DATA REDUCTION IN TELEVISION SIGNALS

FOR BANDWIDTH COMPRESSION

A thesis submitted for the degree of Doctor of Philosophy in the Faculty of Engineering, University of London.

by

Marcus Paul Barton
Beit Scientific Research Fellow, 1962-4

Imperial College, London, September 1967.

ABSTRACT

Television and other picture signals as usually transmitted are extremely wasteful of channel capacity. Both statistical and non-statistical redundancy exist and many schemes for bandwidth reduction have been proposed. The scheme examined during the work reported here employs run-length coding.

Data reduction is the process by which redundant material is removed from the original signal prior to coding for efficient transmission (elastic encoding in the scheme considered). It is here divided into run-end detection and run-length restriction. Consideration of the former is facilitated by the use of a model to represent the signal; several models are looked at but the exponential is selected for detailed consideration. A general procedure for run-end detection is outlined and application of this to the exponential model results in the choice of one of three ways of fitting exponential curves to data by the method of least squares. It is shown that the approximations required to make instrumentation feasible do not introduce serious errors As exponential run-end detection will require conversion of the data into a linear form, a first-order run-end detector was built and assessed.

Consideration of run-length restriction starts with an outline of the need for it and some theory. Methods of carrying out the process are reviewed and that involving the blanking of clock pulses is chosen. Factors concerned in the design of the necessary apparatus are discussed and the apparatus is described. Its performance was quite satisfactory, direct measurements of the effect of run-length restriction being possible for the first time.

The two sections of the work are brought together by a comparison of variable velocity scanning and 1st order interpolation and the design and construction of a first-order interpolator. This used a chain of resistors to reconstruct omitted samples, performance being fair but limited by the readout circuits.

ACKNOWLEDGMENTS

The help and encouragement given to me by my supervisor, Professor Colin Cherry, during all stages of the work are gratefully acknowledged.

Colleagues in the Communications Section at Imperial College and at British Telecommunications Research Limited have also been most helpful; discussions with them have been invaluable. The Management of British Telecommunications Research Limited is thanked for making available facilities for processing the thesis.

Mr. A.A. White and other staff of the Electronics Workshop at Imperial College deserve thanks for their help with the construction of equipment.

Maintenance grants were provided first by the D.S.I.R., and later by the Trustees of the Beit Fellowship for Scientific Research; both bodies are thanked for their support.

Funds for equipment were provided by the N.R.D.C. who are thanked for their generosity.

My wife is also thanked for much patient work in typing the first draft of the thesis and Miss Verline Edsall for preparing the stencils.

CONTENTS

PAGE

PAGE

SYMBOLS AND ABBREVIATIONS

In a few cases, a symbol has been used for more than one purpose. It will be obvious, from the context, which meaning is intended in such cases. This lapse from strict rigour has been permitted because readability and ease of understanding have been considered more important.

A	the magnitude of a transfer function
a	a constant, often the constant term associated with a first order run
B	a bistable element
b	(a parameter of an exponential curve (the slope of a 1st order run
C	(a compression ratio (a channel capacity (a capacitor (capacitance per unit length of a cable
C_m	a maximum compression ratio
CPG	a clock pulse gate
c/s	cycles per second
D	(a diode (a parameter of an exponential curve
DFS	the Detail Function Shaper
dB	decibels
∂	the partial differentiation operator
E, E_1, E_2	voltages
E_{gk}	a grid-cathode voltage (with respect to cut-off)
e	the exponential operator
F	a factorial moment
F_1, F_2 etc.	Threshold units or the values of their thresholds
f	a frequency, usually the Nyquist sampling frequency
G	a gate
g_m	mutual conductance
H	an information rate
h	a parameter of an exponential curve
h(t)	impulse response

I	an inhibit gate
I_b	beam current
i, j	sample values
j	the square root of -1
K	a parameter of an exponential curve found by the method of factorial moments
K_1, K_2	constants
L	the inductance per unit length of a cable
M	a power moment
MMV	monostable multivibrator
m	(a number of samples (a ratio of sampling rates
N	(an average noise power ($\log_2$ of a number of amplitude levels (a number of samples per second
n	a number of samples or points
$\bar{n}$	an average run-length
P(n)	the probability that a random variable will have the value n
p_i	the probability of a sample having the value i
q	(a number of samples (a bandwidth reduction factor due to high-level AM encoding
R	redundancy
R_e	an emitter load resistor
R_L	a load resistor
R_K	the emitter resistor in a "long-tailed pair" circuit
r	a mean run-length
S	(seconds (an average signal power (the true value of a signal sample
SGU	Sample Gating Unit
s	the sum of squared errors
T	a time interval, often the Nyquist interval
t	time
t_o	a group delay

V	a scanning velocity
V_0, V_1, V_2 etc.	Signal samples
VDU	Variable Delay Unit
v	a signal voltage
v_{in}, v_o	input and output voltages
W	a bandwidth
Ω	ohms
X	(a number of samples (a sample of the signal, or its magnitude
Y	the first difference of two signal samples
y	a sample amplitude
Z_o	characteristic impedance
^	indicates an estimated value
β	the current gain of a transistor
Δ	the difference operator
ε	an error
θ	a parameter substituted for $\omega\tau$
ρ	a parameter of an exponential curve found by the method of factorial moments
Σ	the summation operator
σ	a standard deviation
τ	a delay or rise time
τ_D	delay per metre of cable
Ω	ohms
ω	a weight or an angular frequency
ω_c	an angular cut-off frequency

Prefixes

M - mega $\times 10^{6}$

K - kilo $\times 10^{3}$

m - milli $\times 10^{-3}$

μ - micro $\times 10^{-6}$

n - nano $\times 10^{-9}$

p - pico $\times 10^{-12}$

GLOSSARY

Bandwidth

The number of cycles per second expressing the difference between the limiting frequencies of a frequency band (I.R.E., 1948).

Binary

(i) (noun) a bistable circuit so connected as to change state at the arrival of each input pulse.

(ii) (adjective) having the base 2, or 2 states.

Bit

A binary digit.

Data Reduction

The removal of unnecessary data from a signal in such a way that the signal can be reconstructed within satisfactory perceptual limits from the reduced data by the application of certain predetermined rules.

Dot Interlace

A scanning process in which each scanning line is subdivided into a number of dots, adjacent dots being scanned during successive fields.

Facsimile

In this thesis, the term "facsimile" is used to refer to any scheme for transmitting electrically, a single still picture. Usually such schemes operate at high definition (100 lines per inch or more) and use electromechanical equipment.

Factorial Moment

A function defined by:

$$F_m^{(q)} = \sum_{i=1}^{m} \binom{m + q - i}{q} y_i$$

where q is the "order" of the factorial moment and the data, y_i, are assumed to be spaced at unit intervals of x.

Field

One of the two (or more) equal parts into which a frame is divided in interlaced scanning (NTSC· I.R.E., 1948).

Frame

A complete scan of the picture (with the usual 2:1 line interlace, a frame consists of two fields).

Flyback

The operation of returning a scanning spot from the end of a scan (line or field) to the correct position to start the next scan.

Flying-spot Scanner

A pick-up device in which a luminous spot scanning a raster on the face of a cathode-ray tube is focussed on a diapositive of the scene to be transmitted and the transmittance of each element of the diapositive is measured by a photomultiplier. (The scheme can be modified for use with real scenes also.)

Flywheel Synchronisation

An arrangement in which the timebase contains an oscillator the phase of which is compared with that of the incoming synchronising pulses. The oscillator frequency is adjusted by the comparator so as to oppose phase changes. By this means, a short interruption of the synchronising signal, due perhaps to interference can be ignored.

Kell Factor

The ratio of the effective number of scanning lines to the actual number. The difference arises because there is no filter to interpolate between successive lines as would be used if interpolation between samples of the signal were to be carried out.

Luminance

The luminous intensity of any surface in a given direction per unit of projected area of the surface as viewed from that direction. ("Brightness" is the corresponding term relating to sensation.)

Maximum Likelihood Estimator

The likelihood function of n random variables $x_1, x_2, x_3 \ldots x_n$ is the joint probability density of the n random variables $g(x_1, x_2, x_3 \ldots x_n, \theta)$ which is considered to be a function of θ. If $\hat{\theta}$ is the value of θ which maximises $g(x_1, x_2, x_3 \ldots x_n, \theta)$ then $\hat{\theta}$ is said to be the maximum likelihood of θ.

Monochrome

Although not strictly in accordance with approved definitions, the term "monochrome" has been used in this work to denote black-and-white, as distinct from colour, television.

Monoscope

A video generator comprising a cathode-ray tube with an in-built target in the form of the scene it is desired to televise, varying signal levels being produced by varying secondary emission characteristics of the target.

Nyquist Interval

The maximum permissible interval (T) between successive regular samples of a signal limited to a bandwidth W if the signal is to be reconstructed from the samples without error merely by the use of a low-pass filter of bandwidth W. $T = 1/2W$.

Photomultiplier

A light detecting device consisting of a photo-cathode and an electron multiplier in the same envelope.

Pick-up Device

Any equipment for scanning a live scene or a reproduction and generating a video signal therefrom.

Picture Point or Picture Element

A segment of a scanning line, the dimension of which along the line is equal to the nominal line width.

Power Moment

The power moments of a distribution are the expected values of the powers of the random variable which has the given distribution.

Raster

A predetermined pattern of scanning lines which provides substantially uniform coverage of an area. (I.R.E., 1948).

Redundancy (Statistical)

A property given to an information source by virtue of an excess of rules (syntax) whereby it becomes increasingly likely that mistakes in reception will be avoided.

(Non-statistical)

That which is due to the limitations of the observer, either in his inability to perceive all aspects of the received information or in his ability to draw on his store of past experiences.

Scanning

The process of analysing or synthesising successively, according to a predetermined method, the light values of picture elements constituting a picture area.

Signal-to-noise Ratio

The ratio of the value of the signal to that of the noise. In television practice, the ratio of peak signal power to r.m.s. noise power is employed.

Standard Deviation

The root-mean-square value of the deviations of a series of quantities from their mean. (The square root of their variance.)

Unbiased Estimator

An estimator $\hat{\theta}$ is called an unbiased estimator of θ if the expected value of $\hat{\theta}$ equals θ.

Video

A term pertaining to the bandwidth and spectrum position of the signal resulting from television scanning. In current usage, video means a bandwidth of the order of megacycles and a spectrum position that goes with a d.c. carrier.

1. INTRODUCTION

Human life is made up of experiences which are perceived by the senses of sight, hearing, taste, smell and touch. For a great many purposes, however, only some of these senses are used.

Communication can be established using only one of the five senses, although the use of more than one may enhance realism or, by taking advantage of redundancy, may increase reliability. Thus a television programme intended for entertainment usually comprises both picture and sound.

Television is but one example of a visual communication system; it is probably the most frequently cited example because domestic television is so common (in Britain at any rate). By domestic television we are conditioned to think of high quality pictures with little or no visible interference, the successive frames being presented in quick succession so that flicker is not annoying and an impression of motion is obtained.

Facsimile systems are also in use and here the definition is frequently higher than that employed in television, the transmission time for a single picture being correspondingly long. (A system marketed by Muirhead & Co. Ltd. uses 96 lines per inch and requires 4.4 minutes to transmit a picture measuring 18 x 22 inches.) Although there is no requirement here to transmit moving pictures, the quality is still very high.

In Professor Cherry's section at Imperial College, the current thinking is that little attention has so far been given to the possibilities of using channels in which the interference or distortion (or both) are quite severe; it is conceivable that useful pictures could be transmitted in this way. Enjoyable viewing is not always the criterion; in military and commercial applications severely distorted speech is often tolerated - why not use pictures in a similar economic manner?

1.1 Television Techniques

Fortunately, for the sound engineer, many of the characteristics of a sound can be represented by a one-dimensional time-varying signal. A simple way to convert sound energy to electrical energy, and vice-versa, was hit upon quite soon after the invention of the electric telegraph had laid the foundations of electrical communication, and even today there seems to be little need for radical change or the adoption of new principles.

In the case of visual communication, the problem is considerably different, however. Scenes in real life are characterised by distributions of colour and brightness in three dimensions. It is well known, from drawing and photography, that such scenes can well be represented by a two-dimensional picture which, in many cases is quite satisfactory in black-and white (monochrome). Clearly, it is not possible to obtain a single signal which represents the scene directly as in the case of aural communication.

This difficulty is overcome by scanning the picture, the signal which is sent being some measure of the brightness of a small region of the picture. At the receiver, the picture is built up by a similar scanning process, the signal being used to control the brightness of a small spot and the scanning being synchronised to that at the transmitter. The first practical scanning device was the Nipkow disc invented by Paul Nipkow in 1884. The principle of its operation is well known.

Using the Nipkow Disc, pictures having up to 240 scanning lines could be transmitted at a rate of 25 pictures per second and a regular television service using these standards existed in this country during late 1936 and early 1937[35]. Higher definitions than 240 lines are beyond the range of mechanical devices, however, and electronic methods were at this time being developed, culminating in the adoption by the BBC of the Marconi-EMI system[6].

Although, with a fully electronic system, many modes of scanning are easily possible (e.g. spiral, figure-of-eight or more complicated Lisajous' figures, variable velocity[4], little thought seems to have been given to any but the now common system of scanning at constant velocity in a series of horizontal straight lines. This was presumably a result of conditioning by the Nipkow disc and the reading process. Indeed, the only alternative to the present system which appears to have been considered is the "zig-zag" scanning system in which lines are scanned alternatively left-to-right and right-to-left so that no line flyback is involved. Quite reasonably, this was discarded since any slight non-linearity of scan or inaccuracy in synchronisation will cause the break-up of a vertical line. In a system of unidirectional scanning, considerable non-linearity can be tolerated as this results only in a distortion of the geometry of the picture, something which is quite often unnoticeable.

Unfortunately, the unidirectional system now universally adopted is exceedingly wasteful in several ways. Most important is the fact that all possible pictures are treated as being equally likely. A 405 line system, for instance, can reproduce every one of about $10^{339,000}$ pictures (See Appendix 4). No account is taken of whether the pictures transmitted are likely to have any meaning to the observer or of whether the pictures could be generated by scanning natural scenes. (Test cards are artificial scenes and can always be devised to have any desired properties.)

Further waste is caused by the repetitive nature of the scanning. Because of the high incidence of vertical contours in natural scenes, a given scanning line is very much like its immediate neighbours. Successive frames, too, are very similar in the majority of cases. Only a quick "cut" from one scene to another results in a large inter-frame difference. At the other end of the scale, successive picture elements are frequently very similar because brightness does not always change rapidly in a horizontal direction. These correlations between successive picture points, successive lines and successive frames are said to cause redundancy (see later sections).

A feature resulting in less waste is the allocation of a large proportion of the scanning cycle to flyback. This was originally necessary because of the limitations of scanning circuits and scanning coil assemblies. (Although the flyback intervals are used to send synchronising signals, these signals could be sent in much less time with no sacrifice of quality.) With modern techniques and materials, flyback times could be much reduced or, by use of an alternative scanning system, even eliminated. In the British 405 line system, 14 lines are allocated to vertical flyback, and, of the line period, 16.2 - 18.7% is allocated to horizontal flyback. Thus only about 77% of the time is available for picture transmission.

1.2 The Need for Television Bandwidth Compression

Once a method of producing the television signal has been decided upon, the necessary bandwidth can be calculated. Since the resolution of the eye is equal in the horizontal and vertical directions, it is reasonable to aim at this property in a television picture. The vertical resolution is determined by the number of lines per frame. The necessary bandwidth is thus determined by this figure together with the frame repetition frequency (See Appendix 4).

Although the factor now usually attributed to Kell had been mentioned in a paper by Kell, Bedford and Trainer[30] in 1934, this was not taken into account in the design of the British 405 line system. As a result, the horizontal resolution is somewhat better than the vertical. The bandwidth allowance for the video signal is 3 Mc/s. (More on other systems.) Clearly, to transmit this signal by radio, a carrier frequency of several tens of megacycles is necessary. Although, in the 1930's, when all-electronic television was being developed, the VHF bands were little used, there were considerable problems to be overcome. Electromagnetic waves of this frequency are not normally reflected by the ionosphere as well as waves of lower frequencies (e.g. those of the broadcast bands). Thus, transmission was expected to be limited to "line-of-sight" only. Tests showed that this estimate was somewhat pessimistic and that satisfactory reception could be expected over greater distances. Even so, several transmitters have been found necessary to serve the whole of the United Kingdom adequately, special low-powered transmitters being used to supplement the main ones where reception is exceptionally difficult due to features such as large ranges of hills.

Here then is a reason for seeking to reduce the bandwidth requirements of the signal. A narrower bandwidth would allow the use of a lower carrier frequency with correspondingly increased coverage. Alternatively, more programmes could be made available within the same bands.

Use of a VHF carrier also means that direct transatlantic television is not possible. This problem, initially overcome by the "Cablefilm" process which employed slow transmission of recorded material, is now solved by the use of communication satellites. The high cost of the scheme suggests that any means of sending more than one television signal via a satellite at any one time would be desirable. Thus, although measurements indicate that it is not possible to reduce the bandwidth sufficiently to employ a carrier capable of bridging the Atlantic, bandwidth compression may have its place in transatlantic television.

Other difficulties caused by the immense requirements of a television signal include the design of circuits with a satisfactory response * extending essentially from d.c. to several megacycles per second. Transmission by cable over a distance of more than a few miles is precluded and recording requires very special techniques.

---

* As well as a wide bandwidth, circuits must have a linear phase response in this band, in order to reproduce pulse-type waveforms faithfully.

1.3 Redundancy in Television Signals

In television, two types of redundancy are usually distinguished, statistical and non-statistical. The former arises from the fact that elements of a signal are not wholly independent. Thus successive frames are very similar, as are successive lines and successive picture points. An estimate of the amount of statistical redundancy can be obtained by measurements on the signal, the difference between the maximum possible information rate and the actual rate being computed.

Schreiber(47), and Kretzmer(32) have reported measurements to determine the amount of redundancy due to correlation between successive picture points. Both workers allowed 64 brightness levels so that the maximum information rate was 6 bits per sample. Kretzmer, using photographic methods and a small number of scenes found that the redundancy was between 2.4 and 3.4 bits per sample owing to correlation between adjacent picture elements. This indicates that a compression of about 2:1 would be possible if a perfect coding method were available. Schreiber obtained similar figures and, in addition, measured the third order probability density (due to correlation between a given sample and the two previous samples). He found that this introduced a small amount of redundancy but not as much as that introduced by correlation between a sample and its immediate neighbour only. Kretzmer also measured the simple amplitude probability distribution which indicated a one bit per sample redundancy.

Correlation between adjacent lines does not appear to have been investigated at all but Kretzmer gives some figures for frame-to-frame correlation measured from two motion-picture films. The values obtained for autocorrelation were 0.80 and 0.86 which, using a formula due to Elias(16) indicate a redundancy of only just over one bit per sample. This figure is surprisingly low as frame-to-frame correlation is usually thought to account for most of the redundancy in a television signal.

1.4 Statistical Redundancy

Redundancy is a term which is often used very loosely. One (mathematical) definition is: "Unity minus the ratio of the information rate of the source to its hypothetical maximum rate, when encoded with the same set of signs. Broadly, a property given to a source by virtue of an excess of rules (syntax) whereby it becomes increasingly likely that mistakes in reception will be avoided."(7)

To see how this definition can be usefully applied, consider first a source providing discrete signal samples at a rate of N samples per second. If each sample can have one of a range of values 1, 2, ...i....n, the prior probability of a sample having the value i being p_i, then the amount of information associated with each sample is

$$H_1 = - \sum_{i=1}^{n} p_i \log_2 p_i \text{ * } \quad \text{bits/sample} \qquad (1.1)$$

or

$$- N \sum_{i=1}^{n} p_i \log p_i \quad \text{bits/sec} \qquad (1.2)$$

assuming that all samples are independent.

In the more general case the value of any sample depends to some extent on the values of other samples, the simplest example of this occurring when the value of a sample depends on the preceding sample but not on any others. The probability of the second sample having the value j when the first is known to have the value i will be denoted: $p(j \mid i)$. (read "probability of j given i") The probability of the joint event is then

$$p(i,j) = p(i).p(j \mid i) \qquad (1.3)$$

and the information conveyed by the two samples is

$$H_2 = - \sum_{i,j} p(i,j) \log_2 p(i,j) \text{ bits per two samples.} \qquad (1.4)$$

---

*(i) The minus sign is needed because the logarithm of a fractional quantity is negative; thus a positive value is obtained for the quantity of information.

(ii) Taking the logarithm of p_i to the base 2 gives the information in binary digits (bits) per sample. To obtain the rate in other units it is simply necessary to use the logarithm to the appropriate base.

Clearly we can extend this treatment to the case where the value of the q th sample depends on those of the previous q-1.

Then
$$H_q = - \sum_{i,j,\ldots q} p(i,j\ldots q) \log_2 p(i,j\ldots q)$$
bits per q samples. (1.5)

The maximum information rate of the source is obtained when the samples are independent and every value which a sample can have is equally likely. i.e.

$$p_1 = p_2 = p_3 = \ldots\ldots\ldots p_i = \ldots\ldots\ldots p_n. \qquad (1.6)$$

Then $p_i = 1/n$ and from (1)

$$H_{max} = - \sum_{i=1}^{n} \frac{1}{n} \log_2 \frac{1}{n}$$

$$= \log_2 n. \qquad (1.7)$$

Thus, from the definition, the redundancy per sample, R, when q samples are correlated is given by

$$R = \log_2 n - \frac{H_q}{q} \qquad (1.8)$$

1.5 Non-statistical Redundancy and the Properties of Vision

Non-statistical redundancy is that which is due to the limitations of the observer, either in his inability to observe all aspects of the displayed picture or in his ability to draw on his store of past experiences, the brain using its store of prior probabilities to "fill in" where picture information would otherwise be necessary.

Six effects have been listed in the past [27], although this list is not exhaustive and interactions between effects are known to occur. Advantage is taken of some of these effects in present day television systems.

1. Of all the effects, limitation of visual acuity is probably the most important. The fact that the eye does not have infinite resolving power allows us to find photographs and television pictures satisfactory despite their grain and line structure respectively. Thus a scanning process can be employed in television and suitable line standards and brightness can be chosen so that the line structure is not perceivable by observers with normal eyesight seated at a reasonable distance from the display screen.

Since the acuity of the eye is approximately the same in the vertical and horizontal directions, a similar equality is usually aimed at in the design of television systems. Thus a bandwidth requirement can be stated in terms of the number of lines and the frame repetition frequency.

The fact that visual acuity deteriorates as brightness or contrast is decreased(19) could be advantageously used since resolution could be reduced in regions of low contrast or brightness. Such a scheme could easily be combined with the Closed-Loop or the Open-Loop bandwidth compression system (See Sections 1.8 and 1.9).

2. The observer's ability to perceive flicker at the brightness levels used in television makes a field repetition rate of 50 c/s (60 c/s in the U.S.A.)* necessary. Persistence of vision is such, however that 25 pictures per second are quite adequate to convey an impression of motion. 2:1 interlace is therefore used in all present day television systems, alternate lines of the raster being scanned in one field (taking 1/50th of a second on British Standards) and the remainder in the other.

3. The exact amplitude and shape of short transients is poorly perceived by the eye. At present this feature is not utilised in television, but several schemes have been proposed(31, 48, 49) to take advantage of this redundancy. Broadly, the schemes involve the direct transmission of the low frequency components of the signal. The high frequency part is then coarsely quantised and run-length coded (see later sections).

---

* Not because American eyes are better than British, but because it is convenient to have the field repetition rate the same as the power-supply frequency!

4. The resolving power of the eye is not uniform all over the field of vision but decreases towards the edges. This is so because of the distribution of the light-sensitive rods and cones in the retina. A small portion of the retina directly opposite the lens has a very high concentration of cones, the receptors mainly responsible for "daylight" vision. Thus a high resolving power is possible. If television pictures always had sufficient interest at the middle of the screen, less resolution could be tolerated at the edges of the picture.

5. Despite poor perception of the exact shape and amplitude of short transients as mentioned in 3 above, the eye "hooks on" to edges and these form the main items of interest in a picture. The eye is thus very critical of errors in positioning an edge while the exact brightness of large areas is not so important. The requirement of accurate edge positioning is reflected in the fact that precise time synchronisation is required for acceptable reproduction of pictures. Sophisticated techniques such as flywheel synchronisation are therefore commonly employed in television receivers.

6. It has been estimated that the information capacity of the eye-brain channel is about 30-50 bits per second[55]. Although, at first sight, it appears that this figure exposes vast redundancies in television transmitted at (say) 40 Mbits/sec., caution must be employed. Unless the scanning spots at the transmitting and receiving ends of the system can be made to follow the movements of the viewers' eyes, a complete picture must be transmitted in an interval comparable with the bit interval in the eye-brain channel. Additionally, where there is more than one viewer, the needs of all the viewers must be satisfied and unless their eyes can be made to move in synchronism this implies a multiplicity of scanning patterns simultaneously. A means for exploiting this particular redundancy is not easy to imagine.

1.6 Run-length Coding

It is fairly well known, from the Sampling Theorem*, that a signal limited to a bandwidth W cycles per second can be adequately represented

---

* Well explained by Woodward[60]. Probably first derived by Nyquist[41] but more fully explained by Hartley, Gabor and Shannon.[26, 20, 53]

by a set of samples of its amplitude taken at intervals of 1/2W seconds and that the original signal can be recovered with no error by passing the sample train through a low-pass filter of bandwidth W. (See Figure 1.1.) To recover the signal it is not necessary to know the time-origin of the samples.

From examination of the waveforms of some signals it can be seen that the amplitude is not always changing rapidly but is often constant over several sampling intervals or is changing only slowly (e.g. region A-D in Figure 1.1). Intuitively it is felt that such regions of the signal occupy less bandwidth than the full W c/s permitted. Further, where the signal amplitude is constant, it is clear that only the first sample of such a region is necessary, together with a signal to indicate the time for which this amplitude applies. This method of signal representation is known as "Run-length Coding". The earliest reference to the process which I have been able to find is given by Oliver[42] in 1952. It is clear from his paper, however, that Run-length Coding was quite well known at this time.

The saving to be expected from run-length coding can readily be estimated. Consider a signal as shown in Figure 1.2, 2^N amplitude levels being necessary to describe each sample of the signal. The Nyquist sampling rate is 2W samples per second so that the information rate if every sample be transmitted is 2WN bits/sec.

Now suppose that run-length coding is employed and it is found that the mean run-length is r Nyquist intervals. The average number of bits required to specify the length of the runs will thus be $\log_2 r$ bits per run. In addition, however, a "punctuation" signal must be sent at the end of each run; one bit per run is an adequate allowance for this. The average information rate required to specify the amplitudes of the runs will be

$$\frac{2WN}{r} \text{ bits/sec} \tag{1.9}$$

so that the total information rate is

$$\frac{2WN}{r} + \frac{2W}{r} (\log_2 r + 1) \text{ bits/sec} \tag{1.10}$$

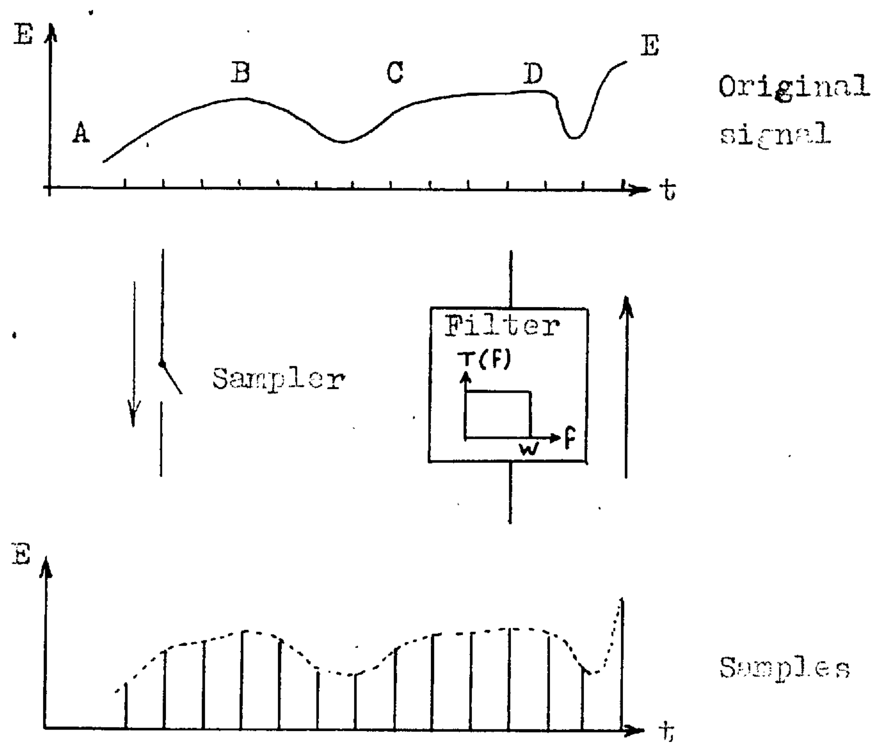


Figure 1.1. The Sampling Theorem

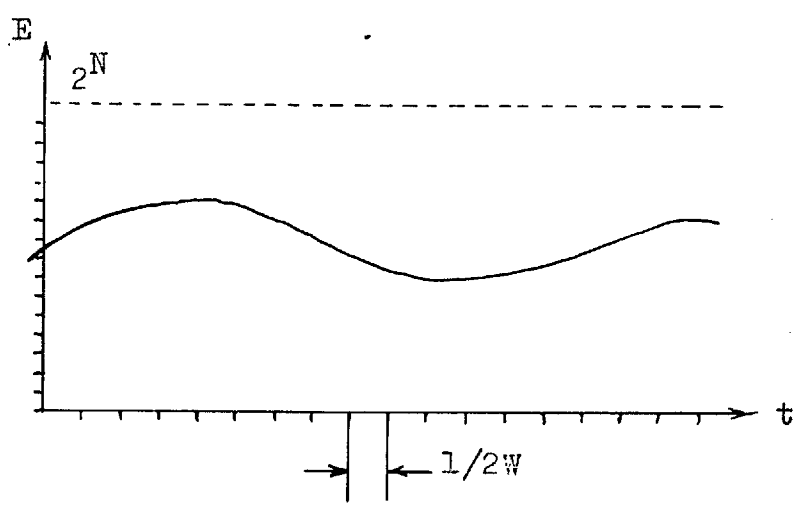


Figure 1.2. Quantisation and sampling

This saving is usually expressed as

$$C = \frac{\text{original info. rate}}{\text{reduced info. rate}} = \frac{2WN}{\frac{2WN}{r} + \frac{2W}{r}(1 + \log_2 r)}$$

$$= \frac{rN}{N + 1 + \log_2 r} \qquad (1.11)$$

Curves of C against r for various values of N are shown in Figure 1.3. It will be noticed that the value of N has little effect on C. With a mean run-length of 4 Nyquist intervals (a reasonable value) the saving to be expected is about 2.5:1. This figure must, however, be considered with caution.

If bandwidth reduction is required, the data has to be transmitted uniformly in time and some loss will inevitably result since no perfect coding process is available for this purpose. In addition, we have assumed that the runs are clearly defined. If a television signal is the subject we are investigating, random noise will inevitably accompany it and there may be difficulty in determining where a run ends. This is discussed further in later sections.

A run-length coding system suitable for use with facsimile systems (where noise presents no problem) has been described by Michel et al.[37]

So far, a run has been considered to be a region where the signal amplitude is constant. A more general, and more useful definition, however, states that a run is a region of the signal where some stated parameter (or parameters) remains constant within certain limits. Thus, runs of constant slope, constant curvature, etc. can be defined. These concepts will be examined further in Section 2.3.

1.7 Proposed Bandwidth Compression Schemes

Bandwidth compression schemes have been reviewed in some detail by Prasada [45] and in less detail by Hill[27]. Little elaboration will therefore be provided in the list which follows. For convenience, the schemes have been divided into those which take advantage mainly of statistical redundancy and those which take advantage mainly of non-statistical redundancy.

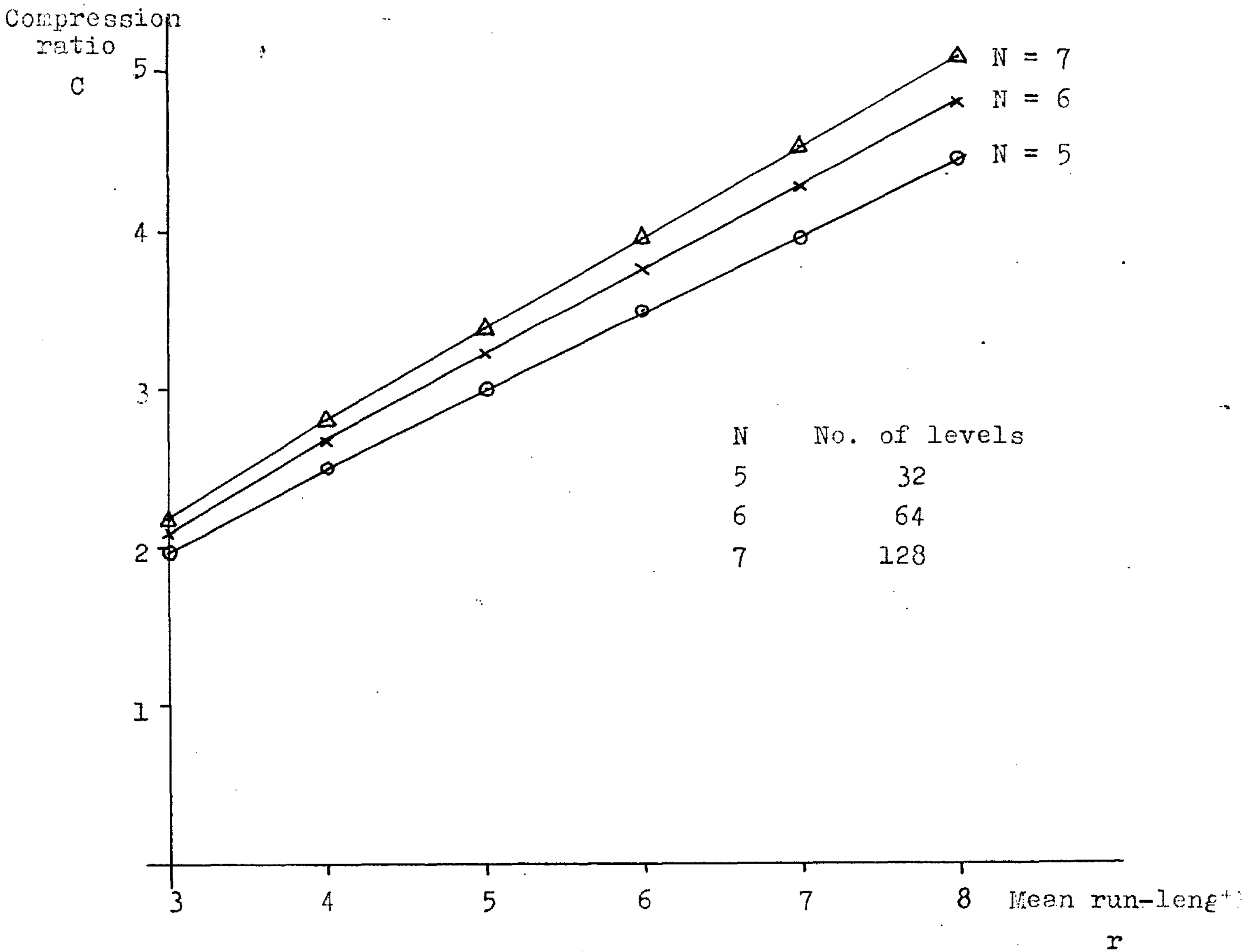


Figure 1.3. Variation of compression ratio with mean run-length and number of amplitude levels

(a) Statistical Systems. (i) for television.

(1) Bell & Howson [5] utilised blank spaces in the frequency spectrum of a television signal to achieve a 2:1 saving by frequency interleaving.

(2) In a theoretical investigation Tsukkerman[58] examined a system in which the co-ordinates of points where brightness changes occurred were transmitted.

(3) Julesz[29] using a computer to process signals at a low speed, reconstructed pictures by linear interpolation, the sampling points being determined by a second differencer.

(4) The contour interpolation scheme investigated by Hill[27] allows the omission of alternate lines, the omitted lines being reinserted at the receiver by interpolating between those on either side.

(5) The only full description of a scheme utilising frame-to-frame correlation appears to be that due to Schröter[50]. The difference between successive frames was transmitted by variable velocity scanning techniques. Considerable storage was required both at the transmitter and the receiver.

(6) Harrison[25] has reported some measurements on linear prediction, the "error power" (difference between predicted and actual signals) being recorded. He found that slope and planar prediction were marginally better than previous value prediction.

(7) One of the earliest schemes to be proposed is that of Cherry and Gouriet[9]. By means of a feedback loop controlling the line time-base the picture is scanned at a speed depending on the amount of detail at any instant. At the receiver the necessary timebase velocity was found by differentiating the received signal.

(8) After some investigation of the previous scheme, Cherry and Beddoes[8] put forward a system in which the scanning velocity at the receiver was controlled by a separate signal.

(9) A run-length coding scheme for television signals has been described by Gouriet[22]. As all run-lengths are permitted in this system, a very high signal-to-noise ratio (about 72 dB) is needed to specify the positions of the run ends.

(10) The Open-Loop System arose from investigations of the system described in 7 and 8 above. It is fully described in Section 1.9.

(ii) For facsimile.

(1) Run-length coding for facsimile material was proposed by Michel, Fleckenstein and Kretzmer[37]. In their scheme the transmission time depended on the complexity of the material.

(2) A scheme instrumented by Treuhaft[57] gave a saving of 12 times by sending outlines only.

(3) By severely reducing standards a signature verification scheme[2] has provided 100 line definition over a small area, each picture requiring 5 secs. for transmission on a 5Kc/s channel.

(b) Non-statistical.

(1) In work reported by Cunningham[13], a computer was programmed to interpolate between "data-blocks", the average signal level in each data-block being sent.

(2) Ando[1] reported promising results of tests on a system in which the h.f. components of a signal were coded as a 2-level pulse train, contour positions, polarities and widths being specified. Standard triangular pulses were used for all the transients in the reconstructed picture.

(3) An engineered system described by Schreiber and Knapp[48] splits the signal into two frequency bands, the high-frequency components being coarsely quantised and run-length coded while the low-frequency components are sent directly.

(4) Band splitting was again used in Kretzmer's reduced alphabet system[31] but here the high-frequency signal was taper-quantised. The system was said to give rise to spurious contours near sharp edges.

(5) "Synthetic Highs"[49] is an extension of the two previous systems, the first-difference of the highs now being taper-quantised. At the receiver, the amplitude and position of transients are determined by the incoming signal while their shape is made the inverse of the shape of the l.f. signal. The scheme was instrumented but many storage tubes were necessary at both the transmitter and the receiver.

(6) Graham[24] has described a system employing tapered quantisation of the first-difference of a signal.

(7) In another paper[23], Graham proposes coarse quantisation (3 bits per sample) of the high-frequency components only.

(8) In an engineered closed-circuit system, Deutsch uses a pseudo-random scan with a frame period of 2 2/3 secs.[14].

Analyses of the problems of bandwidth compression in general terms have been presented by Seyler[52] and Jesty[28]. Deutsch[15] has also given an analytical approach to the use of reduced standards for special applications such as tape recording, transmission by cable. The techniques suggested include dot-interlace, increased line-interlace ratio and the reduction of blanking intervals.

1.8 Investigation of the Variable Velocity System

Preliminary investigation of the Cherry-Gouriet closed-loop system was carried out by McCully who published his findings in 1954[36]. A single line flying-spot scanner with electrostatic deflection was employed and 2:1 compressions were reported for edges produced by masking the trace on the scanner tube. Standards were such that with the control loop open, the channel bandwidth required was 20 Kc/s.

A more detailed investigation by Beddoes[3] used a monoscope tube with a built-in Test Card C target. The uncompressed channel bandwidth was 300 Kc/s in this case but the standards were still scaled down, there being 100 lines per frame at a frame frequency of 10 c/s (field frequency 20 c/s). The possible reduction was increased from 1.7:1 to 3.75:1 by use of a separate channel to transmit position information but serious jitter still persisted. Spatial distortion also occurred due to the use of constant velocity frame scanning. Since, in this system, the line period is not constant but depends on the amount of detail in the line, a ratchet timebase is required for the frame scan. This moves the spot down the screen by a fixed distance at the end of each line.

Prasada[45] completed the research on the variable-velocity system using the full standards of 405 lines, 50 fields per sec (approx.), 2:1

interlaced so that the effect of realistic signal-to-noise ratios could be investigated and the difficulties of engineering at these speeds would be shown up. It was found that compression ratios of 1.8:1 could be obtained but beyond this jitter was so serious as to be completely intolerable. Prasada concluded that this was due to instability in the feedback loop caused by the random noise which inevitably accompanies the signal. The amount of gain in the feedback loop, and therefore the compression ratio, was limited by the noise. It was not considered possible to design an electromagnetic deflection system to have the necessary bandwidth of 20 Mc/s. This could be achieved with an electrostatic system, however.

It was therefore concluded that the closed-loop system was not suitable for use on conventional television standards. It would, however, be well suited to systems working at lower speeds and having higher signal-to-noise ratios such as facsimile systems.

1.9 The Open-Loop System

The study of the closed-loop system suggested that the limitations were centred in the feedback loop so that, if this could be dispensed with, a useful system might result. Thus, the Open-loop system was proposed by Cherry, Holloway and Prasada in 1958[11].

This is a run-length coding system utilising point-to-point redundancies in the signal. It is best explained by referring to Figure 1.4, a very much simplified block diagram of the scheme.

As explained in Section 1.6, the maximum sampling rate necessary where run-length coding is employed is the Nyquist rate, 2W samples per second. The clock-pulse generator (c) is thus made to run at this rate. The signal from the video source (a) is fed to a unit (b) which continuously examines the signal and decides, according to built-in criteria, where the signal requires to be sampled and where samples can be omitted (during runs). The output signal from the "Detail Detector" controls a gate (d), allowing clock pulses to be transmitted to the Elastic Encoder (e) when the detail exceeds a certain level. In the Elastic Encoder the constant amplitude pulses (f) are used to control a switch which routes the pulses to a delay system in such a way that they emerge equally spaced, the

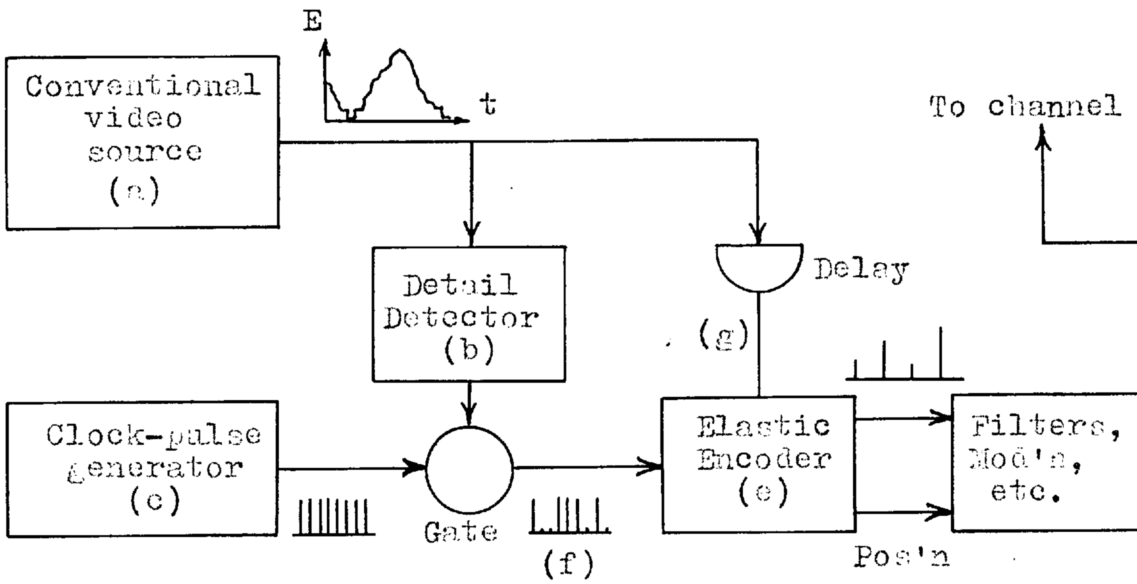


Figure 1.4. The Open-Loop System

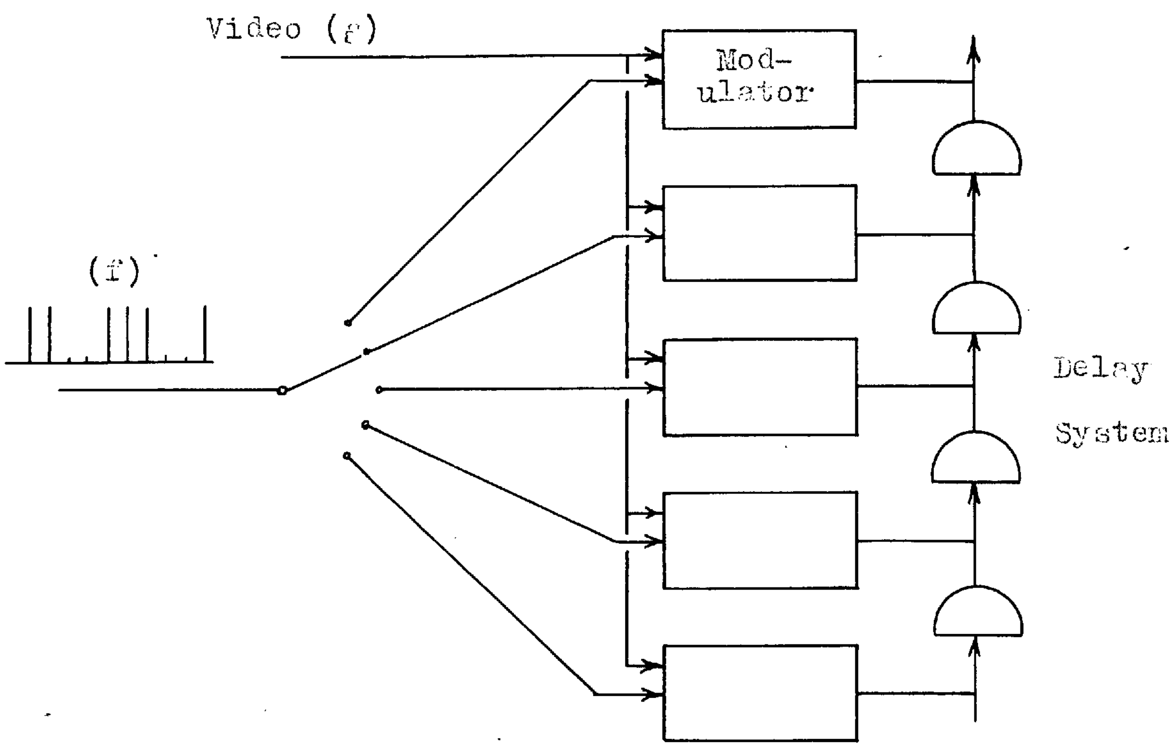


Figure 1.5. The Elastic Encoder

spacing now being larger than the Nyquist Interval. Between operating the switch and proceeding to the delay system, the pulses are modulated by the video signal (g). More details of the Elastic Encoder are shown in Figure 1.5.

As well as samples of compressed video the Elastic Encoder provides a position information signal (also in sample form) which indicates the amount of delay inserted in the path of each sample. This signal is necessary for decoding at the receiver. The two sample trains are now filtered to obtain their envelopes, any further coding or modulation depending on the nature of the channel to be used for their transmission.

At the receiver (Figure 1.6) it is proposed to use variable velocity line scanning to decode the signal; this is attractive on account of its simplicity and would make the system very suitable for applications requiring one transmitter but many receivers.

The line scanning speed is controlled by the "position" signal so that the original spatial distribution of brightness is recovered. The compressed video signal is fed directly to the electron gun so as to control the beam current in the usual way. Because variations of scanning speed would cause unwanted brightness variations on the screen, a correction signal derived from the timebase is applied to the tube.

1.10 A Summary of the Foregoing and a Preview of Later Chapters

In addition to introducing the subject, the foregoing sections have surveyed previous work relating to bandwidth compression of television and similar signals. The remainder of the thesis reports research which forms a natural continuation of previous work and mainly examines two aspects of the Open Loop System.

In order to investigate run-end detection whilst avoiding the need for complex elastic encoding and decoding equipment, Kubba[33] used a sample-and-hold technique to reconstruct a picture from a non-uniform train. A "zero-order" or "jump function" model for the signal was assumed. Because of the limitations of this model, which are discussed further in Chapter 2, examination of a more sophisticated model was thought to be desirable.

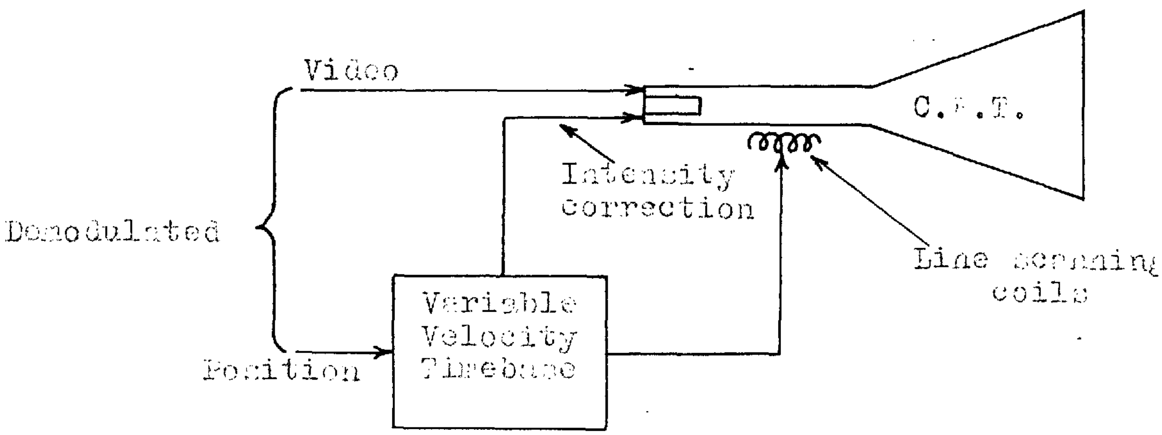


Figure 1.6. The Variable Velocity Receiver

A logical advance from the zero-order model leads one to the first-order model which considers runs as regions of constant signal slope. More elaborate still is the exponential model. However, it is shown in Chapter 2, that the detection of the ends of exponential runs requires knowledge of first-order run-end detection techniques. Thus a simple first-order run-end detector and tests upon it are described.

Chapter 4 deals with the reconstruction of a signal from the data selected by the detector, a first-order interpolator having been constructed and tested.

The restriction of the lengths of runs in the train of reduced data to a small number of standard values is an important feature of the Open Loop System. It is also essential to the reconstruction process described in Chapter 4. Thus it was found convenient to include this topic in the programme of research. Chapter 3 deals with the subject, reviewing the possible methods and describing equipment built. Measurements of the effects of using various sets of standard run lengths complete the chapter.

2. DATA REDUCTION - MODELS AND RUN-END DETECTION

2.1 Data Reduction

The operations involved in a number of bandwidth compression schemes (including the Open-Loop system) can be split up as shown in Figure 2.1.

Data reduction can be defined as the removal of unnecessary data from a signal in such a way that the signal can be reconstructed within satisfactory perceptual limits from the reduced data by the application of certain predetermined rules. (This process must introduce some errors. The amount of data reduction possible is limited by the need to keep the error within acceptable limits.) As well as being part of the run-length coding process, data reduction also occurs in schemes such as the co-ordinate specification schemes of Tsukkerman[58], and in schemes involving coarse quantisation of certain parts of the signal[31, 48, 49].

Bandwidth compression is by no means the only application of data reduction. Once the quantity of information in a given message has been reduced, advantage of the fact can be taken in a number of ways, by the use of suitable coding methods. Thus, for example, the message can be sent in a shorter time using a channel of the same quality as that required for the original signal. A coding device capable of closing up the gaps in the reduced signal would be required for this application. Channel capacity is limited by signal-to-noise ratio as well as by bandwidth, the relationship between the three quantities having been derived by Shannon:[53]

$$C = W \log_2 \left(1 + \frac{S}{N}\right)$$

(2.1) (transmitted signal limited to average power S)

Thus bandwidth can be "traded" for signal-to-noise ratio and vice versa. By first reducing the amount of data to be sent, therefore, a channel of normal bandwidth but with poorer signal-to-noise ratio could be employed. The coding in this case would involve a reduction in the number of signal levels to be distinguished, a correspondingly larger number of samples being transmitted. Taking this process to its logical conclusion results in the familiar binary P.C.M.

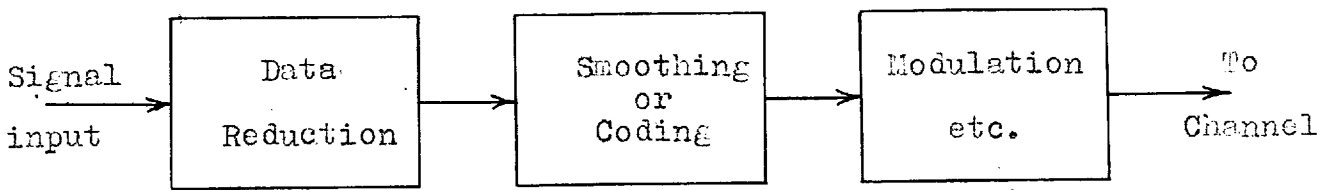


Figure 2.1. Operations involved in bandwidth compression

When applied to run-length coding, data reduction can be divided into two processes - detail detection and run-length restriction. Some aspects of the former are dealt with in this chapter whilst the latter forms the subject of Chapter 3.

2.2 Detail Detection

Since so little is known about the way in which the brain deals with signals it receives from the eyes it is difficult to define detail in a picture in a meaningful way. Further it may be possible to define detail in a manner which is obviously sensible but which does not indicate how data reduction should be achieved. The problem is further complicated by the fact that in television and facsimile, scenes are scanned so that the general form of the objects in the scene is lost. To regain the form when only the signal is available requires considerable storage; at television speeds this is complicated and expensive.

It is clear from examination of pictures that detail is a property of boundaries. Thus it may be that a change of brightness constitutes detail. Alternatively it may be a change in the brightness gradient which attracts the attention of the eye. Obviously the possibilities are endless, some being reasonable and some not so.

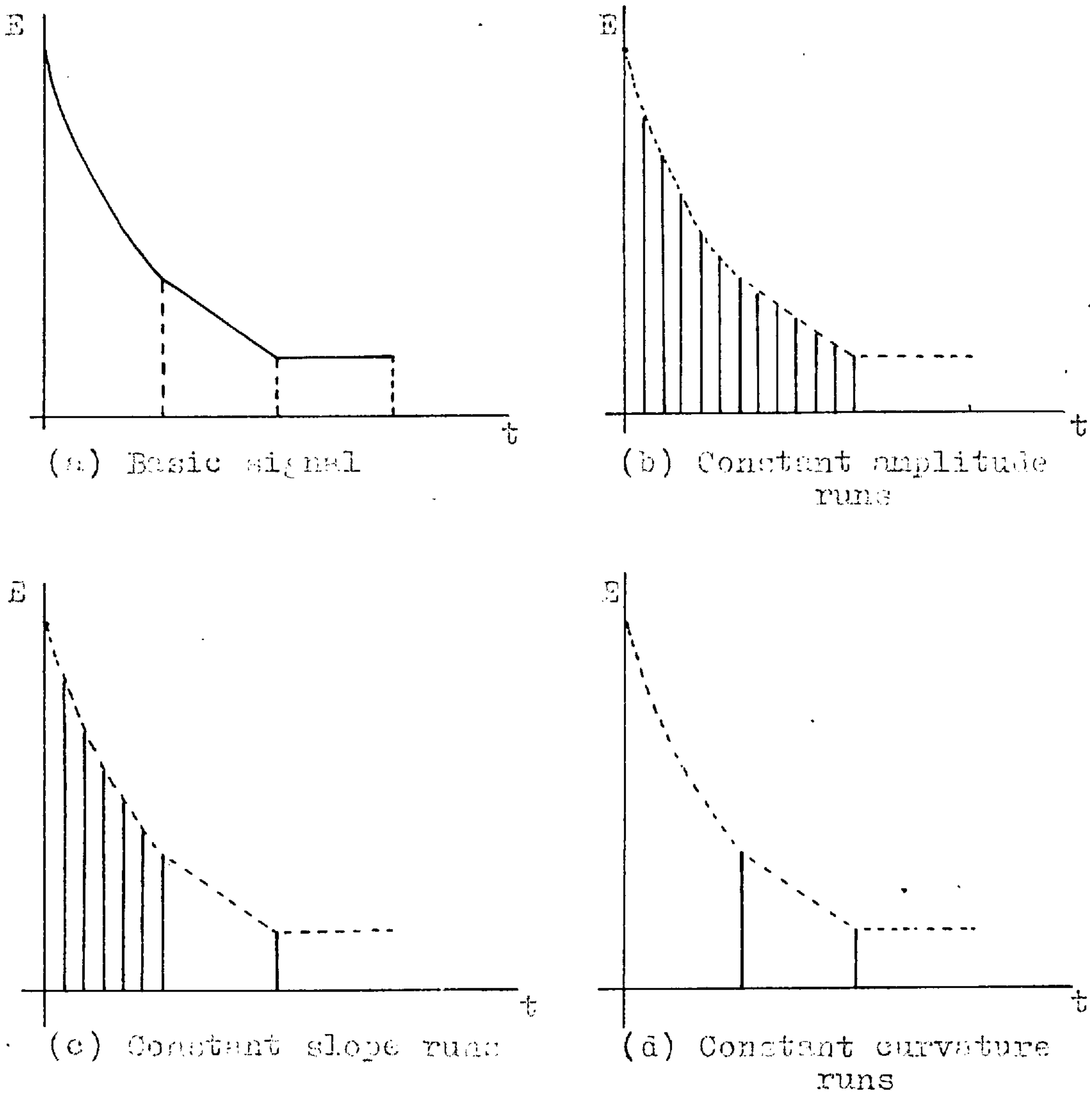
2.3 Run-end Detection

Unless a means of simultaneously transmitting the brightness levels of many picture elements is available, the representation of a scene by a small number (usually one) of uni-dimensional time-varying signals can be regarded as inevitable. Thus what is required is not detail detection but the detection of the ends of runs, a run being specified in such a way as to yield useful data reduction e.g. for a television signal one definition of a run might be "a region in which the signal amplitude remains constant (within certain limits)". It would clearly not be very sensible to use the definition "a region in which the signal frequency remains constant (within certain limits)" although this definition might work well with music - or speech-type signals. Clearly then, one of our problems is to define a run, the definition being such that we can see how to set about detecting the ends of runs.

