A minimum directivity of viewers' antennas, to discriminate against reflected signals, is assumed in planning studies (see Fig. 14.13). Careful positioning of receive antennas may be required to reduce, as far as possible, the effect of local reflections.

12.6 PROPAGATION OVER THE REAL TERRAIN

In practice, the received field strength is usually less than the theoretical free-space value because of obstructions, reflections, etc. On paths over a real terrain, a wide range of different values relative to the free-space value will be experienced. Inevitably, because of the complexity of the overall situation, useful predictions of field strength will depend to a large extent on observations (measurements) over as large a number of actual paths as possible. Curve B of Fig. 12.1 represents an estimate of field strength against distance derived from many measurements in the field. Its interpretation must be understood as follows:

(1) It applies to the broadcast Bands IV and V.
(2) The height of the centre of the transmit antenna above the mean terrain height is 300 m. The transmit reference antenna is a half-wave dipole fed with 1 kW of radio power.
(3) It relates to overland paths across the rolling terrain typical of the UK and parts of Europe.
(4) The height of the receive antenna above the mean terrain is 10 m.
(5) The plotted value of E represents the median value (i.e. exceeded for 50% of the time). For at least 50% of all paths over the given distance, the observed medians are predicted to be greater than or equal to this value of E. (For short, this is known as a '50% of time, 50% of locations' curve.) It will be seen that curve B falls off much more rapidly (faster than $1/D^3$) than the free-space curve A.

Similar curves can be constructed for other transmit antenna heights, different time percentages, sea paths and so on. A representative collection of curves may be found in the texts of Study Group V of the CCIR.

12.7 UHF TELEVISION CHANNELS

Band IV 470–582 MHz, 14 channels (21–34)
Band V 614–854 MHz, 30 channels (39–68)
Vision carrier frequency for channel $N = (8N + 303.25)$ MHz
Wavelengths: mid-band IV = 57 cm
mid-band V = 41 cm

CCIR, *Green Book*, Vol. V, *Propagation in Non-ionized Media*, XVth Plenary Assembly, Report 239, Report 715 (1982)

Index