In the past, the definition of a run has been given little attention. Since sudden changes of brightness are easily noticed by the eye, the definition of a run as a region of constant brightness has been almost universal. For many purposes this definition is quite adequate. Thus, an ideal facsimile signal from a line drawing or a typescript has two levels only and in theory it is necessary to send only the positions of the change-overs, the runs being alternately black and white. This is a case where the form of the signal is very closely controlled, however. In the case of television an immense variety of scenes (and lighting) is encountered so that the form of the signal cannot be as well known. Further, many signal levels are possible instead of the two in facsimile and it is not possible to predict the signal level after a transition from that before the transition. Instead, certain properties of television signals can be used to indicate the way in which a run should be defined. For example, regions of constant amplitude are frequently encountered. This again suggests that the constant amplitude definition for a run should be reasonable. Another type of region which is fairly common is that in which the signal amplitude is changing at a uniform rate - a region of constant "slope". Thus the definition of a run could state that within the run the slope of the signal should remain constant within certain limits. Clearly this definition would include the constant amplitude run since this is merely a special case of a constant slope run in which the slope is zero.

More complex types of region could obviously be considered but with increasing complexity the full extent of the definition would be required less often. For example, although the constant slope definition of a run may be adopted, many of the runs will have constant amplitude. A constant curvature definition is the next in the sequence and would be used to its full extent even less frequently.

The advantage of using higher order definitions of runs lies in the fact that fewer samples can be used to code such runs on the occasions when they do occur. This is illustrated in Figure 2.2. Here a signal is shown having regions of constant curvature (i.e. parabolic), constant (non-zero) slope and constant amplitude. The decreasing number of samples required with increasing run complexity is illustrated in sections (b), (c) and (d) of the figure respectively.



(a) Basic signal

(b) Constant amplitude runs

(c) Constant slope runs

(d) Constant curvature runs

Figure 2.2. Samples required with runs of various orders

The foregoing suggests that a more general definition of a run is required. A flexible, but useful, definition which I have adopted states that a run is a region of the signal in which some stated parameter (or parameters) of the signal remains constant within certain limits. Althoug this definition is flexible, no ambiguity can arise since the parameter which remains constant is required to be stated. Thus runs are known as constant amplitude runs, constant slope runs, etc.

2.4 Zero-and First-Order Models for Television Signals

As stated earlier, the wide variety of scenes (and lighting) from which television signals are produced makes it quite impossible to describe in precise algebraic expressions the form of the signal. The most that can be said is that the signal voltage lies between upper and lower limits (usually 1v. and 0v. for peak white and black respectively) and that frequency components between 25 c/s and 3 Mc/s are present. Any mathematical analysis of a television waveform is thus rendered very difficult.

One way to overcome this difficulty is to propose an easily described model for the signal. To be of any use, the model should be reasonably similar to the signal and it should be possible to approximate to the signal so closely that the errors are not significant when the signal is presented on a monitor*.

It will be realised that the choice of a model is intimately tied up with the type of run it is decided to use in a particular scheme.

A model which has been used in previous work[33, 44] is known as the jump function or zero-order model. Such a function is illustrated in Figure 2.3. It consists of constant amplitude sections with instantaneous transitions, the transitions occurring according to the Poisson distribution and all amplitude levels between certain limits being equally probable. It is clear that this model bears a marked similarity to a signal which has been broken into runs of constant amplitude. It is interesting to note

---

*This is clearly a matter for subjective assessment. Also, the use of a model may, in some cases, introduce considerable errors but these may not be objectionable.

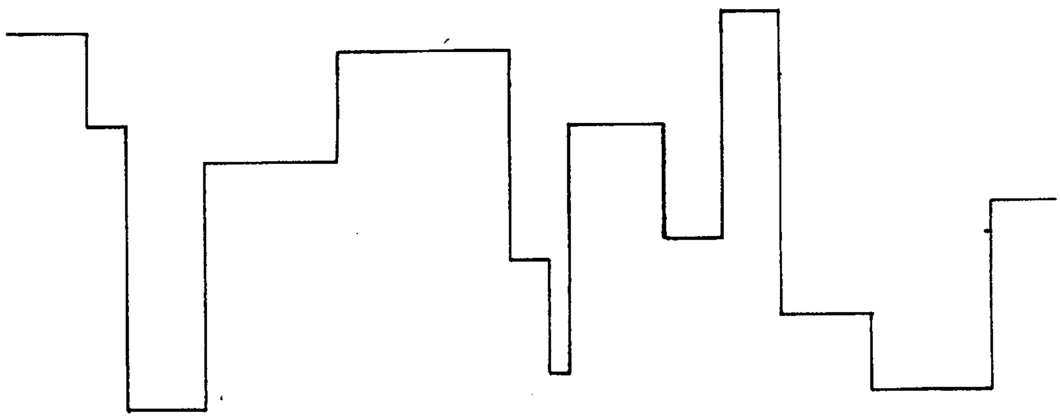


Figure 2.3. A jump function

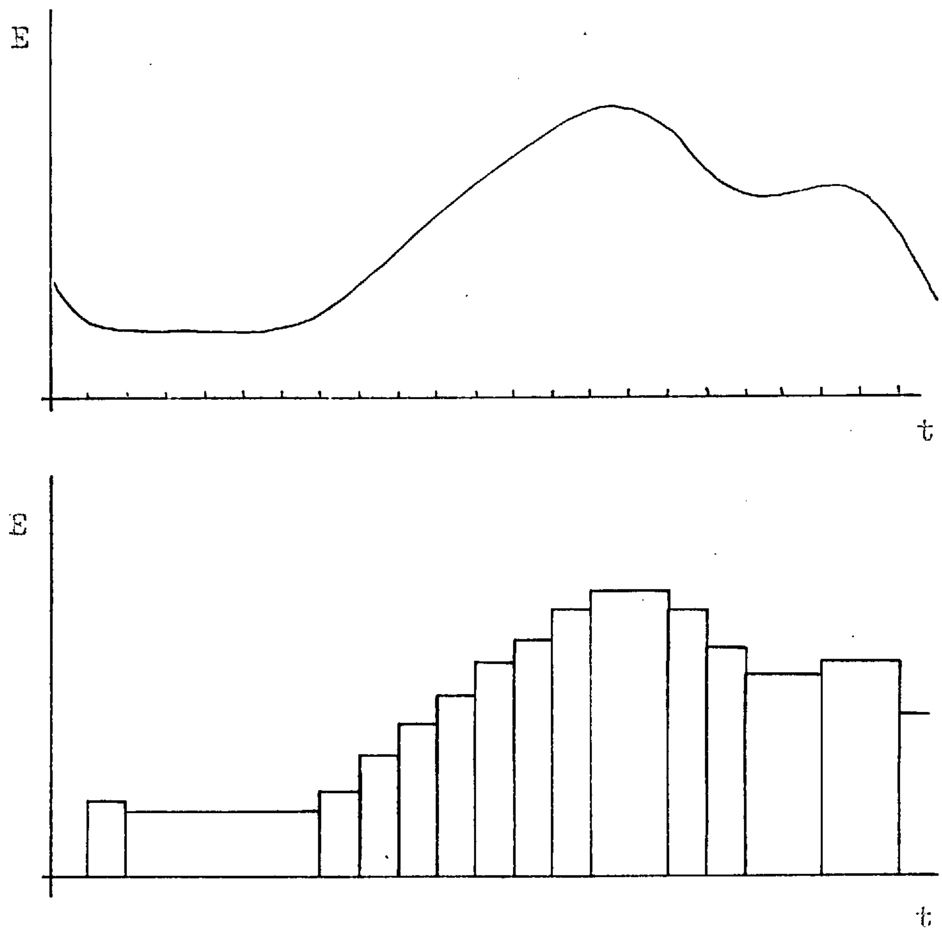


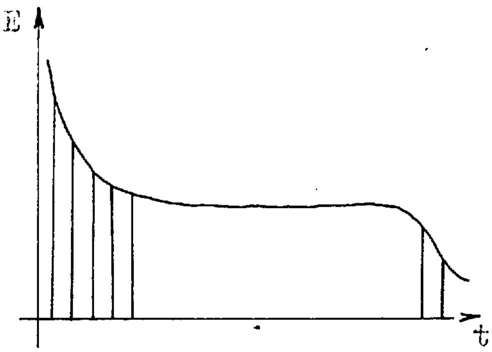
Figure 2.4. Sampling and reconstruction with constant amplitude runs

that the autocorrelation function of a jump function is exponential in shape; Kretzmer's work[32] reveals that the autocorrelation function of television signals is quite closely exponential for a number of scenes.

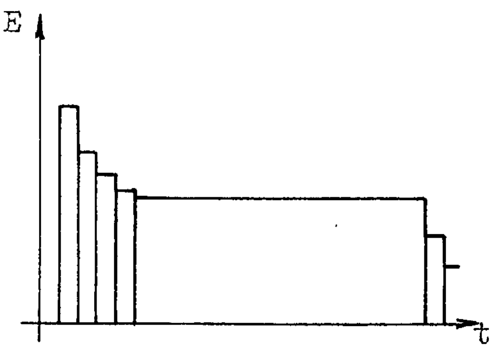
An illustration of the way in which a signal is sampled according to the constant-amplitude run criterion is provided by Figure 2.4. It will be seen from this figure that a signal reconstructed from the samples closely resembles a jump function; the transitions are not exactly Poisson distributed, however, but must occur at the clock pulse instants. In addition the figure shows the waste incurred by the use of constant amplitude runs where constant slope sections of the signal are to be sampled. Another, and probably more important, source of inefficiency in this data reduction method arises from the fact that the amplitude of the reconstructed signal is determined solely by the sample at the commencement of the run; the sample at the end of the run has no influence, except on the succeeding run. Unfortunately it is not possible to assess the magnitude of this waste since first-order reconstruction from the train of samples chosen for zero-order reconstruction would introduce further errors (Figure 2.5); it is essential to have a knowledge of the reconstruction method when the sampling points are being chosen.

It should be mentioned at this point that the brightness distribution corresponding to the signal shown in Figure 2.5 (b) is not presented to the viewer owing to bandwidth and focussing limitations of the display monitor. Thus the transitions are smoothed out and the actual brightness distribution corresponds quite closely to the original signal.

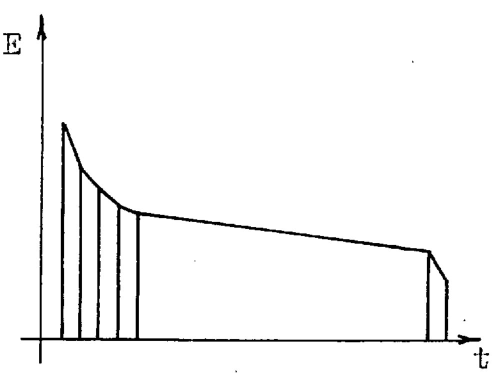
Discussion of the use of constant amplitude runs and the zero-order model thus suggests an attractive alternative, the use of constant slope runs. The corresponding model for the signal is usually referred to as the first-order model. It will be convenient to defer discussion of this model until other models have been mentioned. One point of interest which can be raised here, however, is that instantaneous transitions are never assumed to occur with the constant slope representation as they are when runs of constant amplitude are defined. The most rapid transition which can occur in the original signal could, if the sampling pulses were suitably phased, cause a given sample to have the full amplitude while its successor had zero amplitude. First-order reconstruction from these



(a) Signal and samples



(b) Zero-order reconstruction



(c) First-order reconstruction

Figure 2.5. Introduction of errors by the use of an unmatched reconstruction method

samples would produce a signal of constant but finite slope over that region.

2.5 The Exponential and Other Models for Television Signals

A curve can be represented to any desired degree of accuracy by the use of a polynomial, the number of terms required being determined by the maximum allowable error. This assumes that we have one curve (or perhaps a number of similar curves) and that we wish to fit an algebraic function to it to aid in analysis. In determining a model for a television signal, a somewhat different process is carried out. Here, the type of curve to be fitted is chosen before the actual signal is known, but with a knowledge of the waveforms which are likely to be met. The chosen curve is fitted to the signal over as long an interval as possible. When the divergence of the signal from the fitted curve becomes too great in some sense, the end of the run is said to occur and fitting of the next section of the signal is commenced.

In the zero-order model, only the first term of the polynomial (2.2) is utilised.

$$v(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + \dots\dots\dots \text{ etc.} \qquad (2.2)$$

(The a_r are constants within each run.) a_0 is then the parameter which specifies the run, signal amplitude being the only available parameter in the zero-order case.

If the first two terms of the polynomial are employed, the first order model results and each run has two parameters. As with all straight lines in two dimensions, any two quantities serve to define the line so that a choice of the quantities to be transmitted is available. Thus, depending on the type of channel to be employed, the means of interpolation, etc., it may be decided to send the samples themselves or the slopes of the successive runs.

It is clear that with more complex equipment for run-end detection and interpolation, more terms of the polynomial could be used. The runs thus obtained would, on average, fit the signal over longer intervals but more parameters would be required to specify each run. With more complex models there is a possibility of a further saving of data for it may be found that some parameters remain constant from one run to the next. Thus there is the possibility of coding the signal in such a way that only the parameters which have changed need be transmitted. Naturally this would be complicated and further problems of storage and synchronisation would inevitably arise.

Besides the straightforward polynomial there are many functions which can be used to represent a television signal in order to achieve data reduction. One such function which I have investigated in some detail is the exponential function:

$$V(t) = h + De^{bt} \tag{2.3}$$

where h, D and b are the constants (either positive or negative) which specify the run. A rough idea of the usefulness of this model may be obtained by observing that three parameters are needed to specify each run. Thus for the model to be of value it should be more general than the zero-order and the first-order models which require one and two parameters respectively. This is indeed the case. A zero-order run is seen to be a special case of an exponential run, being obtained when D = 0 so that $v(t) = h$. Similarly, suitable values of D and b can be found to give an exponential curve which is almost linear over the first few Nyquist intervals. (Figure 2.6.) Thus the first order representation is included.

Other models for television signals can readily be conceived. Why then was the exponential selected for further consideration? Before this question can be answered it is necessary to mention some alternatives to the exponential. A hyperbolic model is one such possibility. To be completely general a hyperbola should be specified by four parameters, e.g.

$$\frac{(v - v_1)^2}{v_0^2} - \frac{(t - t_1)^2}{a^2} = 1 \tag{2.4}$$

(v_1, v_0, t_1 and a are constant within a run). Such a curve is sketched in Figure 2.7.

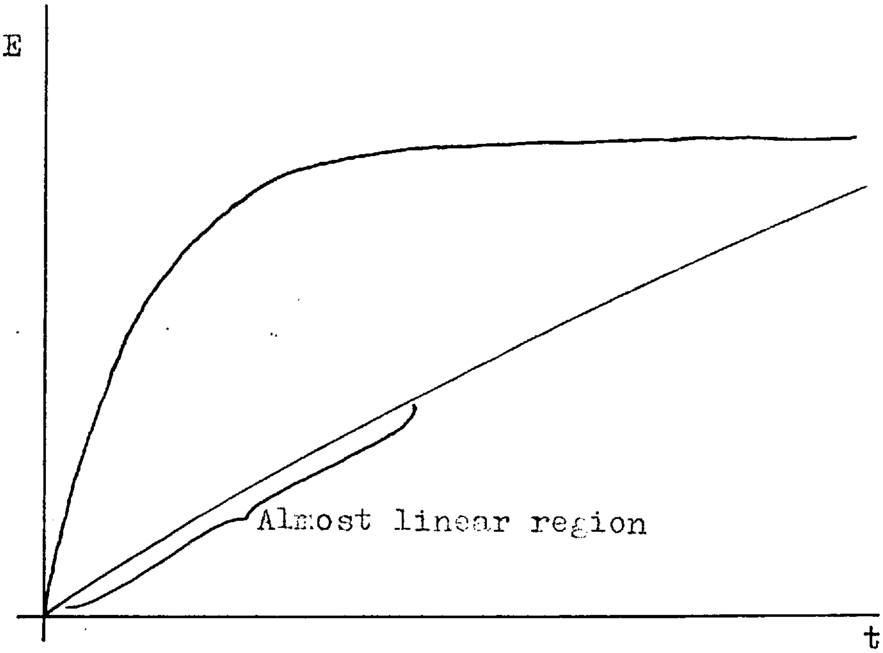


Figure 2.6. Showing that the exponential model can include the first-order model

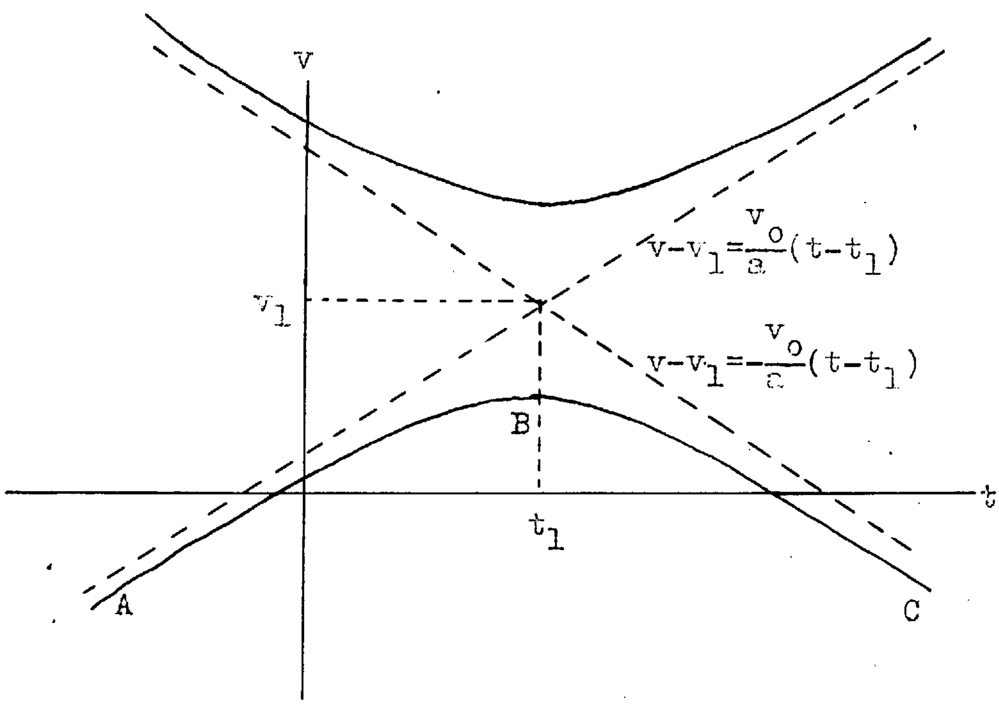


Figure 2.7. The general hyperbola

The hyperbolic model is clearly more complex than the exponential, requiring four parameters in place of the latter's three. If, for simplicity, a three-parameter hyperbolic model were assumed, there would clearly be more waveforms which could not be approximated by the model and which would thus require to be sampled at every Nyquist interval. Thus, for example, if v_1 were discarded, the portion ABC of the hyperbola (Figure 2.7.) would be entirely below the x-axis and portions of the signal with negative curvature would not be able to be represented by the model. A model which can deal with regions of both positive and negative curvature but still require only three parameters is clearly preferable.

The four-parameter hyperbola could alternatively be reduced to three parameters by discarding t_1 thus producing the equation:

$$\frac{(v - v_1)^2}{v_0^2} - \frac{t^2}{a^2} = 1 \qquad (2.5)$$

(Figure 2.8)

This would have the result of making the model usable only with runs commencing with a region of zero slope. The possibility of making either v_0 or a a fixed constant cannot be entertained for in either case this would mean that the curvature of the hyperbola was fixed, thus making the model of very little use indeed.

The rectangular hyperbola, in which the asymptotes are also the axes (Figure 2.9(a)) is, naturally, similar to the conventional hyperbola. For use as a model of a television signal it can be made more general by moving the asymptotes away from the axes, but without any rotation. (Figure 2.9 (b)). The equation then becomes:

$$(v(t) - a)(t - t_1) = c \qquad (2.6)$$

This is a three-parameter curve which is quite useful for it can be fitted to regions of positive or negative slope and positive or negative curvature. However, the fact that it is symmetrical about the 45° line (AB in Figure 2.9 (b)) may limit its application somewhat.

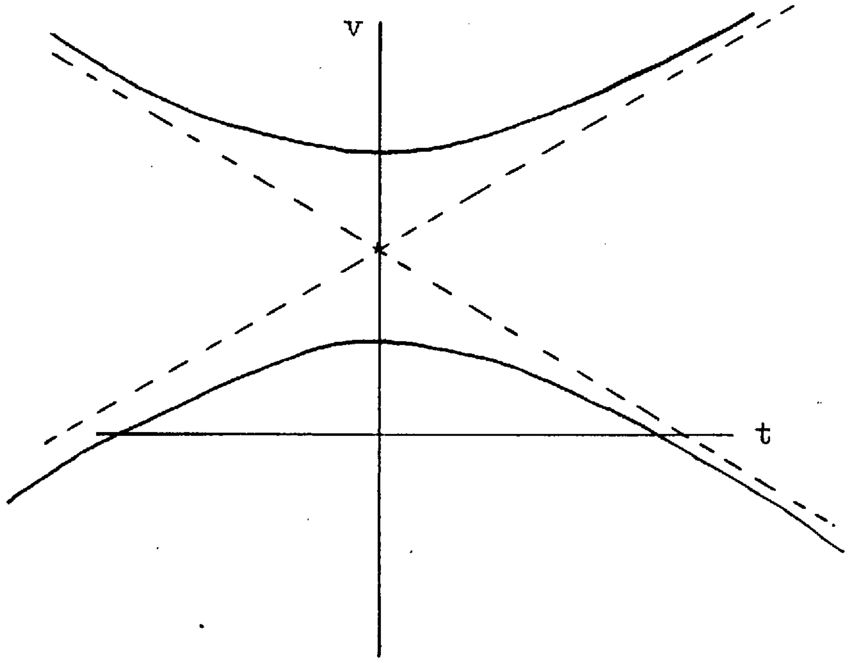


Figure 2.8. The three parameter hyperbola ($t_1 = 0$)

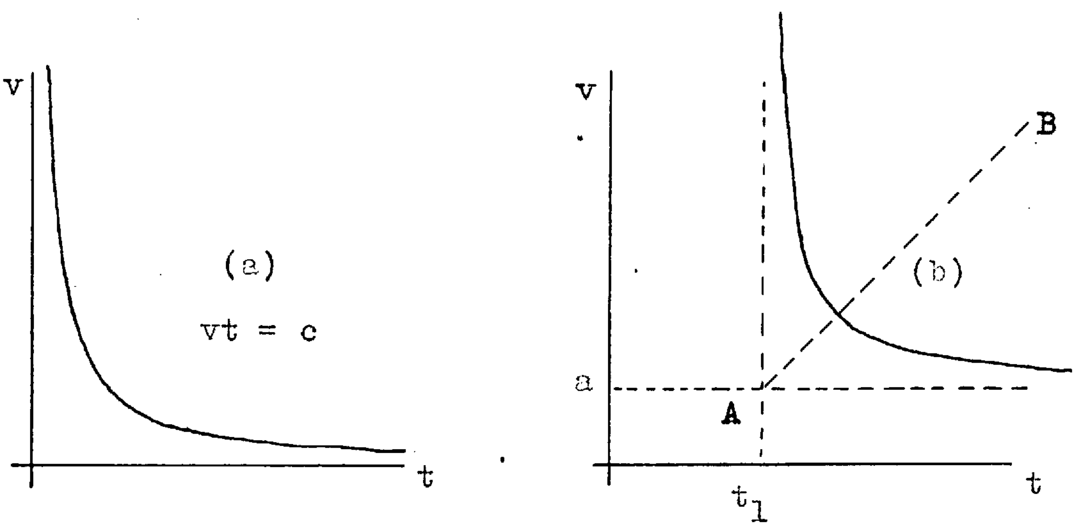


Figure 2.9. (a) The simple rectangular hyperbola
(b) A more general version

The tangent (Figure 2.10) is the last curve which will be considered in any detail; clearly the other possibilities are legion. Here three parameters are again adequate. Moreover the curvature changes monotonically along the whole length of the curve and any value of slope can be had by use of the correct part of the curve. This model thus compares favourably with the exponential in many respects.

In choosing a model, therefore, some functions can quickly be rejected on the grounds that they need more parameters than do other models to achieve comparable results. The chief advantage of the exponential over three-parameter curves of similar utility is, however, that the interpolator which is needed to reconstruct pictures from reduced data can be expected to be very simple in principle, requiring basically a variable rate RC charging circuit. Since the research programme was concerned with data reduction it was felt very desirable to concentrate on run-end detection rather than reconstruction. The possibility of using a simple interpolator was therefore welcomed. (It emerges later in this chapter that the problems of the first-order model should be solved before the exponential model is investigated; the design and construction of an exponential interpolator have not therefore been tackled. An outline block diagram of such an instrument is shown in Figure 2.11, however.)

When simpler models have been under consideration[44], statistics such as the autocorrelation function of the model and of the signal have been invoked as evidence that the model is a reasonable choice. When a more complex model contains one which has previously been tested in this way the statistics must be even more similar for it is possible for the signal and the curve fitted to it to be more nearly identical. In any case, the increased number of degrees of freedom means that very little can be said concerning the autocorrelation function of the more complex model by considering it in isolation.

In summary, therefore, the exponential is a flexible model which can be expected to yield the simplest instrumentation and whose statistics must match those of the signal.

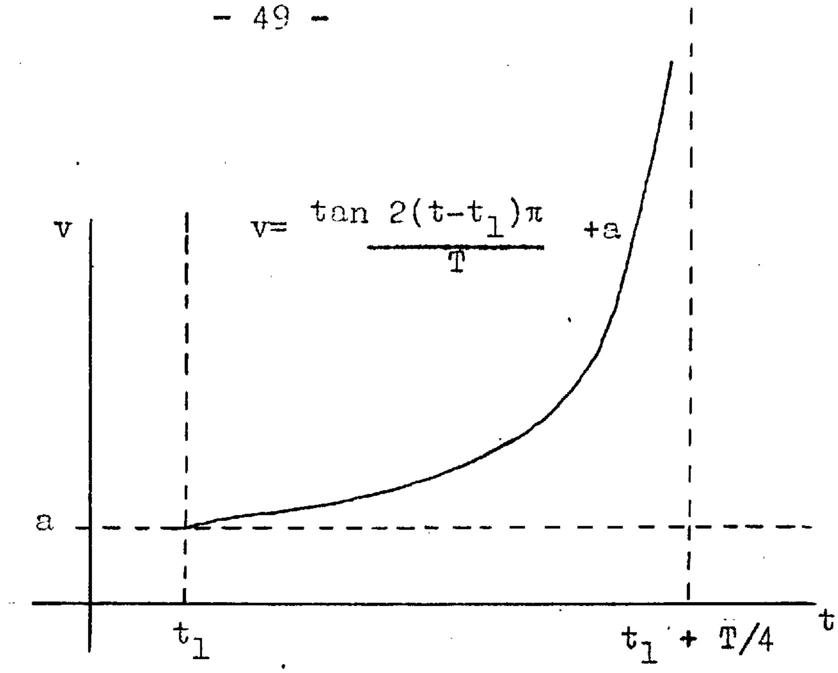


Figure 2.10. The tangent

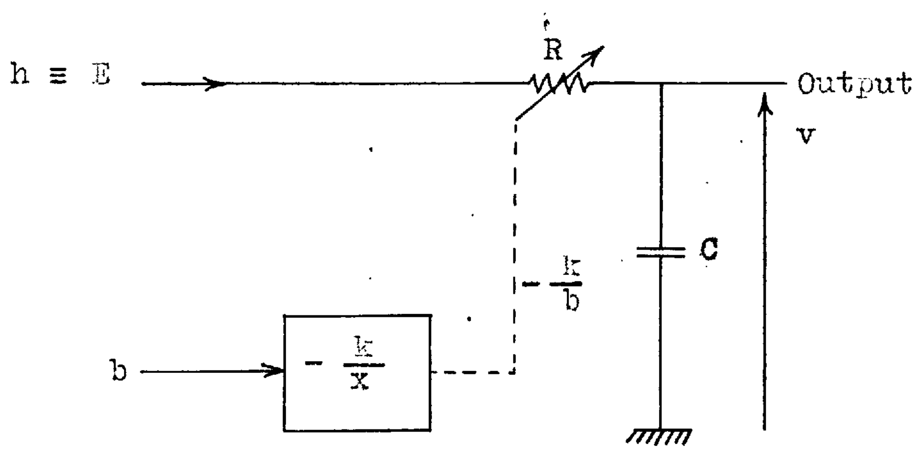


Figure 2.11. Outline block diagram of an exponential interpolator

2.6 Random Noise and Run-End Detection

The production of random noise is inherent in the process of generating television signals. The mechanism of noise production in the various types of pickup device has been reviewed by Pine[44]. Suffice it to say here that in a flying-spot scanner noise is generated partly in the photomultiplier (due to random emission velocities of electrons) and partly by the grain structure of the phosphor of the cathode ray tube (giving rise to a stationary spatial pattern). A flying-spot scanner would typically have a signal-to-noise ratio* of about 40 dB.

In the past, many workers have neglected the random noise and have simulated bandwidth compression schemes by processing the signal at a low speed. The narrow bandwidth which could be employed thus reduced the noise power considerably. Further, the subjective effect due to the noise amplitude at a given point being different in successive frames was absent since only one frame was processed, the result being presented as a single photograph.

In facsimile and other low-speed systems the problem is correspondingly simplified for here again narrow bandwidths are employed. In addition the signal often has two possible levels only so that a signal-to-noise ratio of about 20 dB suffices.

The problems raised by the presence of random noise have, however, been considered theoretically by Voelcker[59] and practically by Kubba[33]. Statistical inference techniques were called in and a run-end detector capable of combating** random noise was designed and tested. The assumption that noise was white, Gaussian and independent of average or instantaneous signal levels was made. As a result of his experimental programme Kubba was able to show that the noise is almost exactly Gaussian and measurements by Pine and Kubba have confirmed that the frequency spectrum of the noise is reasonably flat.

---

*Since a television signal has a d.c. component, it is conventional to express its signal-to-noise ratio as the ratio of peak signal power to r.m.s. noise power.

**The efficacy of this device is disputed by Sekey[51].

It is useful to have some idea of the magnitude of the random noise which accompanies a signal. With a peak signal-to-r.m.s. noise ratio of 40 dB, we have

$$20 \log_{10} \frac{E \text{ sig (pk)}}{E \text{ noise (r.m.s.)}} = 40 \qquad (2.7)$$

thus if E sig = 1v., E noise (r.m.s.) = $\frac{1}{100}$ v. (2.8)

This is the standard deviation of the noise (assuming) a Gaussian distribution) on a voltage scale. If, as suggested by Goodall[21], it is necessary to specify 128 or more brightness levels when quantisation is employed, it is clear that the noise voltage is of the same order as one level; that difficulty will be experienced in detail detection is at once apparent. The fact that about 128 levels are required may itself be due to the presence of the random noise rather than any limitation of the eye; if a system with a lower noise level could be produced more levels might be found necessary.

It is clear that a low-noise source of television signals has yet to be devised. In this direction, some improvement of the present sources could probably be achieved by cooling the photomultiplier of a flying-spot scanner. Thermoelectric devices have recently become available for this purpose. Correction for noise due to the C.R.T. phosphor by feeding back a signal to vary the beam current has also been proposed but promises to be difficult to engineer.

2.7 A Procedure for Run-End Detection

Although a useful definition of a run has been put forward, the all-important process by which the ends of runs may be detected remains to be described. Initially, a procedure which is applicable to any model will be outlined. In later sections the use of the procedure with specific representations of the signal will be dealt with.

It will be convenient to assume that the signal is limited to a frequency band extending from zero to W cycles per second so that samples at intervals of $\frac{1}{2W}$ secs. will specify it completely but at the same time will be independent. (This is a reasonable assumption since most picture signal sources incorporate some form of band-defining filter, one reason for this being the desirability of removing, as far as possible, the high frequency random noise which is inevitably generated with the signal.) The random noise within the band passed by the filter will be assumed to be purely additive (i.e. unrelated to signal amplitude), white and Gaussian.

In order to compare one signal sample with another, the run-end detector will need to incorporate storage. This can take many forms, the various means available having been reviewed by Pine[44]. In the case of a television signal in the familiar analogue form, an electromagnetic delay cable forms a convenient store. This cable is similar to the conventional co-axial cable but has the time delay for a given length increased by winding the inner conductor helically, sometimes around a core of magnetic material. The delay per metre, τ_D, is given by $\tau_D = \sqrt{LC}$ and the characteristic impedance Z_0 by $Z_0 = \sqrt{\frac{L}{C}}$ where L and C are the inductance per unit length and the capacitance per unit length respectively. An increase of inductance thus increases the delay and at the same time enables a higher characteristic impedance to be achieved. The current required from the driving amplifier to achieve a given voltage swing is thus reduced.

The storage of samples by a delay line can be understood by referring to Figure 2.12. If a sample X_1 enters a delay line of time delay T at t_0 it will emerge at t_0+T. If, at this instant, a sample X_2 is entering the line it is clear that the samples X_1 and X_2 must have been taken from the signal with an interval T. Thus several sections of delay line can make a number of samples available simultaneously as indicated in Figure 2.12.

The latest proposals for the Open Loop System[12, 40] suggest that all operations will be carried out on digital signals, an analogue-to-digital convertor comprising the first part of the scheme.

In describing the run-end detection method it will be assumed that previous tests have shown that n samples lie in one run. The n + 1th sample is therefore being tested to determine whether it lies in the same run or whether it is the first sample of a new run. First

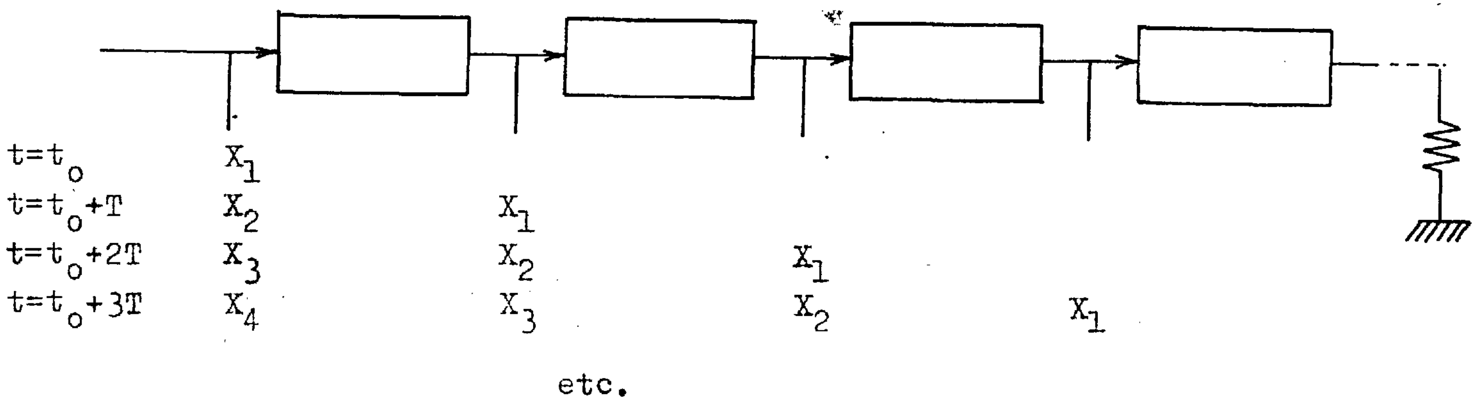


Figure 2.12. The storage of samples by a delay line

it is necessary to estimate the parameters of the run of n samples. This is achieved by fitting a curve to the samples, the type of curve fitted corresponding to the model chosen to represent the signal. Two alternatives are now available:-

(i) The value of the n + 1th sample can be predicted from the parameters of the run of n. A test is then performed to ascertain whether or not the actual value of the sample differs from the predicted value by more than a predetermined amount. If this "threshold value" is exceeded, the n + 1th sample forms the first sample of the next run, whilst if it is not exceeded, a curve is fitted to the n + 1 samples in order to test the n + 2th.

(ii) It may be possible to use the last few samples of the run together with the n + 1th sample to estimate the parameters of a curve for which these comprise the minimum number of samples necessary to specify it uniquely (e.g. the last sample of the run together with the n + 1th will be sufficient to determine the parameters of a straight line whilst two samples prior to the n + 1th would be needed to specify an exponential curve). The parameters thus obtained would then be compared with those of the run in order to ascertain whether any significant change had occurred.

It is plain that for some simple models, such as the first order, the two methods will be identical. When more complex models are employed, however, the second method can be expected to be advantageous for it allows the possibility of detecting which parameters have changed and sending those only. It is also likely that determination of parameters will involve less computation (and thus less instrumentation) than sample prediction since the former is an intermediate step towards the latter.

The procedure thus involves curve-fitting followed by the taking of a decision. The latter is accomplished by feeding a signal to a suitable trigger circuit such as the Schmitt trigger. This circuit exhibits a threshold effect; if the input is above a certain level, the output has one value, if below, another. The output signal can be used to operate a gate controlling the transmission of samples.

The threshold setting is determined by a number of factors and is not rendered necessary solely by the random noise as might at first be thought. Thus the limited amplitude resolution of the eye is one of the factors to be considered in choosing a threshold; there is no point in transmitting samples if the eye cannot detect that a brightness change has occurred. The brightness, or luminance, of a television screen is typically of the order of 30 foot-lamberts. At this level the contrast sensitivity of the human eye (ratio of the luminance to which the eye is adapted to the least perceptible luminance difference between the background luminance and the object luminance) is about 150. This figure suggests that the threshold need not be finer than that necessary to prevent brightness errors of about 1/150 of the full signal excursion.

Again, the availability of a more sophisticated run-end detector could possibly allow the threshold to be relaxed somewhat. To take a simple example, it is possible to envisage a first-order run-end detector which would permit a considerable amount of curvature in the signal without the threshold being exceeded. Thus such a section would appear as a linear section on the reproducing screen. Provided that such regions were similarly treated in successive lines, no degradation of the picture need result; the picture might have a different "texture" however. Further engineering difficulties are raised by such proposals of course.

In practice the run-end detector threshold is set by observation of the reconstructed picture and not by any theoretical considerations. It is axiomatic that improved pictures result from the insertion of more samples. A balance between picture quality and data-reduction has thus to be achieved.

2.8 Curve Fitting by the Method of Least Squares

It is quite obvious that an infinitude of methods can be devised for fitting a curve to a set of data; the value of many of these methods will be exceedingly small, however. Of the methods which are of value, the method of least squares is by far the most common. This method will be used in the run-end detection procedure outlined in the previous section. Before proceeding to details of curve fitting techniques, the reasons for adopting a least squares fit will be mentioned. It will be convenient to assume that the first order representation of the signal is being dealt with; the principles apply to any model, however.

When n samples have been accepted as lying in one run a curve has to be fitted to them but the type of curve cannot be chosen by visual examination of the data as is done when the results of some experiment are plotted on a graph; it has been chosen beforehand. This is not quite equivalent, therefore, to the fitting of a curve to data which is known to obey a linear relation but which has been perturbed by random disturbance (Figure 2.13). Although the form of the curve is known, the parameters remain to be established. In the latter case the errors are often normally distributed and it can easily be shown that a maximum-likelihood fit then reduces to a least squares fit (Mood,[39]). For this case, and for many others, the maximum likelihood estimator provides the ideal solution as it is unbiased and has minimum variance.

With samples of a television signal, conditions such as those illustrated in Figure 2.14 may prevail. A curve is fitted to an increasing number of points until the deviation from the curve is so great that a threshold is exceeded. If the true values of the samples lie on a curve of the type being fitted, the errors will be due solely to noise and the conditions described earlier (in connection with Figure 2.13) will apply. We cannot know this, however, and the most that we can say is that a curvature in one direction is just as likely as a curvature in the other. In a limited sense, therefore, the deviations are random and their average value is zero. The situation is very similar to that occurring when a curve is fitted to perturbed data whose true values do not lie on the type of curve fitted. For example, a quadratic may be fitted to data lying on an exponential curve; the fit obtained is not ideal but may be adequate over the range of interest. There are two reasons for choosing a least-squares fit in this case:

(i) The mathematics of the least-squares method is considerably simpler than that of any other reasonable method. This is of some

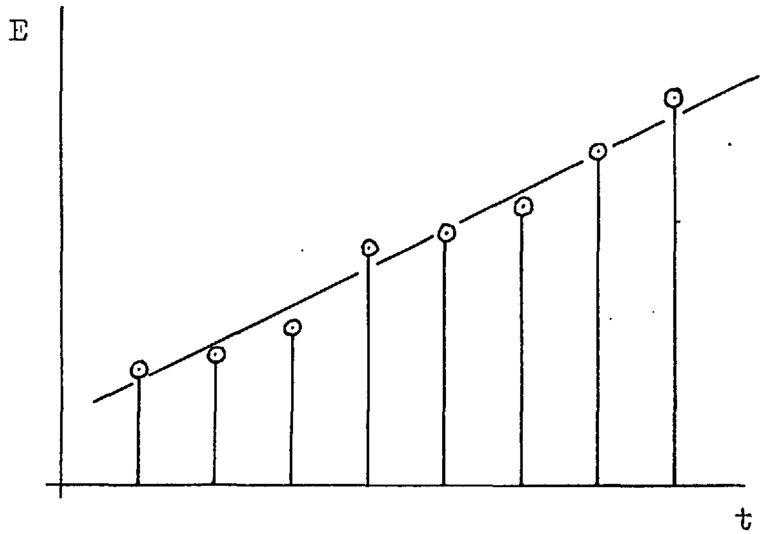


Figure 2.13. Fitting a straight line to noisy data (initially linear)

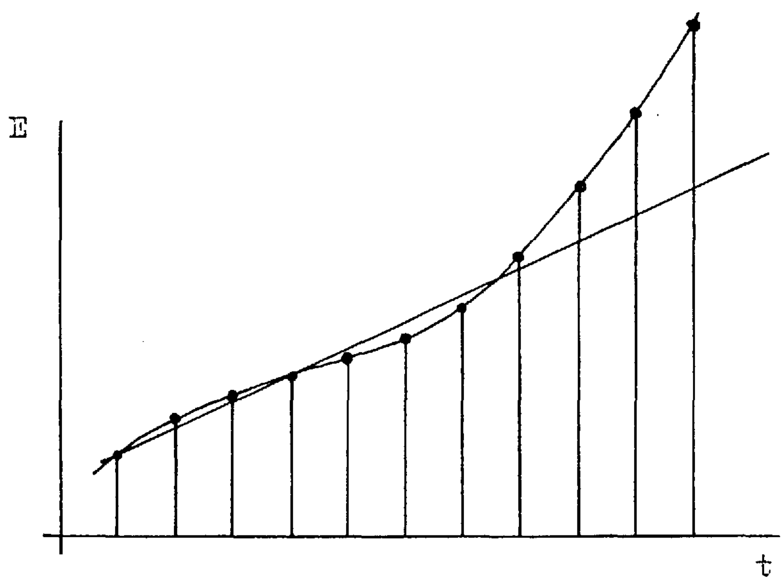


Figure 2.14. Fitting a straight line to data which is not initially linear

importance when the process has to be performed electronically at high speeds.

(ii) If it can be assumed that the errors are randomly distributed with zero mean, the least squares estimators for the parameters of the curve are best unbiased estimators, i.e. estimators which have smaller variance than any other estimators and whose expected values are equal to the true values of the parameters.

2.9 Direct Least Squares Fit of an Exponential Curve

The curve to be fitted to the data will be expressed by the equation

$$S(t) = h + De^{bt} \tag{2.9}$$

where h, D and b are the parameters defining the curve and have to be estimated.

Now, if t can be taken as an integral number of Nyquist intervals, then

$$t = iT \tag{2.10}$$

and

$$S_i = h + De^{ibT} \tag{2.11}$$

$$(i = 0, 1, 2, \ldots\ldots\ldots n)$$

The stored samples can now be denoted by X_i. If the curve (2.9) is fitted to the data, the error, ε at any point is given by

$$\varepsilon = X_i - (h + De^{ibT}) \tag{2.12}$$

and

$$\sum_i \varepsilon^2 = \sum_i (X_i - [h + De^{ibT}])^2 \tag{2.13}$$

This will be denoted by s.

To obtain the values of the parameters which give the minimum value of s we differentiate with respect to h, D and b in turn, the Normal Equations being obtained by setting each differential equal to zero.

$$\frac{\partial s}{\partial h} = \sum_i [-2X_i + 2h + 2De^{ibT}] \tag{2.14}$$

and if $$\frac{\partial s}{\partial h} = 0, \tag{2.15}$$

then $$\sum_i (h - X_i + De^{ibT}) = 0 \tag{2.16}$$

$$\frac{\partial s}{\partial D} = \sum_i [-2X_i e^{ibT} + 2he^{ibT} + 2De^{2ibT}] \tag{2.17}$$

and if $$\frac{\partial s}{\partial D} = 0 \tag{2.18}$$

then $$\sum_i e^{ibT}(h - X_i + De^{ibT}) = 0 \tag{2.19}$$

$$\frac{\partial s}{\partial b} = \sum_i [-2X_i DiTe^{ibT} + 2hDiTe^{ibT} + 2D^2 iTe^{2ibT}] \tag{2.20}$$

and if $$\frac{\partial s}{\partial b} = 0 \tag{2.21}$$

then $$\sum_i iTe^{ibT}(h - X_i + De^{ibT}) = 0 \tag{2.22}$$

Using the three Normal Equations, two of the three parameters can be eliminated in order to find the third in terms of the X_i. In this case, however, the final equation obtained by this process is

$$\left[\frac{\Sigma X_i \Sigma e^{ibT} + \Sigma e^{ibT} \Sigma X_i e^{ibT} - \Sigma X_i \Sigma e^{2ibT} - (n+1)\Sigma X_i e^{ibT}}{\Sigma e^{ibT} \Sigma e^{ibT} - (n+1)\Sigma e^{2ibT}}\right] \times$$

$$\Sigma ie^{ibT} = \Sigma X_i ie^{ibT} \tag{2.23}$$

and it does not appear to be possible to obtain b explicitly. In fact the complexity of equation 2.23 suggests that the aid of a digital computer would be desirable; clearly this method is of little value for run-end detection at high speeds. An alternative was therefore sought.

2.10 Fitting of an Exponential Curve by the Use of Factorial Moments

In statistical calculations it is quite common to take moments of a set of data or of a probability distribution. The type of moment usually calculated should strictly be known as a power moment and is defined by

$$\left.\begin{aligned} M_n^{(0)} &= y_1 + y_2 + y_3 + \cdots\cdots\cdots \quad y_n \\ M_n^{(1)} &= x_1y_1 + x_2y_2 + x_3y_3 + \cdots \quad x_ny_n \\ M_n^{(2)} &= x_1^2y_1 + x_2^2y_2 + x_3^2y_3 + \cdots \quad x_n^2y_n \\ &\vdots \\ M_n^{(q)} &= x_1^qy_1 + x_2^qy_2 + x_3^qy_3 + \cdots \quad x_n^qy_n \end{aligned}\right\} \qquad (2.24)$$

$$\text{or } M_n^{(r)} = \sum_n x^r y \qquad (2.25)$$

where moments are taken about an axis at $x = 0$.

An alternative to the power moment is the factorial moment.

Its derivation will not be reproduced here but is given by Sasuly[46]. The main advantage of the factorial moment is that the moment of any order can be expressed as the sum of the moments of the next lower order so that the moment of jth order can easily be expressed explicitly in terms of the data. Thus, assuming the data, y_i, to be spaced at unit intervals along x, then

$$\left.\begin{aligned} F_i^{(0)} &= y_1 + y_2 + \cdots\cdots y_i \\ F_i^{(1)} &= F_1^{(0)} + F_2^{(0)} + \cdots\cdots F_i^{(0)} \\ F_i^{(2)} &= F_1^{(1)} + F_2^{(1)} + \cdots\cdots F_i^{(1)} \\ &\vdots \\ F_i^{(q)} &= F_1^{(q-1)} + F_2^{(q-1)} + \cdots\cdots F_i^{(q-1)} \end{aligned}\right\} \qquad (2.26)$$

Alternatively,

$$\left.\begin{aligned} F_m^{(0)} &= y_1 + y_2 + \cdots\cdots y_m = \sum_{i=1}^{m} y_i \\ F_m^{(1)} &= my_1 + (m-1)y_2 + \cdots\cdots (1)y_m = \sum_{i=1}^{m} \binom{m+1-i}{1} y_i \\ F_m^{(2)} &= \binom{m+1}{2} y_1 + \binom{m}{2} y_2 + \cdots\cdots \binom{2}{2} y_m = \sum_{i=1}^{m} \binom{m+2-i}{2} y_i \\ &\vdots \\ F_m^{(q)} &= \binom{m+q-1}{q} y_1 + \binom{m+q-2}{q} y_2 + \cdots\cdots \binom{q}{q} y_m = \sum_{i=1}^{m} \binom{m+q-i}{q} y_i \end{aligned}\right\} \qquad (2.27)$$

To fit an exponential curve to data by means of factorial moments we write

$$S = h + De^{bt} \qquad (2.28)$$

and

$$t_i = T(i - 1), \qquad i = 1, 2, 3, \ldots\ldots n \qquad (2.29)$$

$$\left.\begin{aligned} S_1 &= h + D \\ S_2 &= h + De^{bT} \\ S_3 &= h + De^{2bT} \\ &\vdots \\ S_n &= h + De^{(n-1)bT} \end{aligned}\right\} \qquad (2.30)$$

Summing various numbers of terms yields the zero-order factorial moments:

$$\left.\begin{aligned} F_1^{(0)} &= h + K_1 (e^{bT} - 1) \\ F_2^{(0)} &= 2h + K_1 (e^{2bT} - 1) \\ &\vdots \\ F_n^{(0)} &= nh + K_1 (e^{nbT} - 1) \end{aligned}\right\} \quad (2.31)$$

$$\text{where } K_1 = \frac{D}{(e^{bT}-1)} \quad (2.32)$$

The first order factorial moments are obtained by summing right and left-hand members of (2.31) so that the moments of successive orders may be summarised:

$$\left.\begin{aligned} F_n^{(0)} &= \binom{n}{1} h + K_1\left[1\ (e^{nbT} - 1\right] \\ F_n^{(1)} &= \binom{n+1}{2} h + K_1\left[\rho_1 (e^{nbT} - 1) - \binom{n}{1}\right. \\ F_n^{(2)} &= \binom{n+2}{3} h + K_1\left[\rho_1^2 (e^{nbT} - 1) - \rho_1\binom{n}{1} - \binom{n+1}{2}\right] \\ F_n^{(3)} &= \binom{n+3}{4} h + K_1\left[\rho_1^3 (e^{nbT} - 1) - \rho_1^2\binom{n}{1} - \rho_1\binom{n+1}{2} - \binom{n+2}{3}\right] \end{aligned}\right\} \quad (2.33)$$

etc.

$$\text{where } \rho_1 = \frac{e^{bT}}{(e^{bT} - 1)} \quad (2.34)$$

The parameters h, D and b may now be found by using the data, y_i, to evaluate the factorial moments on the left-hand side of (2.33). As there are three parameters to be found, three of the equations in (2.33) will be necessary. Thus, if the first three are used, and h and D are eliminated, we are left with an expression for e^b in terms of n and the factorial moments.

$$\frac{(n+1)F^{(0)} - 2F^{(1)}}{(n+2)F^{(1)} - 3F^{(2)}} = \frac{4}{(n+1)^2}\,\frac{(n+1)(e^{nbT}-1)-2[\rho_1(e^{nbT}-1)-n]}{(n+2)[\rho_1(e^{nbT}-1)-n]-3[\rho_1^2(e^{nbT}-1)-\rho_1 n-\binom{n+1}{2}]} \tag{2.35}$$

As with the calculation in Section 2.9, e^{bT} must now be found by approximation; an explicit solution for e^{bT} does not appear to be obtainable.

An alternative, however, is to eliminate the term $(e^{nbT}-1)$ by multiplying each row of (2.33)by ρ_1 and subtracting it from the following row:

$$\left.\begin{aligned}
F_n^{(1)} - \rho_1 F_n^{(0)} &= h\binom{n}{1}\left[\frac{n+1}{2} - \rho_1\right] - K_1\binom{n}{1} \\
F_n^{(2)} - \rho_1 F_n^{(1)} &= h\binom{n+1}{2}\left[\frac{n+2}{3} - \rho_1\right] - K_1\binom{n+1}{2} \\
F_n^{(3)} - \rho_1 F_n^{(2)} &= h\binom{n+2}{3}\left[\frac{n+3}{4} - \rho_1\right] - K_1\binom{n+2}{3}
\end{aligned}\right\} \tag{2.36}$$

etc.

Elimination of h between the first and second rows of (2.36) now gives

$$\rho_1 = \frac{6F_n^{(3)} - 3(n+2)F_n^{(2)} + \binom{n+2}{2}F_n^{(1)}}{6F_n^{(2)} - 3(n+2)F_n^{(1)} + \binom{n+2}{2}F_n^{(0)}} \tag{2.37}$$

while if h is eliminated between the second and third rows of (2.36), the value obtained for ρ_1 is:

$$\rho_1 = \frac{20F_n^{(4)} - 8(n+3)F_n^{(3)} + 2\binom{n+3}{2}F_n^{(2)}}{20F_n^{(3)} - 8(n+3)F_n^{(2)} + 2\binom{n+3}{2}F_n^{(1)}} \tag{2.38}$$

The agreement of the two values of ρ_1 thus obtained is an indication of the adequacy of the chosen function to represent the data. The value of ρ_1 having been calculated, K_1 and h, and thus D, quickly follow.

The method of factorial moments is not widely known and did not come to my notice until near the end of the research. Meanwhile another method of fitting an exponential curve to data had been investigated in some detail. The theory of this method will be explained in Section 2.11 and resulting work in subsequent sections.

2.11 Fitting an Exponential Curve by transforming the Data into a Linear Form

If the curve to be fitted is

$$S = h + De^{bt} \tag{2.39}$$

and T is the Nyquist interval, then we can write:

$$\left.\begin{aligned} S_0 &= h + D \\ S_1 &= h + De^{bT} \\ S_2 &= h + De^{2bT} \end{aligned}\right\} \tag{2.40}$$

etc.

The first differences are

$$\left.\begin{aligned} S_1 - S_0 &= D(e^{bT} - 1) \\ S_2 - S_1 &= De^{bT}(e^{bT} - 1) \\ &\cdots\cdots \\ &\cdots\cdots \\ S_{n+1} - S_n &= De^{nbT}(e^{bT} - 1) \end{aligned}\right\} \tag{2.41}$$

and if $S_{r+1} - S_r$ is denoted by Y_r, then

$$Y_r = De^{rbT}(e^{bT} - 1) \tag{242}$$

$$\log Y_r = \log D + \log(e^{bT} - 1) + rbT \tag{2.43}$$

It will be noticed that Y_r can have a positive or a negative value; the fact that the logarithm of Y_r is taken may, therefore, cause some concern. This point is, however, resolved in Section 2.13.

From (2.43) it can be seen that the graph of $\log Y_r$ against r is a straight line of slope bT, having an intercept at $\log D(e^{bT} - 1)$ on the $\log Y_r$ axis. This straight line will now be fitted to the data by the method of least squares. However, since logarithms have been taken, the weighting at each point will not be the same. The weighting to be applied to the squared error at r will therefore be denoted by ω_r and for brevity the straight line to be fitted will be written:

$$y = a + brT \tag{2.44}$$

then the error

$$\varepsilon = y - a - brT \tag{2.45}$$

The weighted squared error is given by

$$\omega_r \varepsilon^2 = (y - a - brT)^2 \omega_r \tag{2.46}$$

and the sum of the weighted squared errors is

$$\sum_r \omega_r^2 = \sum_r (y - a - brT)^2 \omega_r = s \tag{2.47}$$

Now $$\frac{\partial s}{\partial a} = 2a\sum_r \omega_r - 2\sum_r y\omega_r + 2bT \sum_r r\omega_r \tag{2.48}$$

and we put this equal to zero to obtain the minimum value of s.

$$\therefore \; a\sum_r \omega_r - \sum_r y\omega_r + bT\sum_r r\omega_r = 0 \tag{2.49}$$

Similarly

$$\frac{\partial s}{\partial b} = 2bT^2\sum_r r^2\omega_r - 2T\sum_r ry\omega_r + 2aT\sum_r r\omega_r \tag{2.50}$$

$$\therefore \; bT^2\sum_r r^2\omega_r - T\sum_r ry\omega_r + aT\sum_r r\omega_r = 0 \tag{2.51}$$

Equations (2.49) and (2.51) are the Normal Equations. ω_r must now be expressed in terms of y.

The weight, ω_f, of a function f is defined as being inversely proportional to the variance, σ^2, of f. Thus, if $y = f(x)$, then

$$\frac{\omega_y}{\omega_x} = \frac{\sigma_x^{\,2}}{\sigma_y^{\,2}} \tag{2.52}$$

If Δx is the error in x at some instant, then, for small Δx,

$$\Delta f = f'(x)\Delta x \qquad \text{(nearly)} \tag{2.53}$$

(and in some cases exactly).

$$\therefore \quad (\Delta f)^2 = \left(\frac{df}{dx}\Delta x\right)^2 \tag{2.54}$$

If we now average over all allowed values of Δf and Δx, we obtain

$$(\sigma_f)^2 = \left(\frac{df}{dx}\,\sigma_x\right)^2 \tag{2.56}$$

In the particular case we are considering, $f(x) = \log x$ so that (2.56) becomes

$$\sigma_f^{\,2} = \left(\frac{1}{x}\,\sigma_x\right)^2 = \frac{\sigma_x^{\,2}}{x^2} \tag{2.57}$$

and from (2.52)

$$\omega_f = \frac{\omega_x \sigma_x^{\,2}}{\sigma_f^{\,2}} = \omega_x x^2 \tag{2.58}$$

The quantity represented by x is here the difference of two sample values. Every sample taken directly from the signal has the same average noise power so all the first differences will also have equal noise powers. Thus ω_x can be taken as unity (since weightings are only relative) and

$$\omega_f = x^2 \tag{2.59}$$

In the Normal Equations (2.49), (2.51), therefore, the weight, ω_r, must be put equal to Y_r^2 since this is the quantity whose logarithm is used. Hence:

$$a\sum_r Y_r^2 - \sum_r yY_r^2 + bT\sum_r rY_r^2 = 0 \qquad (2.60)$$

and $$bT^2\sum_r r^2Y_r^2 - T\sum_r ryY_r^2 + aT\sum_r rY_r^2 = 0 \qquad (2.61)$$

Thus $$a = \frac{\sum_r yY_r^2 - bT\sum_r rY_r^2}{\sum Y_r^2} \qquad (2.62)$$

and by substituting this value into (2.61),

$$b = \frac{\sum_r Y_r^2 \sum_r ryY_r^2 - \sum_r rY_r^2 \sum_r yY_r^2}{T\left(\sum_r Y_r^2 \sum_r r^2Y_r^2 - \sum_r rY_r^2 \sum_r rY_r^2\right)} \qquad (2.63)$$

It is clear that the instrumentation of these formulae would require exceedingly complex equipment. The possibility of obtaining a considerable simplification by neglecting the weighting factors was, however, noticed and it was thought worth while to examine further the effect of using the simplified formulae. These are obtained by putting $\omega_r = 1$ in the Normal Equations (2.49) and (2.51).

Then

$$na - \sum_r y + bT\sum_r r = 0 \qquad (2.64)$$

and

$$bT^2\sum_r r^2 - T\sum_r ry + aTr = 0 \qquad (2.65)$$

(n is the number of points to which the straight line is to be fitted.)

so that

$$a = \frac{\sum_r y - bT \sum_r r}{n} \qquad (2.66)$$

and

$$b = \frac{n \sum_r ry - \sum_r r \sum_r y}{T\left(n \sum_r r^2 - \sum_r r \sum_r r\right)} \qquad (2.67)$$

For given numbers of points the denominators of these expressions are constants and the numerators are merely weighted sums of the y_r (weighting may be positive or negative). Thus instrumentation can be expected to be much simpler than would be the case for the full formulae in which products such as yY_r^2 occur. The use of the approximate formulae is actually equivalent to fitting the "best straight line" to the points obtained by plotting y against r (or time). The errors in fitting to be expected when using these formulae are not easy to derive for the general case. A number of calculations have been worked through for various special cases, however, the cases considered and the results being presented in the next section.

2.12 Calculation of typical errors caused by the use of the approximate formulae

In order to test the approximate formulae, artificial data were employed, a variety of exponential curves being considered. The "true" values of regularly spaced samples were obtained by the use of tables and the addition of random noise was simulated by adding to each sample a quantity obtained from a table of random normal deviates [18].

The unperturbed curves have been drawn in Figure 2.15, the appropriate parameters also being indicated. It will be seen that these curves encompass all the shapes of exponential curve which could be found in television signals, bearing in mind the limited rate of change which is possible. In all cases the signal amplitudes have been normalised so that the signal level is always between 0 and 1. This is entirely arbitrary but fits in with the practice of television studios where video signals are transmitted within the studio on

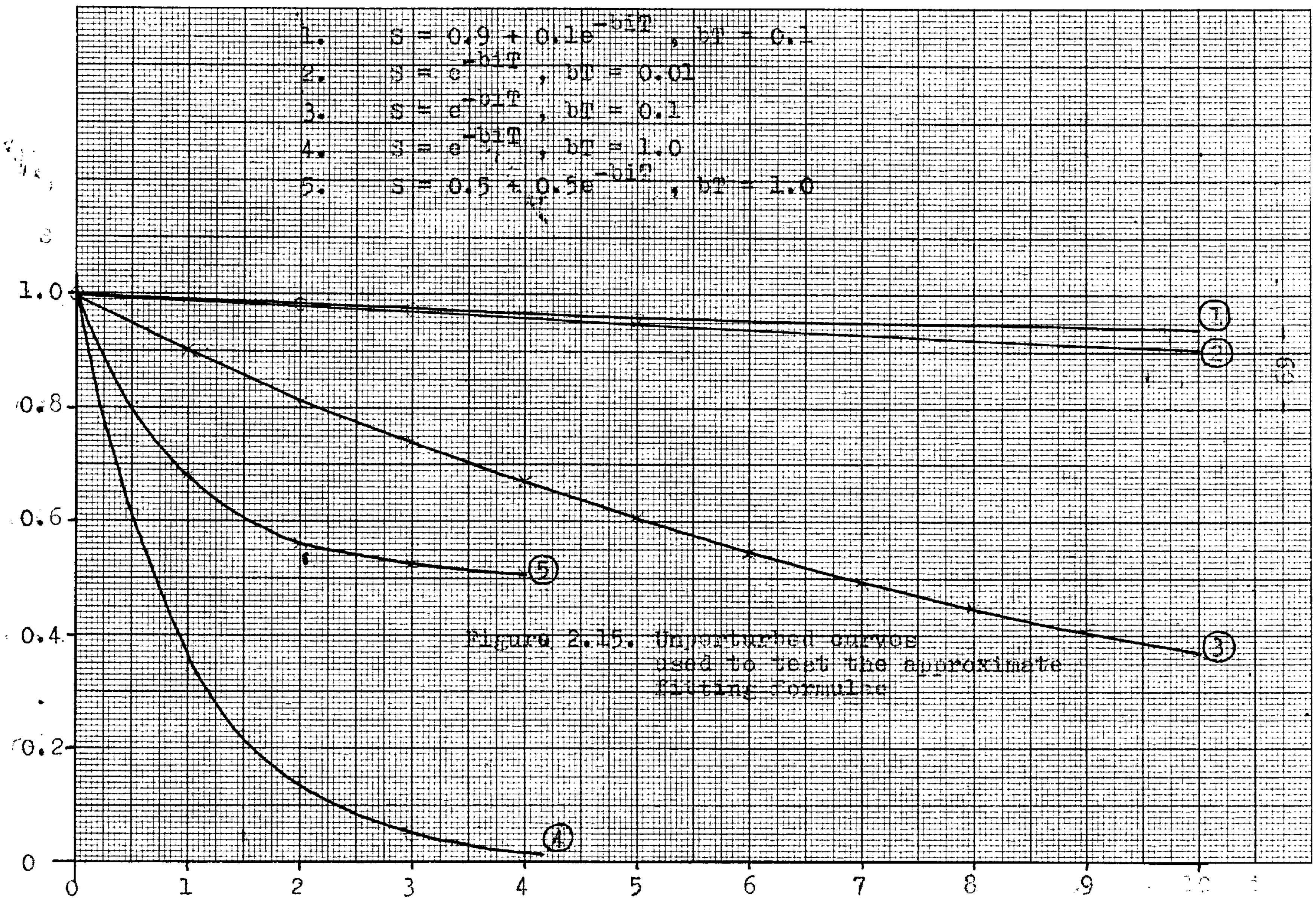


Figure 2.15. Unperturbed curves used to test the approximate fitting formulae

coaxial cables at 75 ohm characteristic impedance with an amplitude of one volt. For calculation purposes it is necessary to choose an amplitude level only in order to derive a noise level. Originally it had been hoped to carry out the calculations for a number of values of signal-to-noise ratio. However, as the calculations were rather more lengthy and time-consuming than had been anticipated, fittings were carried out for only one value of signal-to-noise ratio, 40 dB. The tables of random normal deviates gave figures with a mean of zero and a standard deviation of unity. Each figure was therefore multiplied by the standard deviation of the required noise before adding it to the "true" value of the sample. Thus, in order to simulate a signal-to-noise ratio of 40 dB, each figure was multiplied by 1/100 since

$$20 \log_{10} \frac{1 \text{v}}{1/100 \text{ v}} = 40 \text{ dB} \qquad (2.68)$$

Using various numbers of samples in the fitting process, the parameters of the curves were estimated from the 'noisy' data by use of the approximate formulae. The resulting estimates of bT and D, together with some intermediate results, are listed in Table I.

In some cases, where the slope of the curve was very small, it was found that negative quantities were produced by the first-differencing process. This was due to the noise causing the amplitude of a sample to become less than that of the following sample (also perturbed). With negative quantities the taking of logarithms would have been impossible and, as an examination of the curves showed that the curvature must be very small in such cases, it was decided that under such conditions it would be permissible to fit a straight line, the exponential fitting process being temporarily discarded. Curve No. 2 shows this process occurring; here the first differences oscillated so wildly (in comparison with their magnitude) that the fitting process was taken no further. From Figure 2.15 it is clear that Curve No. 1 would have been an even less worthwhile proposition; no parameter estimation was attempted in this case therefore. Estimates of bT and D were, however, obtained for curves 3, 4 and 5. Eleven points on curve 3 were employed, resulting in 9 estimated

TABLE I

Calculation of Typical Errors

Curve No.	True X	i	Noise	Perturbed X	$\Delta_1 X = Y_r$	$\log Y_r = y_r$	r	n	ry_r	Σy_r	Σry_r	Σr	Σr^2	$\hat{b}T$	$\hat{b}T\Sigma r$	a	e^a	$1-e^{bT}$ *	$\hat{D}$
1	1.0000	0	-.0140																
	0.9905	1	-.0036																
	0.9819	2	+.0130																
	0.9741	3	-.0138																
2	1.0000	0	-.0149	.9851	.0007														
	0.9900	1	-.0056	.9844	.0089														
	0.9802	2	-.0047	.9755	-.0087														
	0.9704	3	+.0138	.9842															
3	1.0000	0	-.0038	.9962	.0905	-2.4024	0	1	0.0000										
	0.9048	1	+.0009	.9057	.0906	-2.4013	1	2	-2.4013	-4.8037	-2.4013	1	1	+.0011	+.0011	-2.4024	.0903	.001	90.3
	0.8187	2	-.0036	.8151	.0873	-2.4384	2	3	-4.8768	-7.2421	-7.2781	3	5	-.018	-.054	-2.3760	.0929	.0178	5.219
	0.7408	3	-.0130	.7278	.0630	-2.7647	3	4	-8.2941	-10.0068	-15.5722	6	14	-.1124	-.6744	-2.3331	.0970	.1060	.9151
	0.6703	4	-.0055	.6648	.0628	-2.7678	4	5	-11.0712	-12.7746	-26.6434	10	30	-.1094	-1.094	-2.3361	.0967	.1033	.9361
	0.6065	5	-.0045	.6020	.0660	-2.7181	5	6	-13.5905	-15.4927	-40.2339	15	55	-.0858	-1.287	-2.3676	.0937	.0824	1.137
	0.5488	6	-.0128	.5360	.0254	-3.6730	6	7	-22.0380	-19.1657	-62.2719	21	91	-.1705	-3.5805	-2.2265	.1080	.1563	.6888
	0.4966	7	+.0140	.5106	.0570	-2.8647	7	8	-20.0529	-22.0304	-82.3248	28	140	-.1242	-3.4776	-2.3191	.0984	.1166	.8439
	0.4493	8	+.0043	.4536	.0552	-2.8968	8	9	-23.1744	-24.9272	-105.4992	36	204	-.0965	-3.4744	-2.3836	.0922	.0920	1.00217
	0.4066	9	-.0082	.3984	.0332	-3.4052	9	10	-30.6468	-28.3324	-136.1460	45	285	-.10485	-4.7183	-2.3614	.0943	.0997	.9558
	0.3679	10	-.0027	.3652															
4	1.0000	0	-.0043	.9957	.6432	-0.4413	0	1	0										
	0.3679	1	-.0154	.3525	.2251	-1.4913	1	2	-1.4913	-1.9326	-1.4913	1	1	-1.050	-1.050	-.4413	.6434	.6501	.9897
	0.1353	2	-.0079	.1274	.0804	-2.5208	2	3	-5.0416	-4.4534	-6.4329	3	5	-1.040	-3.120	-.4445	.6408	.6465	1.4416
	0.0498	3	-.0028	.0470	.0209	-3.8680	3	4	-11.6040	-8.3214	-18.1369	6	14	-1.131	-6.786	-.3839	.6811	.6773	1.0056
	0.0183	4	+.0078	.0261	.0236	-3.7465	4	5	-14.9860	-12.0679	-33.1229	10	30	-0.899	-8.990	-.6156	.5401	.5730	.9108
	0.0067	5	-.0042	.0025	.0083	-4.7915	5	6	-23.9575	-16.8594	-57.0804	15	55	-0.853	-12.795	-.6774	.5081	.5739	.8853
	0.0025	6	-.0083	-.0058	-.0099														
	0.0009	7	+.0032	.0041															
	0.0003	8	+.0178																
5	1.0000	0	+.0022	1.0022	.3323	-1.1017	0	1	0										
	0.6840	1	-.0141	.6699	.1277	-2.0581	1	2	-2.0581	-3.1598	-2.0581	1	1	-.9564	-.9564	-1.1017	.3320	.6156	.5393
	0.5677	2	+.0255	.5422	.0226	-3.7898	2	3	-7.5796	-6.9496	-9.6377	3	5	-.8656	-2.5968	-1.4509	.2343	.5794	.4044
	0.5249	3	-.0053	.5196	.0038	-5.5728	3	4	-16.7184	-12.5224	-26.3561	6	14	-1.5145	-9.0870	-.8589	.4236	.7802	.5429
	0.5091	4	+.0067	.5158	.0003	-8.1117	4	5	-32.4468	-20.6341	-58.8029	10	30	-1.7535	-17.5350	-.6218	.5369	.8269	.6493
	0.5034	5	+.0121	.5155															

Curves 1 and 2 are almost linear; fitting was not therefore carried out.

Curve 5: Curve very flat; results not reliable

* $1-e^{bT}$ was used, rather than $e^{bT}-1$; the reason for this is explained in Section 2.13

values of the parameters. Curves 4 and 5 flattened out sooner than curve 3 so that only five and four pairs of estimates were made respectively. A negative first-difference is shown in Table I in connection with curve 4; further fitting could thus not be attempted. The calculations for curve 5 were not taken quite as far but it is clear that values of $X_{r+1} - X_r$ for r = 3 or 4 were small compared with the expected noise amplitude so that the use of further points would have been of little value.

The results indicate that good estimates of bT and D can be expected when the number of points used is adequate (e.g. five or more) and the slope of the curve is not so small that the first-differences become comparable with the r.m.s. noise amplitude.

A few of the estimates of D and bT have been taken and used to plot the curves shown in Figure 2.16. In all cases the plotting has assumed that the "true" value for h should be retained, the estimated value of De^{biT} being added to this. For curves Nos. 4 and 5 this process has resulted in 'reconstructed' curves very similar to the originals despite the apparently large errors in $\hat{D}$ and $\hat{bT}$. As might reasonably be expected it can be misleading to consider percentage errors in $\hat{bT}$ since this is used as an exponent and a given percentage error in $\hat{bT}$ will not result in the same percentage error in e^{biT}.

The reconstructed versions of curve 3 indicate that the assumption of the "true" value for h is not satisfactory in all cases. However, the two reconstructed curves are very similar to the original curve, vertical displacements being sufficient to produce almost exact coincidence. (The two reconstructed curves were plotted using values of $\hat{D}$ and $\hat{bT}$ which were deliberately chosen for their large differences from the true values of these parameters.) Thus by first using the simplified formulae to estimate D and bT and then estimating a value of h it should be possible to obtain a good fit whatever the shape of curve being considered. These examples have thus shown the adequacy of the approximate formulae in which carying weightings are dispensed with.

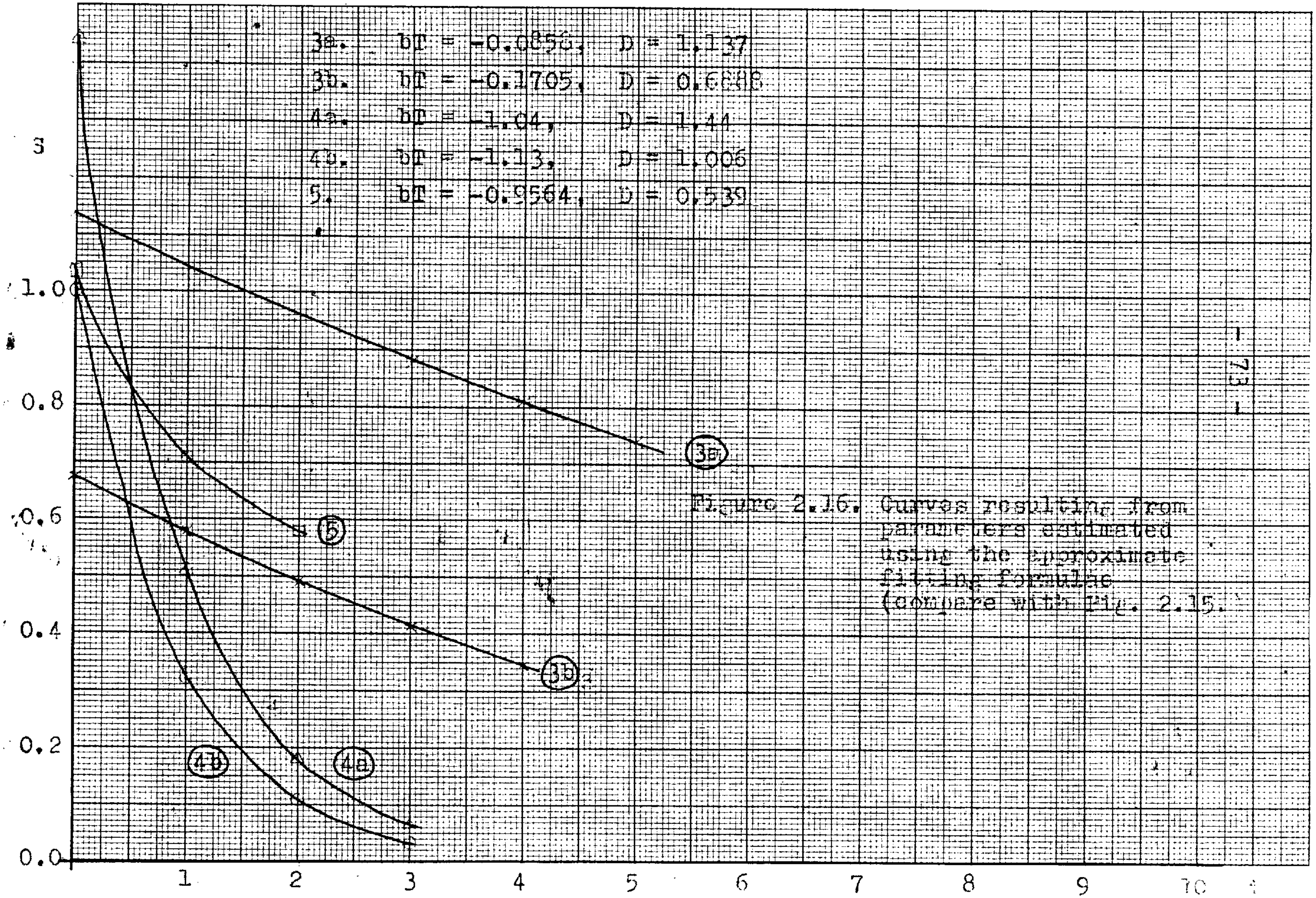


Figure 2.16. Curves resulting from parameters estimated using the approximate fitting formulae (compare with Fig. 2.15.

The estimation of h is a simple matter compared with that of estimating D and bT, for the simplest of least squares fits is sufficient. Reference to Figure 2.17 will show what is required. Suppose that the curve A-B is the curve which best fits the samples. The A-B has the formula

$$x = \hat{h} + \hat{D}e^{i\hat{b}T} \qquad (2.69)$$

The error at any point is given by

$$\varepsilon = X_i - \hat{h} - De^{i\hat{b}T} \qquad (2.70)$$

so that

$$s = \sum_i \varepsilon^2 = \sum_i (X_i - \hat{h} - \hat{D}e^{i\hat{b}T})^2 \qquad (2.71)$$

and

$$\frac{\partial s}{\partial h} - 2\Sigma X_i + 2\Sigma h + 2\Sigma \hat{D}e^{i\hat{b}T} \qquad (2.72)$$

For minimum s,

$$\frac{\partial s}{\partial h} = 0 \qquad (2.73)$$

so that

$$h = \frac{\Sigma X_i - D\Sigma e^{ibT}}{n} \qquad (2.74)$$

This means that the estimate of h is obtained merely by averaging the difference between each sample and the corresponding ordinate of the curve $\hat{D}e^{i\hat{b}T}$, a common sense result.

2.13 The Exponential Run-end Detector

The calculations of Section 2.12 having shown that the approximate formulae can yield acceptable results, more features of the exponential run-end detector can now be considered. It will clearly not be sufficient simply to estimate $\hat{bT}$ or $\hat{D}$ and say that a run has ended when a certain change occurs in either of these parameters, for

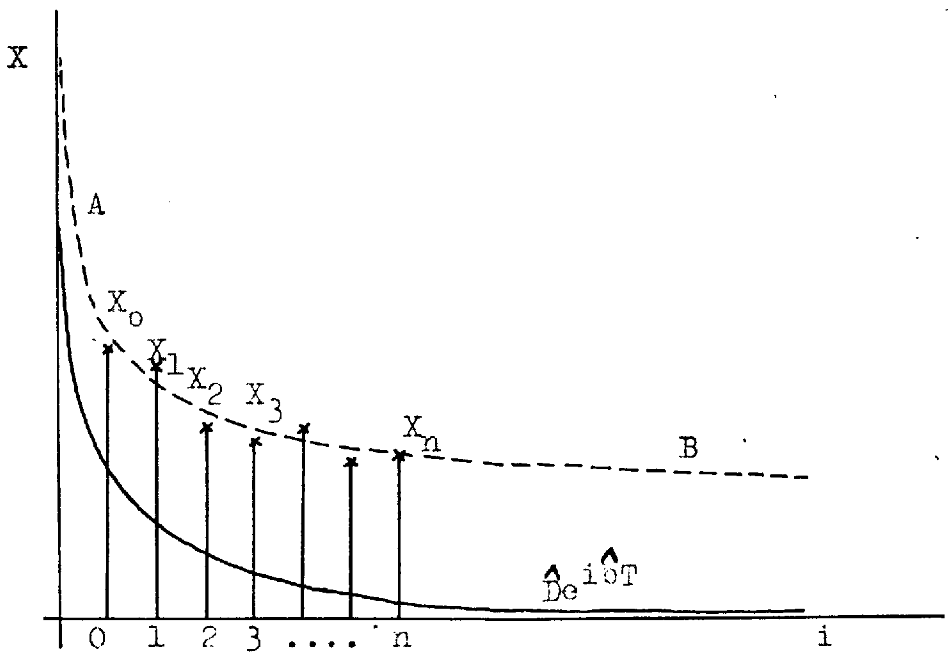


Figure 2.17. Estimation of h

both $b\hat{T}$ and $\hat{D}$ can vary quite considerably as the number of points used in the estimation increases. Further, it is clear from Section 2.12 that estimation of $\hat{h}$ is necessary in some cases. The run-end detector will thus take the form shown in the block diagram of Figure 2.18 which, however, gives a general outline only and few details.

A varying number of video samples is dealt with, depending on the time that has elapsed since the end of the last run. Thus, if the order in which samples arrive is X_0, X_1, X_2, X_3, etc., testing might proceed as shown in Table II

The entering video first undergoes a delay of one Nyquist interval in order that the difference amplifier can obtain the first differences of successive samples (see Equation 2.41 of Section 2.11). The natural logarithms of the first differences are obtained by the use of an amplifier with a logarithmic characteristic and straight line fitting is then carried out by the summation of successive samples of the transformed signal, with the appropriate weightings.

No arrangement for sampling the signal is shown in Figure 2.18. Even using pulses of the $\cos^2$ type, a much larger bandwidth is needed for samples of a signal than for the signal itself. Thus simpler circuits can be employed if continuous signals can be processed in the run-end detector. There is no reason why run-end detection should not be carried out using continuous signals, sampling of the video signals being necessary only at the elastic encoding stage of the bandwidth compression system. It is, however, easier to explain run-end detection using discrete samples.

It is convenient at this stage to explain how the apparent difficulty of taking logarithms of negative quantities is overcome. The four possible types of exponential curve, together with the corresponding polarities of the parameters D and b are shown in Figure 2.19. From Section 2.11 we have:

$$S_{r+1} - S_r = Y_r \qquad (2.75)$$

and

$$\log Y_r = \log D(e^{bT} - 1) + rbT \qquad (2.76)$$

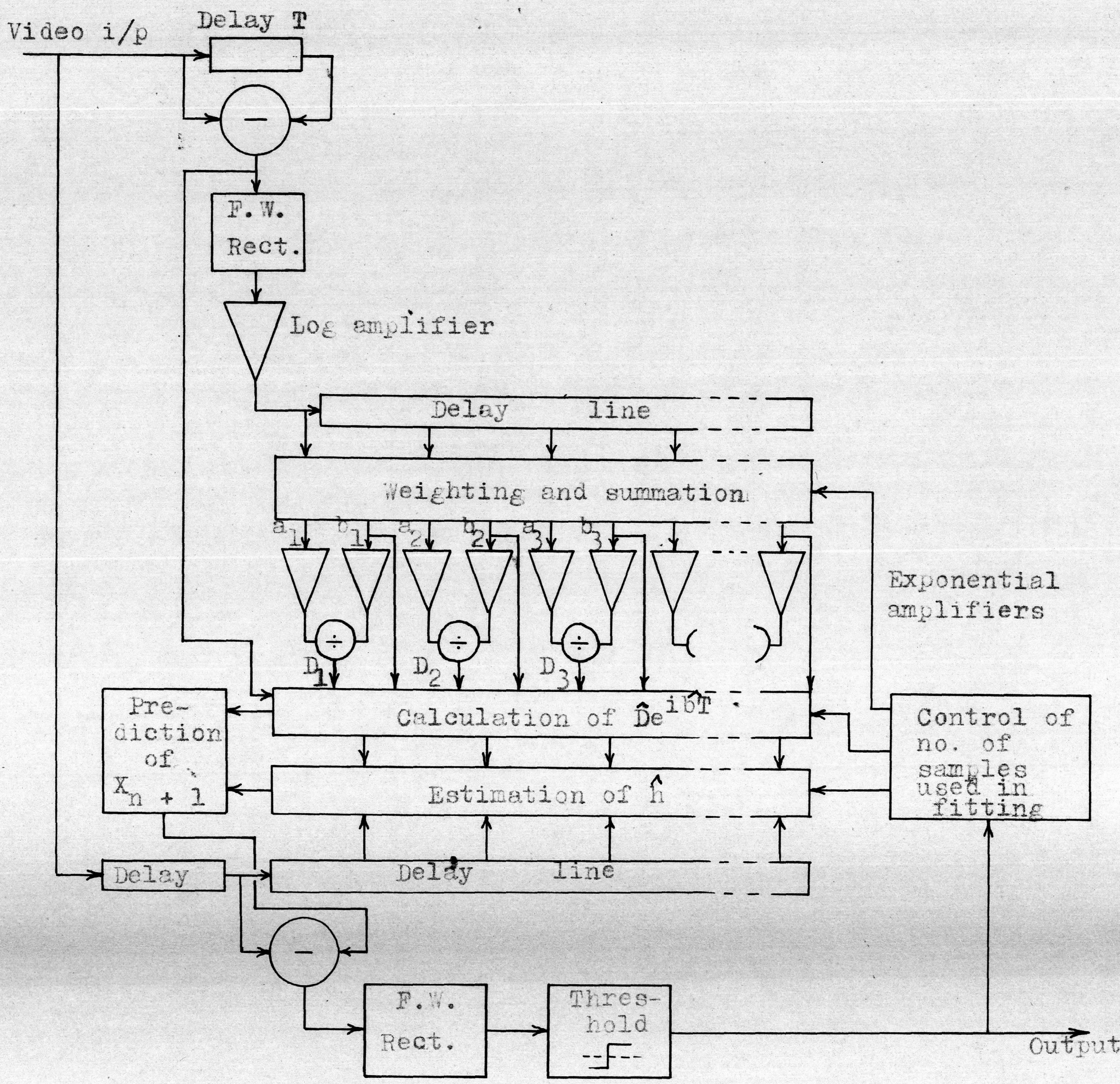


Figure 2.18. Outline block diagram of the exponential run-end detector.

TABLE II

Sequence of Testing Samples

Time Interval	Samples used to estimate a curve	Sample tested against curve	Result
1	X_{-2}, X_{-1}, X_0	X_1	X_1 not in run formed by X_{-2}, X_{-1}, X_0; transmit X_1
2	X_{-1}, X_0, X_1	X_2	Exponential cannot be fitted; transmit X_2
3	X_0, X_1, X_2	X_3	X_3 lies in run: X_0, X_1, X_2, X_3
4	X_0, X_1, X_2, X_3	X_4	X_4 lies in run: X_0, X_1, X_2, X_3, X_4
5	X_0, X_1, X_2, X_3, X_4	X_5	X_5 not in run: X_0 ...... X_4; transmit X_5
6	X_3, X_4, X_5	X_6	Exponential cannot be fitted; transmit X_6
etc.	etc.	etc.	etc.

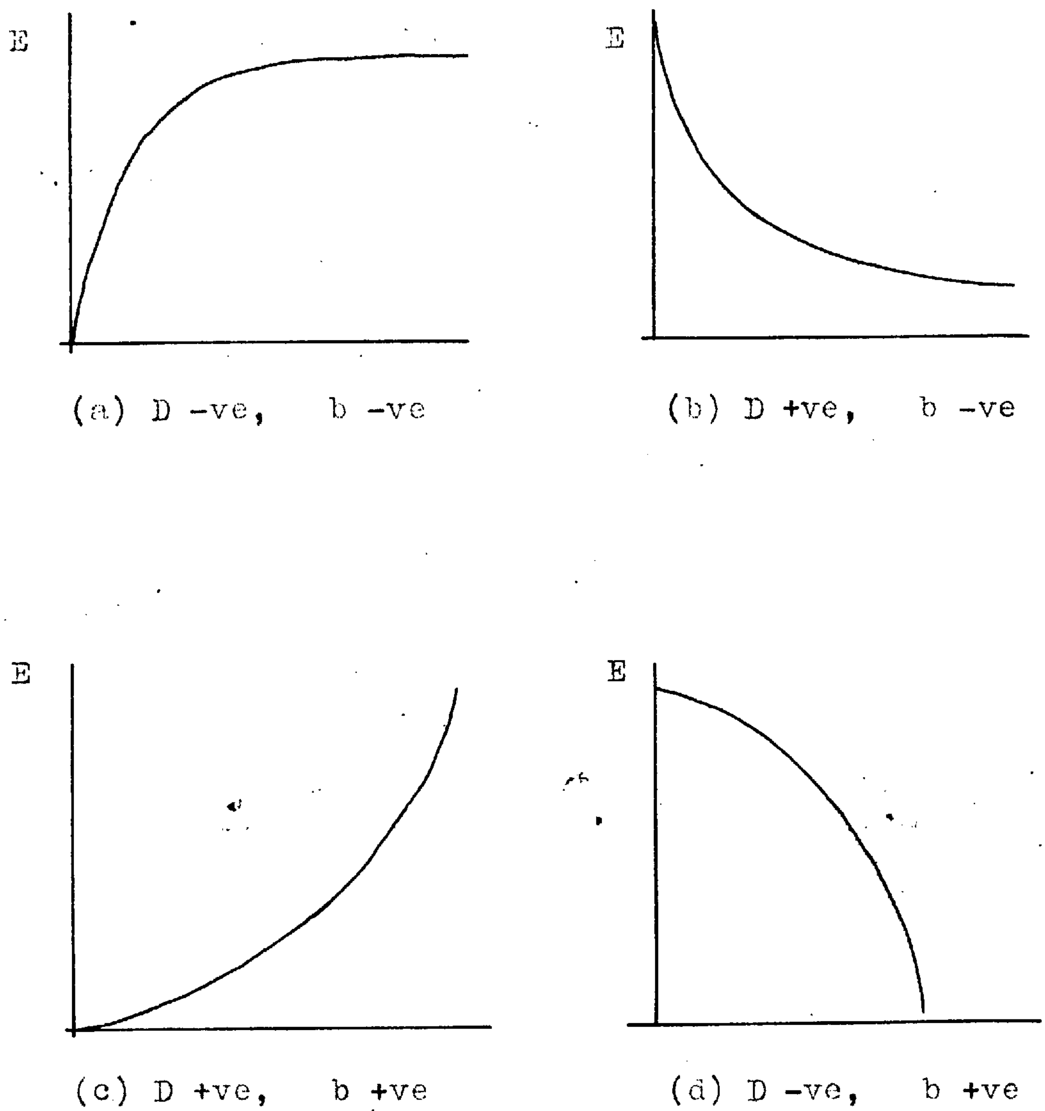


Figure 2.19. The four types of exponential curve found in television signals

Clearly, no difficulty arises in cases (a) and (c) of Figure 2.19 since $S_{r+1} - S_r$ is positive. In both cases the quantity $D(e^{bT} - 1)$ must also be positive and this is consistent with the requirements of Equation 2.76 above. In case (b), however, the negative value of b together with the positive value of D causes $D(e^{bT} - 1)$ to be negative. Fortunately the curve has a negative slope so that $S_{r+1} - S_r$ is also negative and the difficulty is easily resolved by taking $S_r - S_{r+1}$ so that

$$\log (S_r - S_{r+1}) = \log D(1 - e^{bT}) + rbT \quad (2.77)$$

Case (d) is very similar, for here D is negative and b is positive so that $D(e^{bT} - 1)$ is positive; $S_r - S_{r+1}$ is also a positive quantity since the slope of the curve is negative.

The fact that the positive value of the difference between S_r and S_{r+1} is always needed means that a full-wave rectifier can be used prior to the logarithmic amplifier. The fitting of a straight line to the transformed data will then lead to the determination of b directly and the value of D will follow. It should be noted that a knowledge of the polarity of the curve's slope is needed for the determination of D. This is not difficult to obtain from a point preceding the difference signal rectifier.

Estimates of a and b having been obtained, $\hat{D}$ can now be obtained by use of the relation

$$\hat{D} = \frac{e^{\hat{a}}}{e^{\hat{b}T} - 1} \quad (2.78)$$

and the values of $\hat{D}e^{i\hat{b}T}$ then follow. The means of estimating h was derived at the end of Section 2.12; the average of $X_i - \hat{D}e^{i\hat{b}T}$ is found. In Figure 2.18 a delay line is shown for simultaneously supplying successive values of X_i to the averaging circuits.

A curve having been fitted to those samples which, by previous testing, are known to lie in one run, the value of the succeeding sample can be predicted. The actual value of the sample at this sampling instant is now compared with the predicted value by means of a difference amplifier and the modulus of the difference is arrived at by full-wave rectification of the difference signal. The resulting signal undergoes comparison with a number of thresholds, one Schmitt trigger circuit being employed for each threshold value as indicated in Figure 2.18. Which of the thresholds is used in the comparison is determined by the number of points used in the estimation of the run's parameters. Thus, if few points are used, considerable error in prediction is possible and a coarse threshold is switched in. With more points taking part in the estimation process, the probable error is smaller and a fine threshold can be employed. The magnitude of each threshold is dependent not only upon the power of the random noise which accompanies the signal but also upon the amplitude resolution of the observer's eye (as mentioned in Section 2.7). It is not suggested, therefore, that an expression for each threshold value as a function of the noise power can be derived. Rather, the thresholds are to be determined subjectively. In a working system, operating over a channel of a given capacity, threshold values would be adjusted to give the best possible received picture consistent with the limited information rate permitted. An experimental system might require other settings, however. For example, it might be desired to determine the data reduction which could be obtained with no visible degradation in the reconstructed picture. Another alternative is to estimate the information rate required to transmit "just acceptable" pictures of various types. This short discussion thus emphasises that threshold settings cannot be stated mathematically.

The exceeding of any threshold must cause the appropriate sample to be transmitted and must reset the detector so that three samples only are used in the next estimation. All the trigger circuits feed into an OR gate, therefore, outputs being taken to the run-length restriction equipment (see Chapter 3) and to the circuits concerned with estimation.

The work of the past few sections leads to the conclusion that run-end detection for the exponential model will best be carried out by transforming the data into a linear form. A device which fits a straight line directly to data and indicates the points at which the slope changes is a first-order run-end detector, however. Thus, although a more general model for a television signal (the exponential model) has so far been considered it was felt advisable to defer further investigations on this representation until the first-order case had been dealt with. This was, however, found to be a topic of considerable magnitude and it has not been possible to consider the exponential case further. The remainder of this chapter deals with the design and construction of a first-order run-end detector and the results obtained from this instrument.

2.14 First-Order Run-end Detection

The requirements for first-order run-end detection can be derived from the description of the general process in Section 2.7. In this case the parameter with which we are concerned is the slope of the signal.

The slope of the run formed by the n th and n + 1 th sample is to be compared with that of the run formed by the n samples previously tested. The least squares estimate of the slope of a run of n samples is given by Equation 2.67 (see Section 2.11). This equation is now entirely correct since no non-linear transformation is being employed. Unity weighting can therefore be applied to every sample. Thus for n = 2, 3, 4.... the estimates $\hat{b}$ of the slope are given by:

$$\hat{b} = \frac{X_2 - X_1}{T} \qquad (n = 2) \qquad (2.79)$$

$$\hat{b} = \frac{X_3 - X_1}{2T} \qquad (n = 3) \qquad (2.80)$$

$$\hat{b} = \frac{3X_4 + X_3 - X_2 - 3X_1}{10T} \qquad (n = 4) \qquad (2.81)$$

$$\hat{b} = \frac{2X_5 + X_4 - X_2 - 2X_1}{10T} \qquad (n = 5) \quad (2.82)$$

The only estimate possible for the slope of the run formed by the n th and n + 1 th sample is

$$\hat{b}_{n+1} = \frac{X_{n+1} - X_n}{T} \qquad (2.83)$$

Thus the differences of the slope are given by

$$\Delta b = \frac{X_3 + X_1 - 2X_2}{T} \qquad (n = 2) \quad (2.84)$$

$$\Delta b = \frac{2X_4 - 3X_3 + X_1}{2T} \qquad (n = 3) \quad (2.85)$$

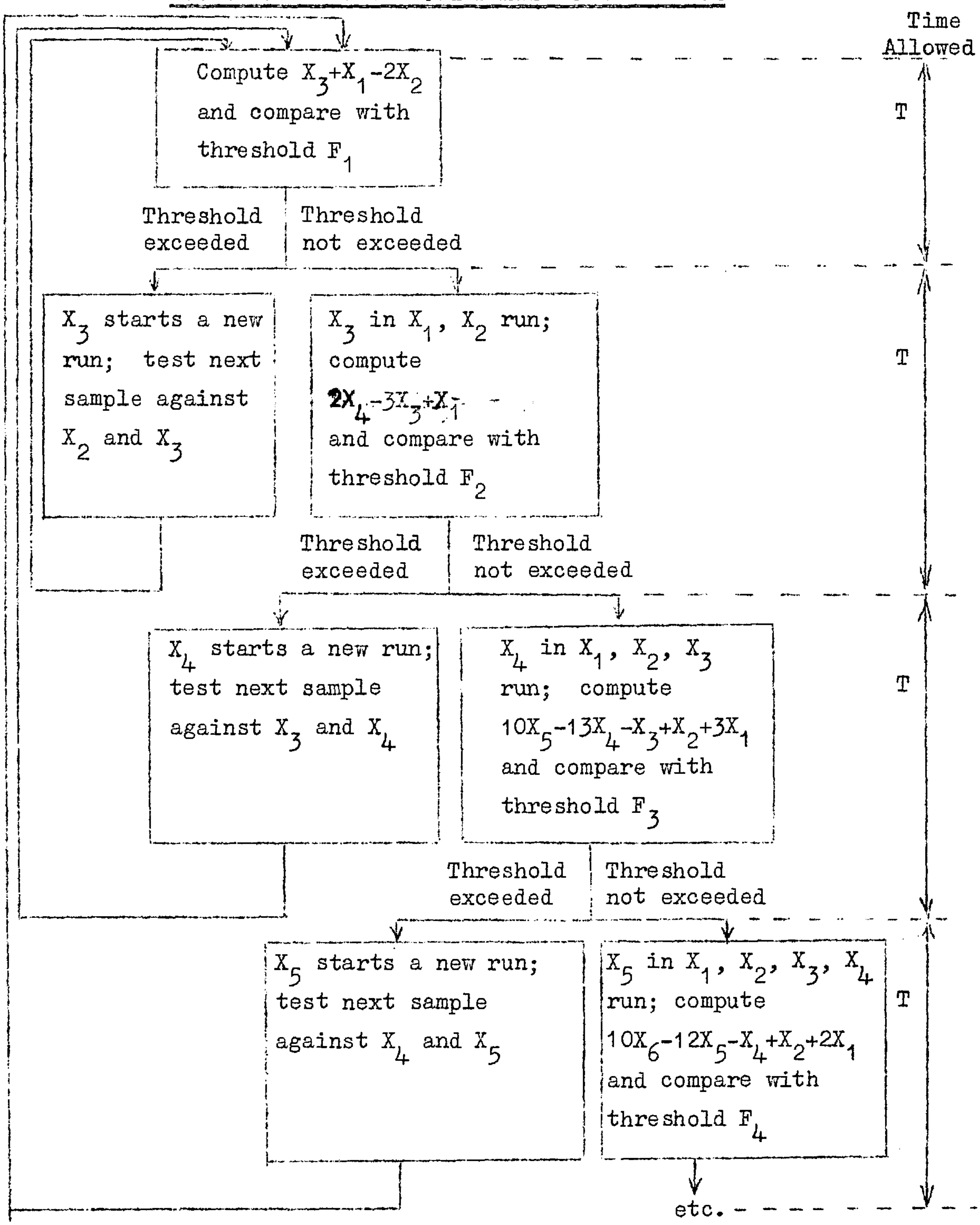
$$\Delta b = \frac{10X_5 - 13X_4 - X_3 + X_2 + 3X_1}{10T} \qquad (n = 4) \quad (2.86)$$

$$\Delta b = \frac{10X_6 - 12X_2 - X_4 + X_2 + 2X_1}{10T} \qquad (n = 5) \quad (2.87)$$

The procedures necessary for first-order run-end detection can now be arrived at; they are best laid out in the form of a chart. (Table III). It will be seen that one computation and one decision are required in each Nyquist interval. So far no upper limit has been placed on the number of samples to be considered in the detector. The use of a larger number of samples in slope estimation results in an estimate with lower noise power than would be the case if few samples were used. Random noise will here cause errors to be made in the decision processes, the probability of error also being a function of the thresholds employed.

TABLE III

Procedures for 1st Order Run-End Detection



An error causing a sample to be inserted where it is not necessary will result in a loss of data reduction whilst an error causing omission of a sample will result in degradation of the reconstructed picture. Prediction of the permissible noise power in the slope-difference signal is thus not easily possible; subjective testing is once again required. As a rough guide to the effect of considering various numbers of samples, however, Figure 2.20 shows the noise power of the slope-difference signal for 1, 2, 3 ......6 samples.

The noise power of such a signal is given by the relation

$$(\sigma\Delta)^2 = \left(\frac{\partial\Delta}{\partial X_1}\ \sigma X_1\right)^2 + \left(\frac{\partial\Delta}{\partial X_2}\ \sigma X_2\right)^2 + \ldots \tag{2.88}$$

provided that the samples X_1, X_2, etc. are independent. Thus, for example, with four samples

$$\Delta = \frac{2X_4 - 3X_3 - X_1}{2T} \tag{2.89}$$

so that

$$\sigma_\Delta{}^2 = \left(\frac{2}{2T}\ \sigma\right)^2 + \left(\frac{3}{2T}\ \sigma\right)^2 + \left(\frac{1}{2T}\ \sigma\right)^2$$

$$= 3\tfrac{1}{2}\ \frac{\sigma^2}{T^2} \tag{2.90}$$

Any length of run can be handled by the detector despite the limited number of samples which can be included in a single computation. Suppose, for example that five samples are the maximum that can be stored and processed. The function involving five samples is

$$\Delta bT = 10\ X_5 - 13\ X_4 - X_3 + X_2 + 3\ X_1 \tag{2.91}$$

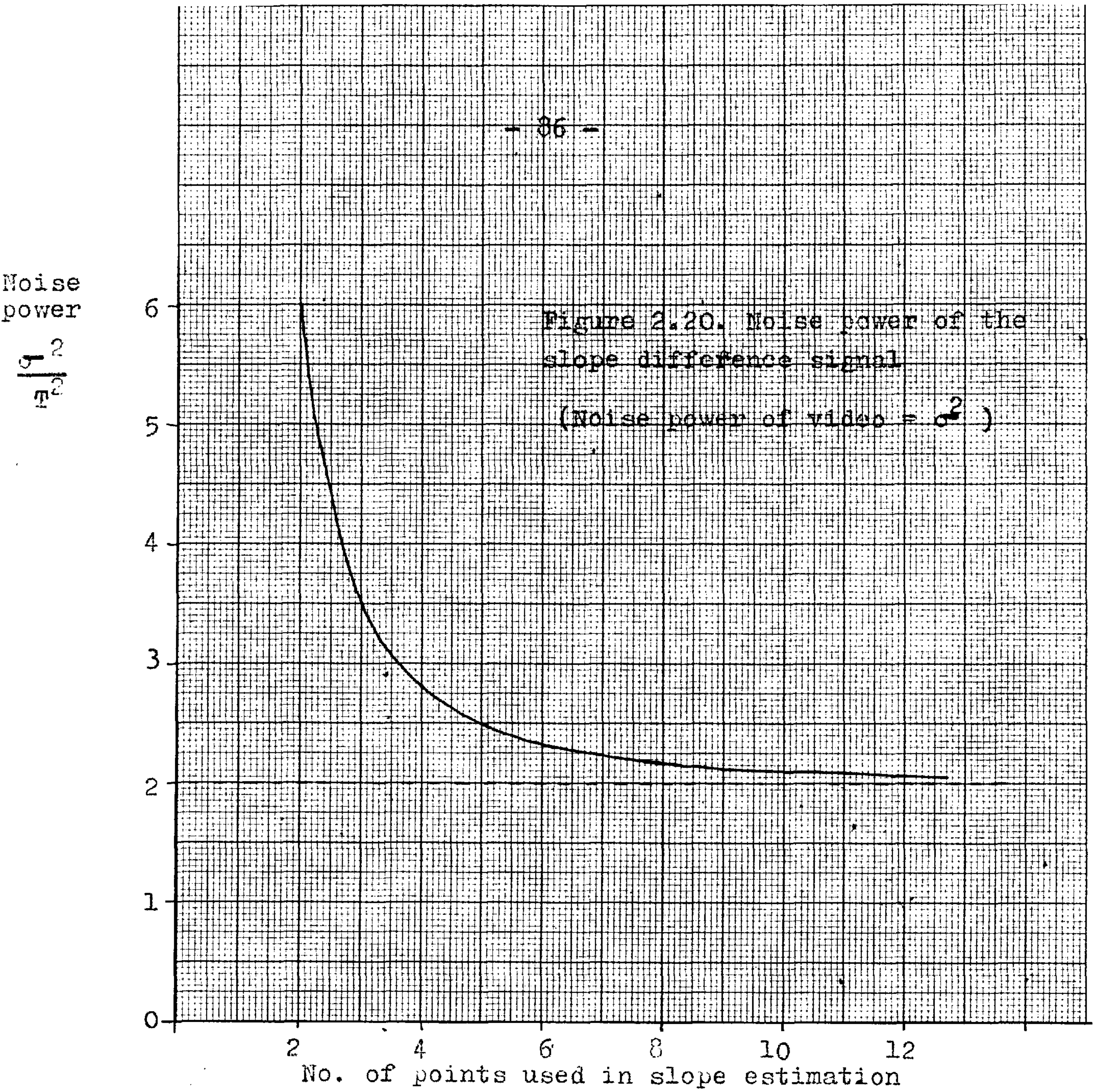


Figure 2.20. Noise power of the slope difference signal

(Noise power of video = σ^2)

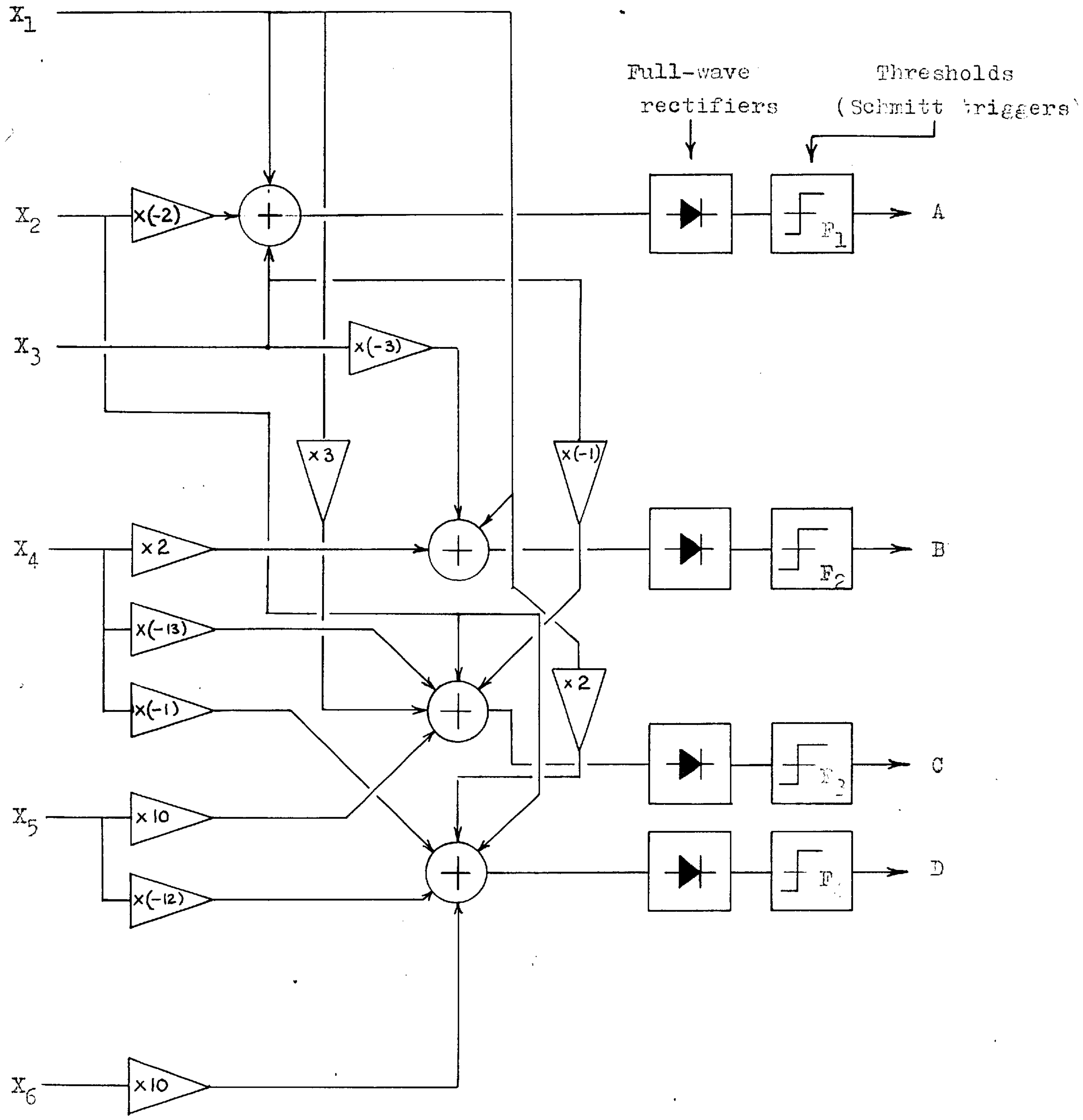
If the threshold (F_3) is not exceeded, the section $X_4 - X_5$ is assumed to lie in the same run as X_1, X_2, X_3 and X_6 must now be tested. This can be done by using the last four samples of the run, X_1 being discarded from the computation. The slope difference is now given by

$$\Delta bT = 10\ X_6 - 13\ X_5 - X_4 + X_3 + 3\ X_2 \tag{2.92}$$

and the noise power of this function will be the same as that when X_5 was being tested against the run X_1, X_2, X_3, X_4. The process can be repeated until the end of the run is found, at which point the detector can revert to the three point computation, as before.

A block diagram of the detector can now be drawn (Figure 2.21(a) and (b). The arithmetic section of the detector will be understood quite readily if examined in conjunction with equations 2.84 - 2.87. The logic section may require some explanation, however. The logic ensures that the sequence of events shown in Table III is adhered to. Suppose that a considerable time has elapsed with no thresholds exceeded. Clock pulses being supplied through INHIBIT gates I1, I2 and I3 will have switched all the bistables B1, B2 and B3 into the '0' state (i.e. the upper output is at the '0' logic level and the lower complementary output is at the '1' level). Thus G1, G2, G2', G3' and G6 are closed while G4 is open. In this condition the exceeding of threshold F4 results in a 1 being output indicating that the run has ended. The exceeding of any threshold causes OR1 to be energised so that a clock pulse passes through AND1 and B1, B2 and B3 are all switched into the 1 state. Under some conditions the exceeding of a threshold could result in one or more of the bistables receiving 'set' and 'reset' pulses simultaneously. A connection from the output of AND1 to I1, I2 and I3 is thus provided so that the 'reset' inputs of B1, B2 and B3 cannot possibly be energised when a pulse is transmitted by AND1.

The operations so far described result in G1 being open while G2, G3 and G4 are closed. Thus the ending of a run has resulted



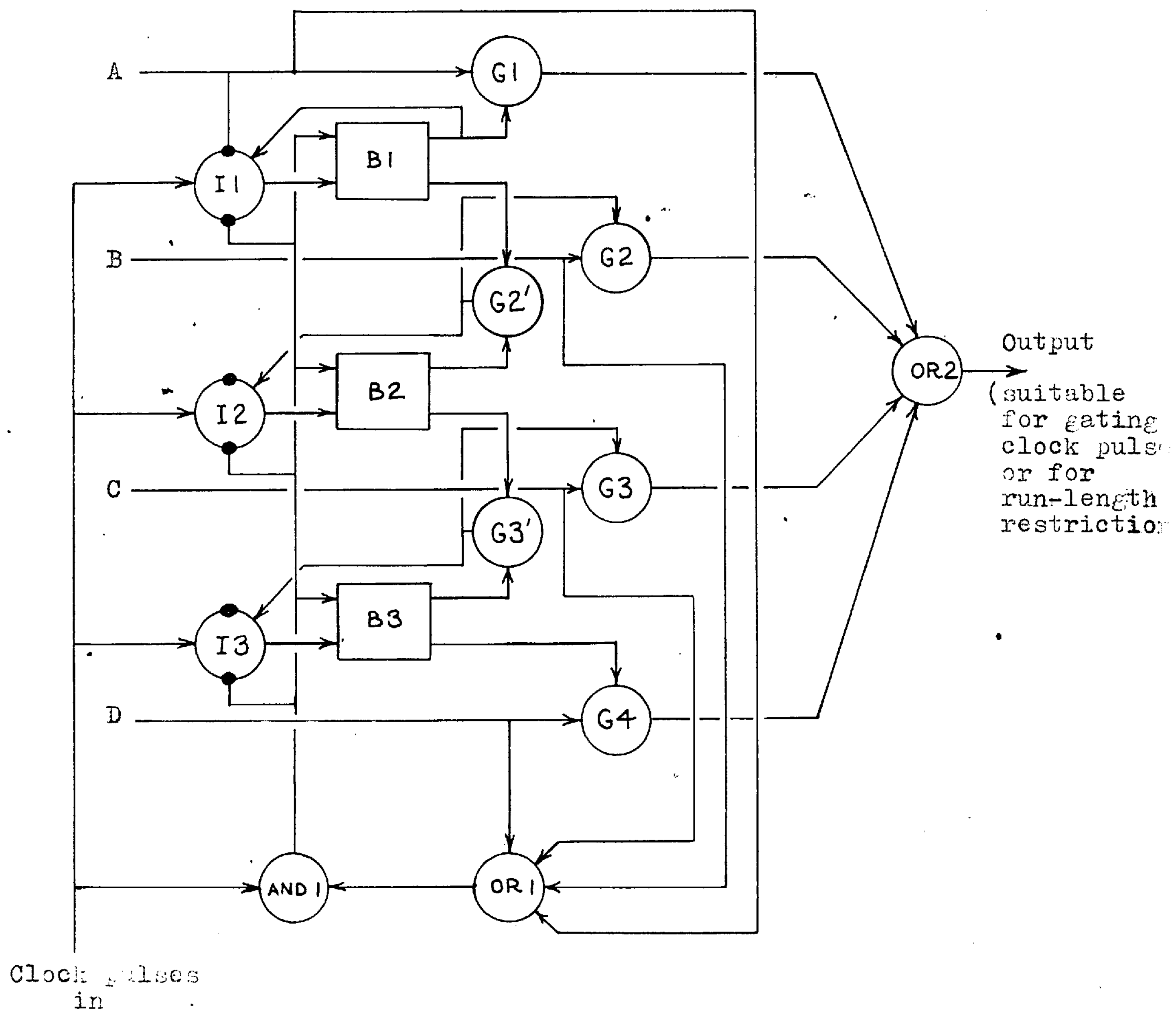


Figure 2.21.(a) The First Order Run-end Detector — Arithmetic Section

(b) Logic Section

in threshold unit F1 being connected to the output gate, OR2 as required by Table III. If F1 is not exceeded during the first Nyquist interval after G1 has opened, a clock pulse passing through I1 switches B1 to the 0 state so that G1 is now closed and G2 and G2' are open. Attention has therefore been transferred to F2 for the next Nyquist interval. The process is repeated at each clock pulse either until G4 is open or until a threshold is exceeded with a resultant return to examination of F1. When G4 is open no further downward steps are possible but this condition remains until F4 is exceeded. Thus the requirements of Table III have been met.

No means of gating sampling pulses has been shown in Figure 2.21(b). The output of OR2 can be used directly for this purpose, however, as it is desired to sample the video signal whenever any threshold is exceeded. If run-length restriction is to be used, however, the signal from OR2 can be sent directly to the Translator (see Section 3.6.2.), gating of the sampling pulses now being performed by the output of the Sample Gating Unit.

2.15 Experimental Investigation of First-Order Run-end Detection

It is clear from Section 2.14 that considerable logic would be necessary for the instrumentation of a detector to handle six samples simultaneously. For an initial attempt at first-order run-end detection, therefore, it was considered prudent to reduce somewhat the number of samples to be examined concurrently. The minimum number required for first-order detection is three and in this case no logic is necessary in the detector; the threshold element provides the output directly. The decision to use only this minimum number of samples was further influenced by the realisation that the zero-order run-end detector constructed by Kubba[33] could, with slight alteration, be used for the purpose.

In the zero-order detector, one of the functions computed was

$$X_3 \sim \frac{X_2 + X_1}{2}$$

In order to determine when the difference between X_3 and $\frac{X_2 + X_1}{2}$ exceeded a given value (either positive or negative) the difference signal was rectified and applied to a Schmitt trigger circuit. The successive samples X_1, X_2, X_3 of the video signal were obtained by the use of a video delay unit, the delay between successive taps being T, the Nyquist interval. (The use of a delay line for simultaneously obtaining samples from various points on a signal is explained in Section 2.7.) From the foregoing it will be seen that the function

$$X_1 + X_3 - 2X_2$$

could be obtained merely by interchanging the connections feeding X_2 and X_3 from the Video Delay Unit to the Run-end Detector. Strictly the signal thus produced is

$$X_2 \sim \frac{X_1 + X_3}{2}$$

The factor of 2 is easily taken care of, however, by using a different threshold setting; there is no need to use a special amplifier for the purpose. A detailed block diagram of the detector forms Figure 2.22; the circuit diagram appears in Appendix 1.

The necessary signals for the Run-end Detector were supplied by a video delay unit which was available in the laboratory; three successive outputs at intervals of T were employed. So that losses in the delay sections should not affect the arithmetic operations, variable gain input amplifiers were provided in the detector. Correct adjustment was achieved by observing the outputs of A_1 and A_3 with an oscilloscope and adjusting the amplitudes of the two signals to be equal. A_2 was then set so that the output of the difference amplifier was zero during blanking periods. This condition was detected by setting the threshold of the Schmitt circuit at a low level and displaying either the Run-end Detector output or the resulting sampling pulses on a monitor.

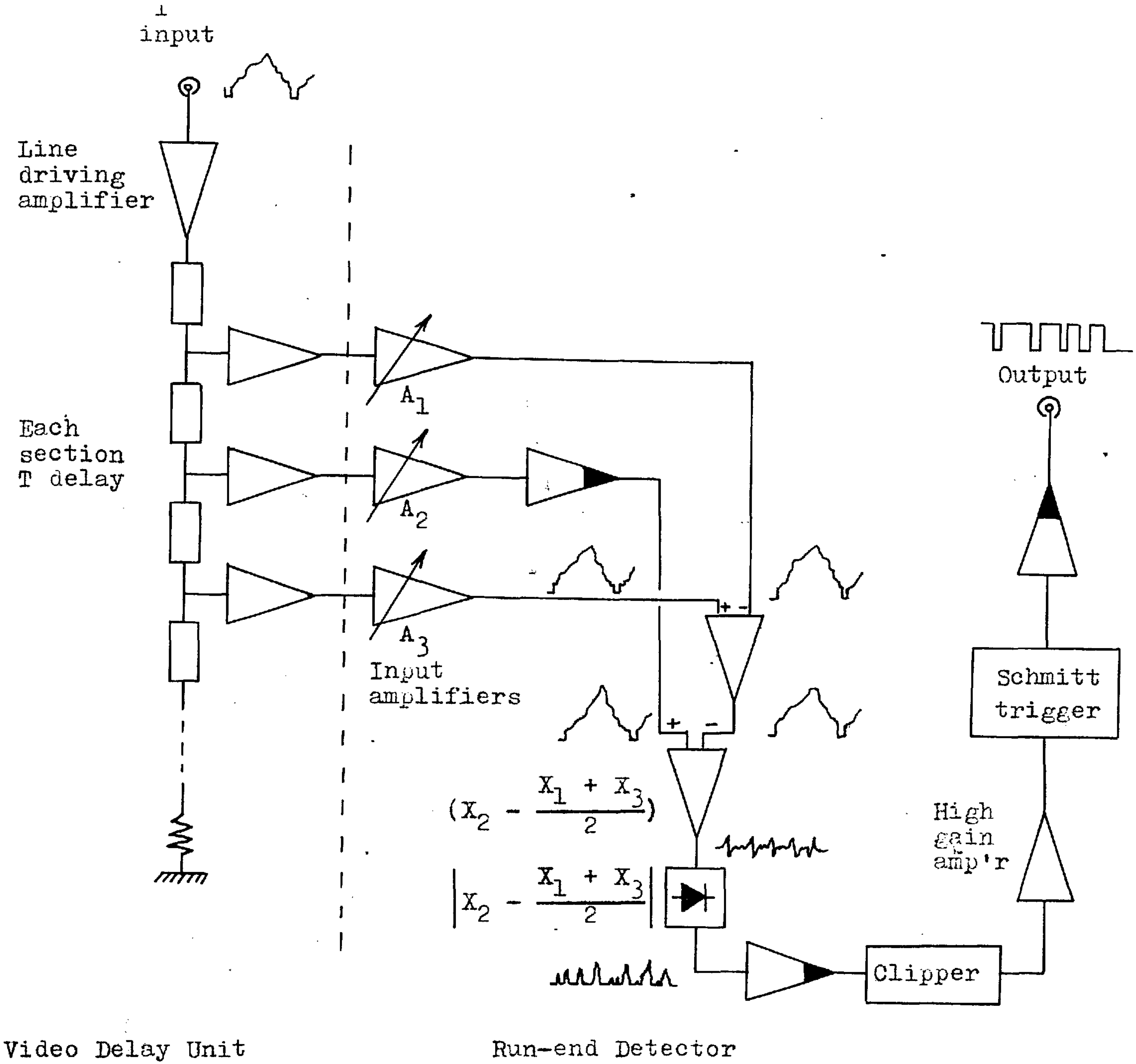


Figure 2.22. Detailed block diagram of actual 1st order run-end detector

The tests used to assess the performance of the detector included the measurement of the frequency response of the linear part of the instrument and a comparison between the characteristic found experimentally and the theoretical curve. To obtain the latter we write:

Output of sum and difference amplifiers =

$$f(t) + f(t + 2\tau) - 2f(t + \tau) \qquad (2.93)$$

where $f(t)$ is the signal at the input to the delay unit

and τ is the delay in each section of the delay unit (here equal to T).

(Constants such as amplifier gains are here omitted and assumed to be independent of frequency).

Now if $f(t) = Ke^{j\omega t}$ (2.94)

then

$$e(t) = Ke^{j\omega t} + Ke^{j\omega(t + 2\tau)} - 2Ke^{j\omega(t + \tau)} \qquad (2.95)$$

and if the frequency response function is denoted by $T(j\omega)$,

then $$T(j\omega) = \frac{Ke^{j\omega t} + Ke^{j\omega(t+2\tau)} - 2Ke^{j\omega(t+\tau)}}{Ke^{j\omega t}} \qquad (2.96)$$

$$= 1 + e^{j\omega 2\tau} - 2e^{j\omega\tau} \qquad (2.97)$$

Now, putting $\omega\tau = \theta$, (2.98)

$$T(j\omega) = 1 + \cos 2\theta + j \sin 2\theta - 2 \cos \theta - 2 j \sin \theta \qquad (2.99)$$

and $$|T(j\omega)| = \sqrt{(1 + \cos 2\theta - 2 \cos \theta)^2 + (\sin 2\theta - 2 \sin \theta)^2} \qquad (2.100)$$

$$= \sqrt{2}\ (3 + \cos 2\theta - 4 \cos \theta)^{\frac{1}{2}} \qquad (2.101)$$

This function has been plotted - Figure 2.23. At frequencies below 100Kc/s it has been necessary to use an approximation since, in this region, the parameter θ is small and $\cos\theta$ tends to unity.

$$g(\theta) = 3 + \cos 2\theta - 4\cos\theta \tag{2.102}$$

$$= 3 + 1 - 2\sin^2\theta - 4(1 - \sin^2\theta)^{\frac{1}{2}} \tag{2.103}$$

At small values of θ we can write

$$\sin\theta \doteqdot \theta \tag{2.104}$$

so that $g(\theta) = 4 - 2\theta^2 - 4(1 - \theta^2)^{\frac{1}{2}}$ (2.105)

$$= 4 - 2\theta^2 - 4(1 - \tfrac{1}{2}\theta^2 - \frac{\theta^4}{2.4} - \frac{\theta^6}{2.4.6} - \ldots\ldots) \tag{2.106}$$

$$= 4 - 2\theta^2 - 4 + 2\theta^2 + \frac{\theta^4}{2} + \frac{\theta^6}{2.6} + \ldots\ldots \tag{2.107}$$

Now as $\theta < 1$, $\theta^6 << \theta^4$,

thus $|T(j\omega)| = \theta^2$ very nearly, at $f < 100$ Kc/s. This agrees with our expectations; the second differencer behaves almost exactly as a double differentiator, except at the upper end of the frequency range where the period of the input signal is comparable with the delay section.

The measured frequency response of the detector is also plotted in Figure 2.23. The gain from the input of the Video Delay Unit to the output of the difference amplifier was found to be 20 dB at 3 Mc/s; this point has therefore been used as a reference for the decibel scale which is otherwise arbitrary. There is, however, no means of fixing the two curves in relation to each other except by fitting together the regions over which they are most similar, i.e. from about 300 Kc/s to 6 Mc/s. Over this region agreement between the predicted and the actual responses is good; at lower frequencies the actual gain of the instrument is higher than predicted. In this region the detector gain is very low, however, so that the measurement of a small signal in the presence of much larger signals is being attempted and errors must be expected. Clearly, breakthrough could

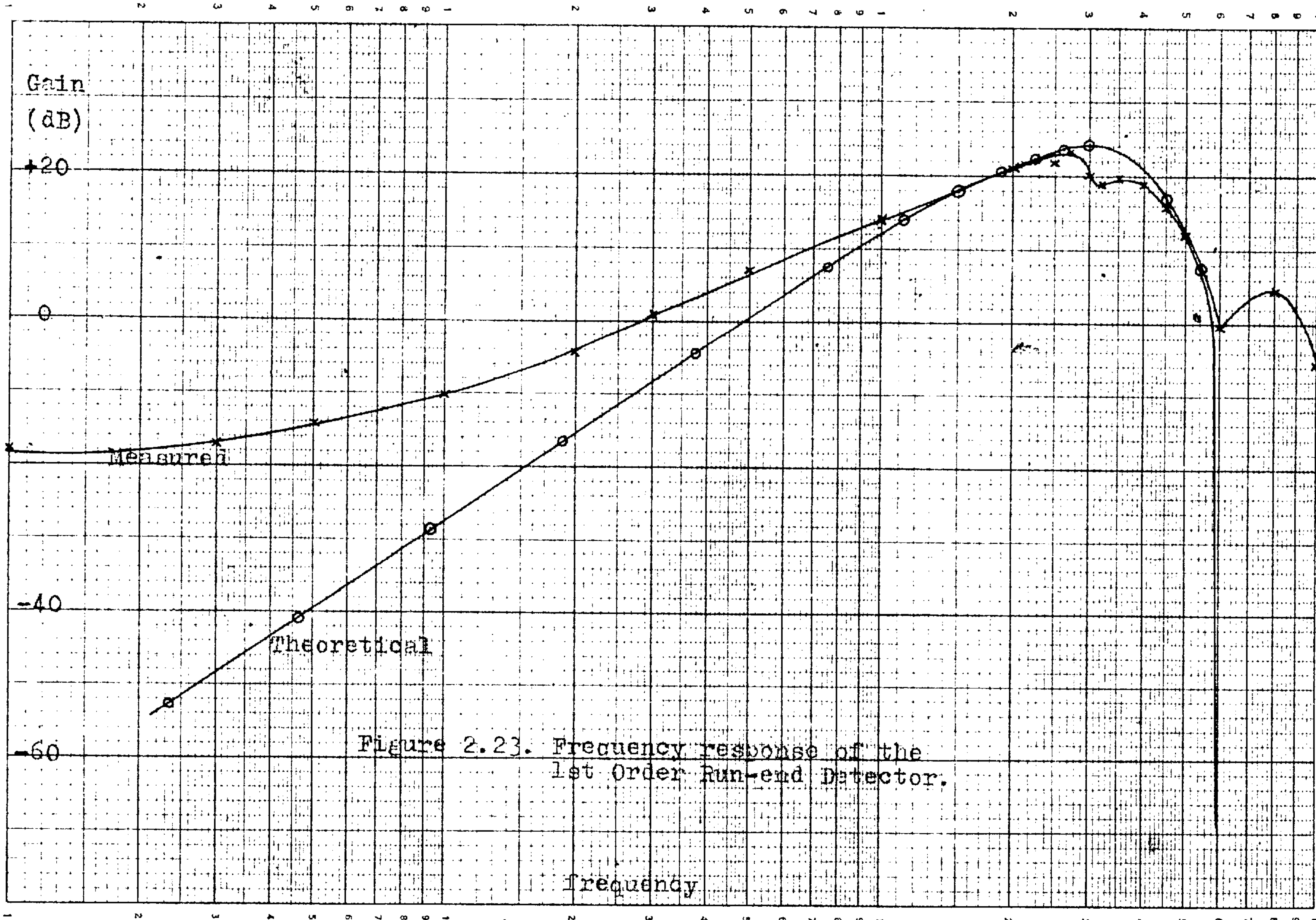


Figure 2.23. Frequency response of the
1st Order Run-end Detector.

be occurring in the detector itself or in the measurement equipment; if the former is the case the error will result in the insertion of more samples than are required when signals with low frequency components are being dealt with. The region over which the errors chiefly occur (from zero frequency to 300 Kc/s) is only about one tenth of the total bandwidth being considered, however, so the errors will be of little importance.

The performance of the detector was also assessed by measurement of the change of signal slope necessary to exceed the threshold at various settings of the potentiometer in the Schmitt circuit. The change of signal slope which it was necessary to detect for the satisfactory reconstruction of pictures could thus be arrived at by processing a video signal, using the detector together with the First-Order Interpolator and observing the setting of the Schmitt control.

A Tektronix oscilloscope provides a sawtooth waveform having the form shown in Figure 2.24. A potential divider circuit was arranged to reduce the amplitude of this signal and to offer an impedance of 75 ohm to the Video Delay Unit (Figure 2.25). The signal has zero slope when its amplitude is constant; thus the change of slope at the commencement of the sawtooth was given directly by observation of the slope of the ramp portion. By detecting the exceeding of the threshold using another oscilloscope, the input signal necessary to just exceed a given threshold could be measured. The resulting calibration curve for the detector is shown in Figure 2.26.

The number of sampling pulses inserted at various settings of the threshold control was measured using video signals from four different scenes ("Jean", "Aircraft", "Rockets" and "Albert Hall"). Photographs showing these scenes as they appeared on a picture monitor are included in Section 4.8.2; the flying-spot scanner used for the generation of the video signals is described in Appendix 2. Measurement of the number of pulses inserted per second was effected by using the output signal from the Run-end Detector to gate clock pulses and counting the resulting pulses with a Hewlett-Packard Type 524C electronic

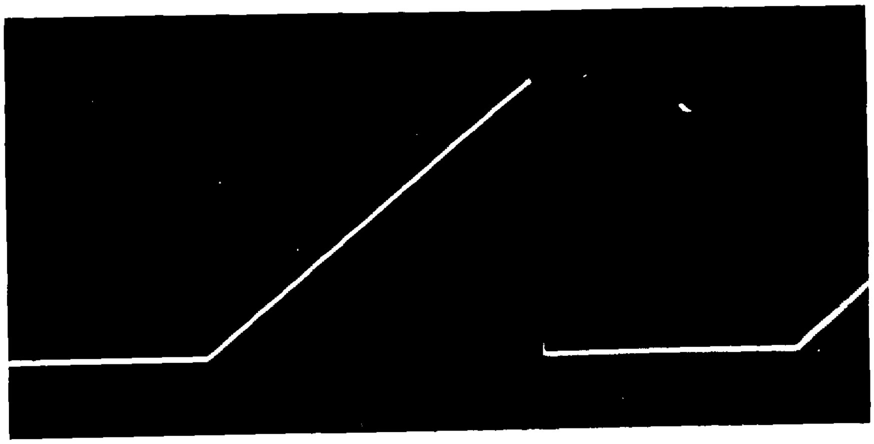


Figure 2.24. Waveform available from the 'Sawtooth Out' terminal of a Tektronix oscilloscope.

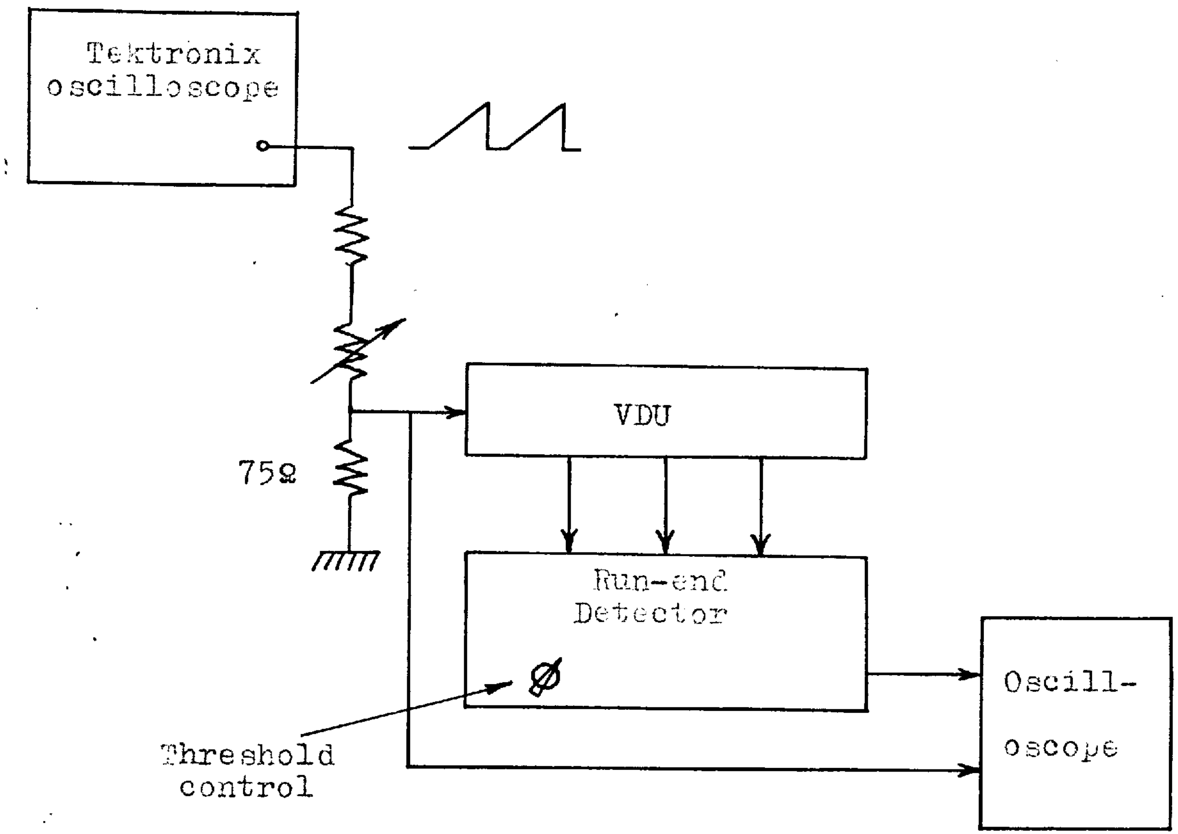


Figure 2.25. Calibration of the 1st Order Run-end Detector

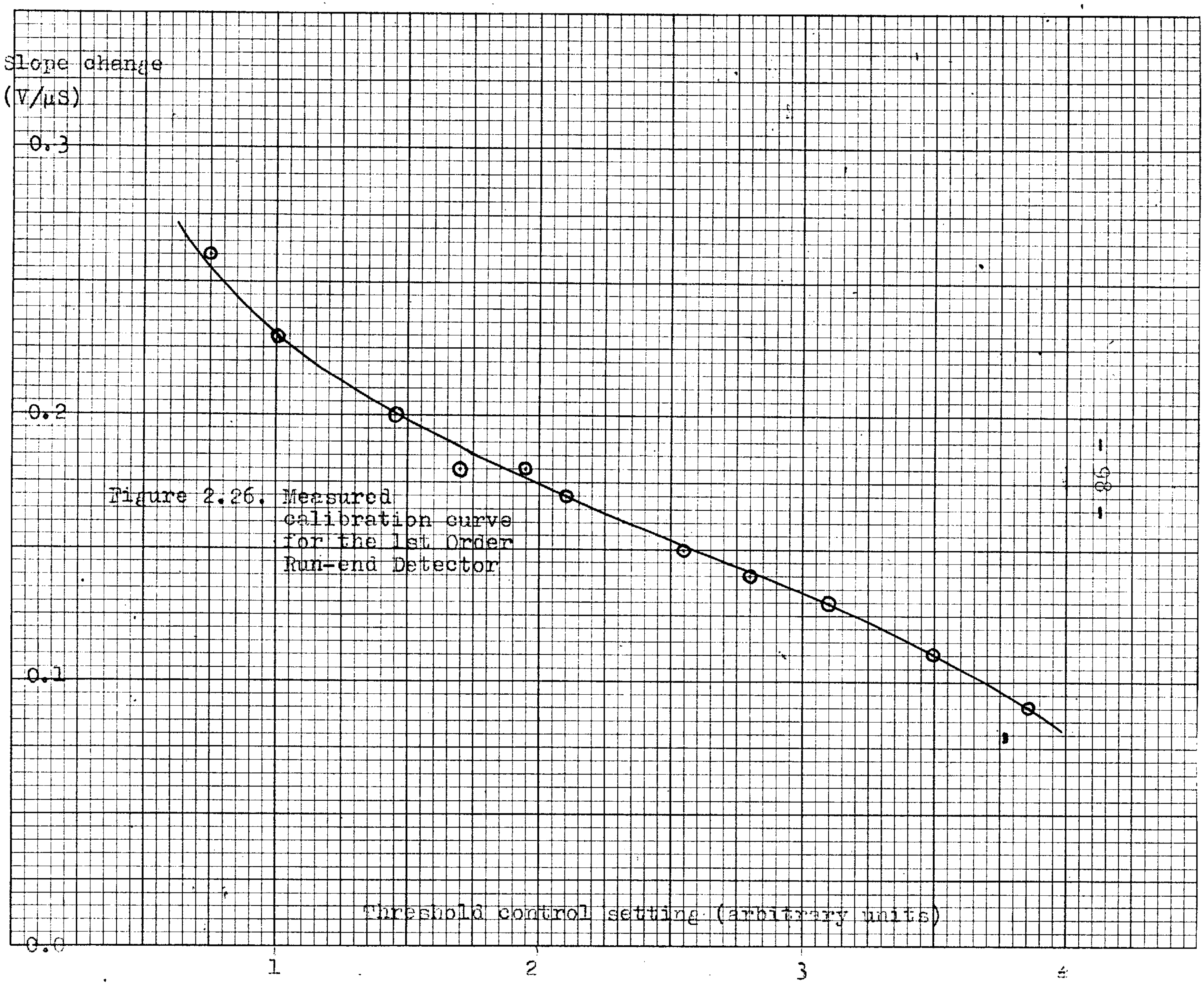


Figure 2.26. Measured calibration curve for the 1st Order Run-end Detector

counter. The results are given in Table IV and Figure 2.27. By examining the reconstructed pictures obtained from the First-Order Interpolator (see Section 4.8.2) and the graphs showing the effects of run-length restriction, the number of samples required prior to restriction, and thus the significant slope change in a given picture can be arrived at. (Table V.)

2.16 Summary and Discussion of Findings on Run-end Detection

The investigations have by no means covered the subject exhaustively; indeed only a few aspects of run-end detection have been touched upon. The early sections of the chapter defined data reduction and laid the foundation for the work by defining a run in a useful but general manner. Discussion of models for television signals then led to the conclusion that the exponential model merited further consideration. Before dealing with the exponential model, however, a general procedure for run-end detection was outlined and it was argued that the method of least squares would form the most satisfactory means of estimating the parameters of runs. Three methods of fitting exponential curves to data then followed - direct least squares fitting, fitting by the use of factorial moments and fitting by transforming the data into a linear form. Of these methods only the third mentioned appeared to be suitable for instrumentation at the speeds necessary for television signals. In addition, the use of approximate formulae was assumed. The results to be expected from such a process were accordingly predicted by fitting curves to points which initially lay on exponential curves but which had been perturbed by random quantities chosen to simulate white noise. The calculations show that the approximate formulae can be expected to yield satisfactory results. Exponential run-end detection will best be carried out by the transformation method, therefore.

The fact that exponential run-end detection will involve the fitting of a straight line to data has led to some investigation of this process and this investigation is reported in the concluding sections of the chapter. The work was in the nature of a pilot

TABLE IV

Unrestricted Sampling Rates

Threshold Control Setting (arbitrary units)	Scene			
	Jean	Aircraft	Rockets	Albert Hall
1.2	130	300	540	
1.5	820	840	770	358
2.0	1490	1400	910	1308
2.5	1760	1640	1120	1590
3.0	1880	1760	1360	1640
3.5	1885	1760	1560	1630
4.0			1630	1700

Kpps (unrestricted)

TABLE V

Significant Slope Change

Scene	No. of pulses per sec. after restriction ($X10^3$)	No. of pulses per sec. before restriction ($X10^3$)	Slope Change (V/μS)
Jean	850	290	0.21
Aircraft	837	280	0.215
Rockets	760	190	
Albert Hall	770	200	0.20

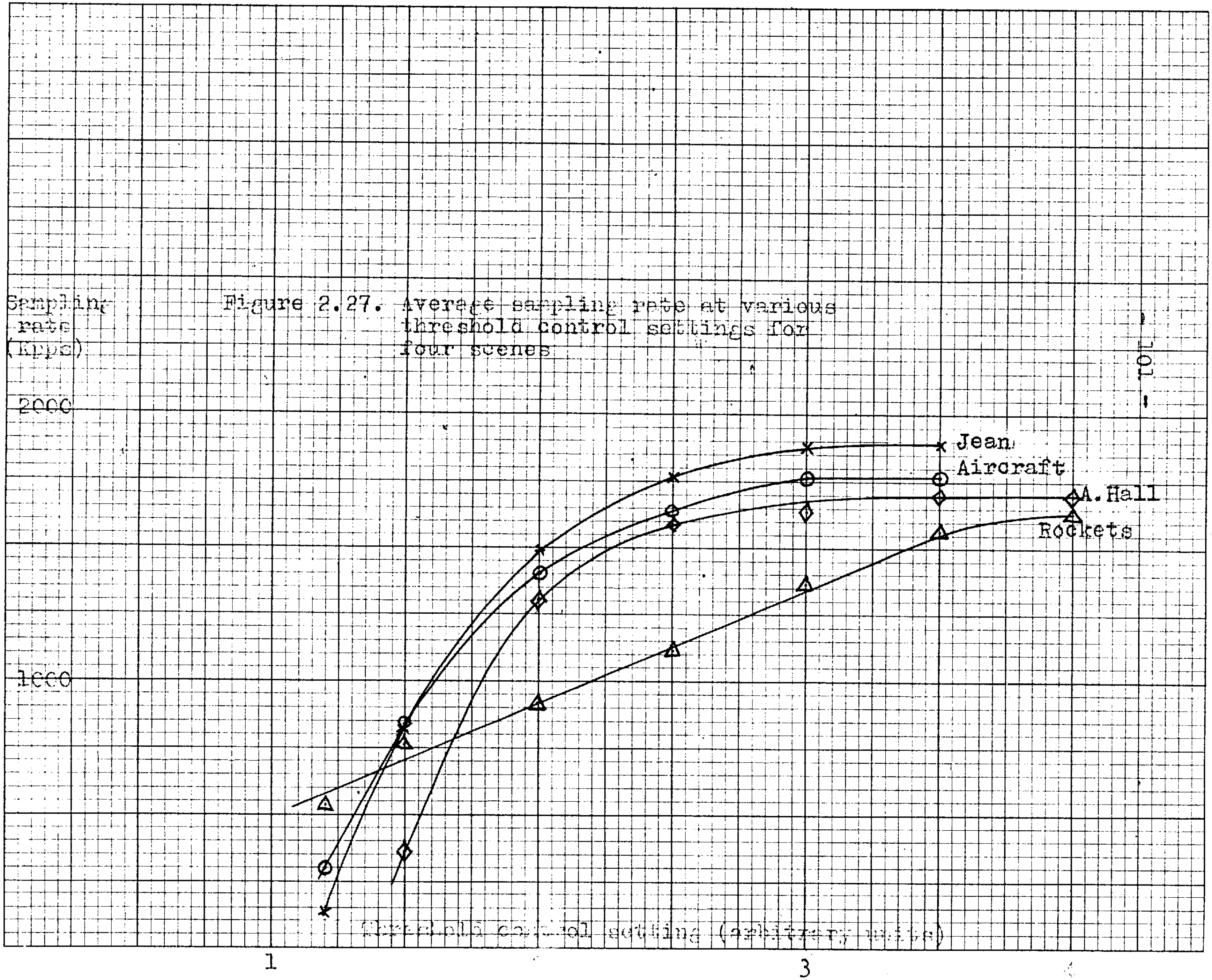


Figure 2.27. Average sampling rate at various threshold control settings for four scenes

study and only the simplest type of run-end detector was constructed. However this operated very nearly as predicted and the results obtained will guide the design of future instruments of this type.

In order to judge the effects of data reduction it was necessary to reconstruct pictures from the reduced data, it being assumed that coding and decoding are processes which taken together need cause no further degradation of the signal. The reconstruction equipment has merited a separate chapter (Chapter 4.) and although not completely satisfactory it has enabled an estimate to be made of the slope difference which must be detected. For four different scenes this is almost constant at 35mV per Nyquist interval or 0.21mV per μS. This is about 1/60 of the maximum possible slope change, a figure which can be compared with the threshold measured in the zero-order case and which was found to be 1/50 of the signal swing from black level to peak white. Thus, in the first-order case we are concerned with detecting a slightly smaller fractional change in the parameter which defines a run than in the zero-order case.

Further comparison with results reported for the zero-order case[33] reveals that fewer samples are inserted prior to run-length restriction in the first-order case, the picture material being of the same nature in both investigations. Two factors are responsible for this but it is not possible to separate their contributions to the reduction. The factors are the use of a model of higher complexity, less sampling at the Nyquist rate thus being required, and the use of run-length restriction thus permitting a slightly lower pre-restriction average sampling rate than the rate which would have been required with no restriction. Further discussion of the effects of run-length restriction is reserved for the following chapter.

3. RUN-LENGTH RESTRICTION

Run-length restriction is an important feature of the Open Loop system but has so far received very little attention.

The process can be briefly defined as the insertion of extra pulses in the non-uniform pulse train fed to the Elastic Encoder so that only a small number of sample spacings remain. The need for this, its advantages and the method of achieving it will be fully described in the ensuing sections.

3.1 The need for run-length restriction

In the system as outlined in section 1.9, the non-uniform train of samples (f) can exhibit any sample spacing, the only restriction being that the intervals between samples must be an integral number of Nyquist intervals. Provided that the Elastic Encoder could deal with a train having all possible run-lengths this would constitute a perfectly feasible system. However, the position signal now has as many possible values as there are run-lengths. If no further restrictions are imposed, this implies an infinite number of possible values for the position signal. The beginning of a case for restricting the permitted run-lengths in some way is already obvious. A simple limitation is to store over one line only (about 600 picture elements) so that all run-lengths from unity to 600 are permitted. Gouriet[22], describing such a system, finds that the position signal would require a signal-to-noise ratio of 72 dB to keep the frequency of errors down to one per line, although various simple means can be used to reduce this figure by a few decibels. Errors in the position channel result in a displacement to the right or left of portions of the scanning line during which the error occurs, resulting in the break-up of vertical or near vertical edges and boundaries in the picture. This is believed to be a very annoying form of picture degradation although little work has so far been carried out to substantiate this belief. Teer[56] has investigated the effect of adding noise directly to the current in the line-deflection coils of a monitor. It was found that the position noise was visible for a peak-to-peak deviation of about one and a quarter times the picture element separation in a 625 line picture. To degrade the picture until it was equivalent to a picture having a conventional brightness signal-to-noise ratio of 30 dB, a peak-to-peak deviation of about $2\frac{1}{2}$ picture elements was required.

Merely adding noise to the scanning current in this way is not at all the same thing as specifying an incorrect run-length, however. Work currently in progress in the Communications Section at Imperial College is aimed towards accurately simulating the effects of errors in specifying run-lengths and determining their subjective effects.

One way in which the effect of noise can be reduced is to recognise only a few values of the position signal. This implies, of course, that only a few run-lengths are permitted. Measurements of run-length statistics, both in our group[44, 45] and outside[29, 61] have shown that short runs are most frequent, the probability distribution being roughly of the form shown in Figure 3.1. For cases where the probability distribution is peaked, as it is here, a statistical code has been derived by Shannon[53] and Fano[17]. In this code, the longest code groups are assigned to the least frequent states. Unfortunately it is not suitable for coding the position signal as a code of constant word length is required. The alternative is to allow only a few standard run-lengths to be fed to the Elastic Encoder, long runs being broken down to multiples of shorter runs in the most economic way. Although this results in some loss of compression, careful choice of the standard run-lengths will minimise the loss and a considerably simplified position signal will result.

3.2 Theory of run-length coding using standard run-lengths

(This section is based largely on the analysis given by Cherry, Kubba, Pearson and myself in a recent paper[10].)

It will be assumed that the run-length statistics for a large number of pictures have been measured.

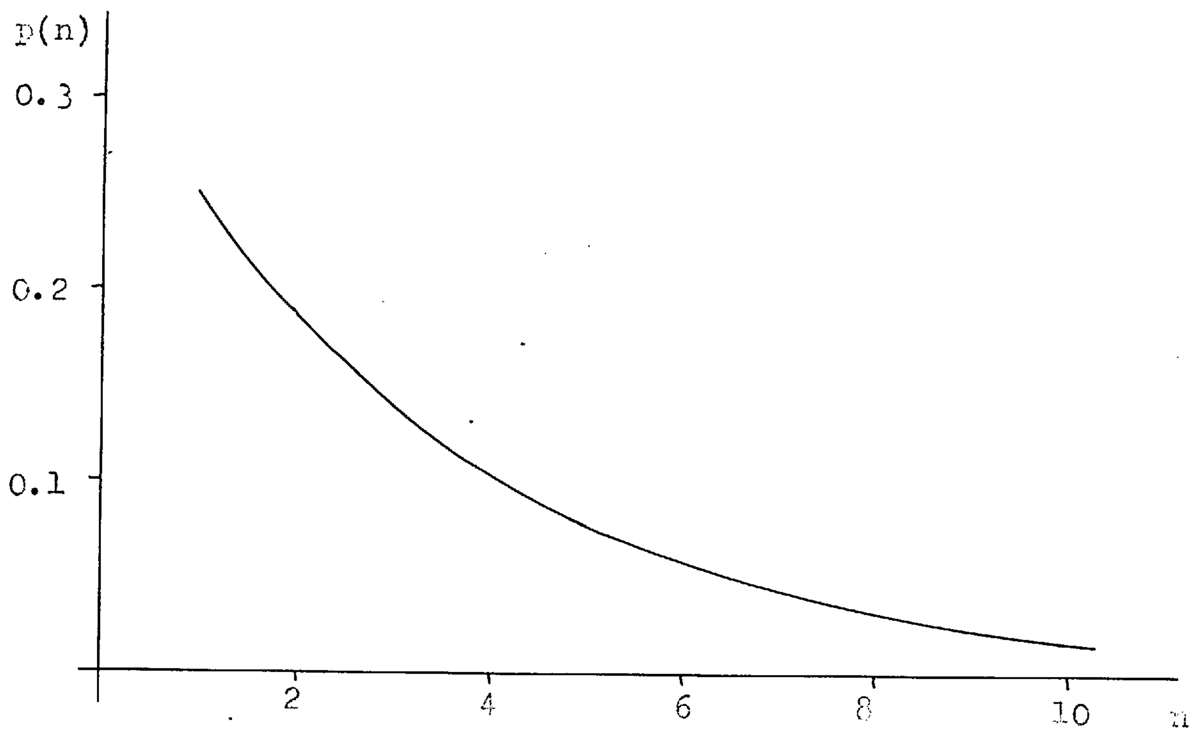


Figure 3.1. Form of run-length probability distribution

The best values for the standard run-lengths depend entirely on the run-length statistics of the pictures it is required to transmit through the system. It is assumed, therefore, that measurements have been made on pictures felt to be representative of these so that an average run-length probability distribution P(n) can be stated. The average run-length, $\bar{n}$ is given by

$$\bar{n} = \sum_{n=1}^{\infty} n\, P(n) \tag{3.1}$$

and if one sample only is needed to specify the amplitude (or other suitable parameter) of the signal during each run, then the maximum obtainable compression C_m with the given run-length probabilities by

$$C_m = \bar{n} \tag{3.2}$$

So far, no restriction has been placed on the length of runs and no position signal has been considered. Suppose that extra samples are inserted so as to break all runs down to the standard lengths, the number of extra samples for a run of previous length n being $X_n - 1$. Thus the total number of samples used to code a run of previous length n is X_n. The compression now becomes C_m' where

$$C_m' = \frac{C_m}{\sum\limits_{n=1}^{\infty} X_n P(n)} \tag{3.3}$$

Clearly, X_n will depend on the standard run-lengths chosen.

In order to specify the run-length, each sample giving the parameter of a run will require to be accompanied by another sample whose possible states (amplitudes) are the same in number as the number of standard run-lengths permitted. The overall compression will thus be

$$\frac{C_m'}{2} = \frac{C_m}{2 \sum\limits_{n=1}^{\infty} X_n P(n)} \tag{3.4}$$

If the number of standard run-lengths is small (3 or 4, say), the position signal will be exceedingly resistant to the effects of channel noise and the possibility of further encoding it can be considered. A signal having a possible levels and a bandwidth W can be encoded so as to have a^q levels when the bandwidth will be W/q. Thus a 3 level signal could have its bandwidth halved by encoding it into 9 levels. Oliver, Pierce, and Shannon(43) give a formula relating the error rate to the signal-to-noise ratio for various numbers of levels. To obtain an error probability of 1 in 10^6 with a 9 level (balanced) signal requires only 28.2 dB.
The overall compression now becomes

$$C_o = \frac{C'_m}{1 + 1/q} = \frac{\sum_{n=1}^{\infty} nP(n)}{(1+1/q) \sum_{n=1}^{\infty} X_n P(n)} \qquad (3.5.)$$

For maximum compression, therefore, the standard run-lengths which made $\sum_{n=1}^{\infty} X_n P(n)$ a minimum must be chosen. X_n is determined by the values of the standard run-lengths. For example, if runs of length 1, 3 and 9 Nyquist intervals only were permitted, a run of length 8 intervals would be broken down into two runs of 3 and two of 1 Nyquist interval and X_8 would be 4. Calculations along these lines have been carried out by Pine(44) and Kubba(34). Unfortunately this work is of limited value for reasons which are discussed in Section 3.8.1.

In view of the shortage of run-length statistics, a programme concerned primarily with their measurement is now under way in the section. It is hoped that the equipment constructed will enable measurements to be carried out quite rapidly so that a large amount of data can be collected. Direct measurement of the decrease in data reduction due to run-length restriction can also be carried out when run-length restriction equipment is available. Such measurements were included in the programme of research and are described in Section 3.8.

3.3 Possible Methods of run-length restriction

So far, no mention of the equipment necessary for run-length restriction has been made. In the previous section it is assumed that each run of previous length n requires $X_n - 1$ additional pulses inserted to break it down into standard lengths so that a total of X_n pulses results. Figure 3.2 shows an example of this process, the standard run-lengths being 1, 2, and 7.

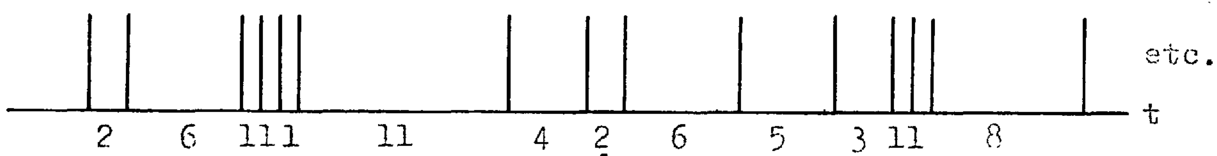
It is obviously wasteful to insert more pulses than are absolutely necessary, for this can result only in a loss of data reduction with the increased probability of overloading in a completely instrumented system. X_n is therefore the minimum number of pulses required to produce the standard run-lengths. Thus, one of the problems in devising equipment for run-length restriction is to ensure that the standardisation is done in the most economical way.

Clearly the ideal way to standardise a run is to ascertain the longest standard run-length which will fit into the unstandardised run. The remainder of the unstandardised run is then examined and the process is repeated. If the longest of the standard run-lengths will not fit into the unstandardised run then the next longest is tried and so on.

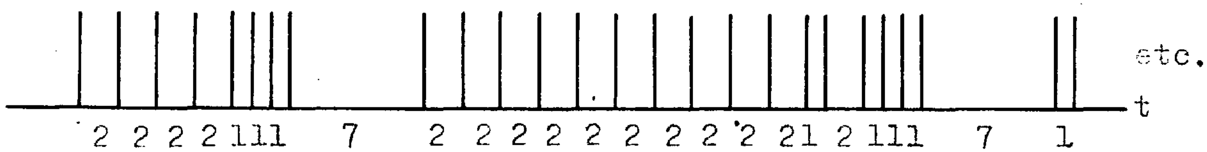
3.3.1 Synchronised oscillators method

The first description of the Open Loop System required three synchronised oscillators to generate the sampling pulses at the three standard rates. A detail detector recognising high, medium, and low levels of detail was envisaged, each of the three rates corresponding directly to a detail level. The detail detector could thus be coupled directly to a switching unit between the three oscillators and the Elastic Encoder as indicated in Figure 3.3.

In this diagram no gates have been indicated in the switching unit; instead the conventional sign for a switch has been used. This is intended to emphasise the fact that two simple AND gates are not sufficient. The combination of logical elements necessary



Run-lengths before standardisation



Standardised run-lengths

Figure 3.2. Standardisation of run-lengths to 1, 2 and 7 Nyquist intervals.

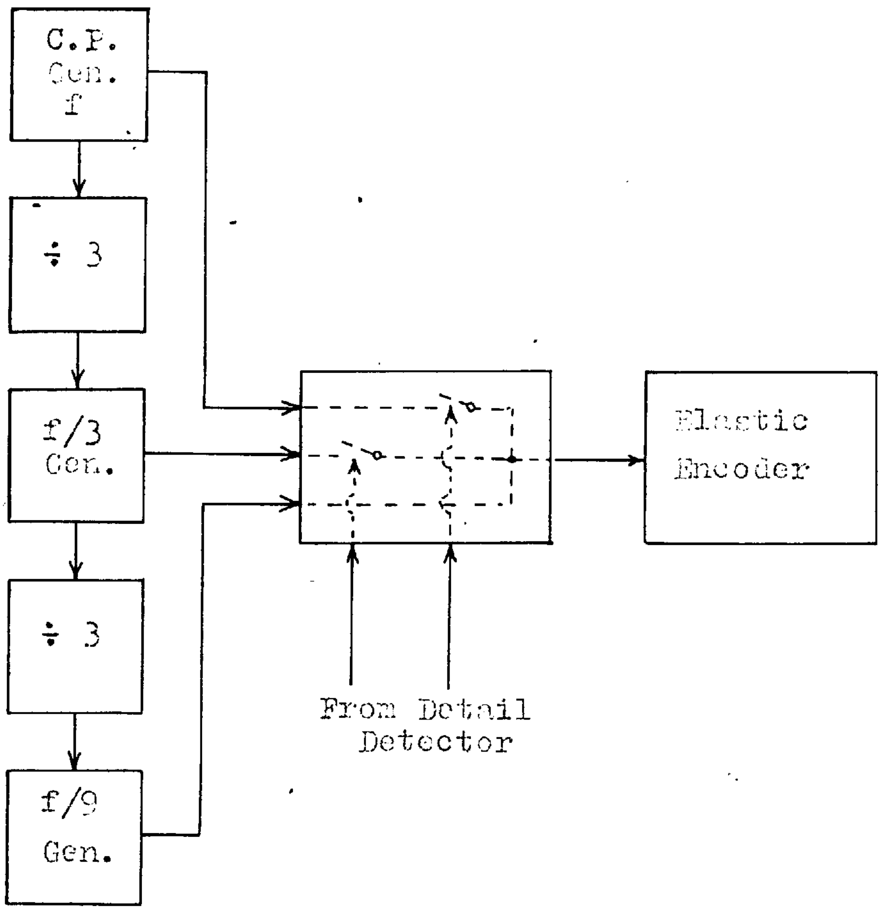


Figure 3.3. Sampling pulse generation by synchronised oscillators

is, in fact, quite complex. The following requirements must be satisfied:-

(i) A change from sampling rate f/m_1 to f/m_2 or vice versa ($m_2 > m_1$) can be made only when there is a pulse present in the train at rate f/m_2 (see Figure 3.4).

(ii) Even the narrowest pulse from the Detail Detector must cause sampling at the full rate to occur, otherwise definition will be lost. This condition may be relaxed if the pulse from the Detail Detector coincides with a pulse at one of the lower rates. (See Figure 3.5.)

A suitable system of logical elements can now be arrived at. One possible scheme is shown in Figure 3.6. To understand the operation of this scheme, suppose first that a region of low detail is being scanned so that neither input to the switching unit is energised. BISTABLES 1 and 2 are both 'off' so that AND 3 and AND 4 are closed. Only f/9 pulses therefore pass through OR 3 to the Elastic Encoder. If now the detail content of the signal increases, the f/3 connection from the detail detector may be energised. Even if the region of increased detail be only short, the pulse stretcher consisting of sections of delay and OR 2 will ensure that there is a signal at AND 2 when the next f/9 pulse arrives. BISTABLE 2 will thus be switched on and f/3 pulses will pass to the encoder via AND 4 and OR 3. By a similar process, BISTABLE 1 can be switched on only when an f/3 pulse allows AND 1 to open.

After a region of detail, one or both of the inputs to the switching unit will become de-energised. The corresponding INHIBIT gate(s) will thus start to pass pulses after a time delay due to the pulse stretchers. BISTABLE 1 will therefore be turned off by the first f/3 pulse after INH 1 opens and BISTABLE 2 by the first f/9 pulse after INH 2 opens.

In this scheme, therefore, requirement (i) has been satisfied by the combination of AND and INHIBIT gates 1 and 2 and BISTABLES 1 and 2, while requirement (ii) has been satisfied by the pulse stretchers consisting of delay sections and OR gates 1 and 2.

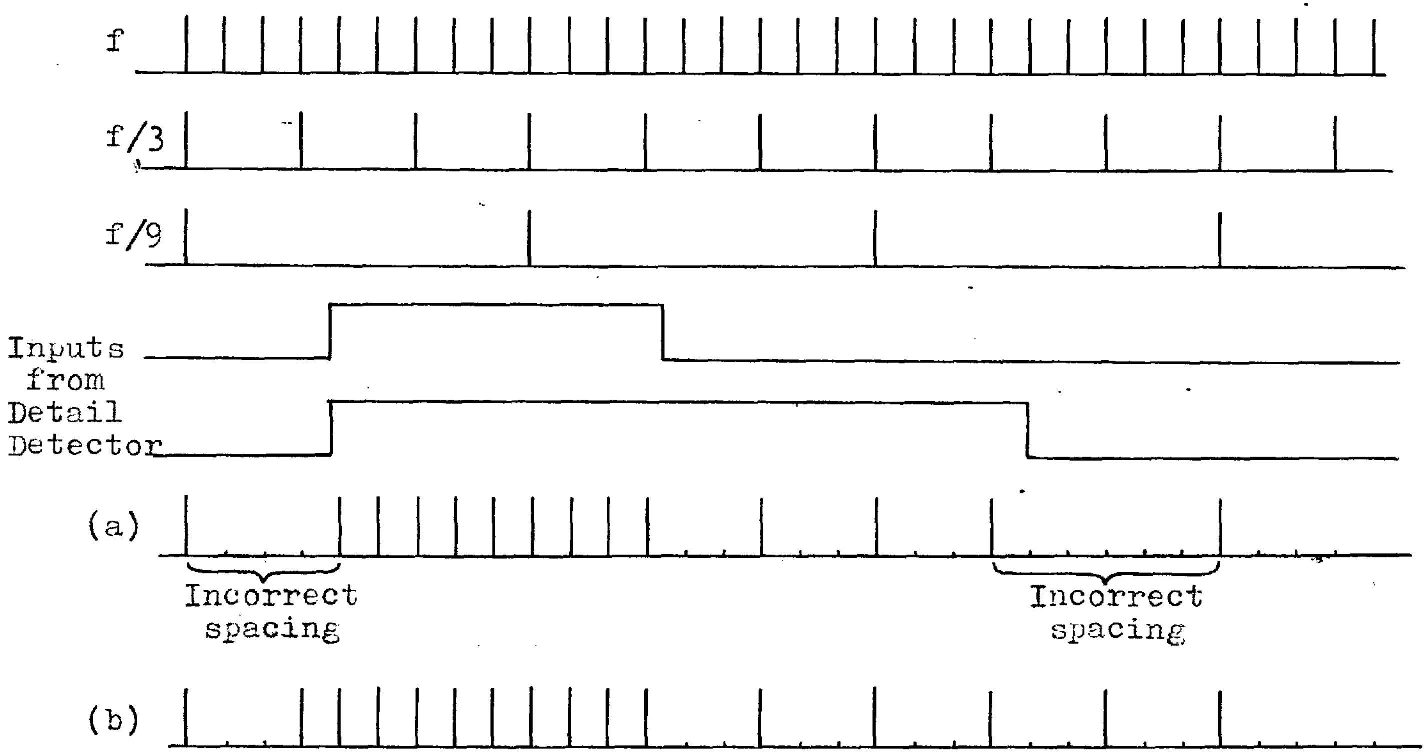


Figure 3.4. (a) Result of condition (i) not being fulfilled.
(b) Result of condition (i) being fulfilled.

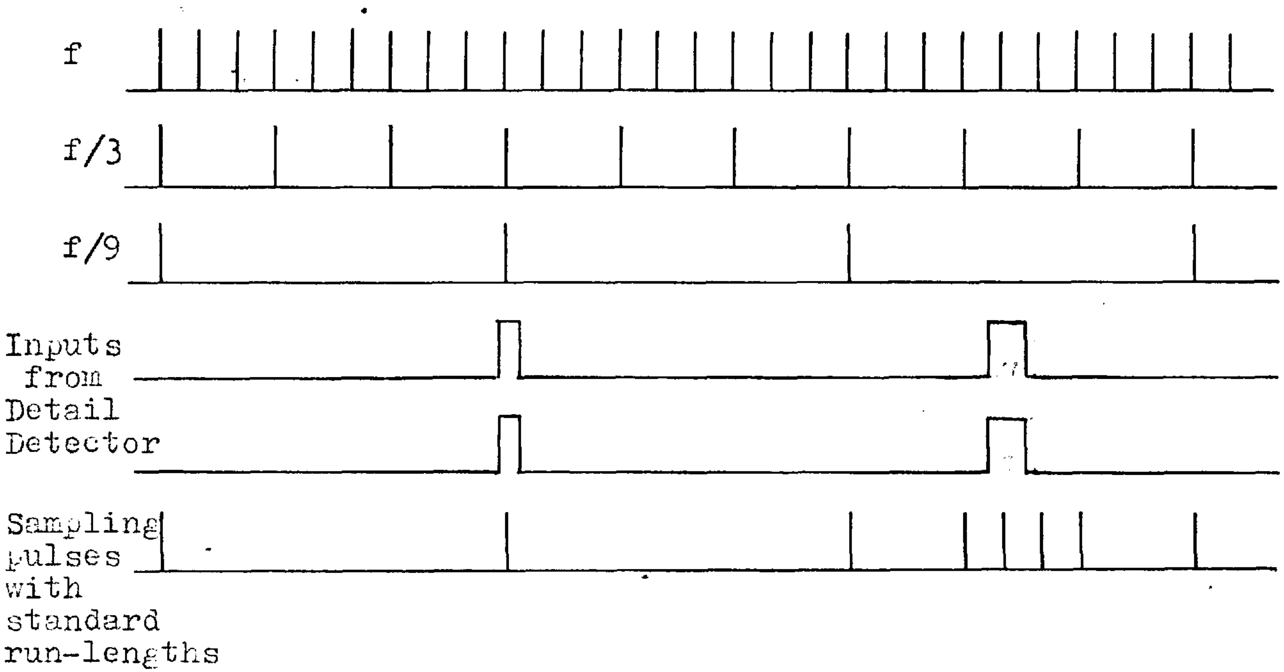


Figure 3.5. Run-length restriction when a pulse from the Detail Detector co-incides with a pulse at one of the lower sampling rates.

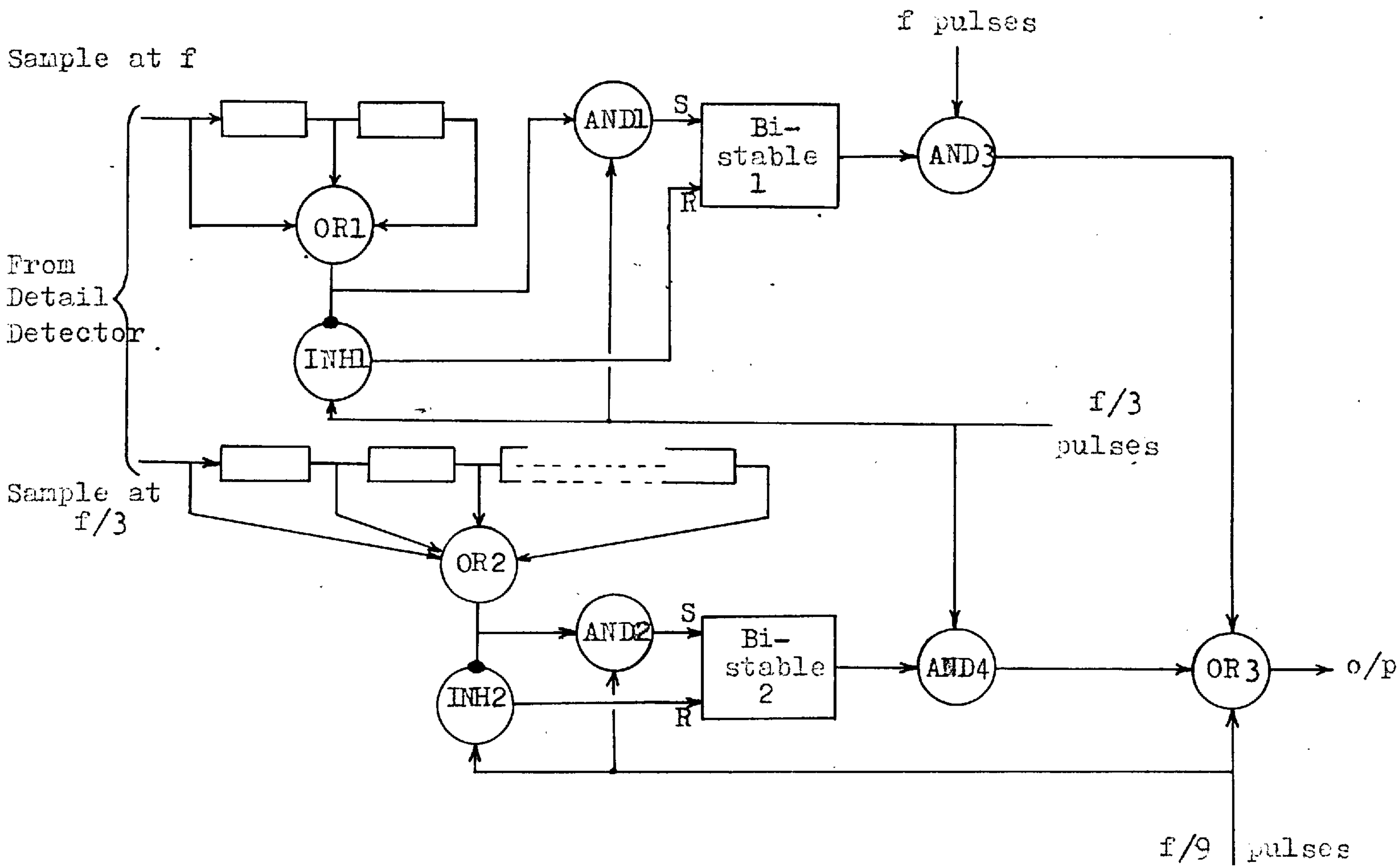


Figure 3.6. Run-length restrictor fulfilling requirements but inefficient.

3.3.1.1 Inefficiency of this Method

Whatever the means adopted for run-length restriction, a change from one sampling rate to another will almost inevitably result in extra pulses being required in order to preserve the correct standard run-lengths at the output. Some examples of this are shown in Figure 3.7, where only two sampling rates, f and f/6 are shown for simplicity. Part (a) of this figure shows what would happen if no precautions to preserve correct sample spacings were employed. In (b), logic has been employed so that sampling at f starts at the end of the last complete 6T interval before the detail commences. In fact this is impossible; it is actually necessary to defer the start of sampling at f until the end of the next 6T interval, the video supplied to the Elastic Encoder also being delayed so that sampling occurs at the correct points. (c) in Figure 3.7 shows the special case where a narrow pulse of detail coincides with a pulse at the lower sampling rate so that no change of sampling rate and no additional pulses are necessary.

The additional inefficiency incurred by the use of synchronised oscillators for run-length restriction is manifested at the ends of regions of detail. Unless the end of such a region coincides with a pulse at the lower sampling rate (Figure 3.8(a)) it is not possible to change to the lower rate and extra pulses have to be inserted at the higher rate until a pulse at the lower rate arrives (Figure 3.8(b)).

Clearly it would be advantageous if the oscillator generating pulses at the lower rate could start afresh at the end of each region of detail, the phase of the pulses in the train no longer being determined by the phase of such pulses before the region of detail. Two methods of achieving this aim suggest themselves:

3.3.2 Gated Counters

Instead of having the f/3 and f/9 generators running continuously as in Figure 3.3 they could be started and stopped as required

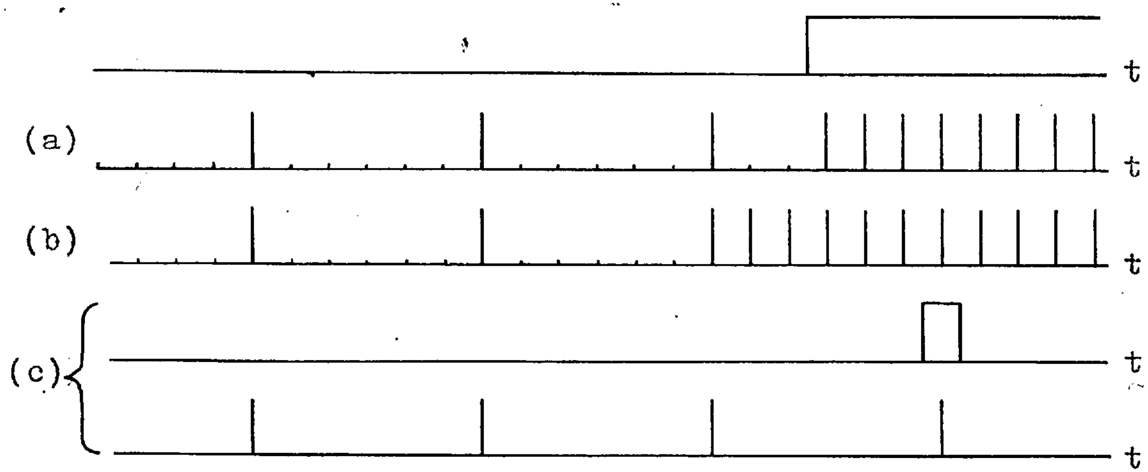


Figure 3.7. (a) Run-length restriction with no precautions to preserve correct sample spacings
(b) With precautions
(c) A special case requiring no sampling rate change

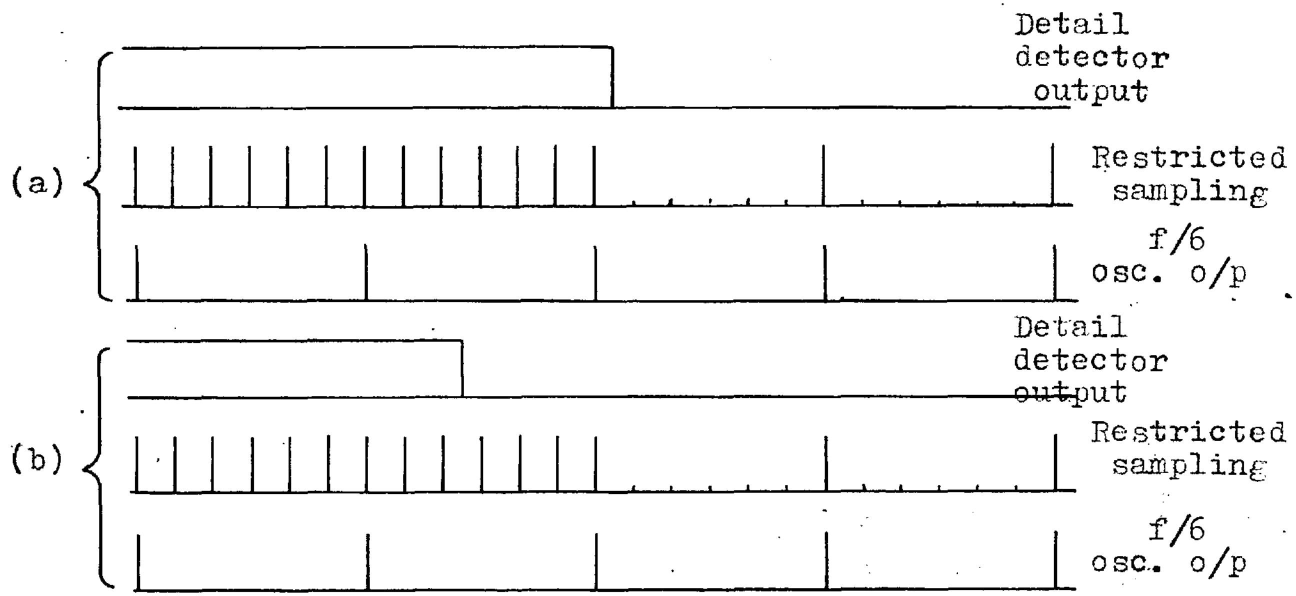


Figure 3.8. (a) Sampling when the end of a detailed region co-incides with a pulse at the lower rate
(b) Necessity for extra pulses when the condition shown in (a) does not obtain.

(provided that a suitable circuit is employed). The simplified block diagram of the run-length restriction equipment would then appear as in Figure 3.9. Each of the generators of pulses at a submultiple (f/m) of the Nyquist rate, f, comprises a counter which provides one pulse at its output for every m at the input. A gate controls the supply of pulses to the input of each counter. As in the case of the continuously running oscillators, it is necessary to ensure that a change from a lower to a higher sampling rate takes place only when a pulse at the lower rate occurs. There is no longer such a restriction on changes in the reverse direction, however. One of the features of this scheme is that no counter resetting mechanism is necessary. Since every transfer from a lower to a higher sampling rate occurs at the end of a sampling interval at the lower rate, the counter will be in the correct state to provide the correct interval before giving out a pulse on return to the rate. An example of a change from f/3 to f and back again is shown in Figure 3.10.

The type of counter used would depend chiefly upon the speed of operation. At television speeds, binary dividers would be essential, suitable feedback loops being employed to give the required division ratios. At lower speeds, counting tubes, such as the dekatron, or Miller rundown circuits could be employed[38].

3.3.3 The Blanking Method

An alternative way of generating the non-uniform pulse train involves blanking out certain pulses from a train of 'clock' pulses at the Nyquist rate (f). For example, if one of the desired sampling rates is f/a, then a-1 pulses out of every a are blanked out. The arrangements necessary for this process are indicated in Figure 3.11. Clearly, the number of blanking pulse generators required will be equal to the number of standard run-lengths to be permitted, with the exception that no generator is required for runs of length T (one Nyquist interval) since all the clock pulses can be allowed to pass in this case. The control logic has two (main) functions:-

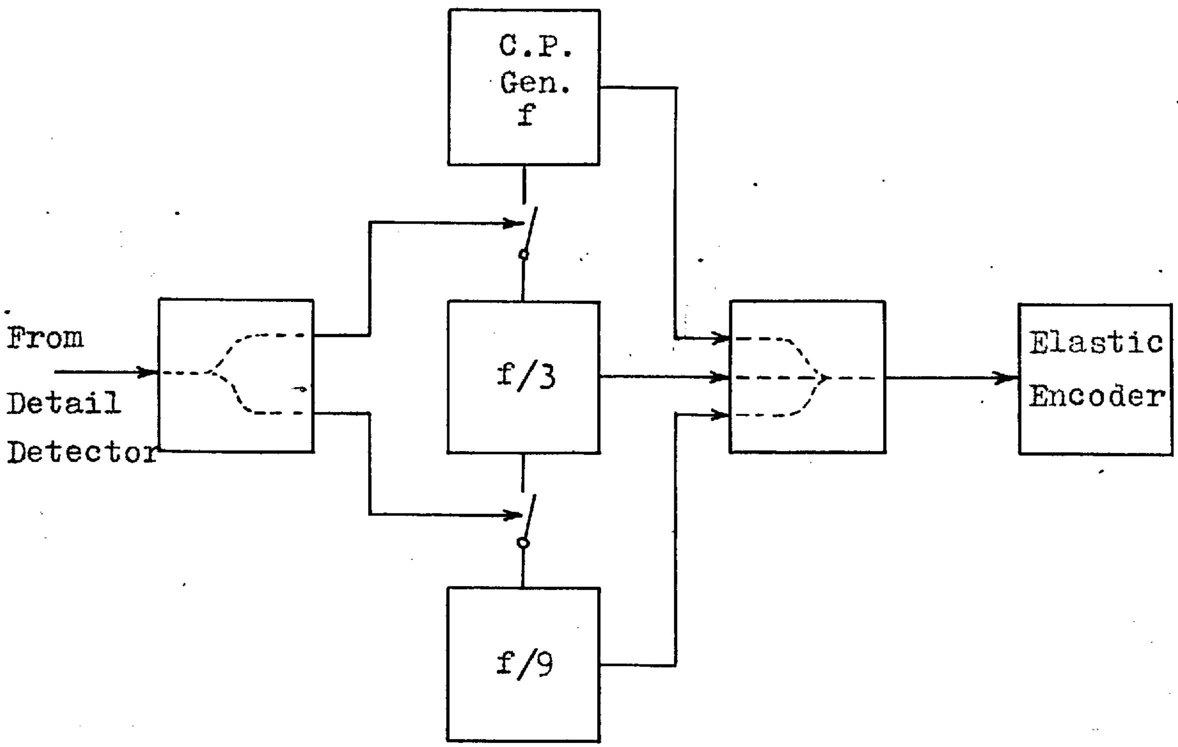


Figure 3.9. The gated counters scheme of run-length restriction.

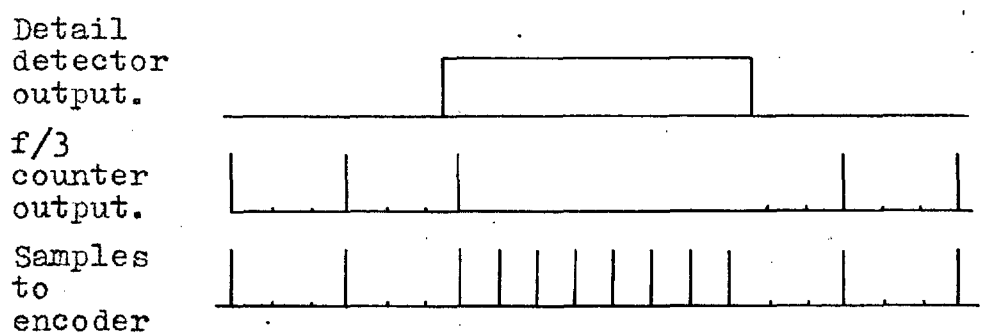


Figure 3.10. Signals associated with the gated counters method of run-length restriction

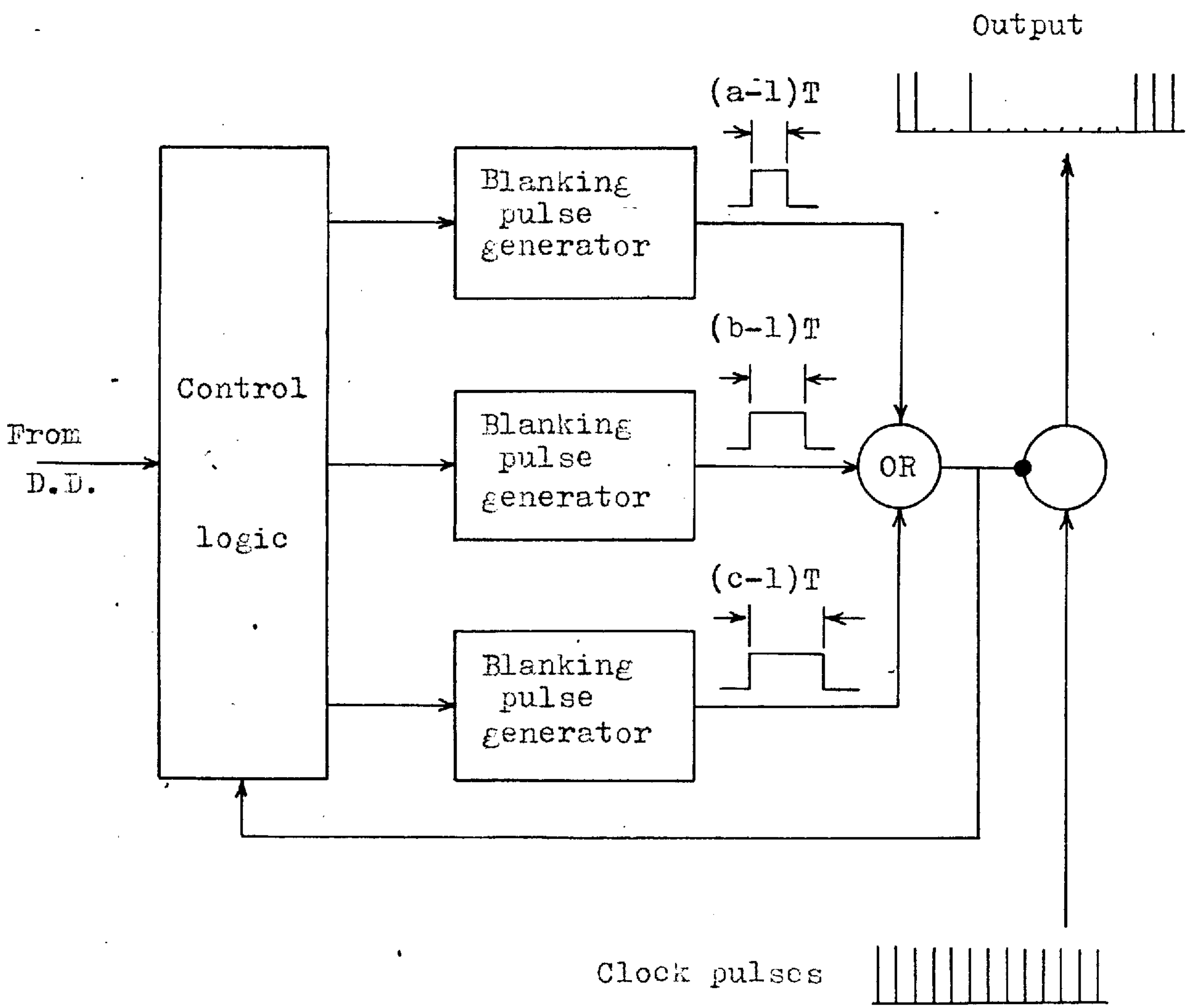


Figure 3.11. Equipment for run-length restriction by the blanking method.

(a) to control the blanking pulse generators in accordance with instructions from the Detail Detector so that runs are broken into standard lengths in the most economical way.

(b) to ensure that no more than one blanking pulse generator is on at any one time. The two requirements of Section 3.3.1 still apply, but in modified form. i.e. (i) No non-standard run-lengths shall appear in the output train, and (ii) sampling at the full rate must be carried out even when the pulse from the Detail Detector is at its narrowest (1 Nyquist interval).

Suitable equipment to achieve these objects is shown in block diagram in Figure 3.12. Here the three sampling rates f, f/3 and f/9 are assumed. This equipment, having been designed recently, is intended to be compatible with a Detail Detector having only one output connection, this being the form of Detail Detector envisaged in the more recent patent proposals on the Open Loop system.[12, 40]

When a threshold in the Detail Detector is exceeded (1) sampling pulses are required to be inserted at the Nyquist rate. If the threshold is not exceeded (0) no sampling is required. Correspondingly, the Run-Length Restrictor must provide sampling at the Nyquist rate when this is demanded by the Detail Detector, relaxing to the lowest rate (for reasons of economy) when sampling is not demanded. The intermediate rate is provided so that runs which do not comprise an exact number of intervals at the lowest rate may be broken down economically. Thus, for example, a run of original length 16 Nyquist intervals would require three extra pulses inserted to break it down into standard runs of 9T, 3T, 3T and T. Had no intermediate rate been available, 7 extra pulses would have been required (Figure 3.13). Sampling at the intermediate rate therefore no longer indicates an intermediate level of detail. The Sample Gating Unit could, however, be used with a Detail Detector of the earlier proposed type; different logic would be necessary in the Translator section of the equipment.

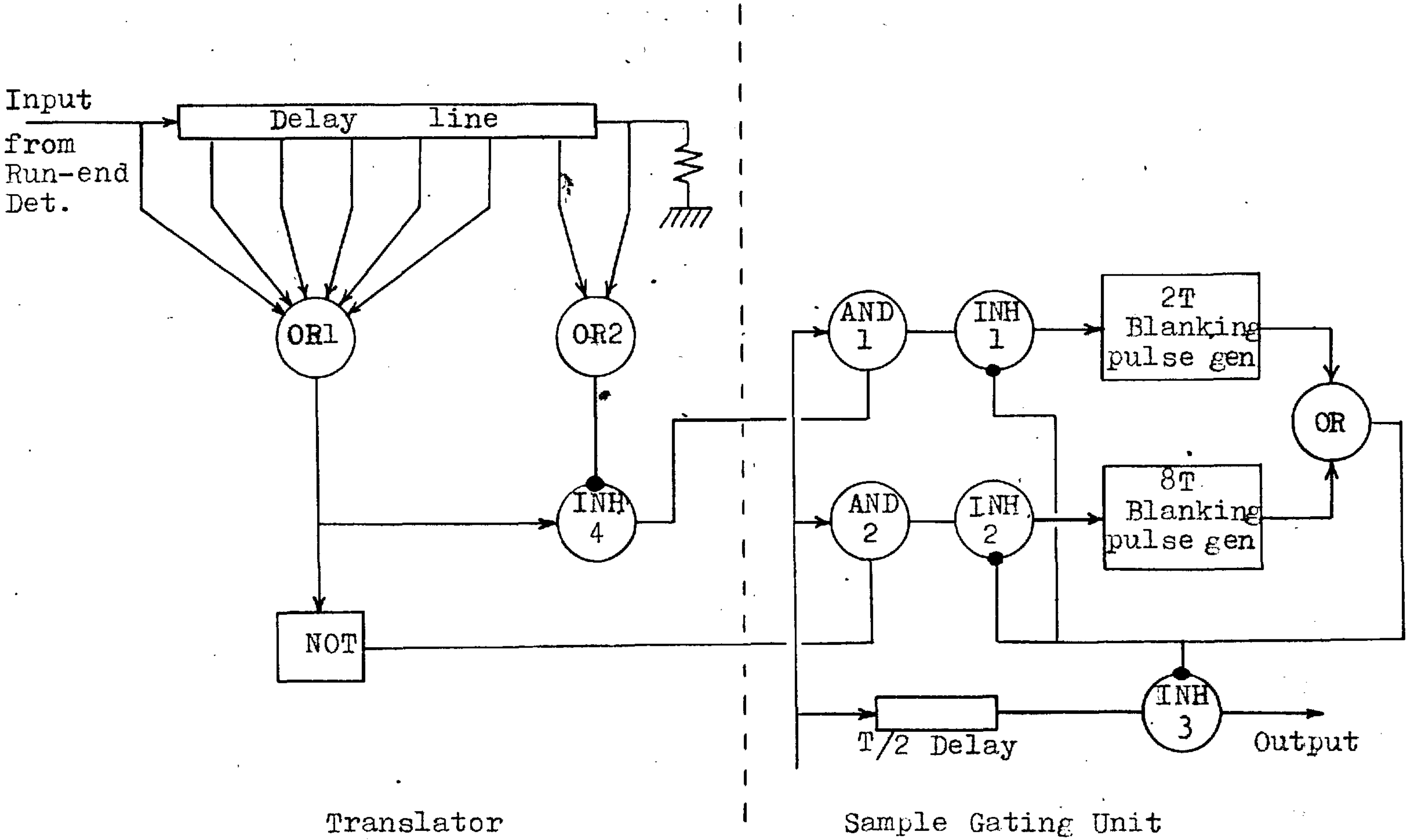


Figure 3.12. More detailed block diagram of the equipment for run-length restriction by the blanking method.

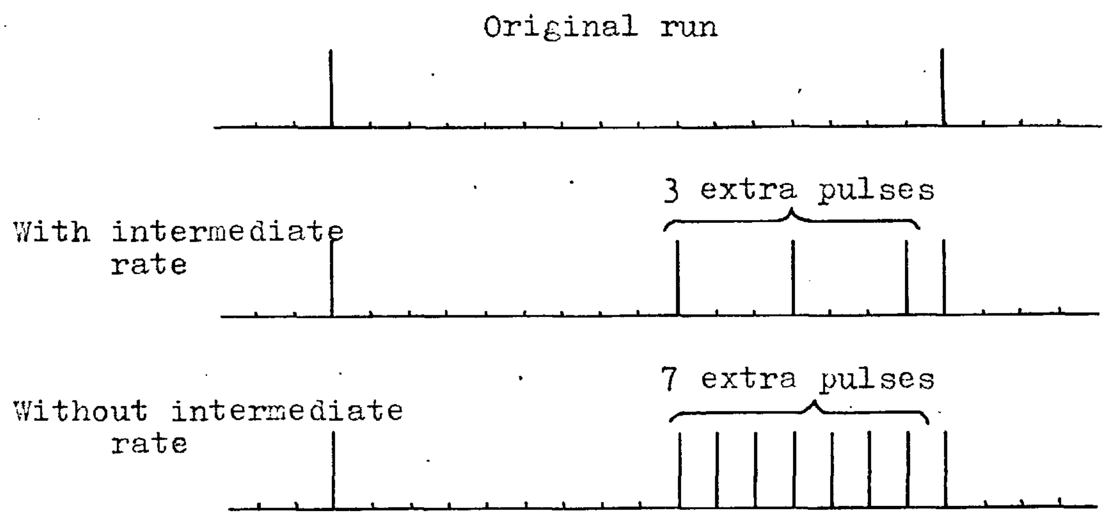


Figure 3.13. Use of the intermediate sampling rate to achieve extra economy in sampling

3.4 Operation of the Translator and Sample Gating Unit

The section of the equipment concerned with blanking out the undesired pulses from a continuous uniform train has been called the "Sample Gating Unit" (SGU). This unit is preceded by the Translator which receives one signal from the Run-end Detector* and converts this into two signals suitable for controlling the blanking pulse generators of the SGU.

The operation of the two units is best understood by referring to Figure 3.12, together with the waveform diagrams in Figure 3.14. Suppose firstly that a region of low detail is being scanned. As long as the threshold of the Run-end Detector is not exceeded, its output will be at the '0' logic level. Assuming that this state of affairs has existed for several picture elements, all the taps of the delay line in the Translator will be at the '0' level. Thus a '0' will be fed to the NOT element and a '1' will be fed from this element to open the gate AND 2 in the SGU. This will allow clock pulses to pass, via an INHIBIT gate, the the 8T blanking pulse generator so that a train of 8T pulses is formed. The application of these pulses to another INHIBIT gate (INHIBIT 3) then results in the blanking out of eight clock pulses from every nine, the correct phasing of the clock pulses being achieved by the use of a T/2 delay.

If now the scanning spot reaches an area of higher detail, the Run-end Detector threshold may be exceeded so that a 1 is fed to the Translator. OR 1 will thus give an output at the 1 level and as the output of OR 2 is still at the 0 level a 1 will pass

---

* The term "Detail Detector" has been used so far in this chapter since it appears in the patents, etc. The name "Run-end Detector" is better, however, and will be used henceforward.

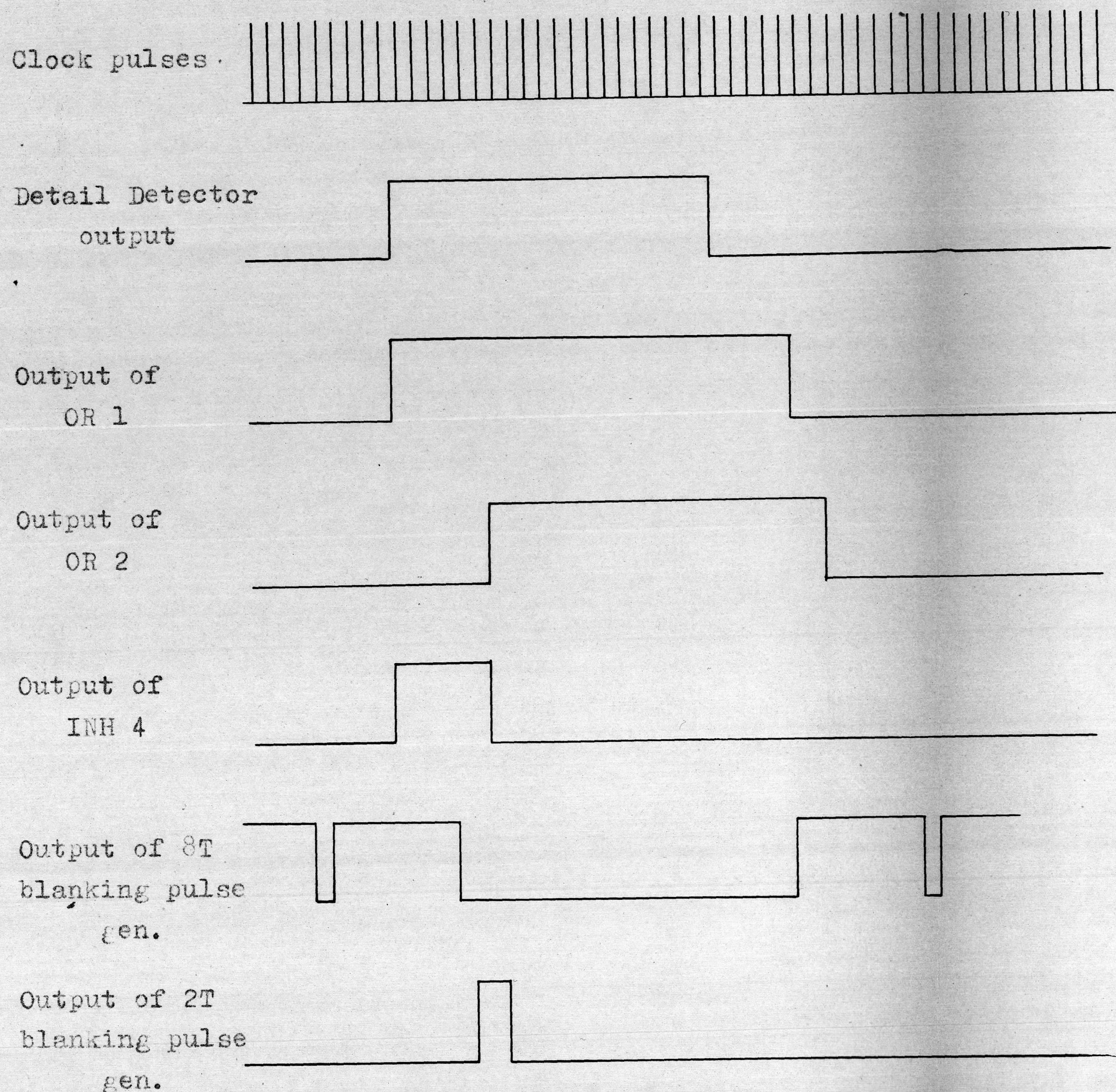


Figure 3.14. Theoretical waveforms explaining the operation of the Translator ahd SGU. Gate numbering refers to Fig.3.12.

through INH 4 to AND 1 of the SGU. Simultaneously AND 2 will close owing to the output of the NOT element going to the '0' level. With AND 2 closed, no more clock pulses will go via INH 2 to the 8T blanking pulse generator so the supply of pulses from this element will cease. Up to this point no clock pulses will have reached the 2T blanking pulse generator since INH 1 will have been held closed by the last 8T pulse. The end of the 8T pulses allows the 2T generator to start, however, so that every third clock pulse is now allowed to pass by INH 3.

When the '1' level in the delay line of the Translator reaches the taps connected to OR 2, the '1' proceeding to AND 1 is inhibited by INHIBIT 4 so that the 2T generator ceases to function and all the clock pulses are able to pass through INH 3. It will be realised that the delay due to the Translator must be matched by an equivalent delay in the video path so that sampling at the full rate is carried out at the correct points of the signal. This delay, together with the delay necessary to compensate for storage in the Run-end Detector is shown in the block diagram of the complete system (Figure 1.4).

At the end of the region of high detail the equipment will revert to its original condition and the 8T blanking pulse generator will restart provided that the low-detail region extends over at least eight picture elements. Should the region be shorter than this, however, one or more of the taps at the beginning of the Translator delay line will be at the '1' level immediately that the last '1' of the previous high detail region has left the delay line. The 2T blanking pulse generator will therefore be energised, as described earlier, and one or two 2T pulses will be generated before INH 4 once again operates to allow sampling of the detailed region at the full rate. Thus regions of low detail are broken down to standard run-lengths in the most economical way.

3.5 Methods of Generating the Blanking Pulses

The blanking pulse generators form the heart of the Sample Gating Unit. Several types of circuit are possible and considerable care was taken to choose the most suitable.

Each blanking pulse generator is required to deliver a train of pulses of constant width for as long as trigger pulses are supplied to it. The intervals between successive pulses must be equal to the Nyquist interval. Thus, in the case of the standard run-lengths 1, 3 and 9 Nyquist intervals, one of the blanking pulse generators must be capable of operating with a mark-space ratio of 8:1. Further, the rise- and fall-times of the pulses must be considerably less than the Nyquist Interval. This can be seen by considering Figure 3.15. Owing to the finite rise- and fall-times of the pulses to be gated, not all of the Nyquist interval is available for the blanking pulse transition. τ could well be as little as 0.6 T. Additionally it is advisable to have a transition time which is small compared with τ so that there is latitude for error in timing and drift. The ease with which the equipment could be altered to other standard run-lengths was also borne in mind, although this consideration is important only in an experimental system.

Circuits considered for this purpose include:

(a) The blocking oscillator.

While this circuit is very simple it is difficult, if not impossible, to obtain transitions which are sufficiently short compared with the pulse width. In addition, adjustment and variation of the pulse width are not simple, involving alterations to a pulse transformer.

(b) Miller rundown circuits.

These would probably be very suitable at low speeds. The equipment to be constructed was required to operate at television

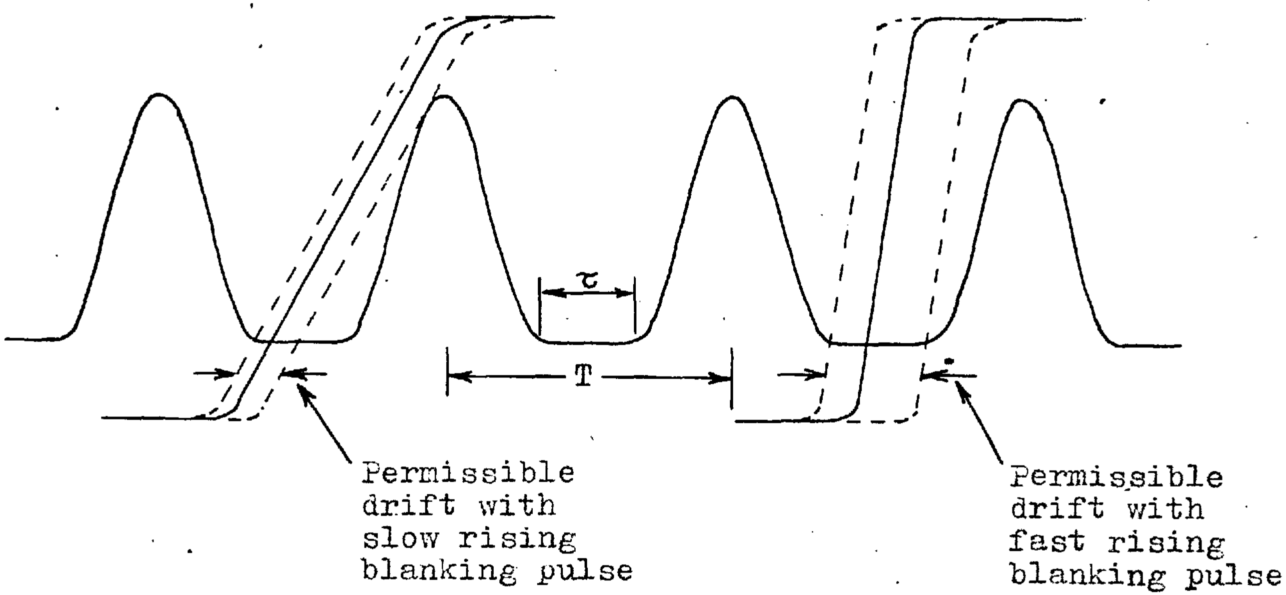


Figure 3.15. Illustrating the importance of fast rise-times of blanking pulses.

speeds, however, the Nyquist interval being 1/6 µS for television on the British 405 line system.

(c) Circuits using delay lines as timing elements.

Electromagnetic delay lines can be made to have good temperature stability but are somewhat bulky. Fine adjustment of their length (and thus of the pulse width) is not easy to engineer.

(d) Counting circuits.

A binary chain would have been able to fulfil the requirements but would have been very expensive, 8 valves or transistors being necessary in the case of the 8T pulse generator. Counting tubes, such as the trochotron are also expensive and require special power supplies in addition to the provision of a magnetic field. Such a tube would probably meet the requirements. Simpler tubes, such as the dekatron have a maximum counting rate of about 5 kc/s and are clearly unsuitable for television work.

Storage counters are reliable at low division ratios but require linearisation (probably using a bootstrap circuit) for ratios greater than about 4:1. Careful control of the input pulse-shape is also necessary. In the majority of cases, a counter merely provides timing information and requires to be followed by some type of pulse generator for this application.

(e) Monostable multivibrator circuits.

The monostable multivibrator (MMV) offers a simple means of generating pulses, adjustment of pulse duration being very simply accomplished by varying a capacitor or a resistor value. This is not wholly an advantage, however, as the pulse width is thus temperature dependent. Also, the recovery time of the circuit can be comparable with the pulse duration unless special precautions are taken.

Of all the means considered for generation of blanking pulses, the MMV was considered to show most promise. This circuit was therefore further examined, the factors affecting recovery time being of particular interest.

3.5.1 Design Considerations for a Monostable Multivibrator

Although other types of MMV can be conceived, the cross-coupled and emitter-coupled versions are usually considered*. In this case it was expected that the emitter-coupled type would satisfy the requirements and that nothing would be gained by the use of a less orthodox circuit. The chief advantage of the emitter-coupled configuration over the cross-coupled is that in the former the output pulse can be taken from the collector of one transistor with no loading effect on the rest of the circuit. The basic circuit of this type of MMV is shown in Figure 3.16.

Whatever the type of blanking pulse generator, it is required to produce a pulse whose length does not vary whether additional trigger pulses (after the pulse which turns it on) are supplied or not. This is easily accomplished here by triggering the collector of T_1 via a diode (D_1). As the collector voltage of T_1 remains low during the pulse period, D_1 remains "off" and no further trigger pulses are passed to the MMV.

Operation of both transistors between the "off" and "saturation" conditions was decided upon as this gives most immunity against pulse width variation due to changes in transistor parameters, R_1, R_2, and the supply voltage. The pulse width is thus determined almost entirely by the values of R and C.

For sampling at f/9 it is necessary to generate pulses of duration 8T separated by intervals of T. The pulse length

---

* A transistorised circuit has been considered here since the final equipment used these devices; the analysis is quite similar for a valved circuit, however.

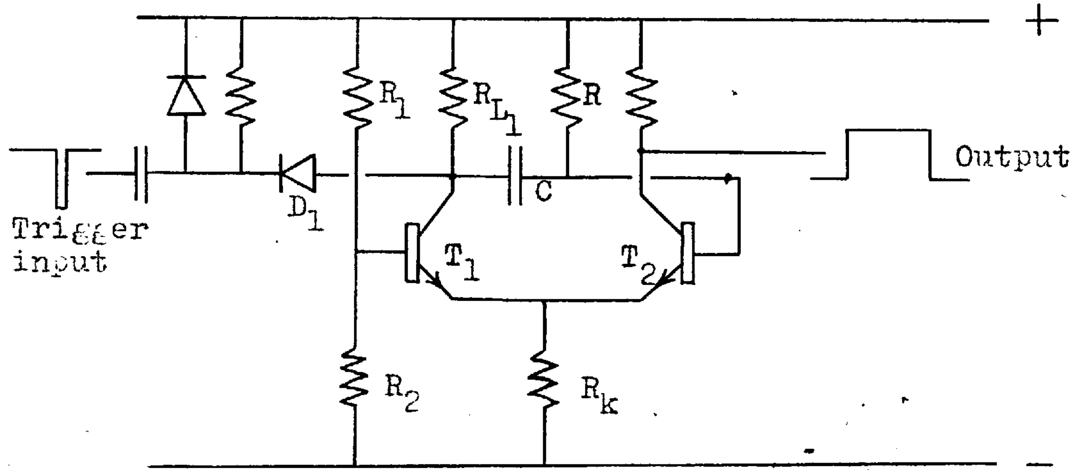


Figure 3.16. Basic circuit of the emitter-coupled monostable multivibrator.

must be the same whether the pulse is the first of a group or not. This requirement is not easy to satisfy, for after each pulse time is required for conditions in the circuit to return to the quiescent state. This arises since the voltage across the capacitor C is not able to change instantaneously. During the pulse period, the voltage at the base of T_2 is rising exponentially as the capacitor charges through the timing resistor R. (See Figure 3.17.) The pulse period ends when the base of T_2 becomes slightly positive with respect to the emitter. T_2 then starts to conduct and the current flow rapidly transfers from T_1 to T_2. The voltage of T_1's collector thus rises, taking the base of T_2 with it and turning T_2 on even harder. Recovery involves the recharging of C until the base of T_2 reaches the quiescent value as determined by R and R_K. If recharging is not complete at the time of the next trigger pulse, the output pulse produced is shorter than it should be since the base of T_2 is not driven sufficiently far negative. The current required for recharging flows through R_{L_1} so that it is desirable to keep this resistor as small as possible if rapid recovery is required. C is also kept as small as possible, the limit being set by the largest value of R which will give consistent operation.

A common way of reducing recovery time is to introduce an emitter follower between the collector of T_1 and the capacitor C as shown in Figure 3.18. It was found necessary to adopt this technique here. The use of the emitter follower reduces the impedance seen by C from R_{L_1} to $\frac{R_{L_1}}{(\beta + 1)}$ in parallel with R_e,

i.e. $$\frac{R_e R_{L_1}}{(\beta + 1)(R_e + R_{L_1})}$$

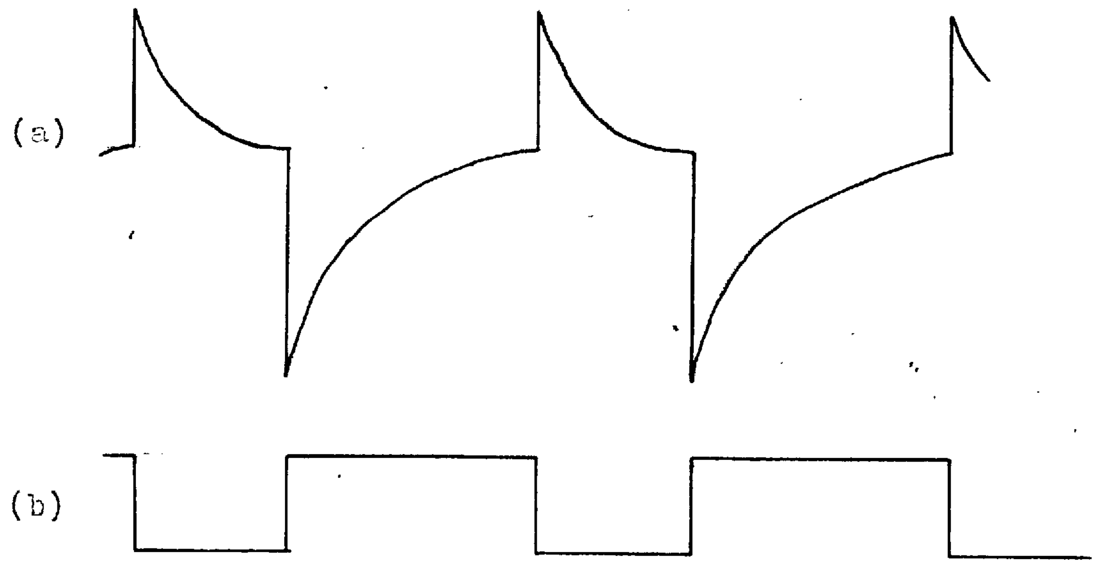


Figure 3.17. Waveforms at (a) base of T_2 , (b) collector of T_2 .

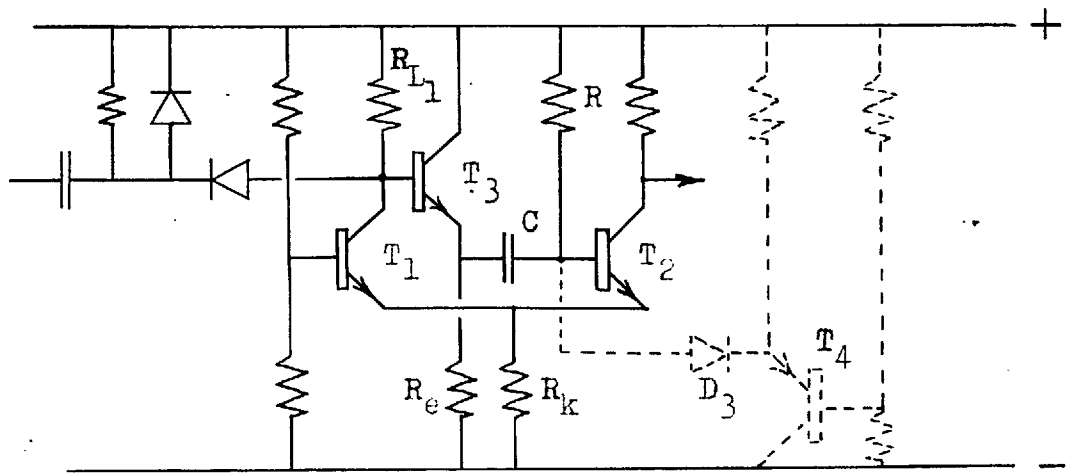


Figure 3.18. Outline circuit of the fast recovery monostable multivibrator

Although the recovery time was thus improved, it was still not short enough. Clearly, C also recharges through the impedance seen looking into the base of T_2. (R is so large that its contribution to the current can be neglected.) The addition of a low-impedance clamping circuit thus further improved matters. The necessary additions are shown by dashed lines in Figure 3.18. The diode D_3 is on only while the overshoot persists; it does not affect the timing during the pulse period. The necessary low impedance is provided by the emitter-follower T_4.

All these measures used together were found to be adequate, the recovery time of the circuit being reduced to about 100 nS for a pulse width of 1.5 µS.

It will be realised that the necessity for recovery limits the mark/space ratio achievable. The problem is thus most acute at the lowest sampling rate, that which has been considered here.

3.6 Description of the Equipment Constructed for Run-Length Restriction

3.6.1. Sample Gating Unit

Work on this section of the equipment was commenced before transistors of sufficient speed were readily available. A fast-recovery MMV was therefore designed and constructed using valves, a recovery time constant of about 11 nS being achieved for a pulse width of 1.34 µS (8T). Despite this apparently good performance it was realised that satisfactory operation as a blanking pulse generator was still not assured. The rise- and fall-times of the pulses were about 50 nS and the heat dissipated due to the use of high currents and small values of anode loads was quite considerable.

The full effect of a slight change in pulse width (possibly owing to poor temperature stability) may not be realised. Reference to Figure 3.19 will show, however, that a small change

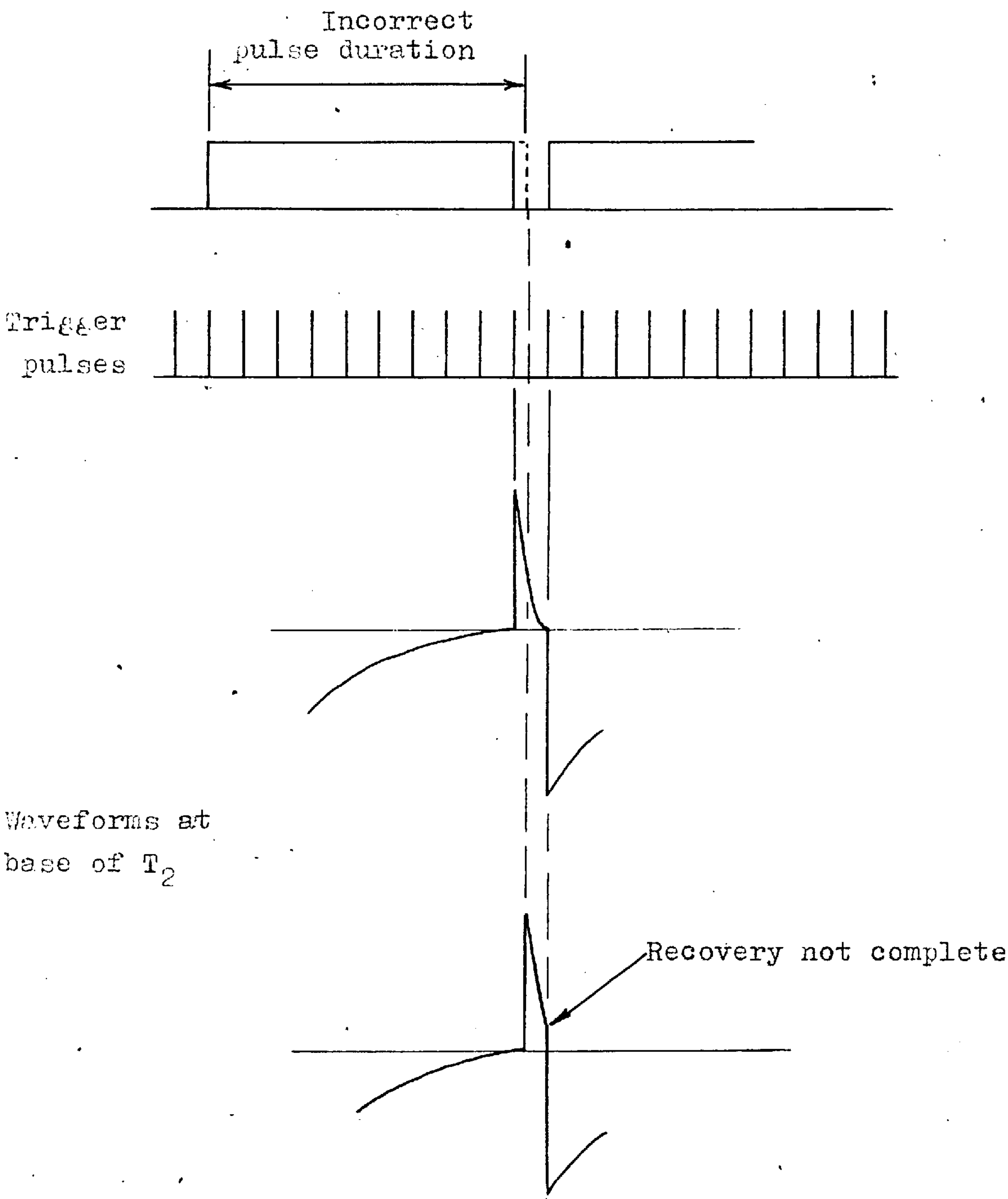


Figure 3.19. The effect of a small change in the duration of the pulse generated by the MMV

in pulse duration may produce a large proportional change in the time between the end of the pulse and the next trigger pulse so that recovery may not be complete. The width of the next blanking pulse will thus be incorrect, with the possibility of incomplete recovery once again. A complete SGU was, in fact, constructed using valves but operation was not good, frequent adjustments being required.

Later in the work, high speed switching transistors began to be readily available at economic prices. The use of these devices in a fast recovery monostable multivibrator was therefore examined. Promising results having been obtained from a single multivibrator, it was decided to go ahead with the construction of a completely transistorised SGU.

A block diagram showing the logic elements necessary in a SGU has previously been given (Figure 3.12). In practice, however, the arrangement of the unit is rather more as shown in Figure 3.20. Here, additional amplifiers, emitter followers, etc. are indicated. A complete description of the circuits is reserved for Appendix 1; some points require explanation here, however:

(i) As will be further explained in Section 4.5, it was considered desirable to gate the clock pulses in the 1st Order Interpolator after further delaying and gating operations had been carried out on the gating signal. The instrumented SGU therefore contains no means for delaying or gating the clock pulses as was previously shown in Figure 3.12.

(ii) The provision of a clock pulse shaper made all my equipment independent of variations in the size and shape of clock pulses on the main laboratory loop. This shaper also fed the 1st Order Interpolator and the Detail Function Shaper (see Section 3.6.3) via variable delay units.

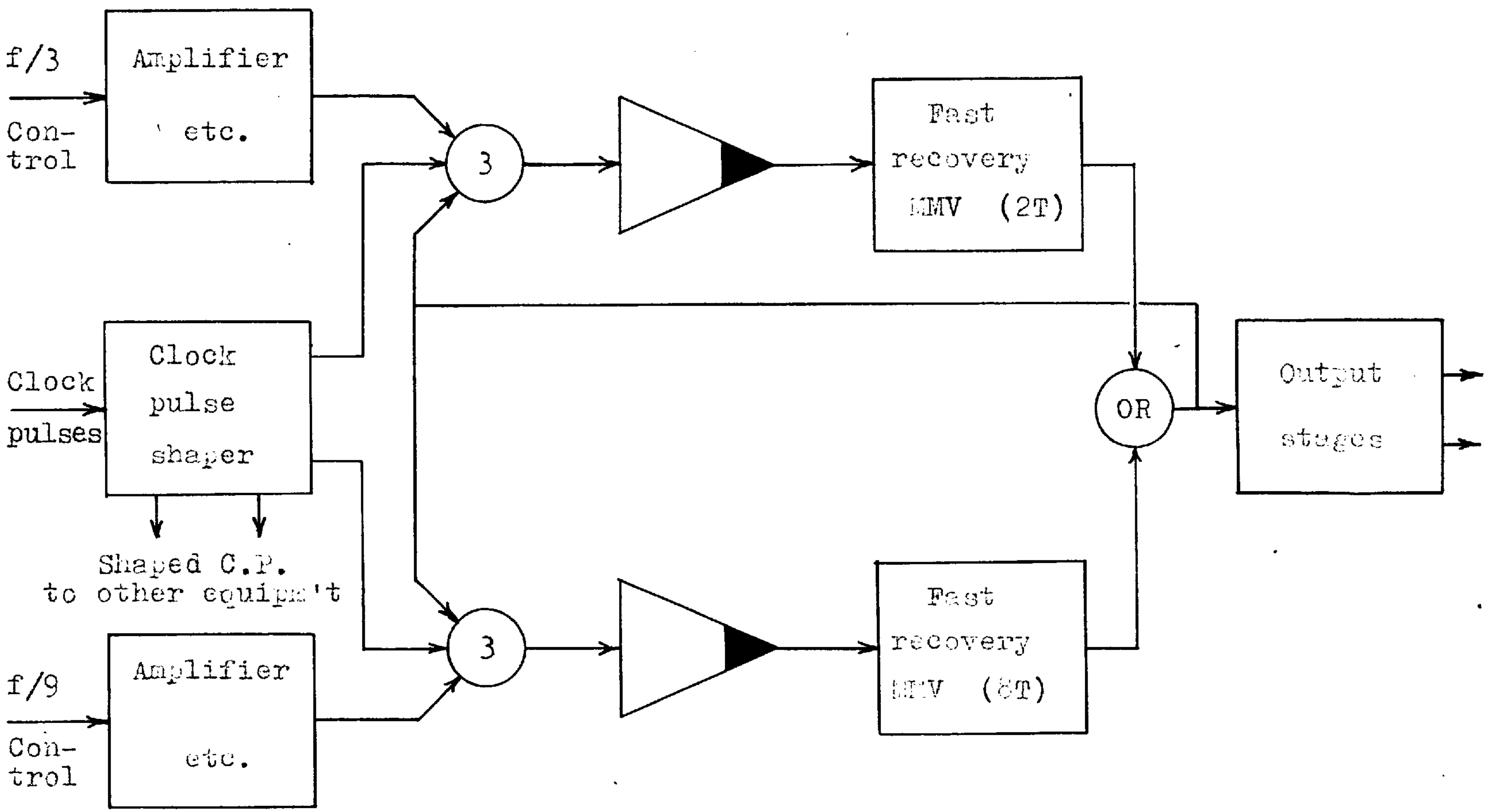


Figure 3.20. Detailed block diagram of the SGU.

(iii) As the Translator used valves, a.c. coupling to the SGU was thought to be desirable. The control signal inputs to the SGU thus required to present a high impedance in order to maintain good l.f. response with capacitors of reasonable size (about 1 μF). Amplifiers with high input impedance therefore preceded the 3-input AND gates.

The logical operation of the SGU has been explained already (Section 3.4); waveform photographs illustrating its performance are held over until the Translator and the Detail Function Shaper have been described.

3.6.2 The Translator

Reference to Figure 3.12 will show that the Translator comprises a number of delay elements, each of 1/6 μS (for the British 405 line television system), 3 gates and an inverter.

Methods of delaying signals at television speeds have been reviewed by Pine[44], although this review did not include the shift register, since suitable transistors were not readily available at the time; a valved shift register is almost unthinkable - it would be very bulky and would require considerable power. In the Translator the incoming signal has one of two levels only so that a suitable shift register would be quite simple and no analogue-to-digital conversion would be required. Even so, it was felt that delay cable was cheaper and simpler so this was employed.

As fast transitions are required, the high-frequency response of the delay system is all-important. In this respect, the most satisfactory cable described by Pine is the Telcon Type Z14M. It was expected, however, that in the time elapsing between Pine's work and my own, some improvements in techniques would have produced something superior. This indeed was the case. A sample of Hackethal Delay Cable, Type HH 1500 α was obtained for testing. This cable has an inner conductor wound helically round a core of magnetic material (see Section 2.7); its delay per foot is 79 nS. This compares favourably with the 66 nS per foot of the Telcon type Z14M. Frequency response functions for 1/6 μS lengths of each cable are plotted in Figure 3.21, the circuit used for the measurements being outlined in Figure 3.22.

The two frequency response curves for 1/6 μS lengths of cable will be seen to be quite similar, unavoidable reflections causing the curves to oscillate at the upper end of the frequency range. The oscillation commences at a slightly lower frequency with Z14M than with HH 1500α but does not become serious until a frequency of about 15Mc/s is reached. These remarks apply only for a 1/6 μS length, however. In the Translator a delay of 1 1/3 μS, (i.e. eight sections of 1/6 μS) is required. Fluctuations of the frequency response will be eight times as large as for 1/6 μS therefore and will become important at about 2.5Mc/s in the case of Z14M but not until 6Mc/s in the case of HH 1500α.

Because of the improved performance and also because of the higher characteristic impedance of HH 1500α (thus requiring less current for a given voltage swing) it was decided to use this type of cable in the equipment, signal pickoff at intermediate points in the cable being accomplished by circuits with the highest possible input impedance.

In this unit, unlike the SGU, valves were expected to be quite adequate. Further, suitable power supplies were available for this type of equipment so that it was constructed using valves. Figure 3.23 shows (a) simplified and (b) complete block diagrams of the unit, the large number of cathode followers necessary being clear from (b). As well as a high input impedance these circuits have low output impedance (about 80 ohm) in the case of an E88CC valve) and are thus ideal for driving the diode-resistor gates employed. Care should be taken when interpreting these diagrams (and also the complete circuit diagram (on Page 230). Owing to

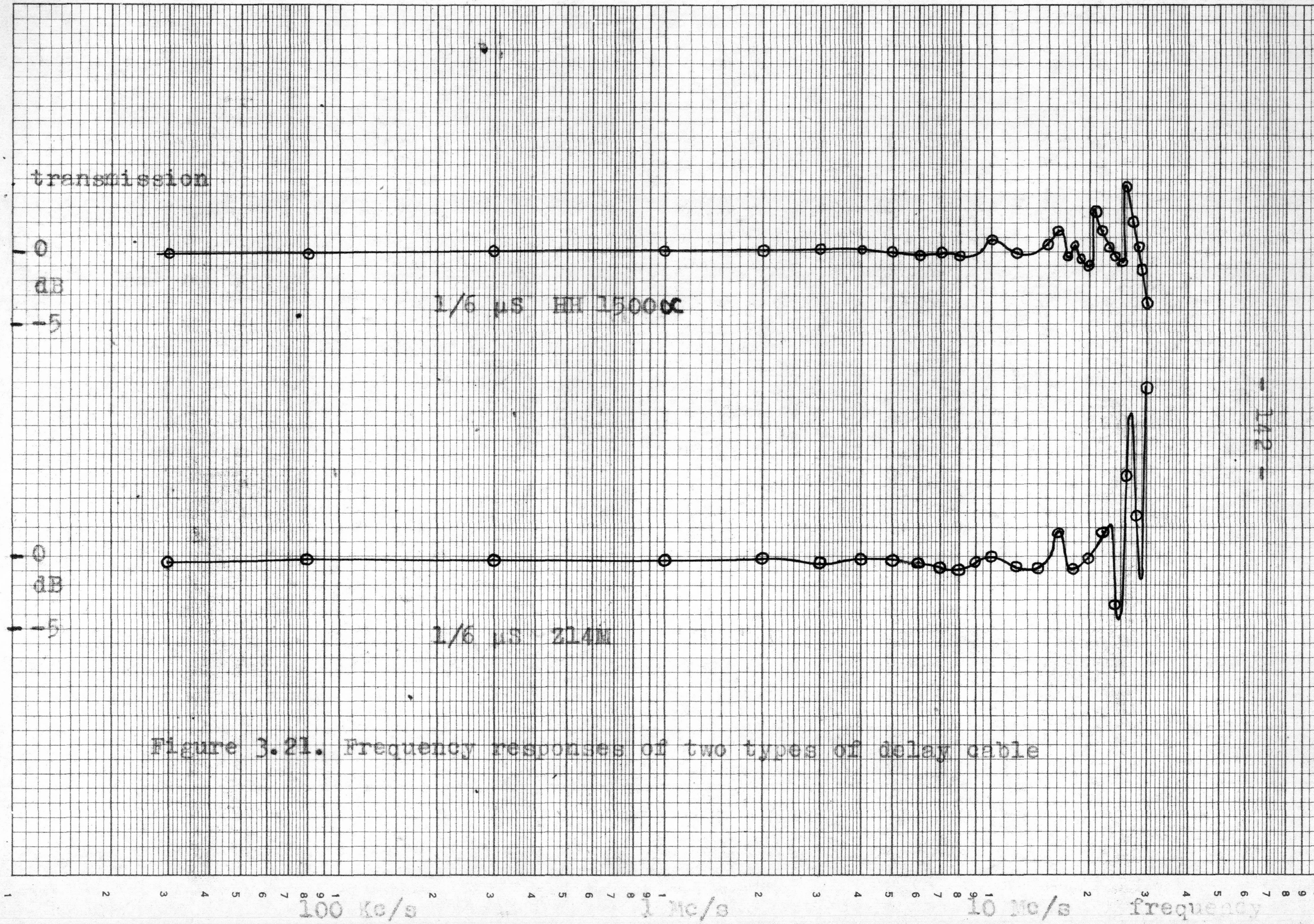


Figure 3.21. Frequency responses of two types of delay cable

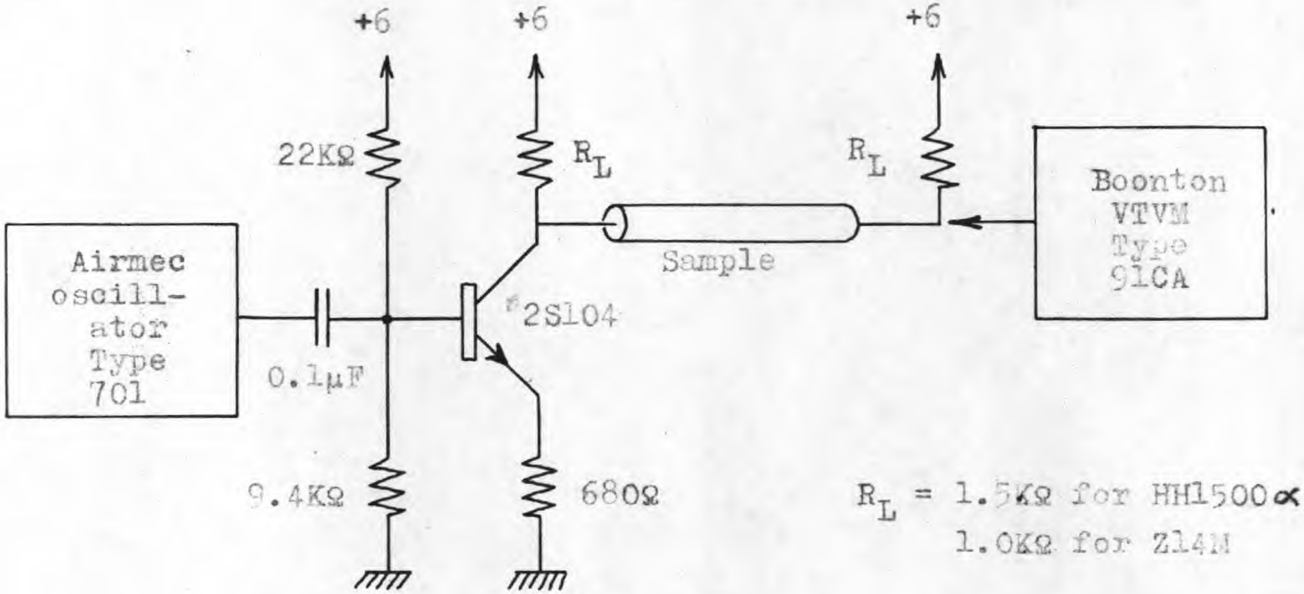


Figure 3.22. Circuit used for frequency response measurements on delay cable

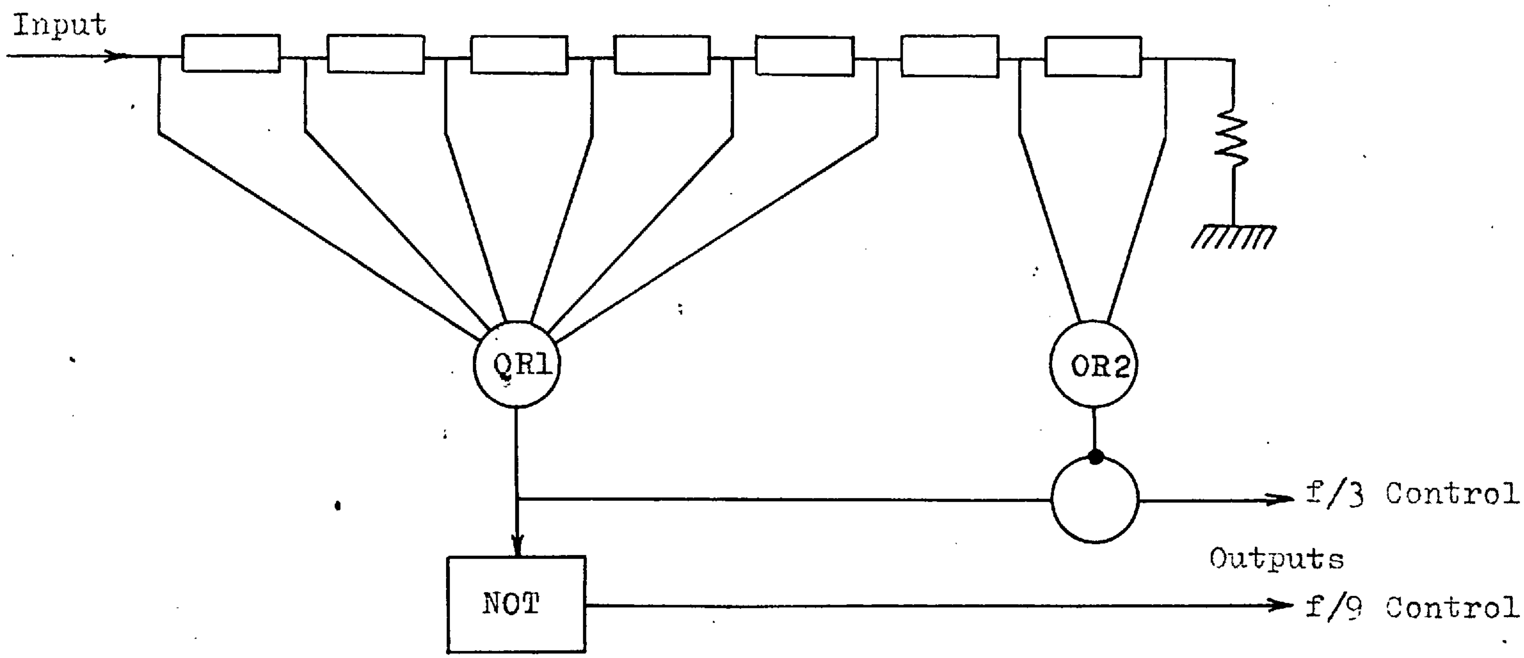


Figure 3.23(a). Simplified block diagram of the Translator

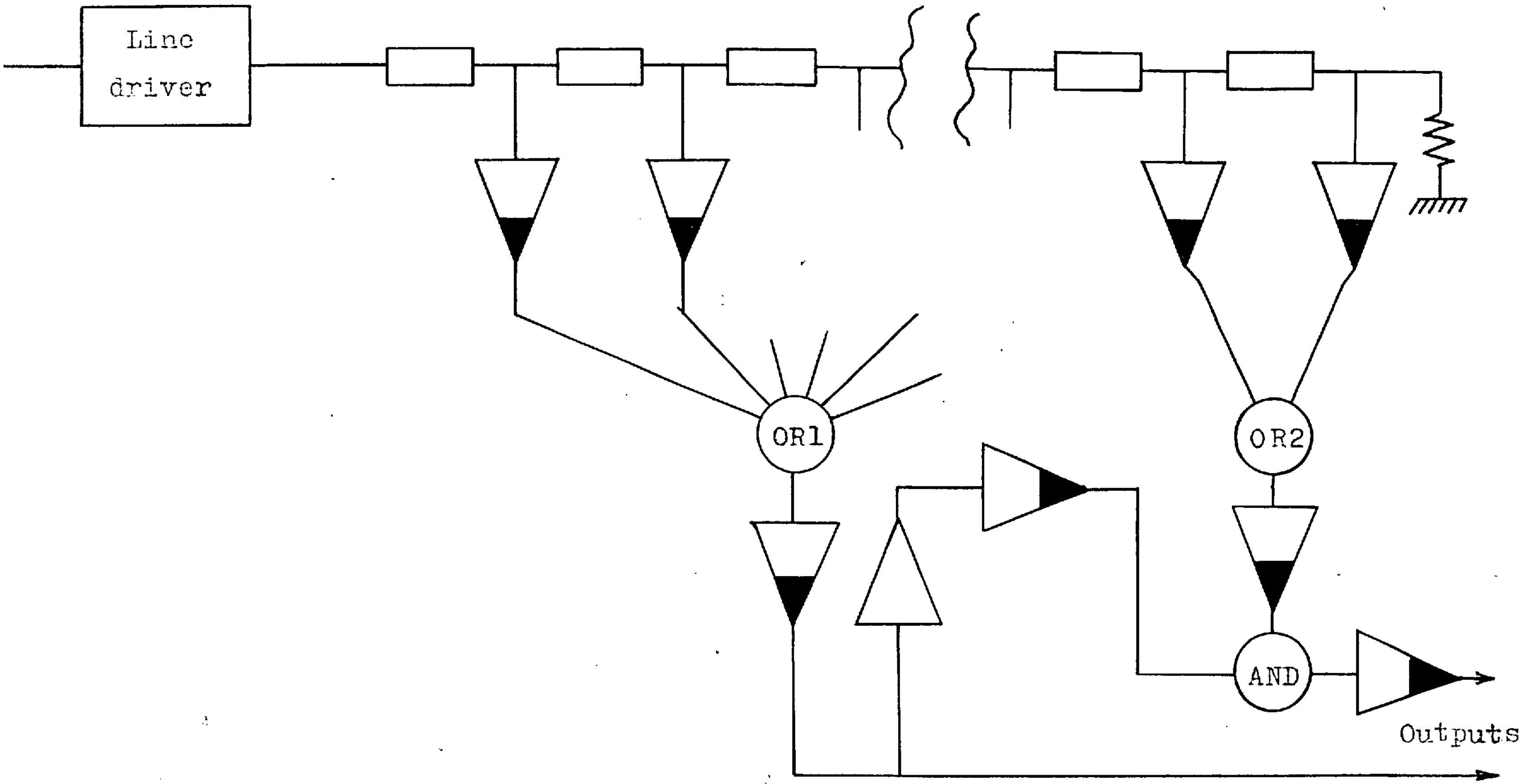


Figure 3.23(b). Detailed block diagram of the Translator

the use of inverting amplifiers for such purposes as driving delay lines, signals sometimes have their '1' level at a more positive voltage than their '0' level and sometimes vice versa. Thus a gate which is an AND gate for one type of signal will be identical to an OR gate for the other type and vice versa. This matter is further explained in Appendix 3. In Figure 3.23(b) no NOT circuit is necessary at (i) since the signal was previously inverted by the line driver. Also, at (ii) the configuration is not quite the same as that of the simplified block diagram. Because of the inversion of the line driver, the INHIBIT gate of Figure 3.23(a) becomes an AND gate; the other input to the AND gate has to be re-inverted, however.

A description of the logical operation of the Translator-SGU combination is to be found in Section 3.4.

3.6.3 The Detail Function Shaper

Initially the Translator was operated directly from the Run-end Detector, or with a Schmitt trigger circuit inserted between the two units in order to improve the waveform applied to the Translator input. Tests on the functioning of the Translator-SGU combination were carried out, the units being driven either by a Cintel Double Delayed Pulse and Sweep Generator or by the Run-end Detector. Although operation was observed to be satisfactory most of the time, occasional errors were noticed. These took the form of the generation by the SGU of blanking pulses of incorrect width. Examination of the triggering pulses fed to the multivibrators suggested the reason for the incorrect operation. Ideally, the signal from the Run-end Detector has transitions which occupy an infinitesimal time. In any physical system, however, the rise time of any signal must clearly be finite and in the case of the Run-end Detector output signal the rise time is about 50 nS. Even if the Run-end Detector be followed by the Schmitt trigger circuit mentioned earlier in this section the rise time is still of the same order. (The effect of the Schmitt trigger circuit is, in fact, to "clean up" the waveform so that the signal has two levels only.)

So far, no restriction has been placed on the phase of the transitions with respect to the Clock Pulses. Thus if a train of Clock Pulses is gated by the detail signal as it comes from the Schmitt trigger circuit, "part pulses" can easily be generated (Figure 3.24) It was found that a critical size of "part pulse" could be formed at the input to the multivibrators in the SGU, sufficient to switch the multivibrator on but insufficient to ensure an output pulse of the correct duration.

The Detail Function Shaper overcomes this difficulty by ensuring that transitions in the control signal applied to the gates of the SGU occur midway between clock pulses. This mode of operation was achieved by the use of a bistable multivibrator. Figure 3.25 is a simplified block diagram of the unit.

Clock pulses are supplied to the DFS via a variable delay unit (VDU 2), the circuit of which is given in Appendix 1. The operation of the unit is as follows:

(a) When a region of high detail commences, the signal from the Run-end Detector takes up the higher of two voltage levels (denoted "1"). AND gate 1 is thus opened and clock pulses are passed to input 1 of the bistable. Meanwhile AND gate 2 is closed since a NOT circuit precedes it. The first pulse at input 1 switches the bistable into state "1".

(b) At the end of a region of detail the input voltage moves to a lower voltage level (0) so that AND gate 1 is closed. AND gate 2 is opened, however, since the NOT circuit inverts the 0 to give a 1. The first clock pulse passed by AND gate 2 switches the bistable into the 0 state.

Thus the transitions of the output signal from the DFS co-incide with the clock pulses fed to it. These clock pulses emanate from a delay unit, however, and their phase (with respect

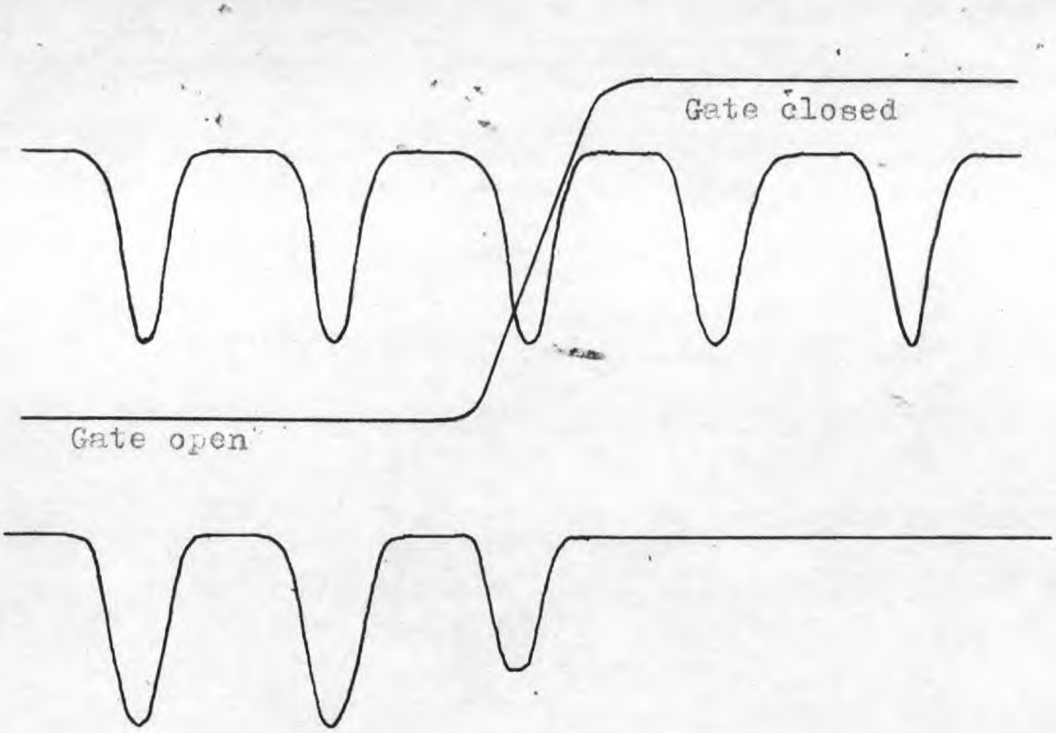


Figure 3.24. The generation of part pulses.

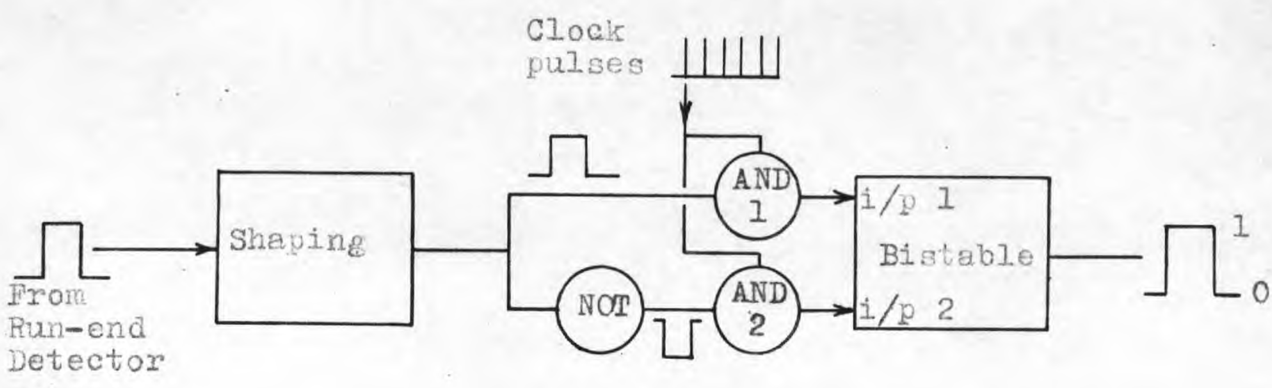


Figure 3.25. Simplified block diagram of the DFS.

to the pulses triggering the multivibrators of the SGU) can be varied. The transitions of the gating signal operating AND 1 and AND 2 of the SGU (Figure 3.12) can therefore be adjusted to occur midway between clock pulses as required. The delay necessary in VDU 2 should theoretically be T/2. This does not take into account the various delays in amplifiers, gates, binaries, etc. Thus it is found simpler in practice to adjust VDU 2 until the gating of the trigger pulses in the SGU is observed to be "clean".

Figure 3.26 is a block diagram showing the interconnections between the various units concerned with run-length restriction.

3.7 Performance of the Run-length Restricting Equipment

Apart from the necessity for allowing a warm-up period of about 10 minutes, the run-length restricting equipment was found to be stable and reliable. Initial drift in the SGU was cured simply by reversing this unit so that it protruded from the front of the rack. Here it was cooler than in its previous position, immediately above a number of power supply units. The drift appeared as a slow change in the periods of the multivibrators. It would probably have been possible to effect an alternative cure by substituting timing capacitors having low temperature coefficients for those used.

In order to ensure that the equipment operated correctly for all proportions of detail in the picture, a simple but effective test was devised. An output socket on the SGU gave a signal which stood at zero when samples were required to be blanked out and made a positive excursion when samples were to be inserted. This signal was fed to a conventional picture monitor synchronised by the composite synchronising signal used in the laboratory. (This synchronising signal was locked to the clock pulse generator, the flying-spot scanner also being synchronised by it (See Appendix 2)). A pulse generator synchronised by line synchronising pulses or by "single frame trigger" pulses (also explained in Appendix 2) was

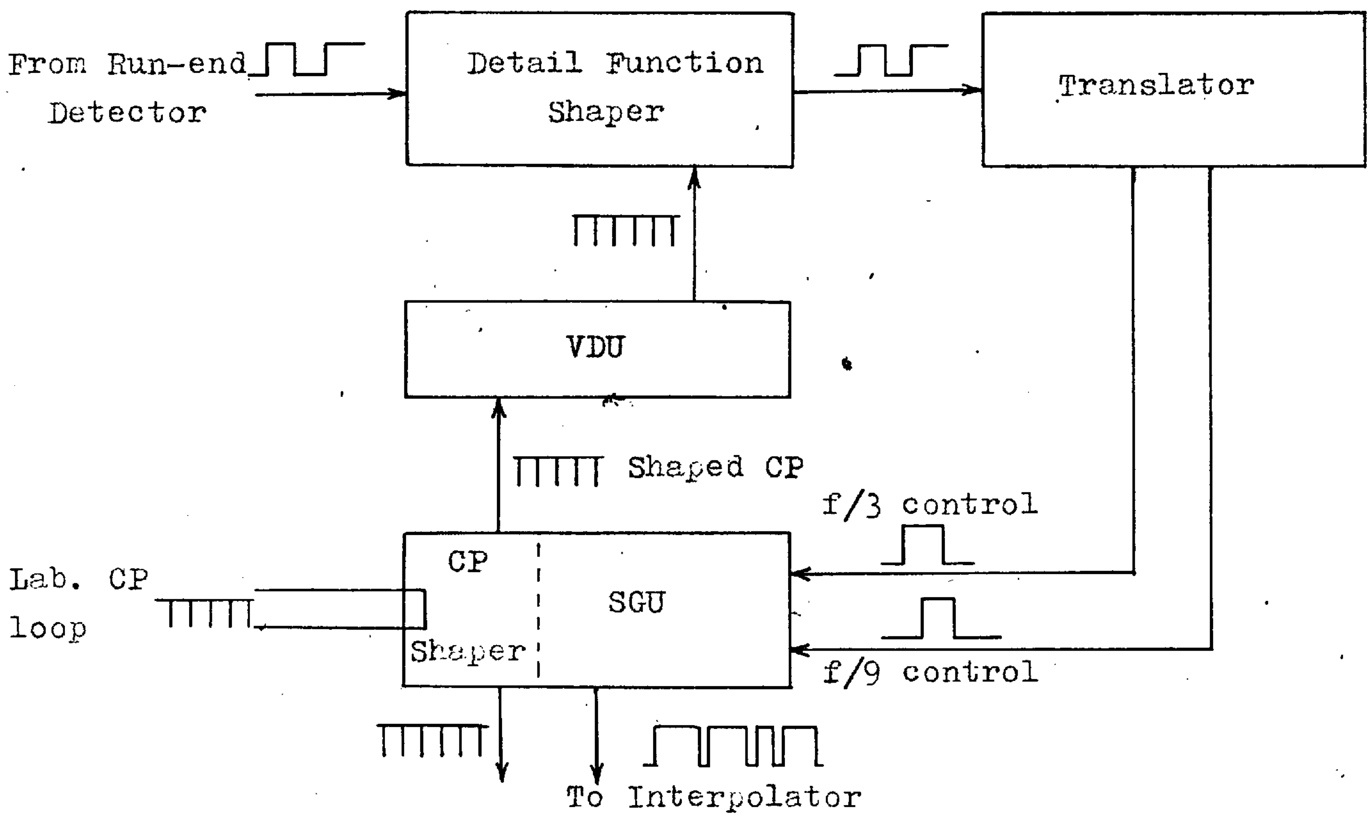


Figure 3.26. Interconnection of the units concerned with Run-length restriction

now used to provide a signal simulating a run-end detector output. Thus a steady, noise-free input to the run-length restricting gear could be obtained and any instability or incorrect standardised run-length could easily be observed on the monitor. Thus, for example, with the pulse generator operating at the line scanning frequency and synchronised by the line synchronising pulses, a pattern which repeated in every line was obtained. Variations of the pattern therefore showed up very well and the proportion of "detail" in the picture could easily be varied by changing the width of the pulse produced by the pulse generator. With the equipment in its final form perfectly steady patterns were obtained and would remain steady for long periods (several hours).

Examination of waveforms at various points in the equipment also showed it to be operating correctly. Photographs of some of the waveforms are shown in Figure 3.27. At (a) the traces, from top to bottom are the SGU output, the DFS output and the input to the DFS. The slight delay between the input and output of the DFS can be seen; the time scale is 0.5 μS per major division. Sections (b), (c) and (d) of Figure 3.27 show, in each case, waveforms at the SGU output, the f/9 connection from the Translator to the SGU, the f/3 connection from the Translator to the SGU and the input from the pulse generator. The onset of a region of detail is simulated, the differences between the three pictures lying in the different phase relationship between the train of 8T pulses from the SGU and the input signal. Thus 0, 1 and 2 2T pulses are generated in (b), (c) and (d) respectively. At the end of a region of detail, sampling at f/9 can recommence immediately with no transition period at f/3. This is, of course, one of the main features of the SGU. The fact that this occurs is illustrated in Figure 3.27(e), where it will be seen that the f/3 connection from the Translator to the SGU is not energised at all.

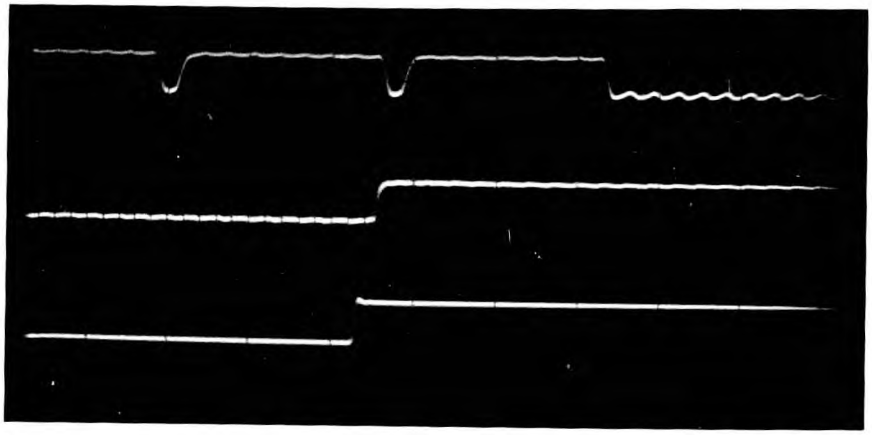


Figure 3.27. (a) Top: SGU output, 5V/major div.
Centre: DFS output, 5V/major div.
Bottom: DFS input, 10V/major div.
0.5μS/major div.

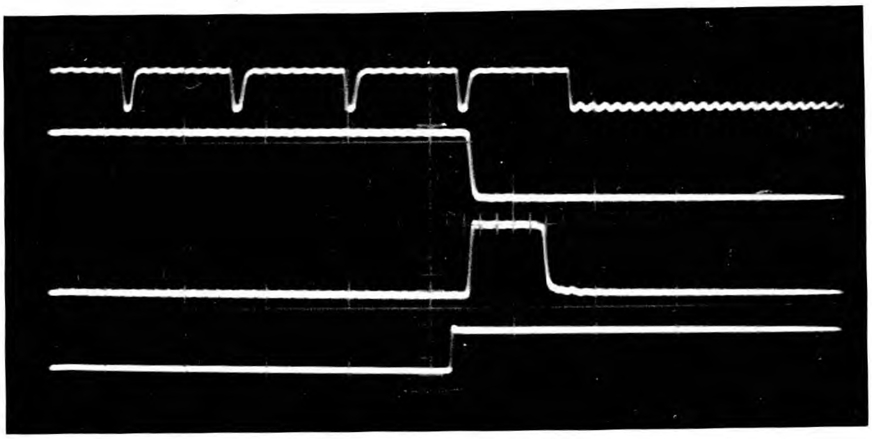


Figure 3.27. (b) Top: SGU output, 5V/major div.
2nd: Translator f/9 output, 10V/major div.
3rd: Translator f/3 output, 10V/major div.
Bottom: DFS input, 10V/major div.
1μS/major div.
These waveforms illustrate run-length restriction when the detail commences in such a phase that no 2T pulses are required.

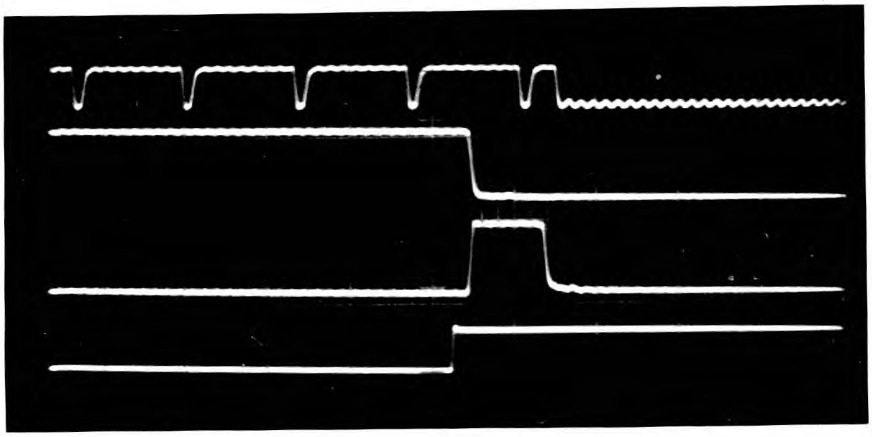


Figure 3.27. (c) As Fig. 3.27(b) but with one 2T pulse

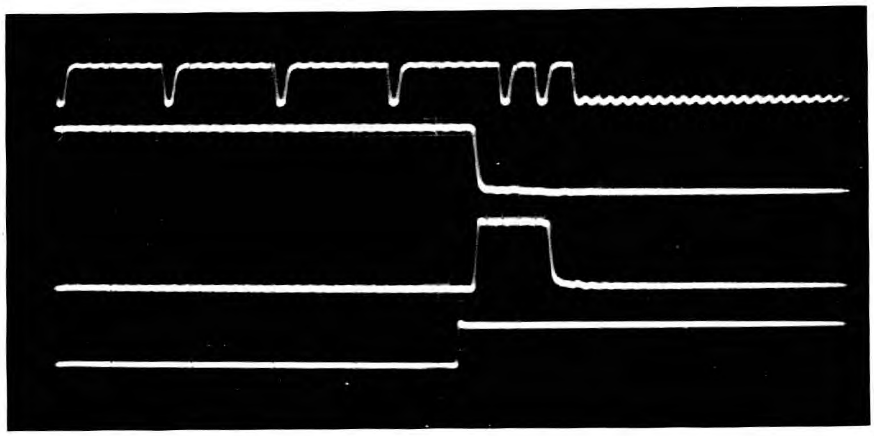


Figure 3.27. (d) As Fig. 3.27(b) but with two 2T pulses

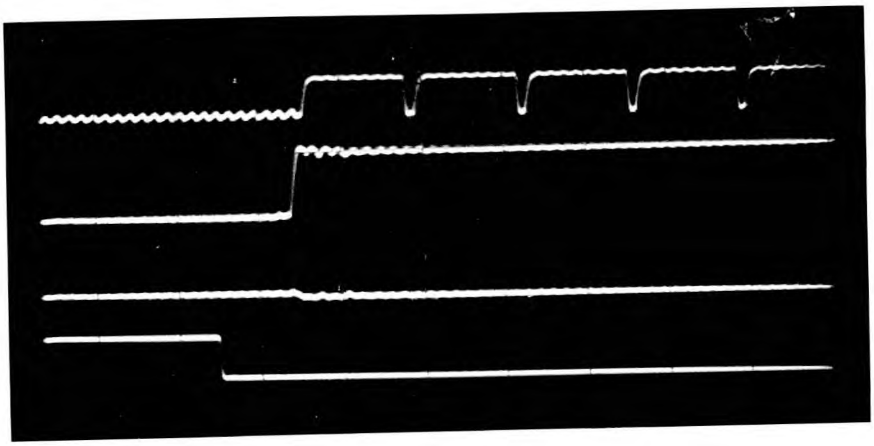


Figure 3.27. (e) Top: SGU output, 5V/major div.
2nd: Translator f/9 output, 10V/major div.
3rd: Translator f/3 output, 10V/major div.
Bottom: DFS input, 10V/major div.
These waveforms illustrate the operation of the run-length restriction equipment when detail ends.

Further photographs showing the effect of the run-length restricting equipment are shown at the end of Chapter 4. The functioning of the equipment when used with a run-end detector is there illustrated.

3.8 Experimental determination of the loss of data reduction caused by various sets of standard run-lengths

Reference has previously been made (Section 3.2) to measurements of run-length statistics from which optimum standard run-lengths may be calculated. The availability of run-length restriction equipment makes possible the direct assessment of these parameters. Moreoever, this method allows the effect of varying the run-end detector threshold to be quickly investigated. This operation can be expected to change the run-length statistics with a consequent effect upon the optimum run-lengths.

The measurements were made using the "COMMON" channel of the read-in section of the 1st Order Interpolator (see Section 4.5) to gate clock pulses. The number of pulses per second in the resulting train was counted by a Hewlett-Packard Type 524C electronic counter. The average of several counts, each of 1 second duration was recorded in order to eliminate variations due to random noise. Before each measurement the pulse train was observed with an oscilloscope and VD1 was adjusted until "clean" gating was achieved. For measurements with run-length restriction in operation, the DFS, Translator and SGU were connected as described earlier in this chapter; for unrestricted operation the DFS output was fed via an inverting amplifier, to the "gating" input of the interpolator. In order to prevent errors due to the inexact resetting of the Run-end Detector threshold control, counts of both unrestricted and restricted pulse trains were taken before advancing the threshold control to the next desired position. The pictures used for the measurements were those shown in Section 4.8.2.

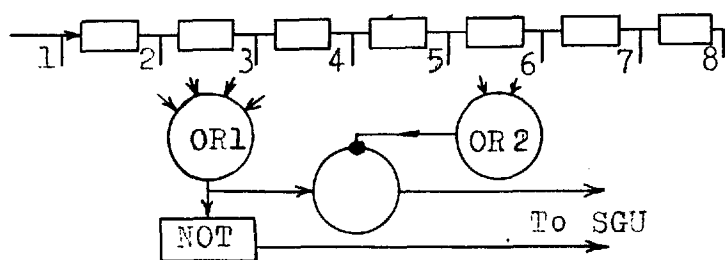
In order to change from one set of standard run-lengths to another it was necessary to adjust both the SGU and the Translator. In the case of the SGU the adjustment consisted simply of varying the lengths of the pulses generated by the monostable multivibrators; alteration of the values of the timing capacitors resulted in the desired effects. In the Translator it was necessary to alter the connections between the delay line and the OR gates; the required arrangements are shown in Table VI.

Measurements were carried out over a wide range of settings of the Run-end Detector threshold, including that producing the best reconstructed pictures. The range of the measurements was limited by the time available, however. For each picture and each set of standard run-lengths the number of pulses required with restriction was plotted against the number of pulses required without restriction. It was quickly evident that almost identical results were obtained with a given set of standard run-lengths for all scenes used. It was therefore decided to average over the four scenes by drawing only one line through the points for each set. The curves thus obtained have been collected together and form Figure 3.28.

3.8.1 Discussion of Results

The insertion of more samples by the Run-end Detector is bound to result in more samples after restriction; thus the curves have a positive slope. Over the range of sampling rates considered, the standard run-lengths of 1, 4 and 9 Nyquist intervals clearly give most economy; the trend of the curves indicates that 1, 3, 8 may be preferable at very high average rates, however. This is to be expected; as more pulses are inserted there will be more short runs and fewer long ones. The apparent lack of order in the arrangement of the curves should cause no misgivings. For instance it may be thought that the standard run-lengths 1, 3, 9 are necessarily more economical than 1, 3, 8. However, with 1, 3 9 a run of 8 Nyquist intervals would require to be broken into two

TABLE VI



Standard run-lengths	OR 1 fed from taps	OR 2 fed from taps
1,2,6	1,2,3,4	5
1,2,7	1,2,3,4,5	6
1,2,8	1,2,3,4,5,6	7
1,2,9	1,2,3,4,5,6,7	8
1,3,6	1,2,3	4,5
1,3,7	1,2,3,4	5,6
1,3,8	1,2,3,4,5	6,7
1,3,9	1,2,3,4,5,6	7,8
1,4,8	1,2,3,4	5,6,7
1,4,9	1,2,3,4,5	6,7,8

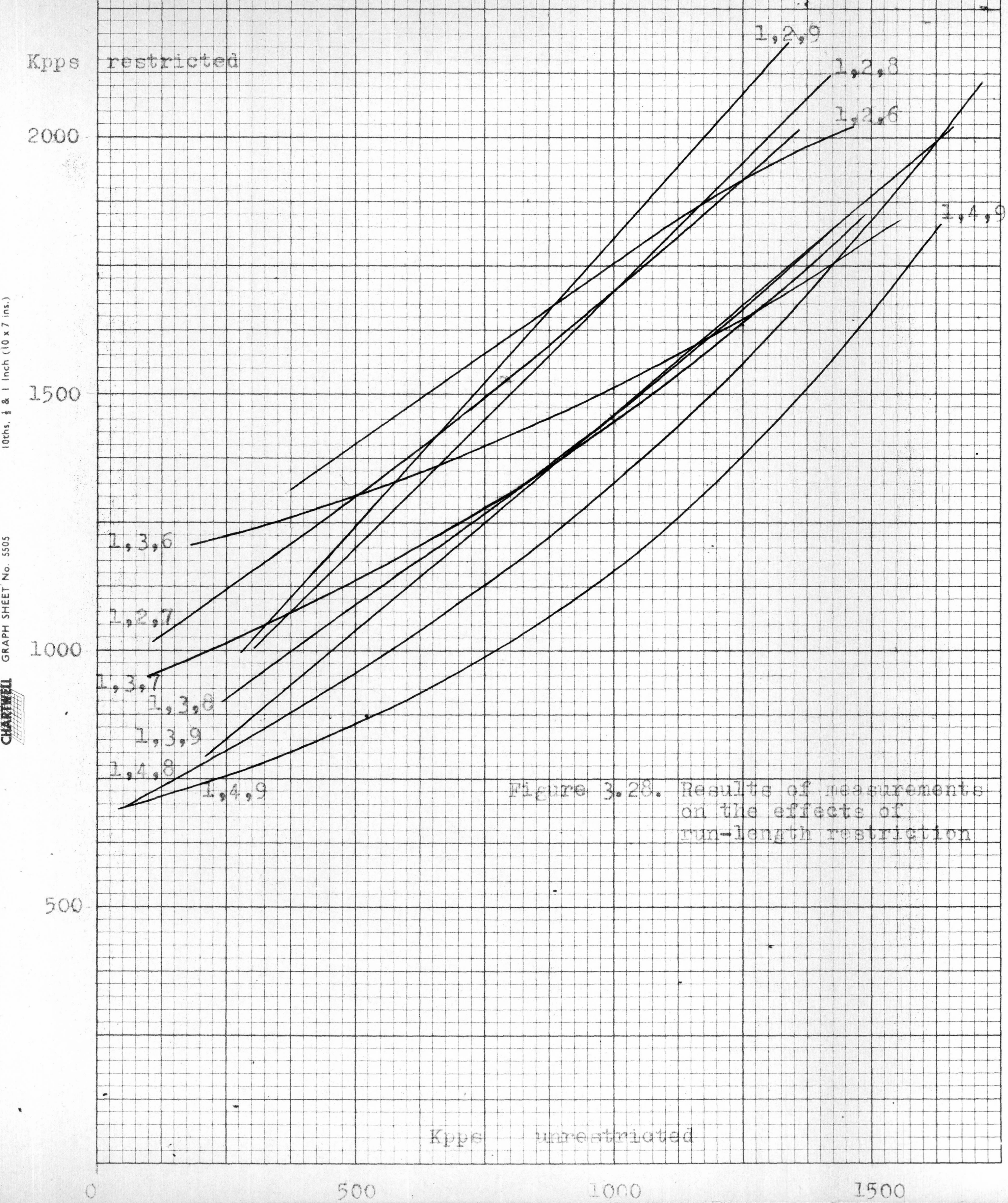


Figure 3.28. Results of measurements on the effects of run-length restriction

runs of three and two of one Nyquist interval thus using four samples in all. With 1, 3, 8 this run would have been coded by one sample. The order of preference thus varies with the run-length statistics and these change with the Run-end Detector threshold (or the number of samples with no restriction).

These results should be compared with standard run-length values calculated from measured run-length statistics. Such measurements have been made by Prasada, Pine and Julesz[45, 44, 29]. Prasada's work, however, used Test Card C as the only scene, so that the results cannot be regarded as meaningful, Test Card C being entirely artificial.

Although Pine also included Test Card C in the four scenes from which he derived statistics, the other three were natural scenes and results of some validity should have been forthcoming. Errors in the method of measurement have been pointed out, however,[33] and it would be optimistic to use the results (best standard run-lengths 1, 2, 6 Nyquist intervals) as any more than a guide to the magnitude which may be expected.

Full run-length statistics were published by Julesz for only one scene, a portrait of a girl. Calculations using these statistics suggest 1, 2 and 5 Nyquist intervals as the optimum restricted lengths but once again there are reservations. Chief among these is the fact that the maximum (unrestricted) run-length permitted was 16 intervals and all runs longer than this were broken down into one or more runs of length 16 units together with a shorter run. The run-length probability distribution was thus distorted and once again the standard run-lengths calculated from it can be only a guide to the true optimum values.

Thus a hard-and-fast comparison of the measured optimum values with those predicted from run-length statistics cannot be achieved. It is submitted, however, that considerably more reliance can be placed on directly measured figures than on those derived from statistics having various shortcomings.

4. THE COMPLETE 1st ORDER SYSTEM

4.1 Introduction

Of the two aspects of data-reduction which have so far been discussed, only run-length restriction can be completely assessed by the methods described. Before the efficacy of a detail detection procedure can be determined it is necessary to re-form a picture from the reduced data. Clearly, in the complete Open-Loop system this state of affairs does obtain. The instrumentation of such a complete system is, however, a considerable task involving the construction of an elastic encoder together with some means of decoding the signal. Since these two processes together ideally have effects which produce a signal identical to that existing before the encoder it should be possible to omit them and to re-form the video signal from the reduced data. Such a process has indeed been employed with considerable success in the case of the zero-order representation of the signal[33]. Here the amplitude of each sample in the reduced train was stored on a capacitor until the succeeding sample arrived. Although the principle is simple, realisation brought out considerable technical problems.

A somewhat different technique is required for 1st Order reconstruction. A run in this case has been defined as a region of constant slope and the Run-End Detector has been designed to detect changes of slope. Thus reconstruction requires the signal amplitude to change linearly between the two samples at the beginning and end of each run (Figure 4.1).

Although this was expected to be considerably more difficult that the equivalent zero-order process it was felt that it would be very worthwhile to design and construct a 1st Order Interpolator*. In addition to its value in data reduction of visual image signals,

---

*Interpolate: Insert (intermediate term) in series, (Concise O.E.D.)

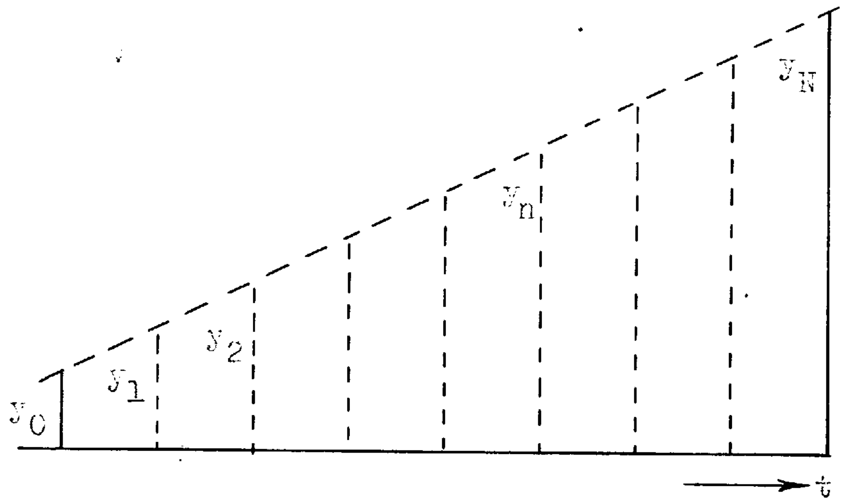


Figure 4.1. The principle of 1st Order reconstruction

such a device would have application in other fields where interpolation with moderate accuracy is required and where the use of a digital computer is not possible or would be too expensive.

It is clear from Figure 4.1. that the samples from both ends of a run must be available in order to re-form the signal between them. Thus it will be necessary to store the amplitude of one sample until the other is available for comparison. This implies, in turn, that the maximum run-length must be limited, for storage of high speed data is not easy over long periods*. Run-length restriction, then, is a necessity before 1st order interpolation can be undertaken; the two programmes fit well together.

As well as using a more general model, 1st order interpolation possesses the virtue of giving an approximation to the signal which would be produced if variable velocity scanning were used at the receiver (as proposed for a complete compression scheme). This is further discussed in section 4.2. The remainder of the chapter reports my investigations into 1st order interpolation for data-reduced television signals.

4.2 A Comparison between 1st Order Interpolation and Variable Velocity Scanning

Section 1.9 briefly mentions variable velocity scanning as a means for decoding the reduced bandwidth signal at the receiver. The operations involved in this process are best explained with the aid of Figure 4.2.

The velocity with which the spot on the face of the receiving CRT moves is controlled by a signal derived from the Elastic Encoder.

*The terms "high speed" and "long periods" are, of course, relative. In general, storage with easy, rapid access is limited to about 100 times the fastest rise-time of the signal. Storage is a subject in itself, however.

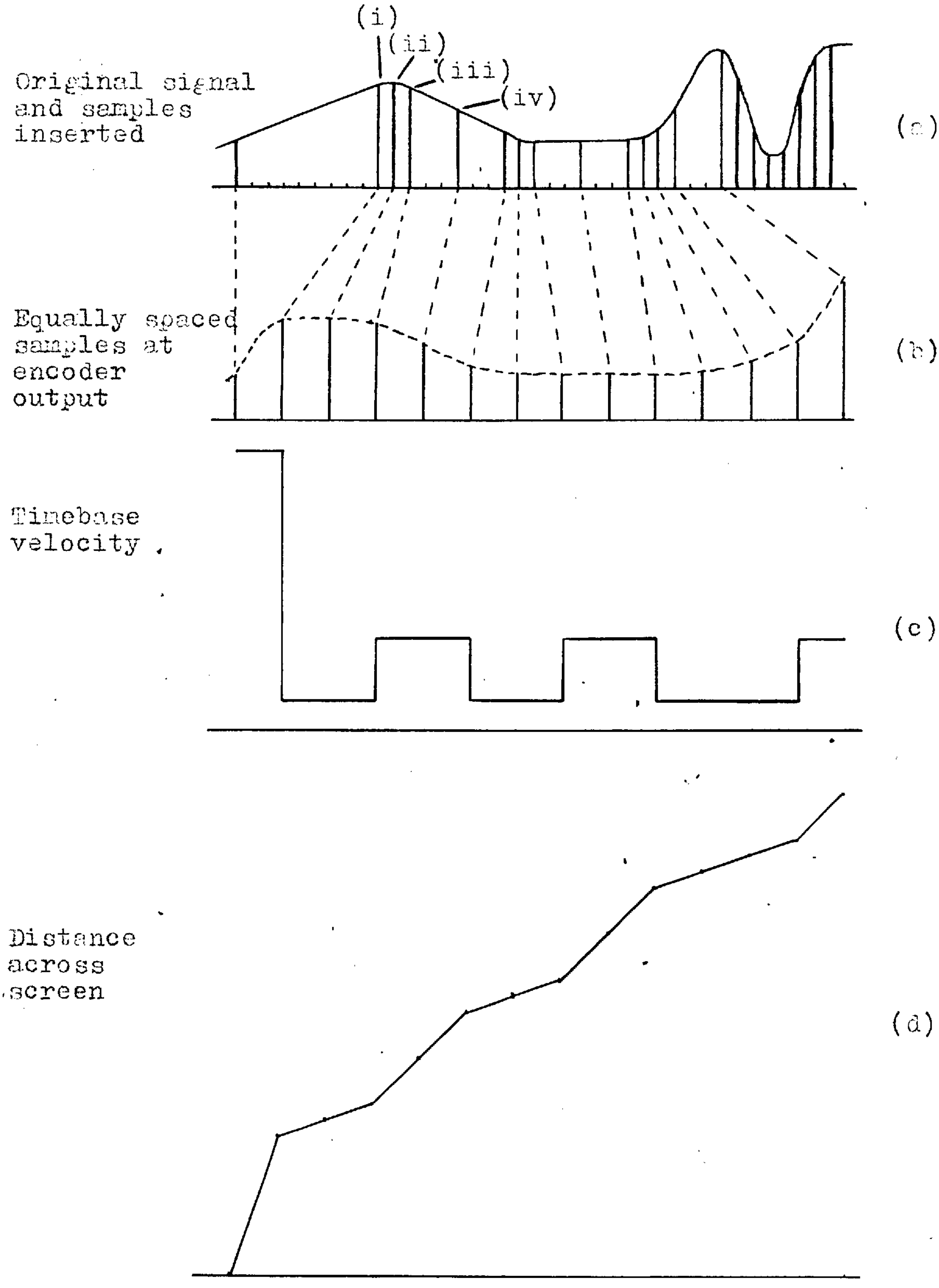


Figure 4.2. Decoding by variable velocity scanning

Samples which were originally close together (e.g. (i) and (ii) in Figure 4.2(a)) have their spacing in time increased by the Elastic Encoder whilst those with large initial spacings e.g. (iii) and (iv) have this reduced. The envelope of the samples at the Encoder output is formed by a low-pass filter and it is this signal that is used to modulate the beam current of the receiver's CRT. In order to cause two points which have their time spacing increased to appear close together on the screen, the receiver timebase scans slowly. Conversely, points whose time spacing has been reduced are moved apart spatially at the receiver by use of a higher speed of scan. Ideally the velocity changes take place instantaneously; if this condition does not obtain, the decoded picture is expected to be distorted but the distortion is expected to be similar in neighbouring lines and vertical edges should not be broken up.

For the purpose of comparing 1st Order Interpolation and Variable Velocity Scanning, instantaneous velocity changes are assumed; the time scale of the waveform interpolated between the uniformly spaced samples leaving the Elastic Encoder can thus be merely "stretched" or "compressed" to yield a plot of the displayed signal on a linear distance scale (Figure 4.3). Further, the "stretching" or "compressing" will be by a factor which remains constant over the interval between each pair of samples at the Encoder output.

The impulse response of the low-pass filter is ideally described by the function:

$$h(t) = \frac{A\omega_c}{\pi} \; \frac{\sin\omega_c(t - t_o)}{\omega_c(t - t_o)} \qquad \text{(Figure 4.4)}$$

where $h(t)$ is the response to a Dirac pulse

- A is the magnitude of the transfer function in the pass band
- ω_c is the angular cut-off frequency
- t_o is the group delay

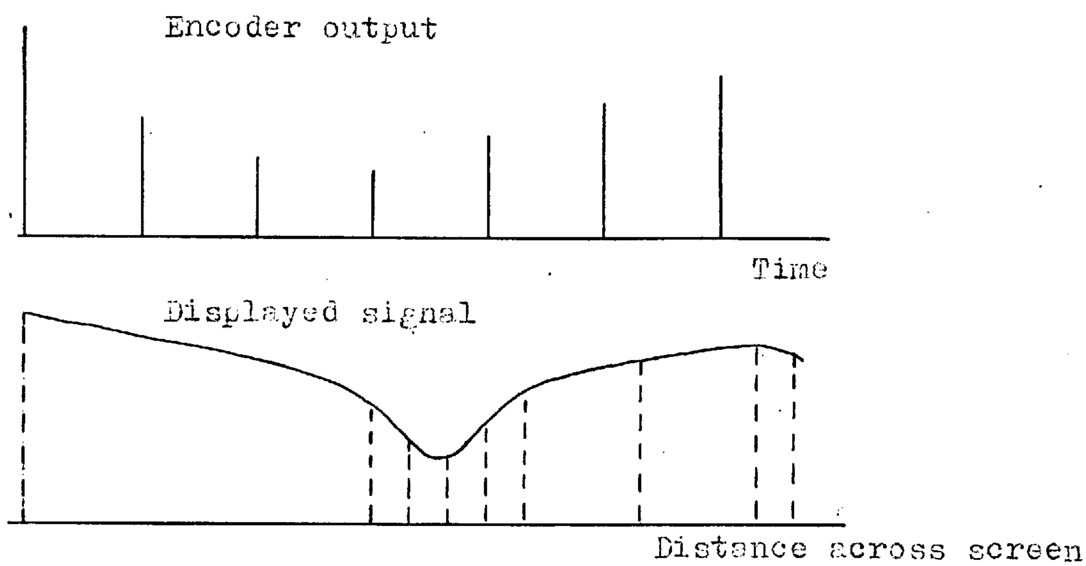


Figure 4.3. An example of decoding by variable velocity scanning

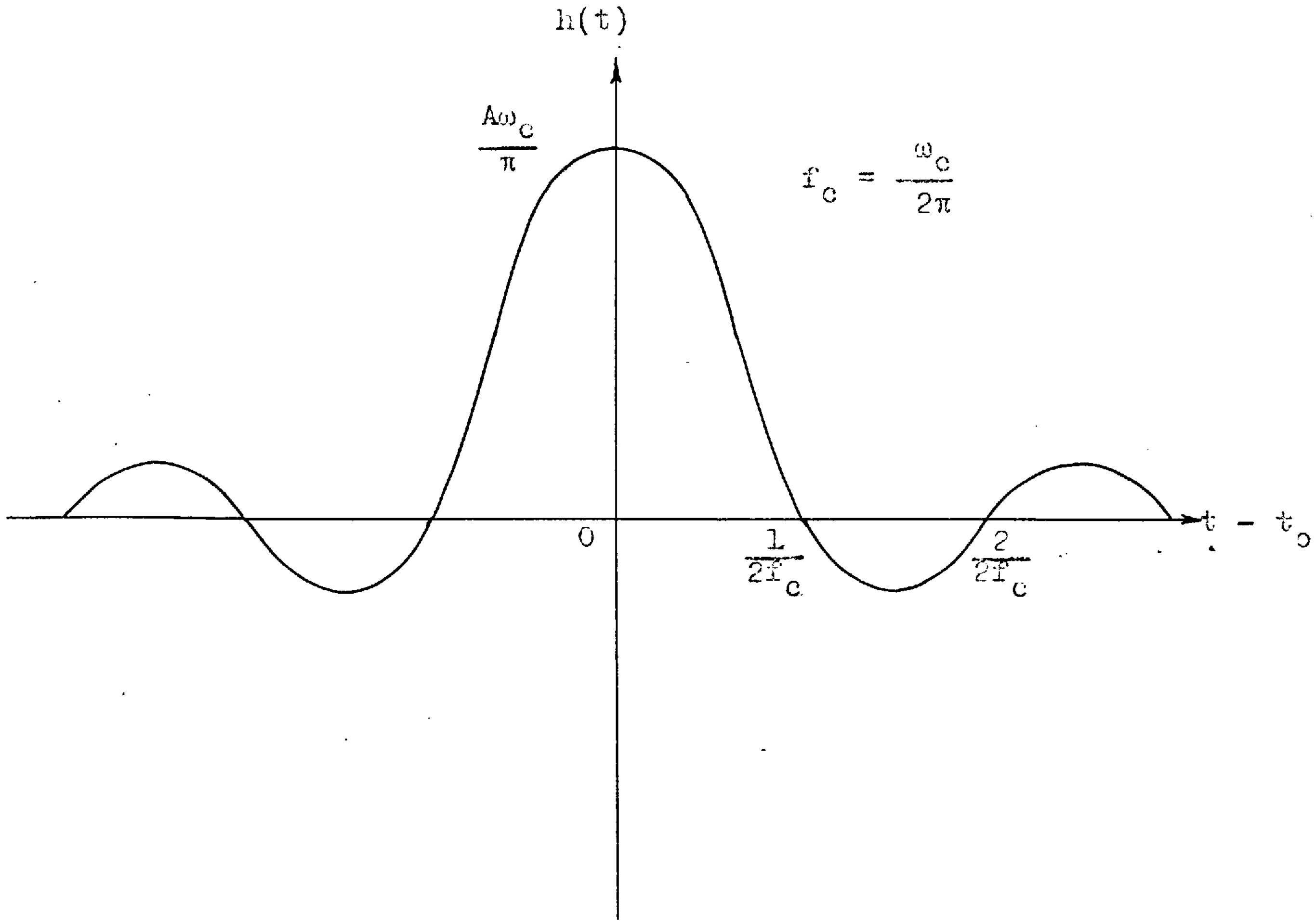


Figure 4.4. Impulse response of the low-pass filter

Clearly, if the samples to be filtered have a spacing of $1/2f_c$, each sample will generate a contribution independent of that from other samples. The samples leaving the Elastic Encoder have spacings of CT so that the filter bandwidth is given by

$$\frac{1}{2f_c} = CT \quad \text{or } f_c = \frac{1}{2CT}$$

where C is the compression ratio.

Knowing the impulse response of the filter, the output waveform can be calculated from any given train of pulses leaving the Encoder. In addition, the time scale between any two consecutive samples can be multiplied by a suitable factor (which remains constant over the interval between samples if instantaneous velocity changes are assumed) in order to arrive at the brightness distribution presented on the receiver screen. Consider, for example, the train of pulses shown in Figure 4.5(a). The quantities

$$X \; \frac{\sin \omega_c(t - t_o)}{\omega_c(t - t_o)}$$

for some of the samples are drawn in Figure 4.5(b) and their sum at any instant traces out the curve of Figure 4.5(c).

Suppose now that scanning velocities of V/3, V and 3V are employed in the receiver as indicated in Figure 4.5(d). (These are the velocities which would be required if restricted run-lengths of 1, 3 and 9 Nyquist intervals had been employed; the compression ratio would then be 3:1.) The brightness distribution on the receiver screen thus has the form shown in Figure 4.5(e). It is this curve which we wish to compare with the curve formed by linearly interpolating between the tips of the samples at their pre-encoding spacing. This can be deduced by using the velocity information to indicate the original spacings of the pulses shown in Figure 4.5(a) e.g. As scanning

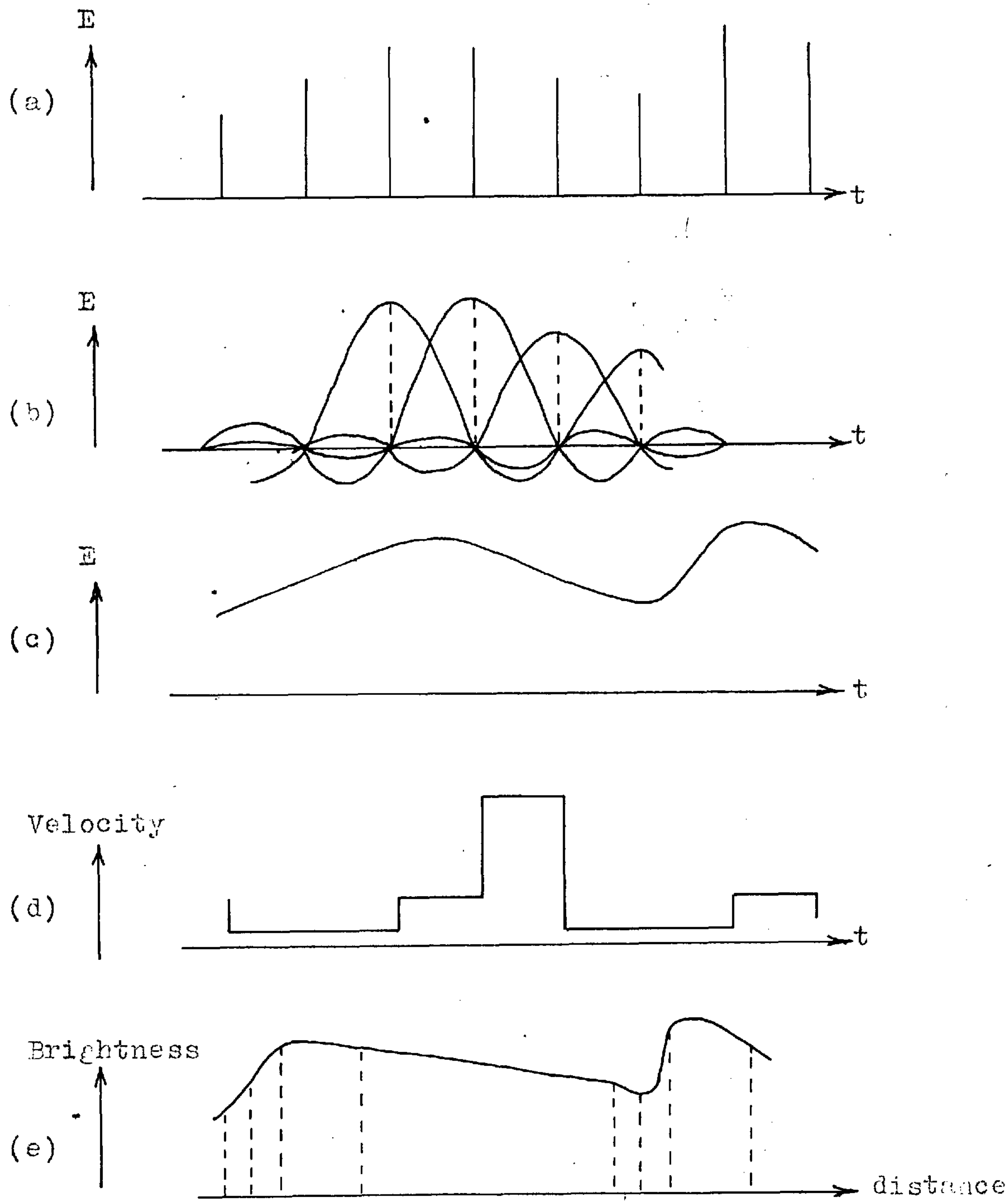


Figure 4.5. Filtering and variable velocity scanning

over the time interval t_1 t_2 is required to be $V/3$, the samples must have had a spacing of T before coding. The samples, although they have actually been removed by filtering, can be drawn in Figure 4.5(e).

Both linear interpolation and filtering give waveforms which can, by suitable scaling, be drawn through the tips of the samples; the former method thus produces a waveform which, although it differs from the variable velocity reconstruction between samples, is identical in form at the sampling instants. The comparison of 1st Order Interpolation and Variable Velocity Reconstruction consists, therefore, in estimating the difference between the two waveforms at some point between two samples. For convenience, the mid-point of a sampling interval is chosen although this is not the point of maximum error.

The amplitude of the quantity

$$\frac{\sin x}{x}$$

midway between zero crossings is given by

$$y_n = \frac{\sin(2n + 1)\ /2}{(2n + 1)\ /2}\ , \quad n = 0,\ 1,\ 2\ldots\ldots$$

$\sin(2n + 1)\pi/2$ is, however, equal to +1 and -1 alternatively so that

$$y_n = \frac{(-1)^n}{(2n + 1)\pi/2}$$

The filter output voltage midway between two samples depends not only on those samples but on all the others preceding and following them since each sample generates a contribution which is added algebraically to that from all other samples to produce the final output. The general case is illustrated in Figure 4.6. where is shown the train of uniformly spaced samples leaving the Elastic Encoder. Suppose that it is required to estimate the difference between the two forms of reconstructed signal between the samples 'a' and 'b'. Midway between a and b, the signal produced by 1st Order Interpolation has the value $(a + b)/2$. A $(\sin x)/x$ function so scaled as to have its maximum at the tip of sample 'a' gives a contribution of $(a \sin\pi/2)/(\pi/2)$ at the same point, where 'a'

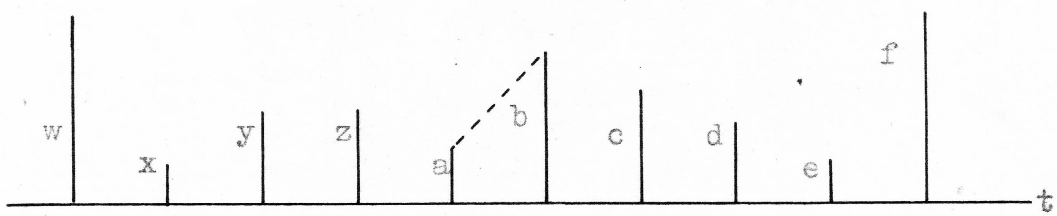


Figure 4.6. A possible train of uniformly spaced pulses as output by the Elastic Encoder

represents the amplitude of that sample. Similarly, the contribution from b will be $(b \sin \pi/2)/(\pi/2)$, that from c will be $(c \sin 3\pi/2)/(3\pi/2)$ and so on, as indicated in Table VII. As many samples as desired may be taken into account and their contributions added to give the resultant waveform. As all the samples entering the filter in theory affect the output, they should all ideally be included in the calculation. Samples far away from a and b add very little, however, and this, together with the fact that successive samples give alternate positive and negative contributions at a given mid-point means that there is likely to be little error in neglecting them.

An interesting and important special case occurs when the samples c, d, e, f, etc. are all of the same amplitude - say k_1, z, y, x, etc. being of another constant value, k_2. The reconstruction midway between a and b then has the value

$$\frac{a + b}{\pi/2} + \left(\frac{-2k_1}{3} + \frac{2k_1}{5} - \frac{2k_1}{7} + \ldots\right) + \left(-\frac{2k_2}{3} + \frac{2k_2}{5} - \frac{2k_2}{7} \ldots\right)$$

$$= \frac{2}{\pi}(a + b) + \frac{2}{\pi}(k_1 + k_2)\left(-\frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \ldots\right)$$

Now the series $1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + , \ldots$

(known as Leibnitz's Series) converges to $\pi/4$ so that, as the number of terms either side of a and b tends to infinity, the amplitude of the filter output at the mid-point tends to

$$\frac{2}{\pi}(a + b) + \frac{2}{\pi}(k_1 + k_2)\left(\frac{\pi}{4} - 1\right)$$

and the difference between this and the voltage produced by

TABLE VII

Sin x/x Interpolation

Sample	Contribution		
a	$2a/\pi$	=	0.637a
b	$2b/\pi$	=	0.637b
c	$-2c/3\pi$	=	-0.212c
d	$2d/5\pi$	=	0.127d
e	$-2e/7\pi$	=	-0.091e
f	$2f/9\pi$	=	0.071f
etc.			
z	$-2z/3\pi$	=	-0.212z
y	$2y/5\pi$	=	0.127y
x	$-2x/7\pi$	=	-0.091x
etc.		etc.	

1st Order Interpolation is

$$\left(\frac{2}{\pi} - \frac{1}{2}\right)(a + b) + \frac{2}{\pi}(k_1 + k_2)\left(\frac{\pi}{4} - 1\right)$$

$$= 0.136\ (a + b - k_1 - k_2)$$

This condition arises when a small region of detail has on either side of it regions of constant amplitude, a situation which occurs frequently in practice.

If there is a simple change from a brightness level k_1 to a level k_2, a will be equal to k_1 and b equal to k_2. Thus the error becomes

$$0.136\ (k_1 + k_2 - k_1 - k_2) = 0$$

i.e. There is no error midway between a and b in this case. The two reconstructed waveforms are not quite identical, however, but are as indicated in Figure 4.7.

Another special case of importance arises when the samples (other than a and b) have large and small values alternately. The extreme case - alternate 0's and 1's is shown in Figure 4.8. All contributions between a and b are here positive so that the difference between the two reconstructions can reach a considerable value. Fortunately the fact that samples are alternately large and small means that the region is highly detailed and the eye will thus be tolerant towards amplitude errors.

Although the word "error" has been used frequently in this section, its significance may not be immediately realised. It may at first be thought that the reconstructed signal formed by filtering and variable velocity decoding must be the ideal, 1st Order Interpolation being an approximation which introduces errors. However, the form of the Run-End Detector was arrived at by assuming a first-order model for the signal and 1st Order Interpolation is the best means of reconstituting such a signal.

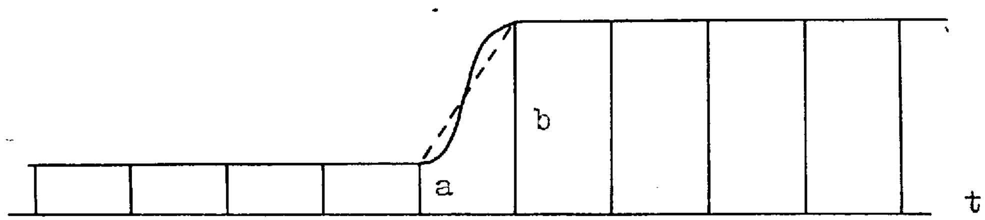


Figure 4.7. The difference between 1st order interpolation and variable velocity reconstruction when a single change in brightness occurs

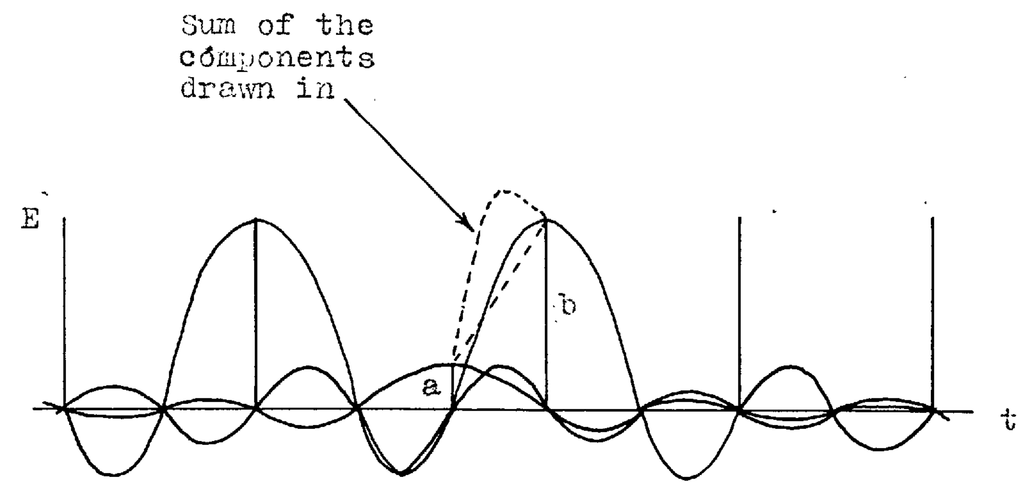


Figure 4.8. The difference between 1st order interpolation and variable velocity reconstruction when the samples have large and small values alternately

Variable velocity decoding of an elastically coded and filtered signal is a process not matched to the Run-End Detector. Thus it is this process which is responsible for the errors. The fact that the original signal was not actually of the 1st Order form but was band-limited suggests that a more accurate result would be achieved by restricting the interpolator output to the same bandwidth. In practice, this condition tends to be obtained as the frequency response of the monitor used to display the reconstruction deteriorates above 3Mc/s, thus providing an approximation to a low-pass filter.

To summarise, therefore, it may be said that this section has outlined the method of estimating an error which is due to the use of a reconstruction method not matched to the run-end detection method. From the calculations the error appears to be small; an estimate of its importance can be obtained only by subjective testing, however.

4.3 Possible Methods for 1st Order Interpolation

So far, the process of 1st Order Interpolation has been thought of as involving the generation of a voltage which changes uniformly with time and has such an amplitude and slope that its **amplitudes** at the ends of a run are equal to those of the samples at those points.

However, in a system where the non-uniform train of samples has been obtained from a band-limited signal it is not essential to derive the complete ramp; the values of the intermediate samples would serve just as well. The envelope of these samples could later be generated by use of a low-pass filter.

Both of these possibilities were considered for the 1st Order Interpolator.

4.3.1 Ramp Generation

To generate a ramp electronically it is necessary to use a circuit of the Miller Integrator type to integrate a voltage which remains constant during a run. Ramps of different slope can be obtained by varying the magnitude of the voltage.

A television signal is continuous since it is band-limited (transitions in 1/6μS, the Nyquist interval, will not be considered to be discontinuities). Thus, the signal amplitude at the end of a run must be the same as that at the beginning of the next run. No "jumps" in the output of the integrator will therefore be called for (Figure 4.9) and there will be no need to add a constant voltage at its output in order to set the ramp to the correct level. However, errors could well be cumulative with this system unless special precautions are taken to correct them. For example, if a section of the signal consists of runs alternately of length 9T and 3T, the 9T section having a negative slope and the 3T a positive, the state of affairs shown in Figure 4.10 could arise. Here, an error in some adjustment causes the 3T section to have too steep a slope while the slope of the 9T section is correct. The resulting error in the reconstructed signal is thus cumulative unless some means of "forcing" the amplitude to the correct value is used. Feedback might be used to ensure that errors did not occur but this would not be easy to arrange since

(a) there are several sampling rates and thus several voltage multipliers (or amplifiers) to be adjusted.

(b) the error would not be apparent until the end of a run, by which time it would be too late to introduce the correction.

The alternative to the use of feedback is to clamp the signal voltage to the sample at the end of each run (Figure 4.11).

In view of the probable difficulties of ramp generation, means of estimating the amplitudes of the omitted samples were considered.

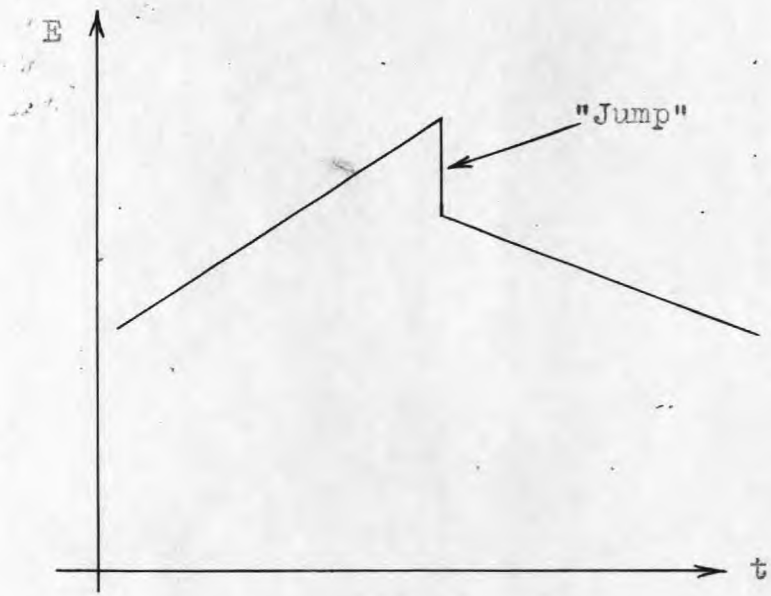


Figure 4.9. A first order signal with a "jump." This is not required when dealing with television signals

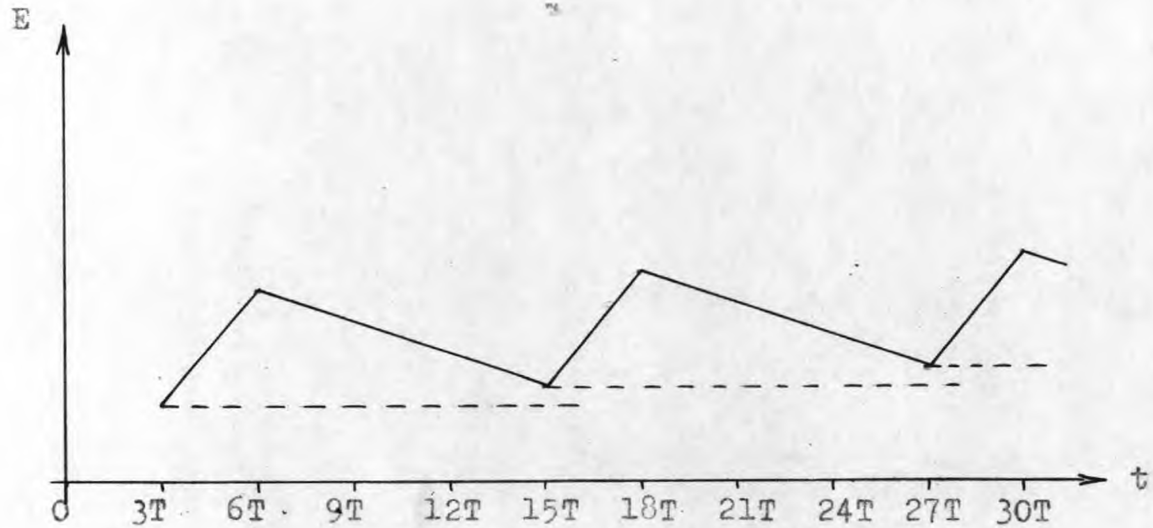


Figure 4.10. Possible effect of an error in the slope of the reconstructed signal over each 3T run

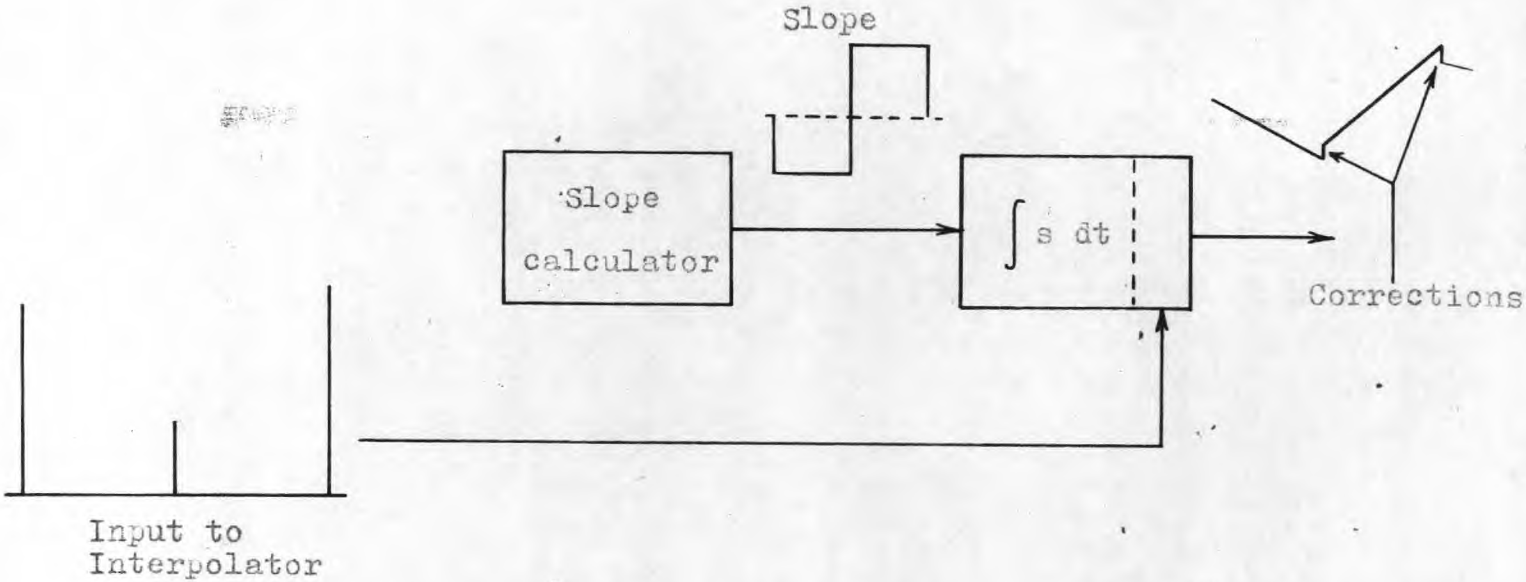


Figure 4.11. Correction of an integrating interpolator output by clamping at the end of each run

4.3.2 Intermediate Sample Generation

A chain of resistors forms the basis of this type of interpolator (Figure 4.12).

Suppose that a sample of amplitude V_0 be applied to one end of the chain, the subsequent sample, V_n, being simultaneously applied to another point on the chain such that the number of resistors between the two points is the same as the number of Nyquist intervals in the run (n). (V_0 will have to be delayed in order to accomplish this.) Then, provided that there is no loading on the intermediate points in the chain, it is clear that the voltages at these points will lie between V_0 and V_n according to a linear relation:

$$V_r = V_0 + \frac{r}{n}(V_n - V_0)$$

If these intermediate voltages, which are generated simultaneously, can now be read out sequentially, the desired aim will have been achieved.

4.3.2.1 Use of the Resistor Chain at Multiple Sampling Rates

In the data-reduction scheme under consideration, sampling of the signal is carried out at several rates. Provision must thus be made for feeding signals into various portions of the resistor chain. For example, if the standard sampling intervals, T, 3T, and 9T are to be employed, the arrangement indicated in Figure 4.13 must be used.

Whenever a sample arrives at the right-hand end of the chain from the common channel, there must be a sample at one (or more) of the points B, C, D, since T, 3T and 9T are the only permitted intervals. If the samples are fed to the chain from sources having low output impedance, a sample at the point D will have little effect on the voltage at C or B. The low impedance at C and D will, however, prevent correct interpolation over an interval of 9T, when linear interpolation between the points D and COM is required. Some means of disconnecting the intermediate sample feeds is thus needed.

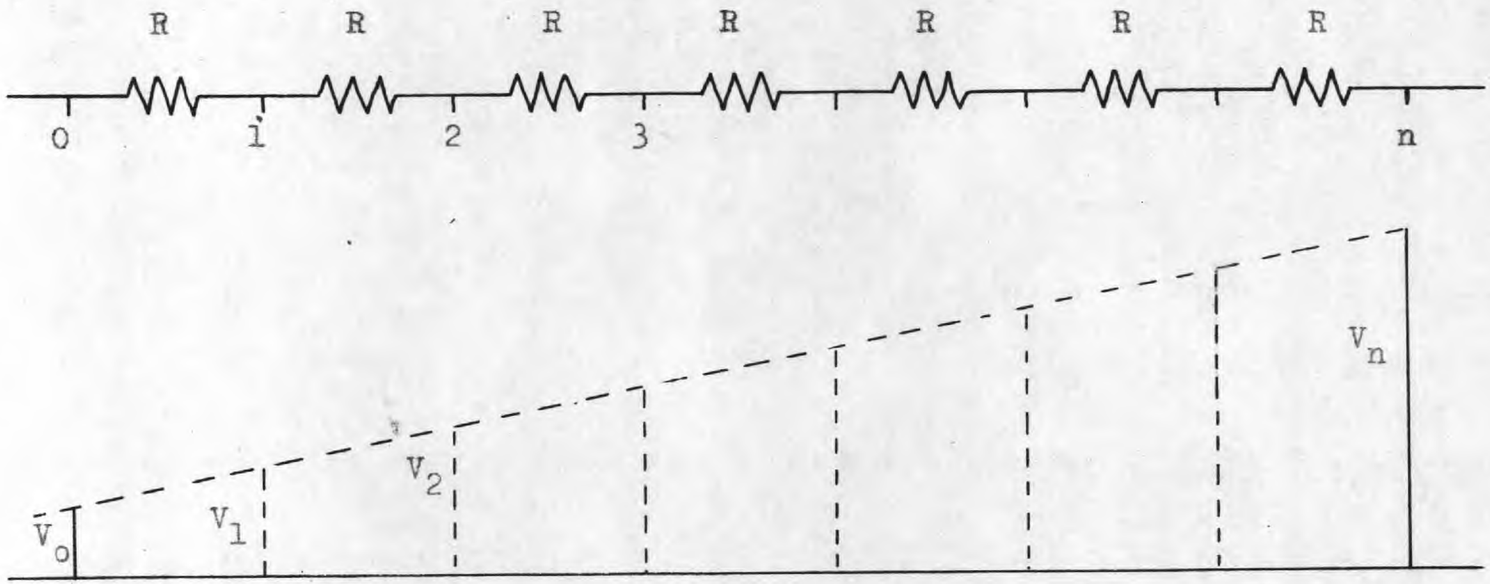


Figure 4.12. Principle of the resistor chain interpolator.

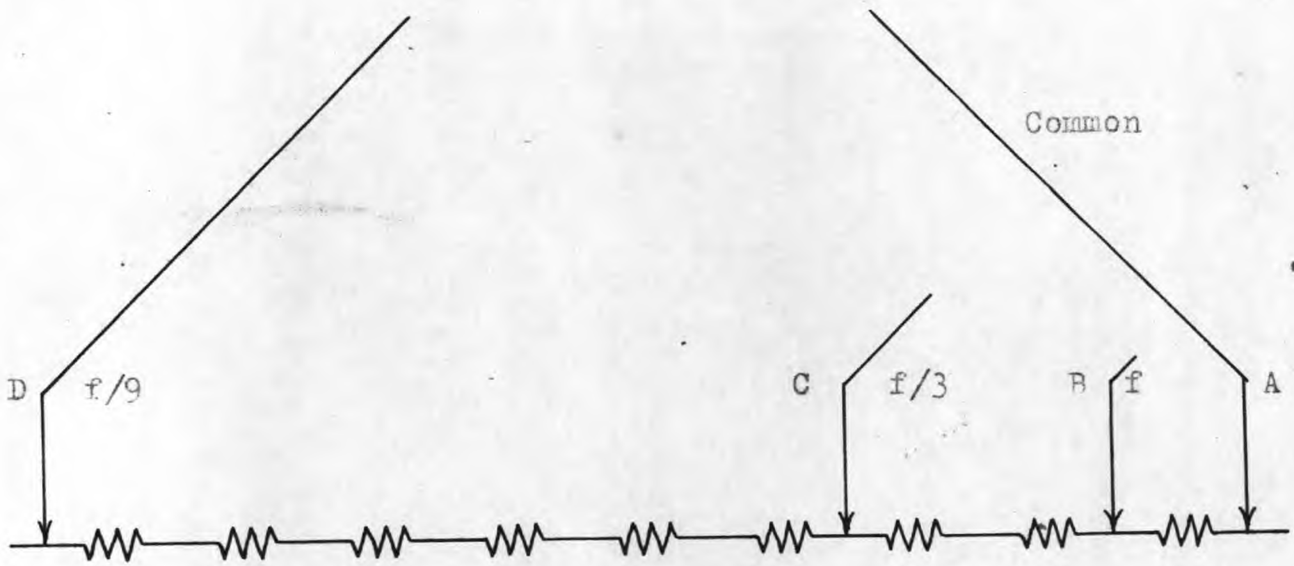


Figure 4.13. Feeds to the resistor chain for reconstruction from a sample train with three rates.

A circuit which fulfils this requirement is outlined in Figure 4.14(a). It will be seen to be a type of double clamping circuit. Normally, the two diodes at the feed point are biased to E_1 amd E_2 as indicated in Figure 4.14(b), E_2 is slightly more positive than the maximum voltage ever occurring on the resistor chain and E_1 is slightly more negative than the minimum voltage. Thus, both diodes are normally OFF and no loading of the chain occurs. The samples fed to the diodes are negative-going from E_2 in the case of D_2 and positive-going from E_1 in the case of D_1. Thus, when a sample is inserted, the point P is clamped at the signal voltage and linear interpolation results between P and any other point on the chain which is also receiving a sample.

4.4 The Video Delay Unit

The resistor chain interpolator requires the two samples from which intermediate values are estimated to be available simultaneously. A means of storing the earlier of the two until the later one has arrived is thus required. This is not the only problem, however. In order that the samples are routed to the correct points on the resistor chain, they will be required to operate some logic. Thus, some degradation of the samples can be expected, due to their passage through the various gates in the logic. Also the logic may not operate well if pulses of varying amplitude are to be used. A similar problem was encountered by Pine in the design of the prototype elastic encoder. His solution was to use constant amplitude pulses to operate the logic, modulation of the pulse amplitude being carried out later. The same procedure has been adopted here. Sampling of the video is carried out immediately before the samples are applied to the resistor chain. Thus four identical sampling circuits have been necessary. The problem of storage has not been completely removed, however, It is still necessary to sample two points of the waveform simultaneously but now it is necessary to store only the video signal and not samples of it. The bandwidth requirement of the storage device is thus reduced by about one order of magnitude.

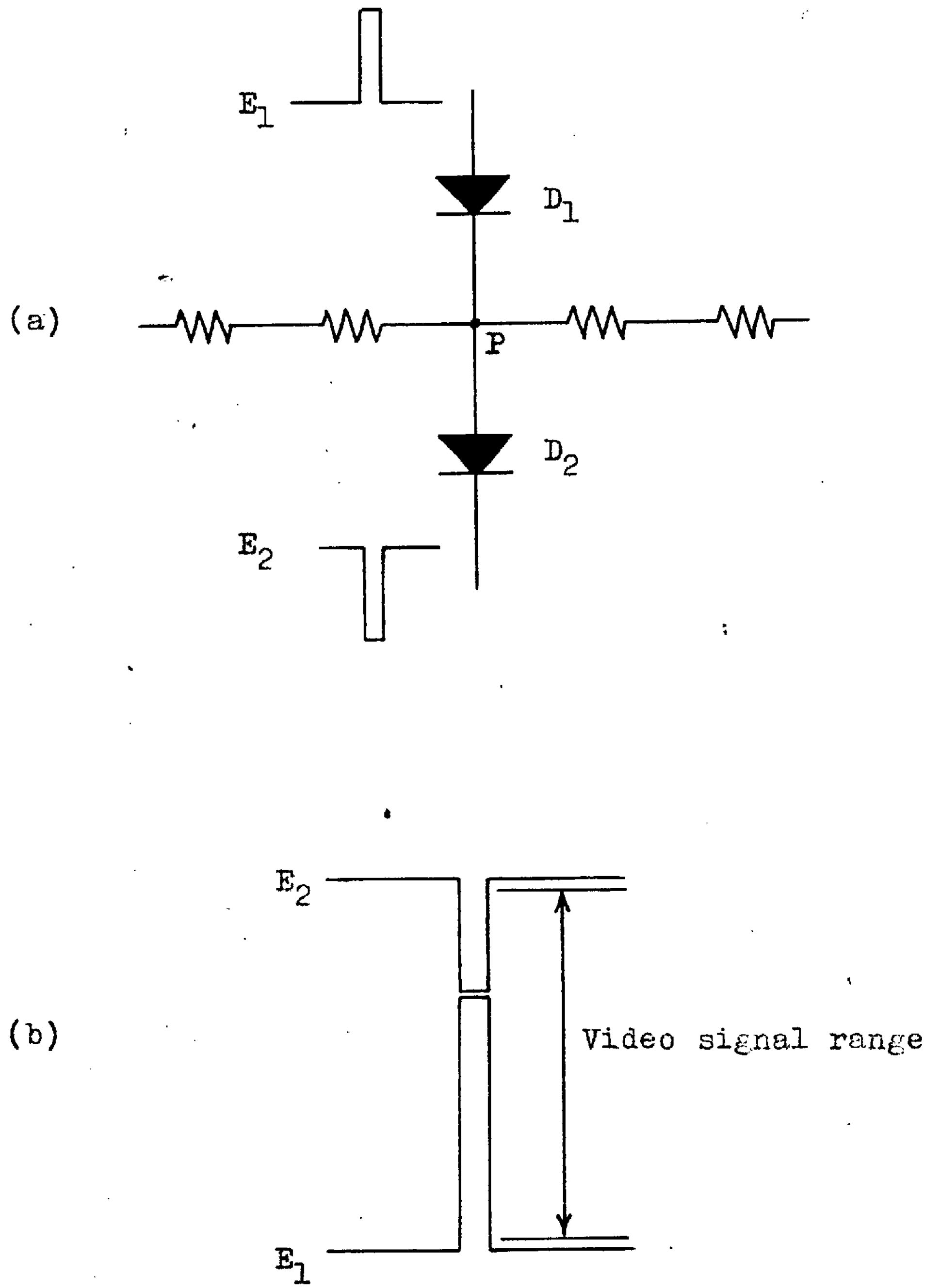


Figure 4.14. The double clamping circuit and explanatory waveforms

In a system having a maximum standard sampling interval of 9T, the video signal must be stored for this length of time. An electromagnetic delay line once again forms a convenient means of achieving this storage.

A video delay unit having a total delay of about 18T, with taps at intervals of T had been constructed by Kubba, the first four sections having been employed in his inferential run-length detector, This unit employed Telcon delay cable, Type Z14M, however. The shortcomings of this material have been described in Section 3.6.2. It was quite clear that a delay of 9T could not be satisfactorily obtained by its use. I therefore decided to construct a completely new video delay unit using Hackethal delay cable type HH 1500a. This unit took the form shown in Figure 4.15.

Although the delay cable was cut into nine sections each of 1/6μS, only four of the taps were equipped with amplifiers to pick off signals, since this was all that was necessary for the first-order reconstruction scheme. By the use of two transistors in the Darlington configuration, a high input impedance was achieved so that loading of the delay line was kept to a minimum and reflections were not troublesome. The output stage associated with each pick-off point was designed to drive 75Ω coaxial cable with a positive-going signal about 4v in amplitude. This figure was later relaxed slightly, the gain of the amplifiers being reduced in order to improve their stability. In order to achieve satisfactory low-frequency response the delay line was directly coupled to the collector of the driving transistor. A.C. coupling would have needed a vast capacitor here owing to the low characteristic impedance of the delay cable (1500Ω). A complete circuit diagram of the unit is given in Appendix 1.

The unit's performance can be assessed from a graph of its frequency response (Figure 4.16) and the waveform photographs shown in Figure 4.17.

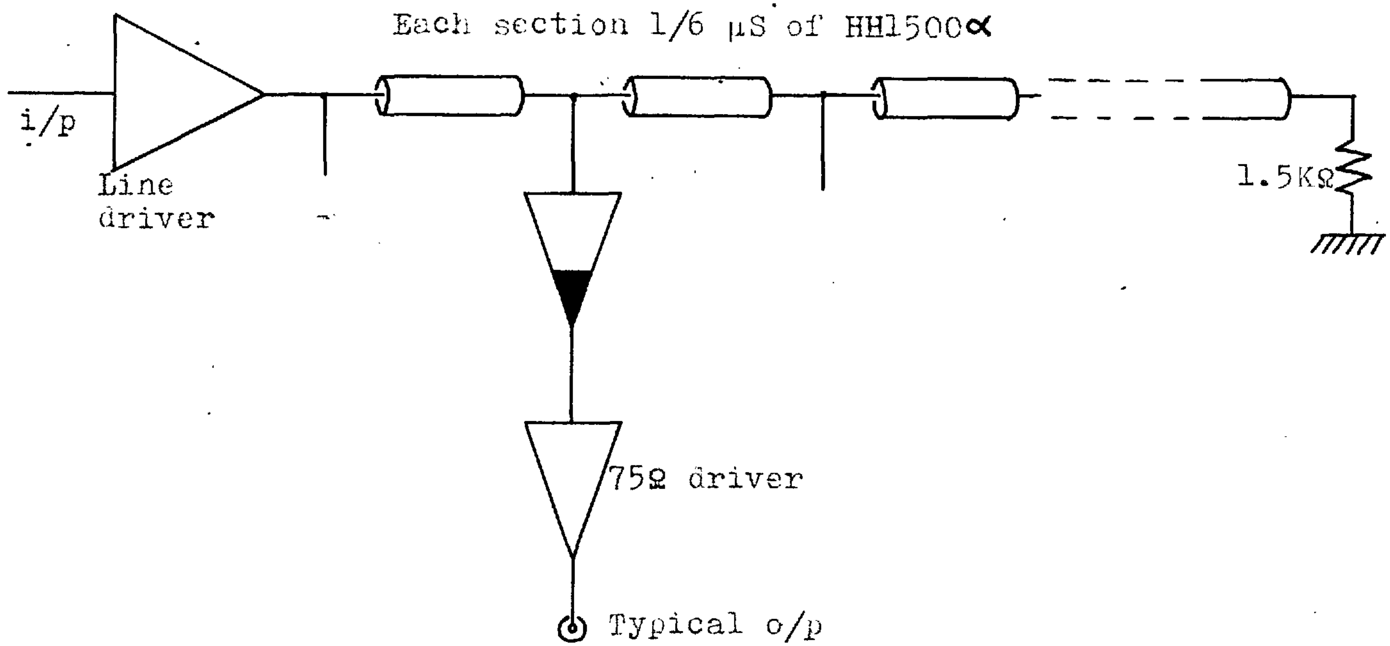


Figure 4.15. Block diagram of the video delay unit.

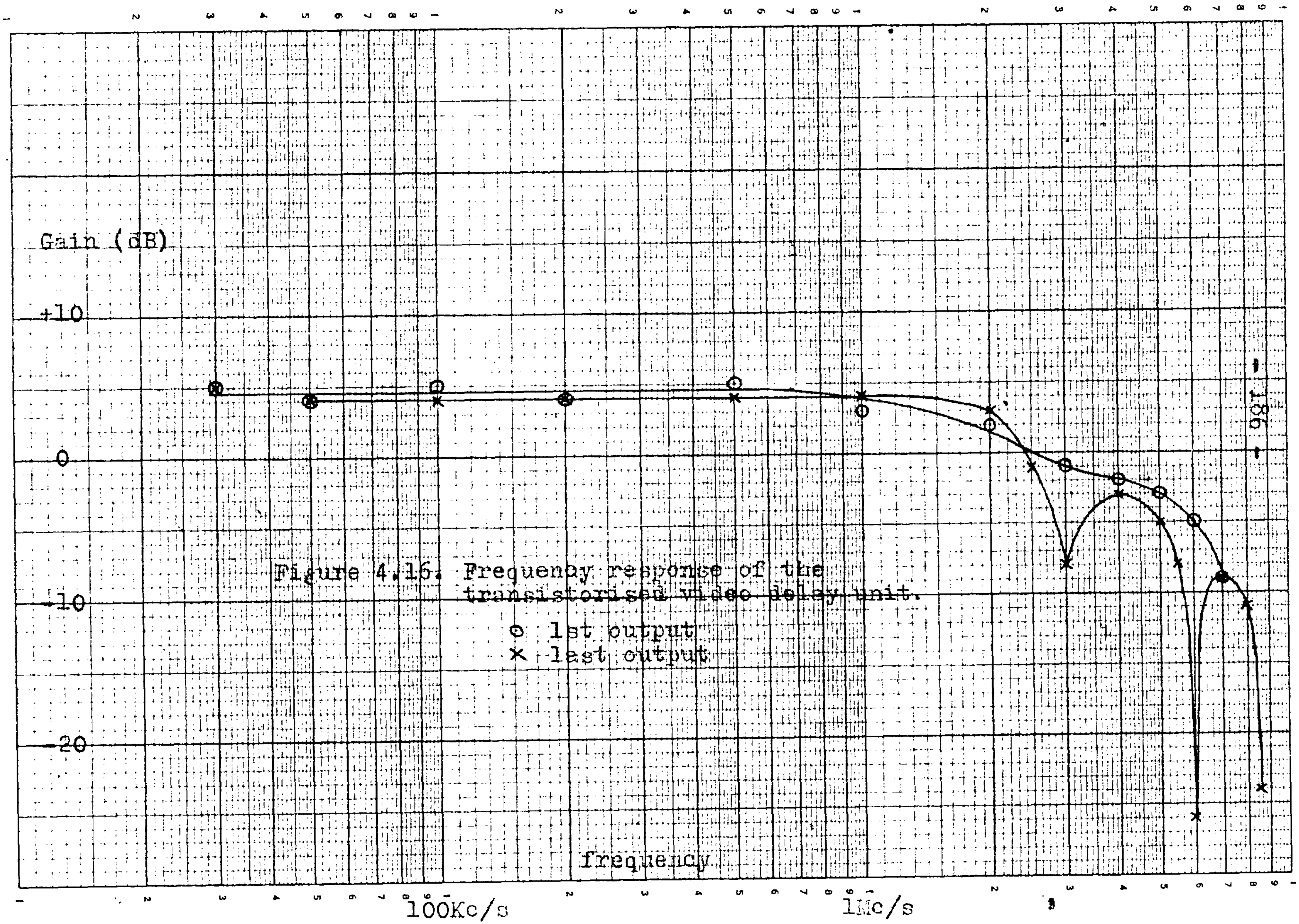


Figure 4.16. Frequency response of the transistorised video delay unit.

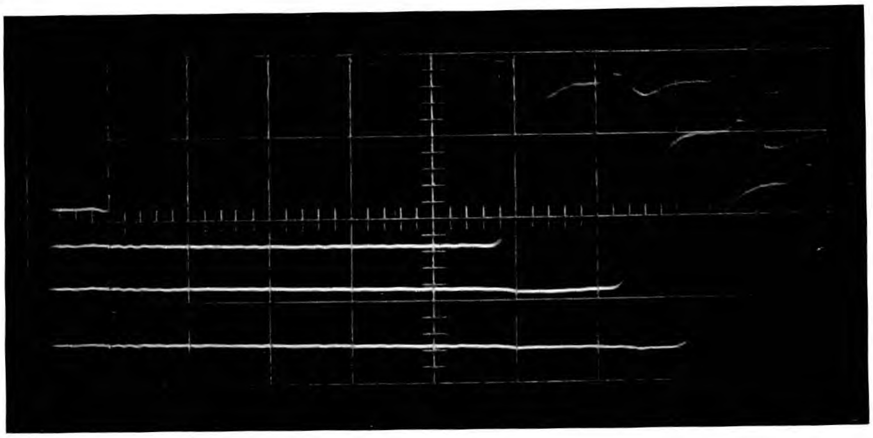


Figure 4.17. (a) Outputs of the Video Delay Unit at delays of 0, 6, 8 and 9 Nyquist intervals. 1V/major div., 0.2 μS/major div.

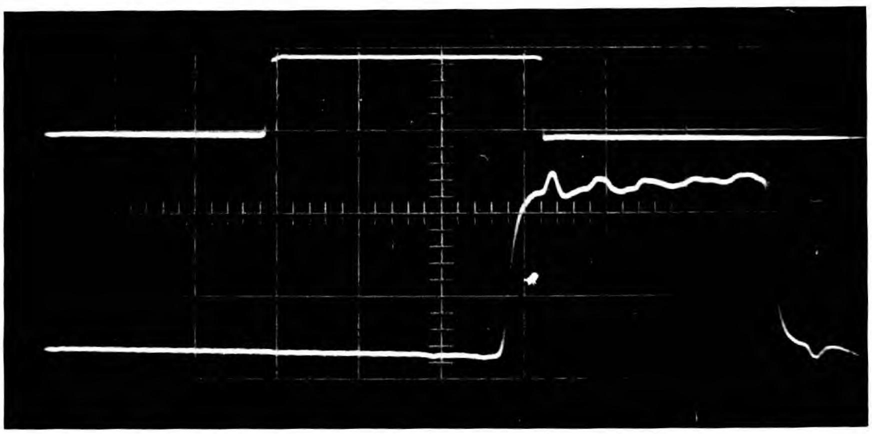


Figure 4.17. (b) Top: Input to the Video Delay Unit
Bottom: Last output of the Video Delay Unit (delay 9 Nyquist intervals)
1V/major div., 0.5μS/major div.

4.5 Read-in Logic

Before the two samples at the ends of a run can be applied to the resistor chain, it is necessary that they shall be available at the same instant. Further, since runs of different lengths have to be dealt with, switching arrangements must be available to apply the samples to the appropriate points in the chain. These two requirements can be satisfied by the arrangement of delay lines and gates shown in Figure 4.18.

It was realised that, owing to the inevitable distortions introduced by the delay lines and various gates it would be necessary to regenerate the pulses before sending them to the samplers. In view of this it was decided to gate the clock pulses after the run-length measuring logic and to send the gating signal from the SGU to the measuring logic. The revised arrangement is shown in Figure 4.19.

The signal from the SGU is of the form shown in Figure 4.20. The two levels can be conveniently denoted by '1' and '0', a '1' opening the gates in the interpolator, permitting the video to be sampled. Since certain run-lengths only are permitted by the sampling Interval Restrictor, there will be a '1' at one or more of the points B, C, D, whenever a '1' is at A. To correspond with these three points the three cases will be denoted b, c, and d.

Case b If, at some instant when there is a '1' at A, there is also a '1' at B, sampling at the Nyquist rate is being demanded and the f channel is energised since the AND gate G 1 supplies a '1' to the clock-pulse gate CPG 1. Clock pulses therefore pass to the shaping circuit of the f channel and thence to the sampler.

Case c When '1' 's are simultaneously at points A and C, two conditions are possible:

(i) A 2T blanking pulse has been generated by the SGU, i.e. sampling at f/3 is required. In this case G 2 and CPG 2 open and clock pulses pass to the f/3 channel.

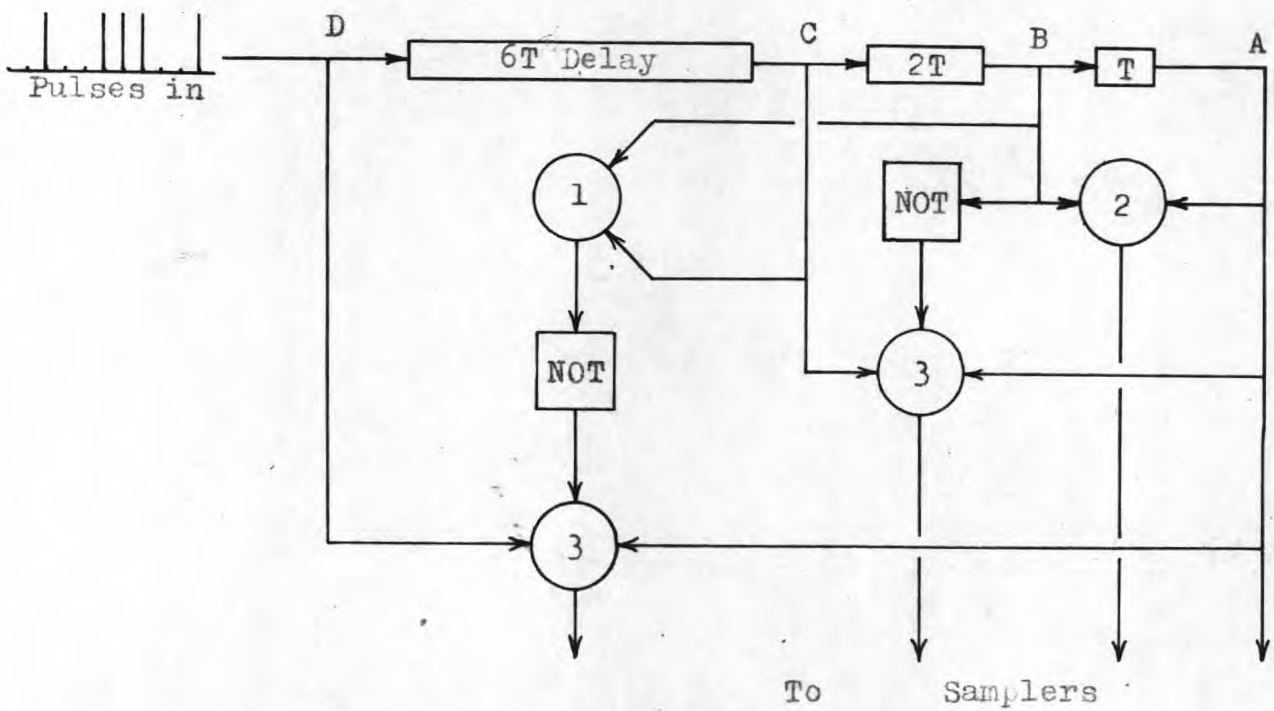


Figure 4.18. Read-in logic suitable for handling sampling pulses

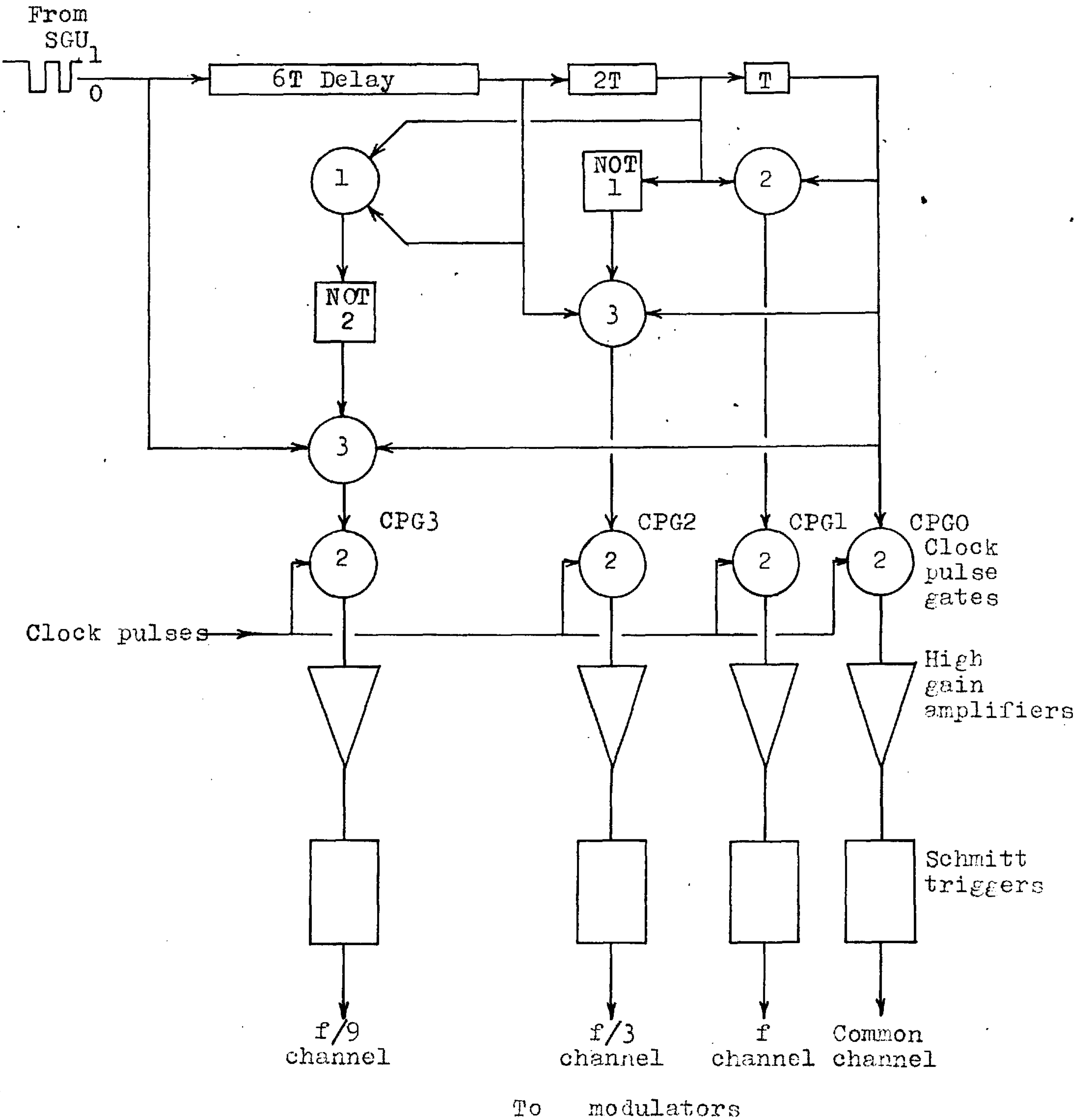


Figure 4.19. Revised interpolator read-in arrangements

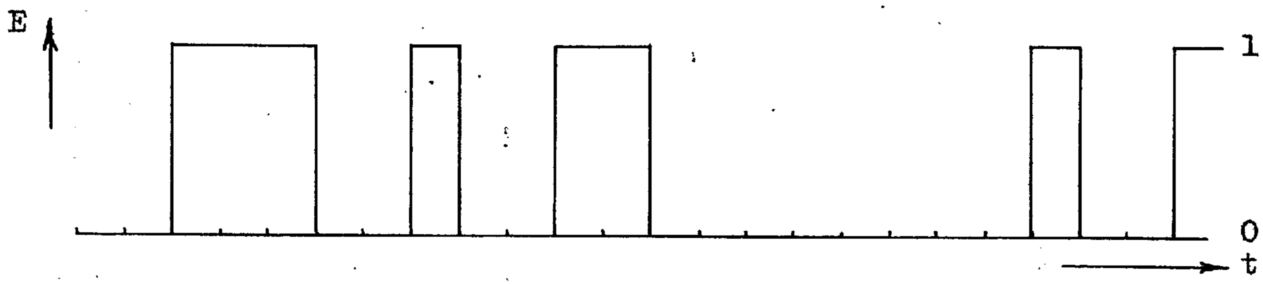


Figure 4.20. Form of the output signal from the Sample Gating Unit

(ii) Alternatively, there is a '1' at B (and incidentally there must also be a '1' in the middle of the 2T delay line since 2T gaps are not permitted) so we revert to case (b). In order to prevent pulses being sent to the f/3 sampler, thus upsetting circuits later in the Interpolator, NOT 1 is provided. A '1' at B results in a '0' at the output of NOT 1 so that G 2 and CPG 2 remain closed.

Case d In this case, the presence of 1's at A and D indicates that one of three conditions is fulfilled:-

(i) An 8T blanking pulse has been generated by the SGU indicating that sampling at f/9 is required. G 3 and CPG 3 therefore open and pulses are passed to the f/9 channel.

(ii) Again, there could be a '1' at B (together with some other 1's between B and D) so that case (b) applies. Under this condition, a connection through OR 1 feeds a '1' to NOT 2 which therefore has a '0' at its output and holds G 3 and CPG 3 closed.

(iii) A further possibility is that there is a '1' at C. Unless there is also a '1' at B, we revert to Case (c) and the other connection via OR 1 and NOT 2 holds G 3 and CPG 3 closed.

In all three of the above cases the '1' at A opens CPG 1 allowing pulses to pass to the COMMON channel. The logic therefore ensures that when there is a pulse from the COMMON channel there is a pulse from one, and only one, of the others.

Each of the clock-pulse gates was followed by a high-gain amplifier and a Schmitt trigger circuit which together produced pulses of a constant shape and amplitude, suitable for driving the modulators coupled to the resistor chain. The high-gain amplifier was biased so as to ignore signals below a certain level (Figure 4.21); an irregular baseline on the sample train from the clock-pulse gate thus caused no trouble. The use of a large gain allowed Schmitt circuits with appreciable hysteresis to be employed. Thus the Schmitts did not require tedious or recurring adjustment to maintain their hysteresis near zero.

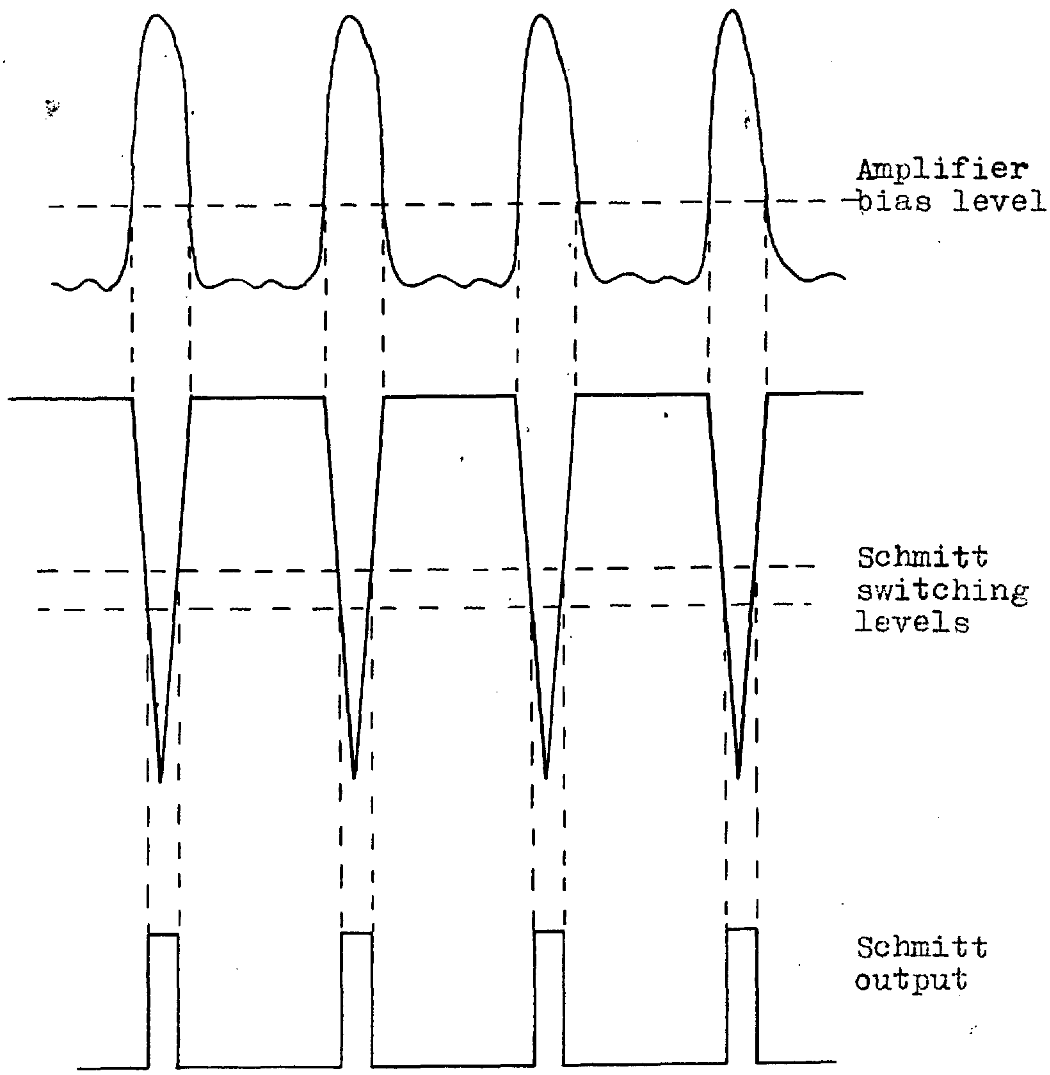


Figure 4.21. Pulse regeneration by high-gain amplifiers and Schmitt triggers

Outputs from both anodes of each Schmitt circuit were employed, as positive- and negative-going versions of each sampling pulse were required.

Waveform photographs illustrating the operation of the sections of the Interpolator so far described appear in Figure 4.22.

4.6 The Modulators and the Resistor Chain

As explained in Section 4.3, both positive-going and negative-going pulses are required by the double-clamping circuit which feeds the resistor chain. The pulses are not inverted versions of each other, however, but both have the modulating signal in the same polarity (Figure 4.23). Two modulator circuits are thus required at each point on the chain where signals are to be fed in. All the modulators require video of the same polarity; this is obtained from the various outputs of the Video Delay Unit, while the necessary positive and negative pulses are derived from the two anodes of each Schmitt circuit.

The conventional diode-resistor AND gate is linear in operation (provided the signals are large compared with voltage changes across a forward-biased diode) and this circuit forms the basis of all the modulators. (Figure 4.24.)

Cathode followers provide suitable low-impedance sources to drive the modulators and also give a high-impedance input. Thus a.c. coupling can be employed where necessary, without recourse to immense capacitors in order to maintain satisfactory l.f. response. D.C. conditions can also be conveniently set up by adjustment of the grid potentials of the cathode followers. Each modulator is also followed by another cathode follower, this providing the necessary high-to-low impedance transformation for driving the double clamp. The three potentiometers associated with each pair of modulators provide for:

(i) adjustment of the d.c. level of the train of positive pulses so that the pulses are satisfactorily modulated by the video.

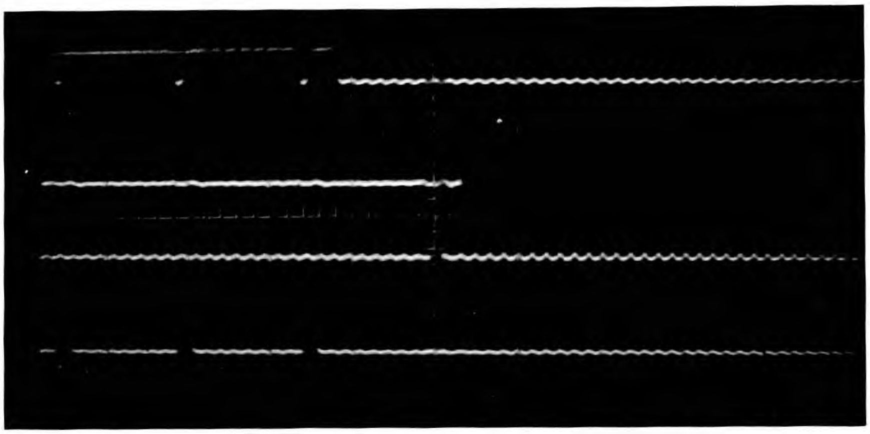


Figure 4.22. (a) Top: SGU output, 5V/major div.
2nd: f channel of Interpolator read-in section, 10V/major div.
3rd: f/3 channel, 10V/major div.
Bottom: f/9 channel, 10V/major div.
These waveforms illustrate the Interpolator's operation at the commencement of detail.
1μS/major div.

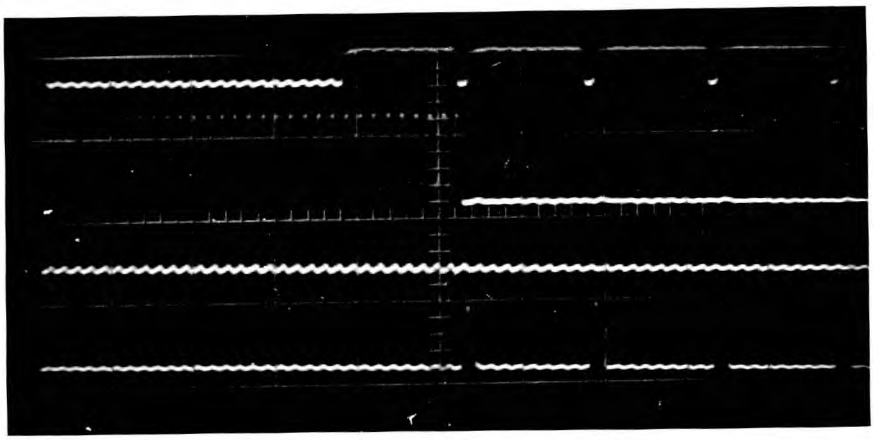


Figure 4.22. (b) As (a) but showing waveforms at the end of a region of detail.

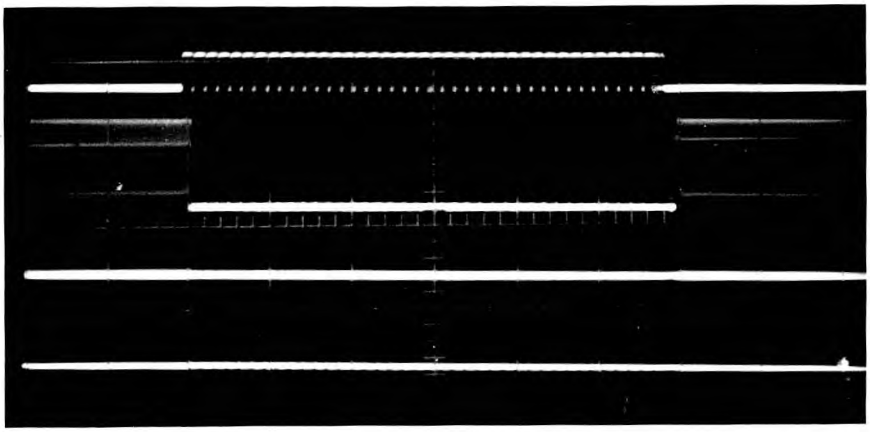


Figure 4.22. (c) As (a) but showing waveforms when detail first ends then recommences. Time scale changed to 10μS/major div.

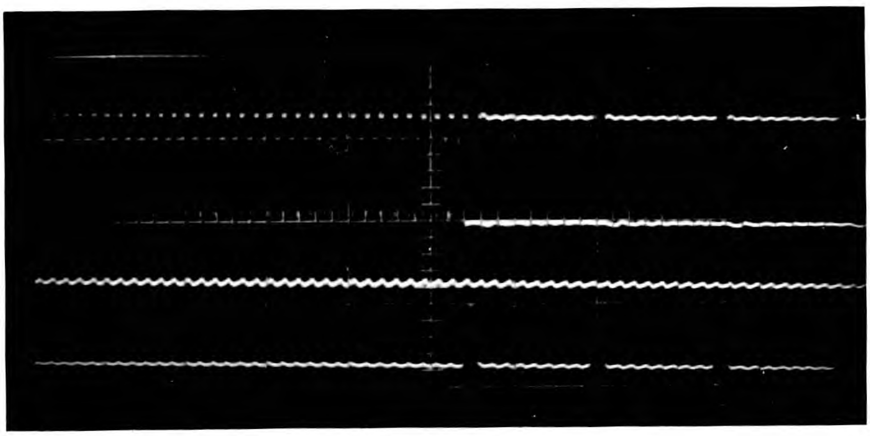


Figure 4.22. (d) Waveforms at the Common, f, f/3 and f/9 channels of the read-in section of the Interpolator when detail ends. 10V/major div., 1μS/major div.

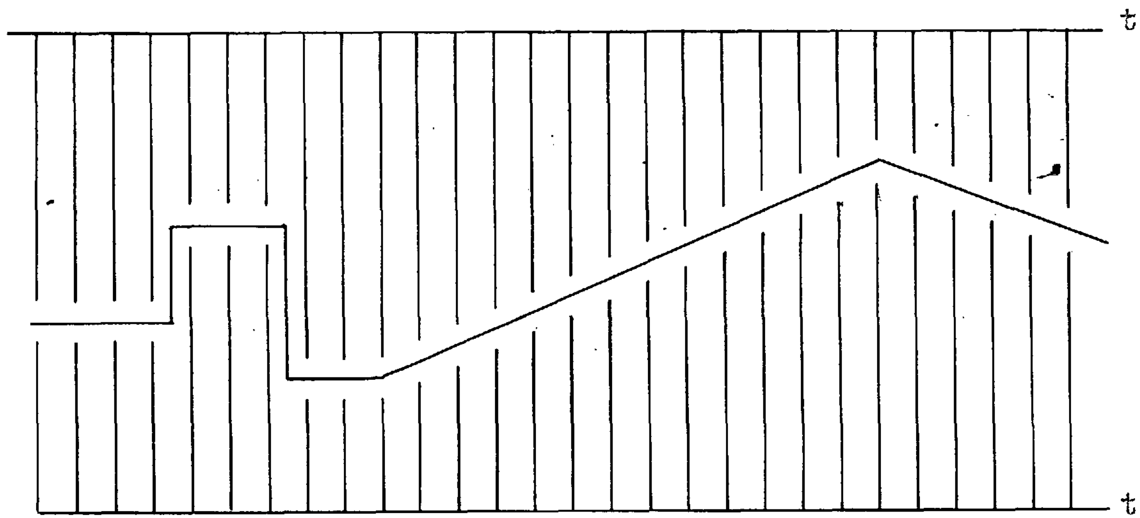


Figure 4.23. The two pulse trains and the corresponding video signal

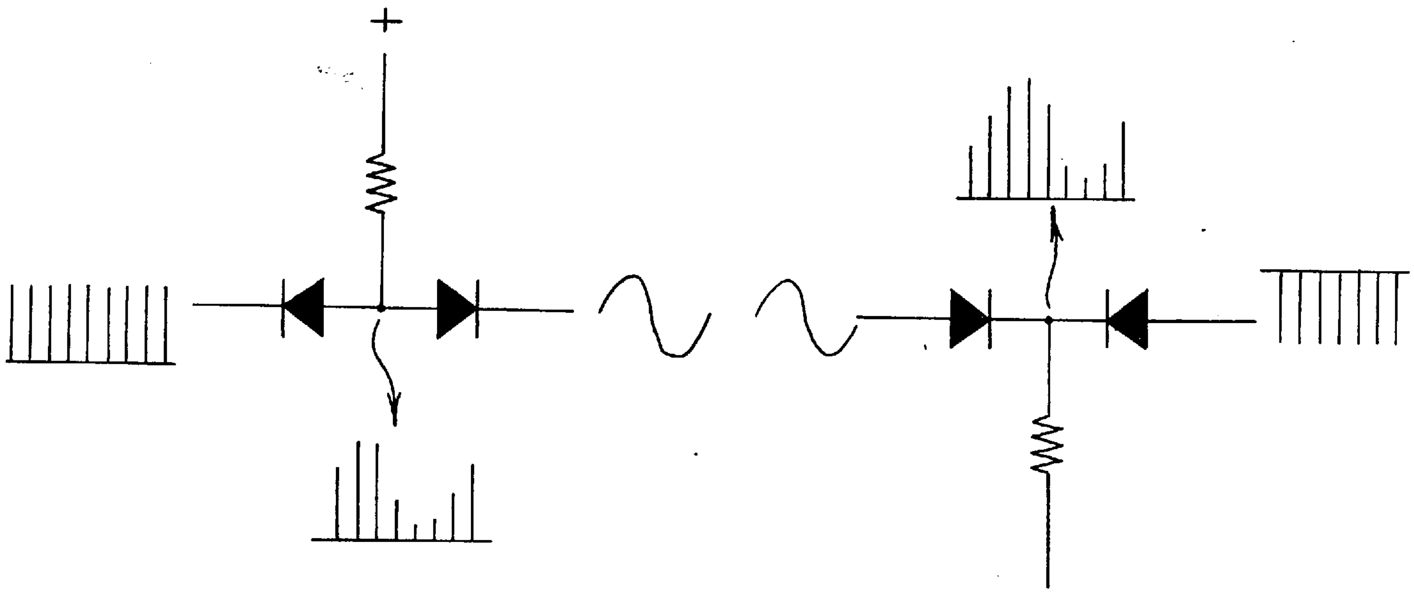


Figure 4.24. The modulators

(ii) a similar adjustment for the train of negative pulses.

(iii) adjustment of the d.c. level of the modulated positive pulses so that their tips are at the same voltage as the tips of the modulated negative pulses.

Figure 4.25 gives a block diagram of the driving equipment for one point on the resistor chain whilst a complete circuit diagram is shown in Figure 6.6.

The selection of a suitable value for the resistors in the interpolating chain is a subject to which considerable attention was given. Owing to the presence of stray capacitances and to the loading caused by the readout amplifiers it seemed desirable to use as low a resistance as possible. On the other hand the use of a high resistance is desirable in order to avoid errors due to loading of the diodes and cathode followers feeding the chain. The output impedance of a cathode follower is equal to the parallel combination of $1/g_m$ and R_L where g_m is the mutual conductance of the valve employed and R_L is the value of the cathode load resistor. For half of an E88CC having a mutual conductance of 12.5mA/V, $1/g_m$ is very much less than the value of R_L used (15 KΩ). Thus the output impedance is almost exactly $1/g_m$ (80Ω). The diode used in the double clamping circuit had a resistance, in the forward-biased condition of about 145Ω. Thus the impedance of a double-clamping circuit in the 'ON' condition is

$$\frac{80 + 145}{2} = 112.5\Omega$$

At the highest sampling rate (f), the loading on the double clamps is greatest since the number of resistors between the two clamps which are 'ON' is a minimum (one). It is reasonable to expect that some loading can be allowed here, however, since the eye is tolerant of amplitude errors in regions of fast transitions. Even so, we should clearly aim to make the value of each resistor in the chain large compared with 112.5Ω.

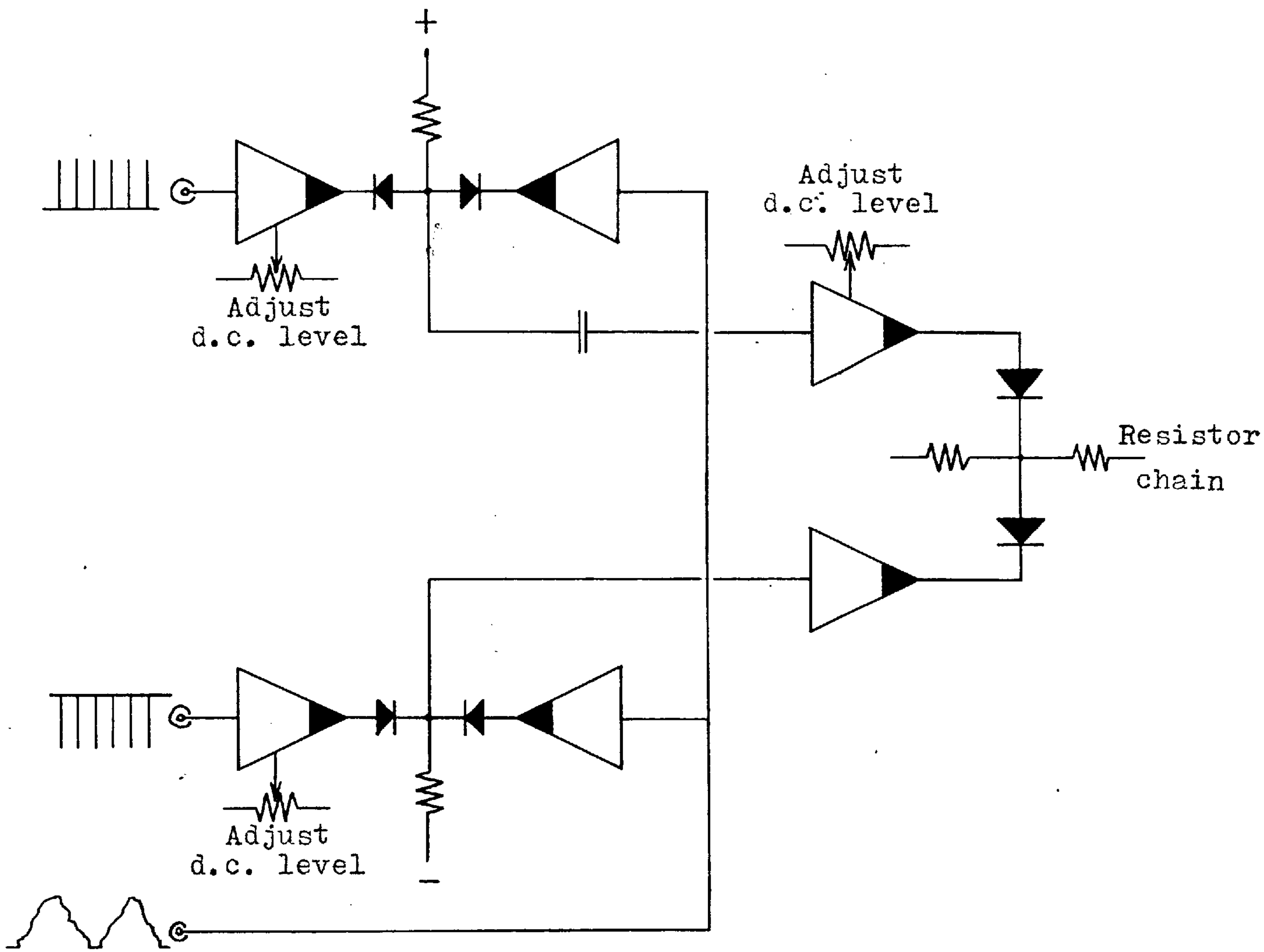


Figure 4.25. Driving arrangements for one point on the resistor chain

The maximum permissible resistor value is arrived at by considering sampling at the lowest rate (here f/9). The number of resistors now in circuit is nine (Figure 4.26).

Clearly, capacitive loading will be present at every point in the chain. Most serious, since it is furthest from the double clamps, is capacitance near the middle of the chain. Such a capacitance is indicated by C in Figure 4.26. If all other capacitances can be neglected, and if C is assumed to be 10pF, then the rise-time of the signal at P, for a step input will be given by

$$\tau \doteqdot \frac{C(4.5R + 112.5)}{2}$$

Now τ must be of the order of 20nS. Thus

$$R \doteqdot \frac{\dfrac{2\text{x } 20 \text{ x } 10^{-9}}{10 \text{ x } 10^{-12}} - 112.5}{4.5} = 863\ \Omega$$

As all the capacitances loading the chain except one have been neglected, this must be the maximum conceivable value for R. Consideration of the two conflicting requirements led to the adoption of a value of 220Ω.

4.7 Video Phasing Delay Unit

It is clearly important that the video arriving at a modulator has the correct phase with respect to the sampling pulses at the same modulator. If this condition were not observed, regions of high detail might be sampled infrequently, those of low detail being allocated a surfeit of samples. The two signals arrive at the modulator from quite different paths and Figure 4.27 shows the delay associated with each unit in these paths (Amplifiers, inverters, gates, etc. are here assumed to contribute no delay). It is clear that for correct sampling, extra delay of 9 Nyquist intervals (i.e. $1\frac{1}{2}\mu$S) is required in the video path.

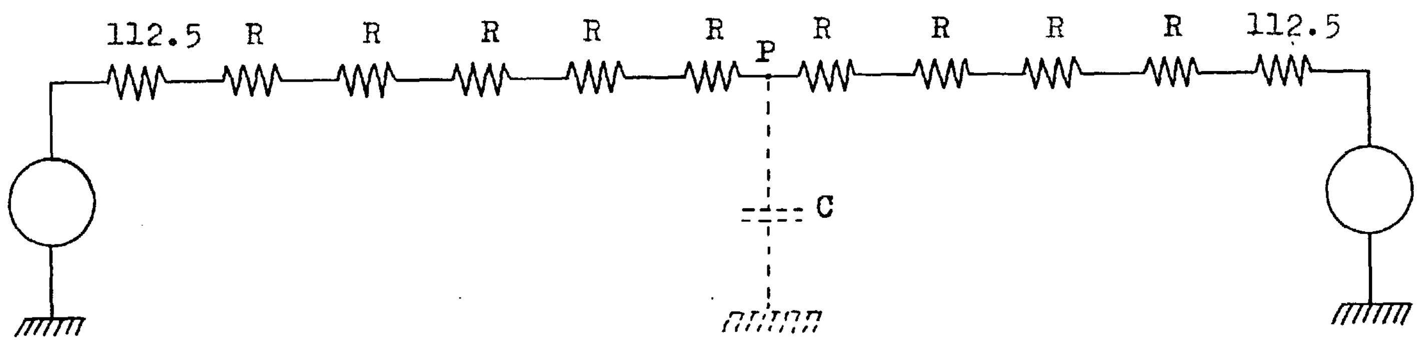


Figure 4.26. Calculation of the maximum allowable resistor value.

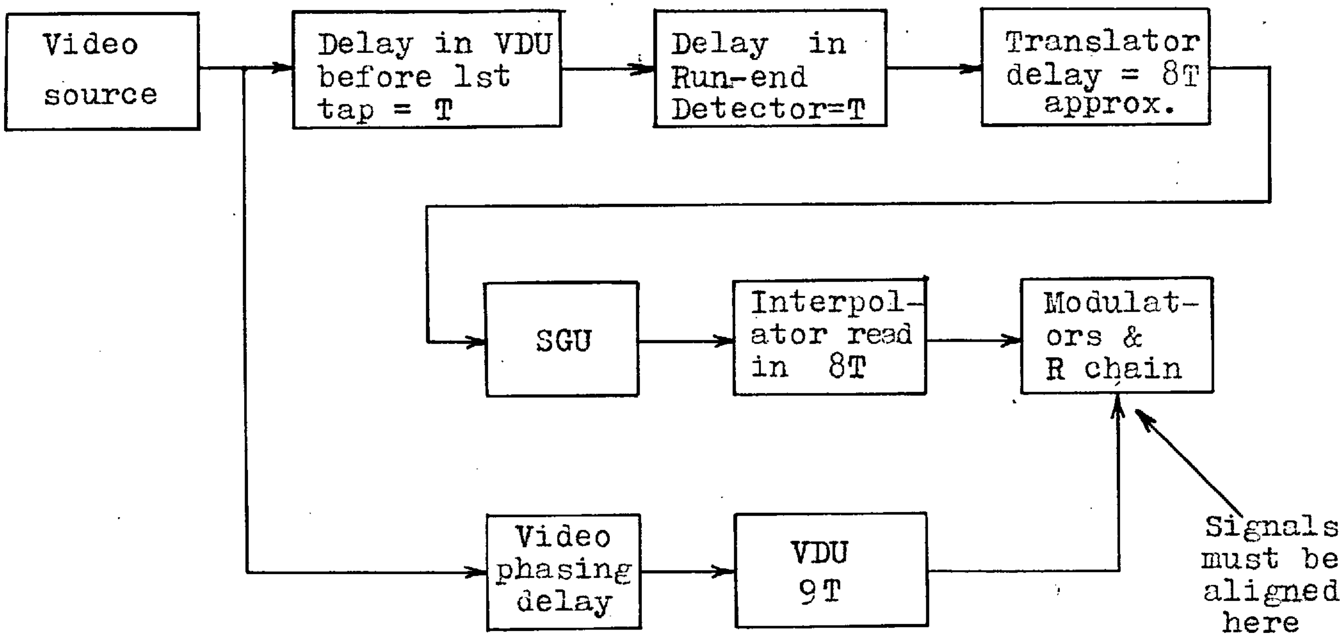


Figure 4.27. Approximate delays

A commercially available L-C delay line (AD-YU Type 605C2) was used to delay the video entering the video Delay Unit by the required amount. A 60-position switch allowed the delay to be adjusted in steps of 75 nS. In order to prevent the delay line loading the video loop which was also used for other equipment in the laboratory, a driving amplifier having a fairly high input impedance and an output impedance of 37.5Ω was constructed. Its circuit is shown in Appendix 1. This value of output impedance enabled the delay line to be driven at the tapping point whilst being terminated in its characteristic impedance (75Ω) at both ends. Frequency response curves of the amplifier and of the amplifier and delay line together are shown in Figure 4.28. As the amplifier is directly coupled, there is no deterioration in response at the low-frequency end of the spectrum; high frequency cut off starts at about 2Mc/s, however, and the response is 3dB down at 3.4Mc/s. Waveform photographs which demonstrate the high- and low-frequency responses and also the response to actual video comprise Figure 4.29. From part (a) of this figure it is clear that the 605C2 delay line degrades the rise-time of a fast pulse to about 100nS; the unit is therefore just adequate for video of 3Mc/s bandwidth. (b) shows that a 25c/s square wave acquires no tilt in this part of the system and (c) shows that video waveforms are identical (except for amplitude) at various points.

All the photographs and frequency response measurements were made with the delay setting used for normal operation of the equipment i.e. with 26 of the 60 sections of the delay line switched in. To find the correct delay, a pulse generator was run at line frequency (10101 c/s - see Appendix 2) and was synchronised to the line synchronising pulses available in the laboratory. The pulse generator output was used in place of video for the entire equipment. The edges of the pulse were detected by the Run-End Detector and this was adjusted so that a

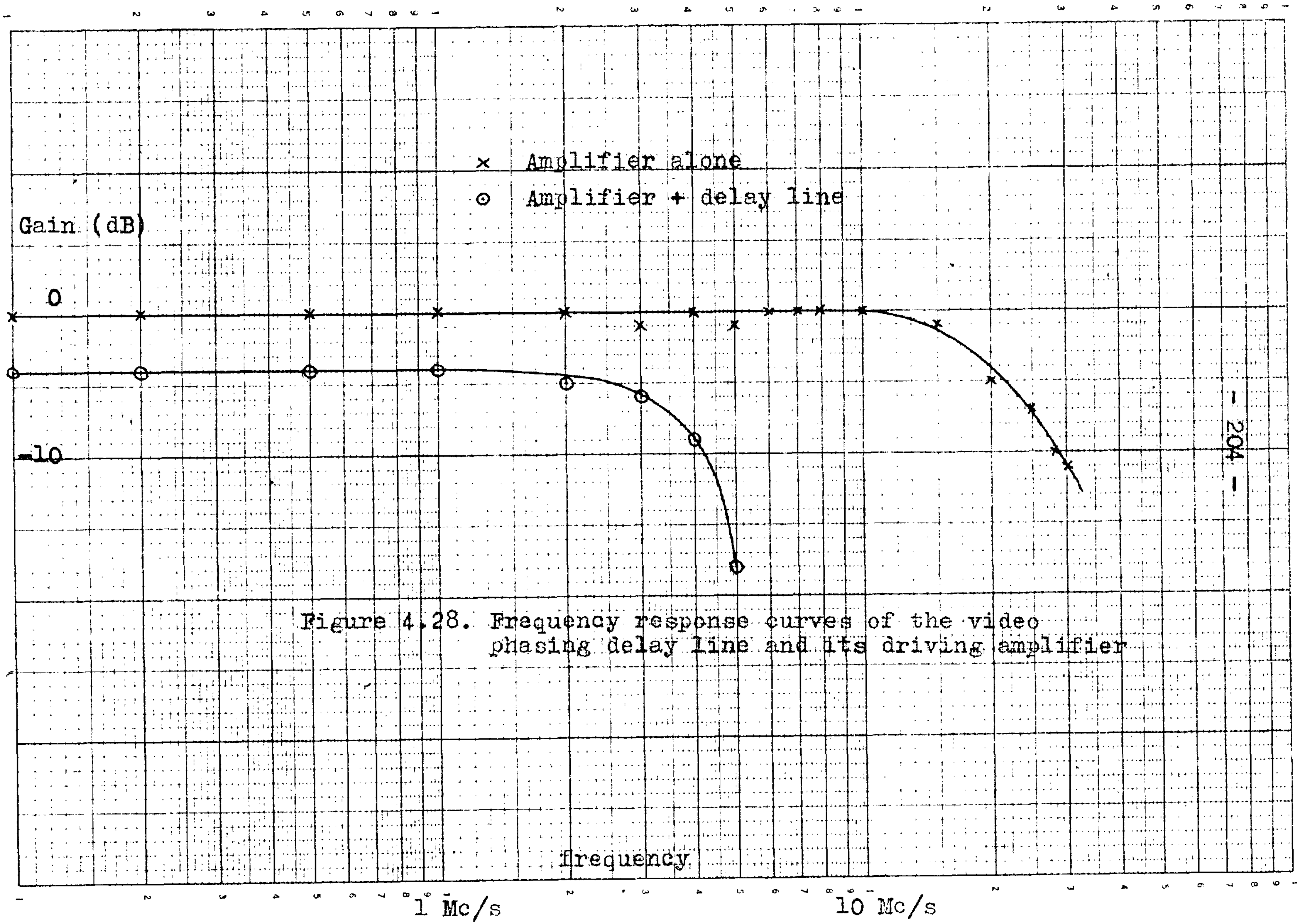


Figure 4.28. Frequency response curves of the video phasing delay line and its driving amplifier

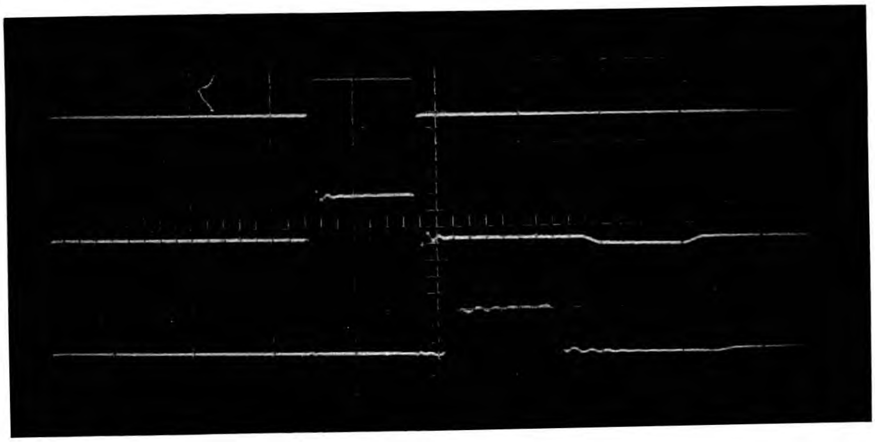


Figure 4.29. (a) Top: Input from pulse generator, 2V/major div.
Centre: Output of amplifier driving 605C2 delay line, 2V/major div.
Bottom: Output of 605C2, 1V/major div.
1μS/major div.

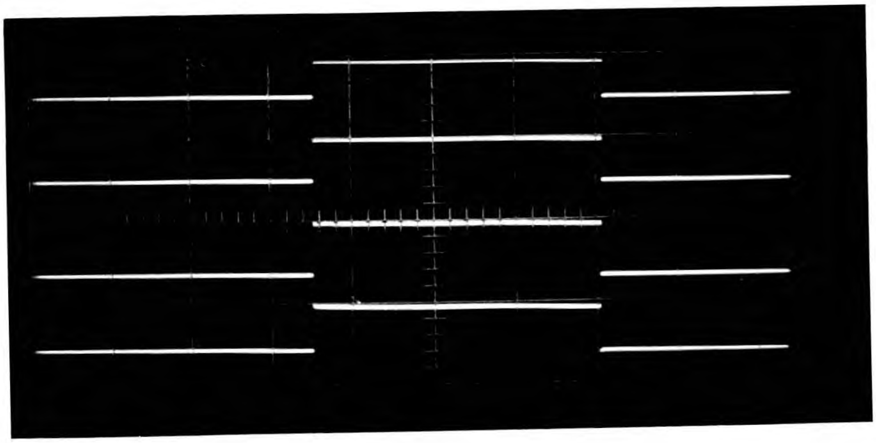


Figure 4.29. (b) Top: Input from pulse generator, 2V/major div.
2nd: Output of 605C2 driver, 2V/major div.
3rd: Output of 605C2, 1V/major div.
Bottom: Last output of Video Delay Unit, 2V/major div.
5mS/major div.

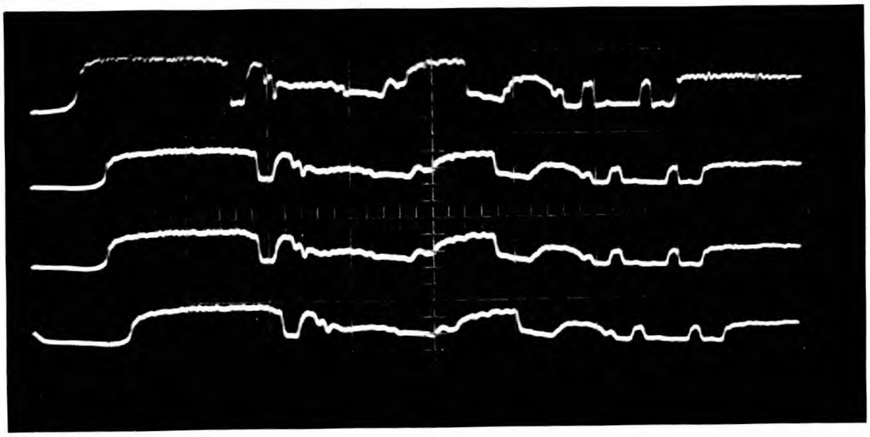


Figure 4.29. (c) Top: Original video, 1V/major div.
2nd: Output of 605C2, 1V/major div.
3rd: 1st output of Video Delay Unit, 2V/major div.
Bottom: Last output of Video Delay Unit, 2V/major div.
5μS/major div.

minimum of pulses (2) at the Nyquist rate was inserted by the f channel of the Interpolator read-in section (see Section 4.5). The modulated pulses emerging from the f channel modulator were then observed using an oscilloscope and the delay line was adjusted until one pulse sampled the upper level of the "video" while its neighbour sampled the lower level (Figure 4.30). The resulting delay of 1.95μS is slightly more than that predicted from Figure 4.27 but the estimate did not take into account delays in amplifiers, gates, etc.

4.8 Sequential Readout from the Resistor Chain

The way in which linearly related samples can be generated has been explained. (Section 4.3.2). Two problems remain, however.

(a) Between sampling instants, the voltage on the resistor chain is not defined. As all the diodes feeding it are reversed-biased, there is, ideally, no connection to the chain at any point. In any practical system, however, there must be some leakage so that the voltage on the chain can be expected to drift. This drift may be appreciable, particularly if sampling is at a low rate. It was therefore considered advisable to interrogate the resistor chain only at the instants when sampling occurred. This was achieved by the use of AND gates fed from the resistor chain by high input-impedance amplifiers. (cathode followers). Once again diode-resistor AND gates were employed in order to achieve linearity.

(b) On the resistor chain, all the samples concerned with a particular run are generated simultaneously. The readout system must therefore incorporate some form of parallel-to-serial conversion. Two possibilities are available for analogue signals:

(i) Storage on a set of capacitors, readout being accomplished by connecting each capacitor to the output point in turn.

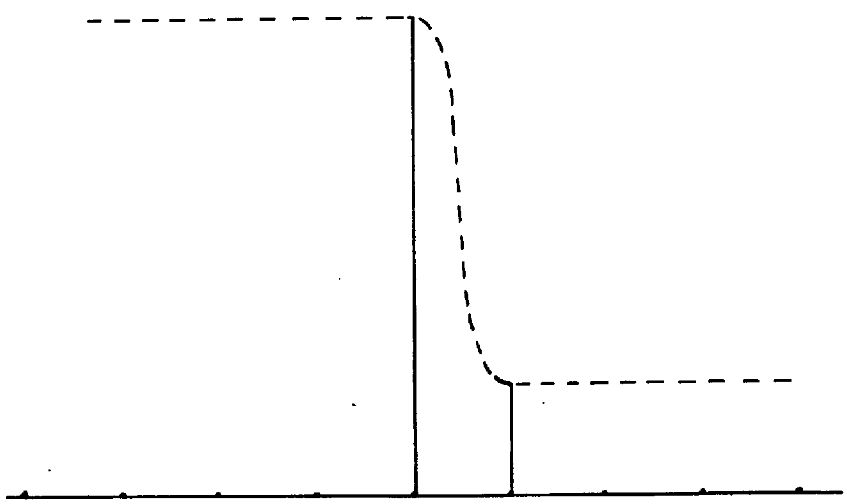


Figure 4.30. Correct phasing of pulses to sample video

This was expected to be expensive and complicated since each capacitor would require a gate to control the read in and another to control the readout. In addition, some sort of ring counter would be needed to operate each of the readout gates in turn.

(ii) Delay line spacing of the samples. The two possible configurations for this method are shown in Figure 4.31 (a) and (b).

The layout of Figure 4.31(a) is attractive since each regenerated sample passes through only one driving amplifier. The possibilities for the combining amplifier include a summing circuit and a linear OR gate. This arrangement promises to be easy to adjust since all the channels are independent; alterations of gain and d.c. level in one channel should not affect the others in any way.

Unfortunately this arrangement requires considerably more delay cable than does the other (36 Nyquist intervals, as against 8) and, in view of the considerable cost of the cable, it was decided to use the alternative configuration. Here, each line-driving amplifier also acts as a 2-input combining amplifier receiving one signal from the previous section of delay line and one from the resistor chain via a linear AND gate. The set-up is somewhat similar to that used by Pine[44] in his Elastic Encoder. The requirements are somewhat more stringent, however, since

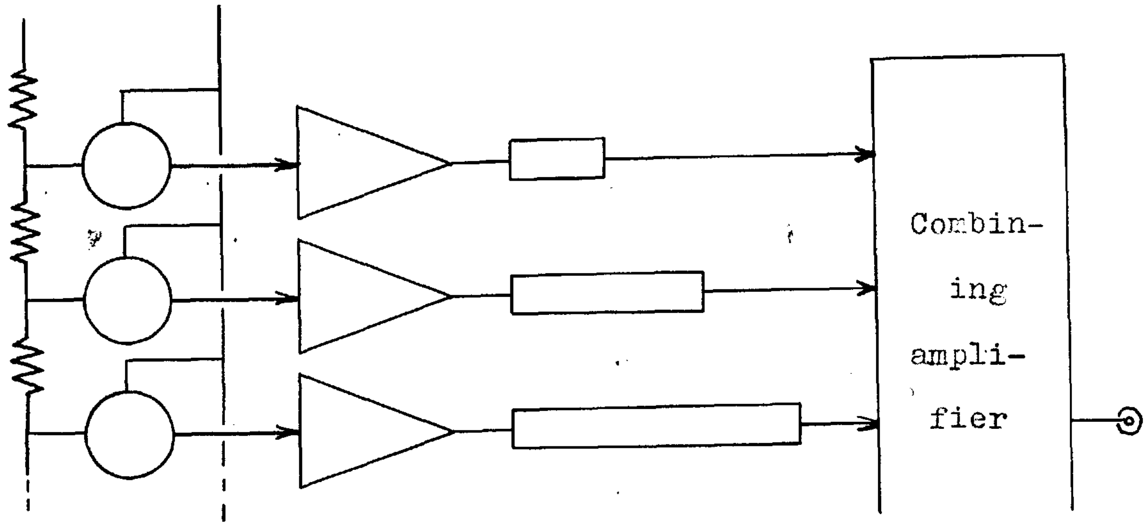
(i) the results are to be presented as pictures.

(ii) the delay between successive pulses is only 1/6μS and not 1/3μS as with Pine's gear; pulses will thus have to be narrower in order to prevent overlap.

(iii) there will be 9 pulses to be combined, instead of 6.

The number of amplifiers required made economy an important consideration in their design. Although transistorised units would have been easier to construct and use I decided to use valves

(a)



(b)

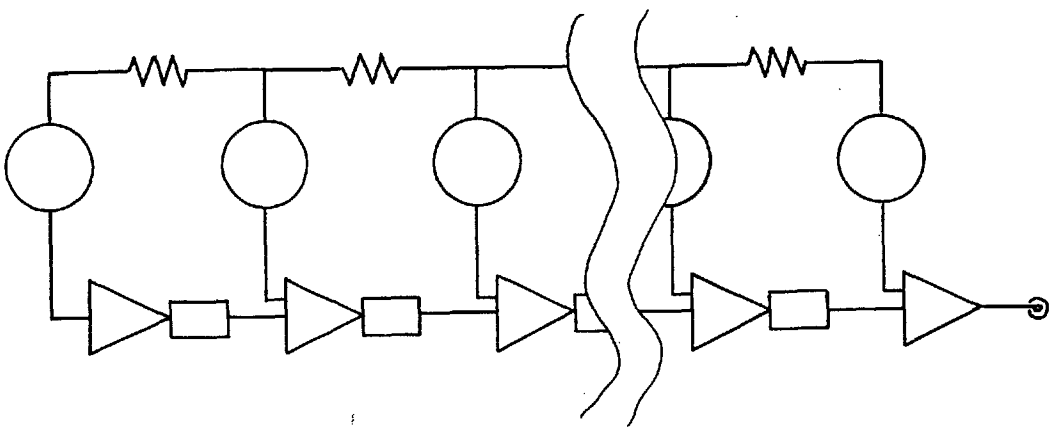


Figure 4.31. The two possible configurations for the readout equipment

since

(a) suitable power supplies were available for valves but not for transistors.

(b) the output of the resistor-chain unit was at a d.c. level of about +125v. Even if a.c. coupling to transistorised amplifiers had been used, elaborate precautions to protect the transistors against surges would have been necessary.

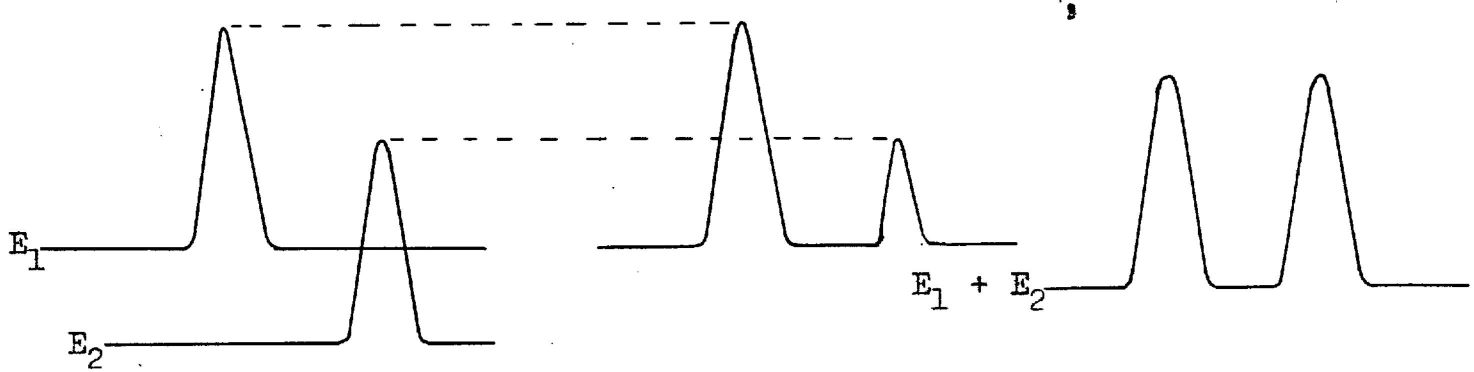
Each amplifier (except the first and the last) is required to combine two signals and to drive a section of delay line, the signals at the inputs and the output being of the same polarity. To keep within reasonable expenditure and the limits of the available power supplies, two valves only were used for each amplifier, various combinations of double triodes and pentodes being investigated. (A double triode is here classed as one valve.)

As with the configuration of Figure 4.31(a), either a summing circuit or a linear OR gate can be used to combine signals. It was realised that the summing method would allow discrepancies in the d.c. levels of the two input signals to be disregarded as shown in Figure 4.32. This was thought to be very desirable; thus further examination of the technique was made.

4.8.1 Summing Amplifiers

The restriction of the circuit to two valves only limited the possible arrangement somewhat; the circuit used by Pine could not be repeated since this required three pentodes. As it was thought desirable to use a pentode as a line driver, the problem was reduced to that of deciding on a suitable valve and circuit for combining the input pulses.

Adding circuits can be formed by using two valves with a common anode load, two valves with a common cathode load or one valve in the operational configuration. The use of a common cathode load is restricted to small-signal amplifiers only since



(a) Pulses to be combined (b) OR-gate combination (c) Combination by addition

Figure 4.32. Comparison of the two pulse combining methods.

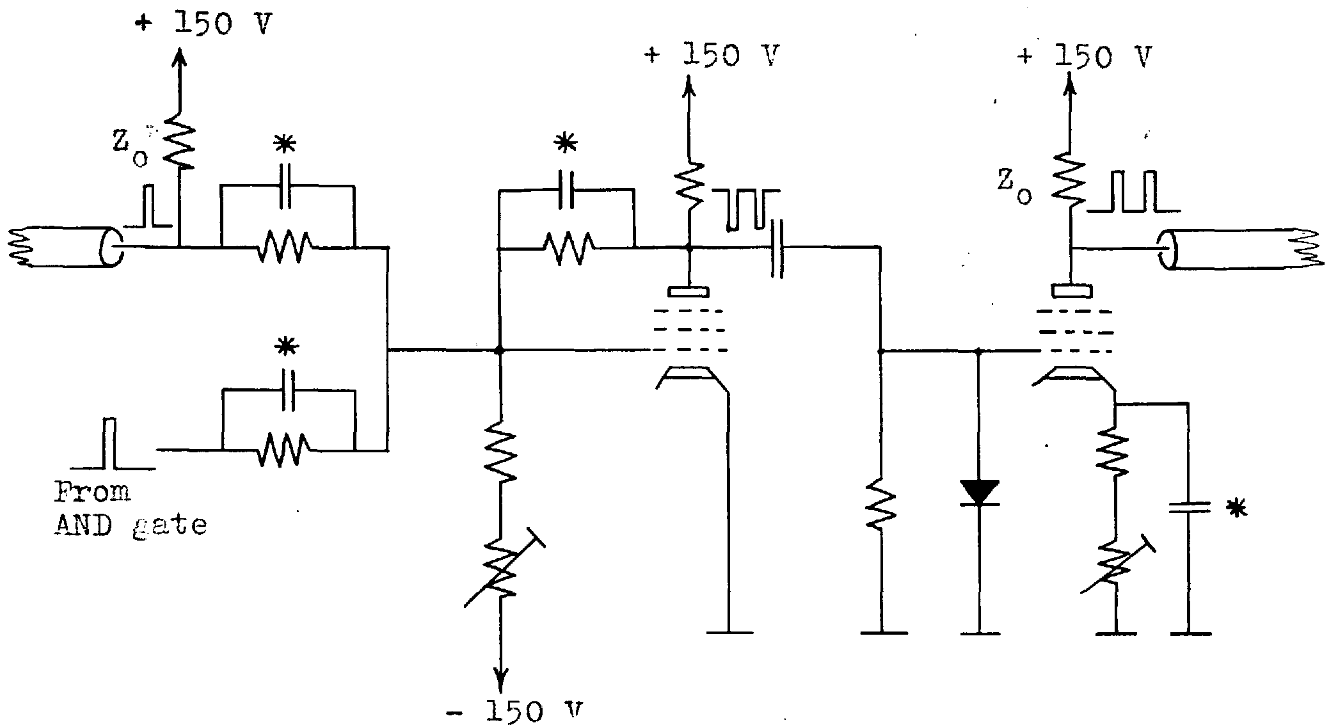
non-linearities are introduced if the two inputs differ by more than a small amount. The common anode load circuit does not suffer from this restriction but is best used with pentodes since the current in these valves is almost independent of anode voltage under normal operating conditions. The operational amplifier circuit was therefore chosen. This has the additional advantage that the inputs can be directly coupled even when the signal has a large d.c. component, as was the case here. Although it was originally hoped to use direct coupling throughout the whole chain of amplifiers, some thought suggested that this procedure would make thermal stabilisation very much more critical. A.c. coupling between each adding amplifier and the following line driver was thus installed. The circuit is shown, in outline form, in Figure 4.33.

It should be noted that the delay required from each section of delay line was rather less than 167 nS in this section of the Interpolator. The delay in each amplifier was found to be of the order of 30 nS so the delay due to each delay line with its amplifier was measured and adjusted as necessary.

4.8.1.1 Performance of the Interpolator with Summing Amplifier Readout

Setting up of the equipment involved adjustments first with continuous sampling at f/9 and then with a pulse generator replacing the Detail Detector. It was found convenient to operate the pulse generator with a repetition rate equal to line frequency (10.101 Kc/s), the pulse width being set so that approximately half the picture was sampled at the full rate and half at f/9. When video was needed this was obtained from a Flying Spot Scanner. (Described in Appendix 2.)

Operation of the Interpolator was fair with sampling at a constant rate, although setting up was difficult and tedious. This involved the adjustment of nine amplifiers so that the gain from the input of each amplifier to the end of the chain was the same in all cases.



* Adjusted for best h.f. response

Figure 4.33. Outline circuit of the summing amplifier

The use of several sampling rates showed that summing could not be expected to be satisfactory and led to the use of OR gates to combine the reconstructed pulses.

Incorrect operation is caused by the broadening of pulses as illustrated in Figure 4.34. If a pulse from a certain point on the resistor chain is broadened by passing through a delay section of limited bandwidth, the combination of this pulse with the pulse from the next point on the resistor chain will have a greater amplitude than is correct.

This effect was found to result in white flashes appearing on the reconstructed picture in areas where sampling was at the full rate. The equipment was so constructed as to allow continuous sampling at the full rate if desired; under this condition the white flashes would disappear as the a.c. coupling of some of the circuits allowed conditions to change slowly.

4.8.2 OR-Gate Readout

The use of a linear OR gate for the mixing of a number of pulse trains overcomes the limitation described at the end of the previous section, for the output signal now follows the input signal which has the greatest amplitude at any instant. This is illustrated for the case of two input signals in Figure 4.35.

Unfortunately, discrepancies in the d.c. levels of the two input signals are not disregarded by this circuit (see Figure 4.32). It was therefore necessary to be able to adjust the d.c. level at one input of each of the OR gates. Adjustment of the gain of each stage was effected, as before, by variation of the cathode resistor of the line driver.

The limited number of valves which could be employed, together with the fact that the overall stage between successive sections of the delay line had to be non-inverting, restricted the available configurations to that shown in Figure 4.36.

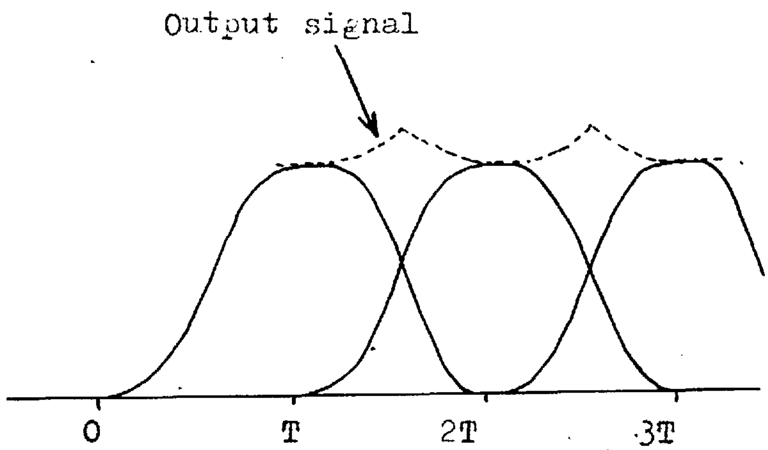


Figure 4.34. Incorrect output amplitude due to addition of broadened pulses

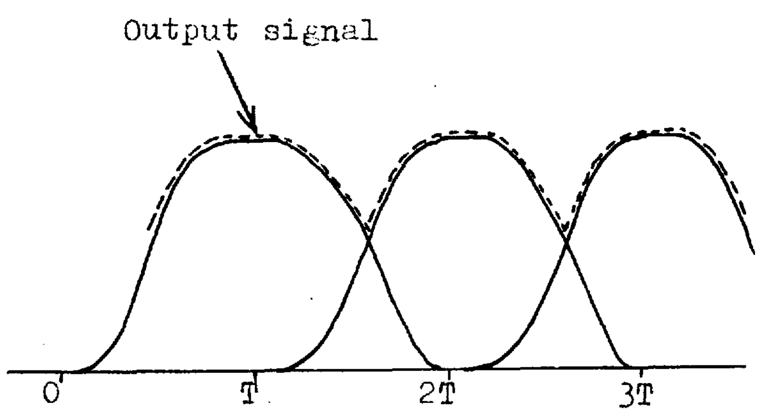


Figure 4.35. Following of input signals by the output signal in the case of OR-gate combining

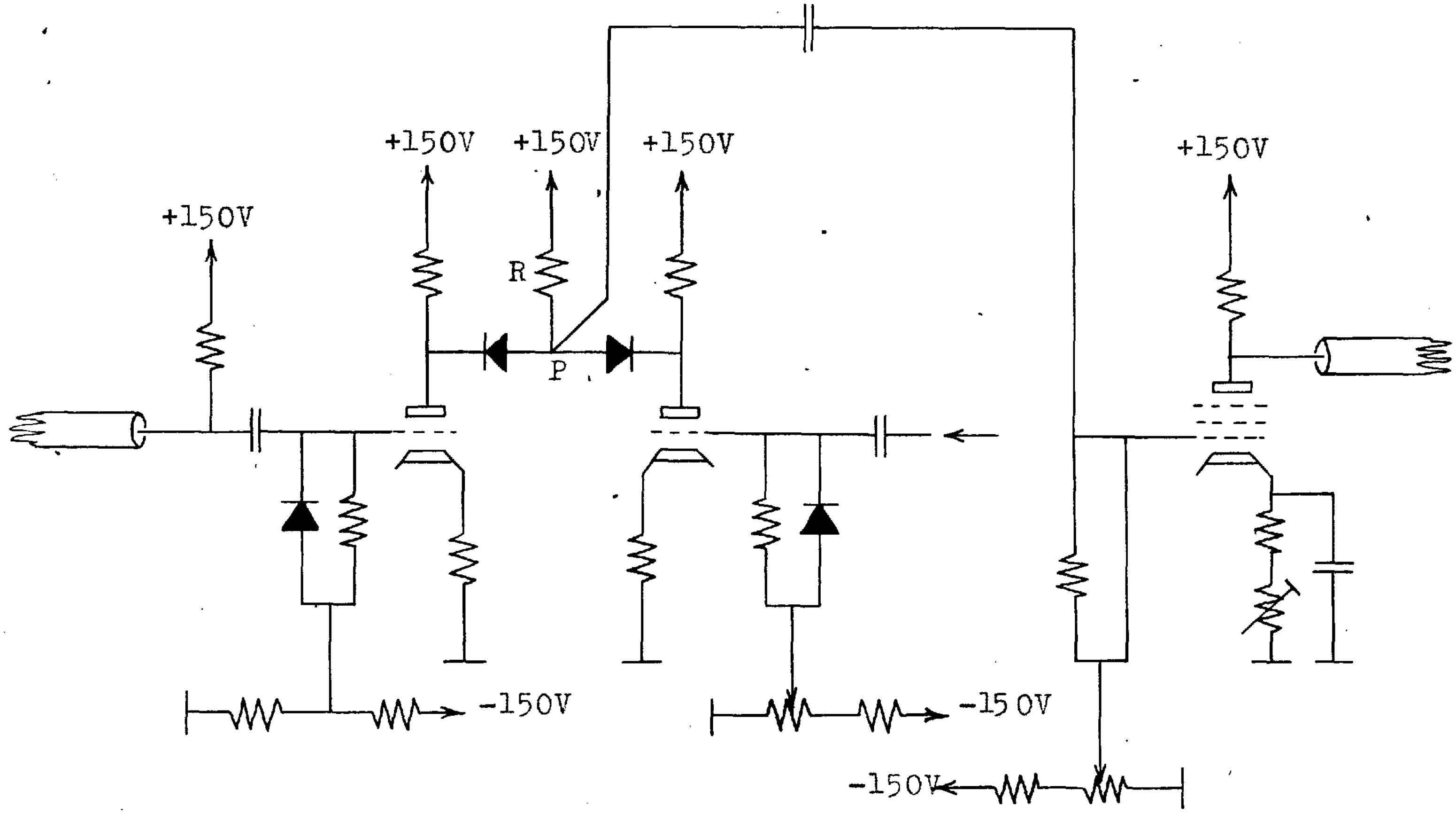
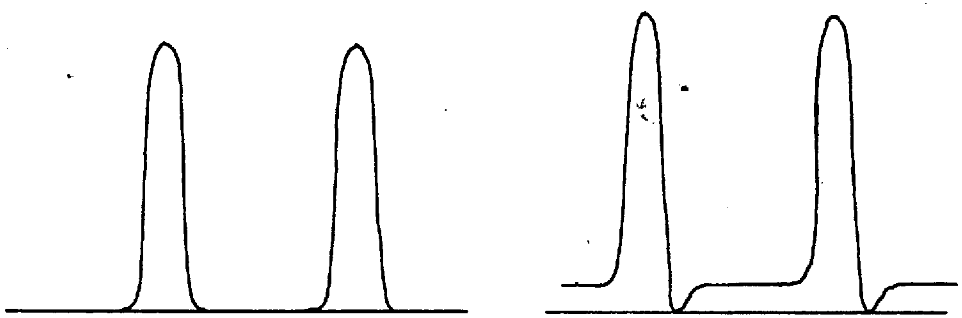


Figure 4.36. Outline circuit of the OR-gate combining amplifier

In order to obtain a sufficiently fast rise time at the point 'P' it was necessary to limit the value of the gate resistor R to 1 KΩ. Ideally, the output impedance of the points driving the gate should have been much less than this, therefore. The limited number of valves available precluded the insertion of cathode followers, however, so that the output impedance of each half of the double triode was dictated by the size of anode load resistor, this in turn being chosen to give the required gain. Since the gain of the E180F line driver was limited owing to linearity considerations, it was necessary to use 470Ω anode loads on the E88CC. The use of such a low value for R also meant that a considerable proportion of the signal appeared across the diodes and was thus wasted.

Despite the use of an OR-gate, some adding of pulses was found to occur, due to the anodes of the E88CC's being loaded by the gate resistor R. The switching on of one half of the E88CC(a) would remove part of the load from the other half (b) thus allowing its gain to increase. The diode associated with (b) would thus tend to come "on". With both diodes "on" the circuit becomes equivalent to an adding circuit formed by the use of two valves with a common anode load.

D.c. restoration is clearly essential in this type of circuit for the pulse trains at the grids of the double triode must maintain the same d.c. levels whatever the proportion of pulses in them. Unfortunately the fact that the pulse trains sometimes had poor baselines caused further small errors in combination due to the "riding up" of the pulses on the irregular baselines. (Figure 4.37.) Irregularities of the baselines were caused largely by reflections in the delay lines preceding a given stage. Although the lines were terminated in their characteristic impedance at both ends, such reflections could not be entirely eliminated.



(a) (b)

Figure 4.37. (a) d.c. restoration of pulses with 'clean' baseline
(b) 'riding up' due to an irregular baseline

Baselines could have been cleaned up by biasing the double triodes beyond cutoff but this would have resulted in the loss of part of the pulse at each stage and it would have been necessary to keep this loss extremely small and carefully controlled in order to avoid losing a wanted part of the pulse. This was not considered to be feasible without the introduction of considerably more equipment.

The interpolator was completed by a two stage output amplifier of conventional design (see Appendix 1). It had originally been hoped to conduct subjective tests using a variety of observers in order to assess the quality of the reconstructed pictures. It was, however, felt that the pictures obtained were not of adequate quality to make this procedure worthwhile. Photographs of some reconstructed pictures with various data reduction ratios are therefore presented as an indication of the performance obtained (Figure 4.38).

4.9 The Findings of this Chapter Summarised

The results obtained from the 1st Order Interpolator indicate that while the problems of run-length measurement and intermediate sample generation have been solved, further modifications to the readout section are necessary before fully satisfactory operation can be achieved. Such modifications would have included the addition of more valves together with the necessary power supplies, the existing supplies being fully loaded. The expenditure of considerable time would also have been required. It was therefore decided that the available time would be better spent in carrying out the measurements reported in Sections 2.15 and 3.8. Also influencing this decision was the fact that in a system of such complexity, a considerable portion of the time available is spent in diagnosing and curing faults. (These occur, not only in the equipment itself but also in test and measurement gear.) Any increase in complexity could thus be expected to increase the incidence of such faults.

Figure 4.38 (a) "Jean" A reconstructed picture using 850Kpps

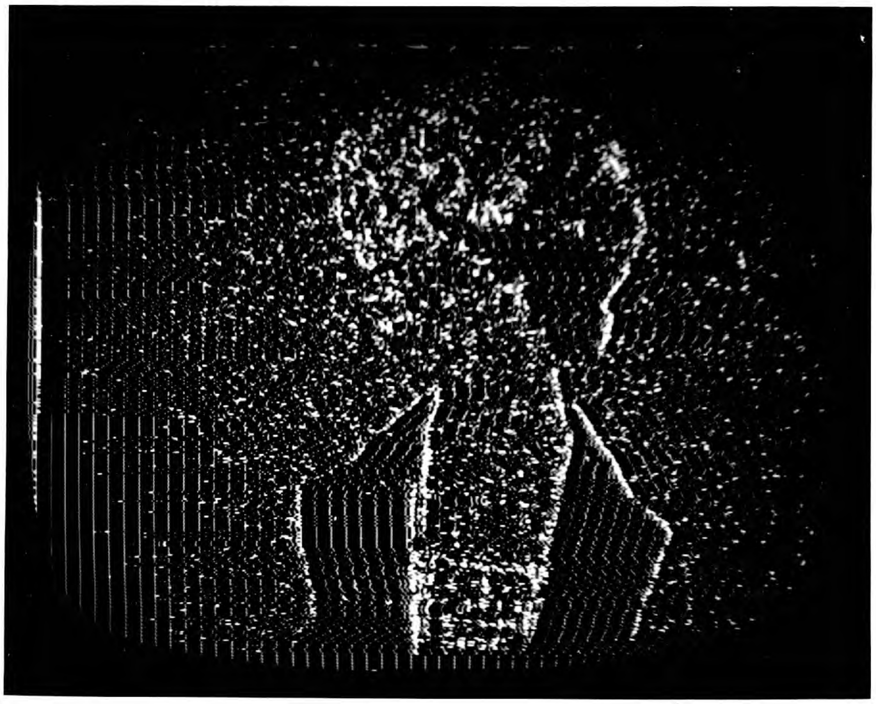


Figure 4.38 (b) Sampling pulses (after run-length restriction) 765Kpps

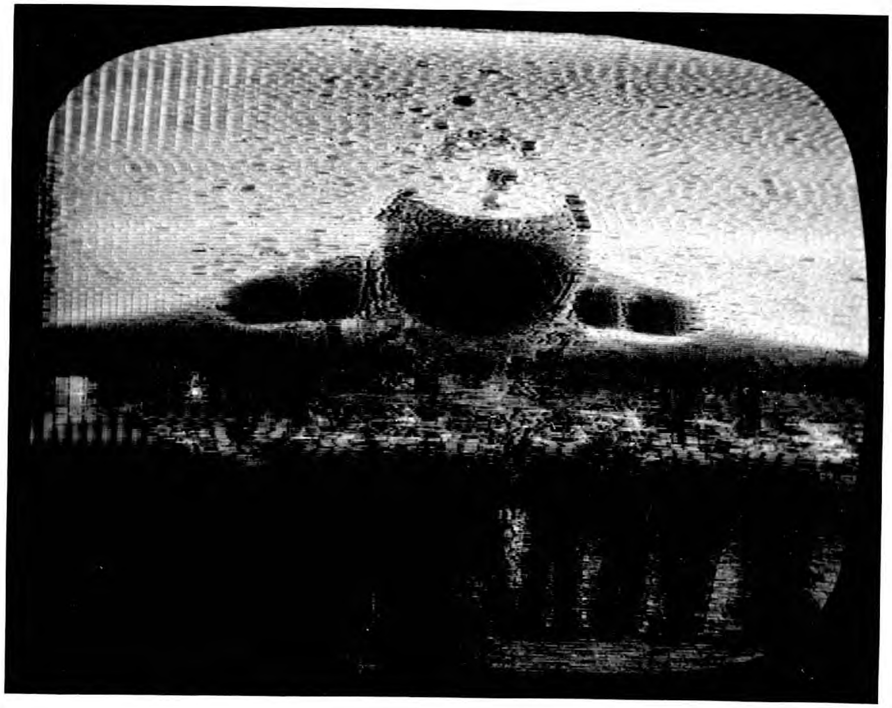


Figure 4.38. (c) "Aircraft" Reconstructed at 837 Kp

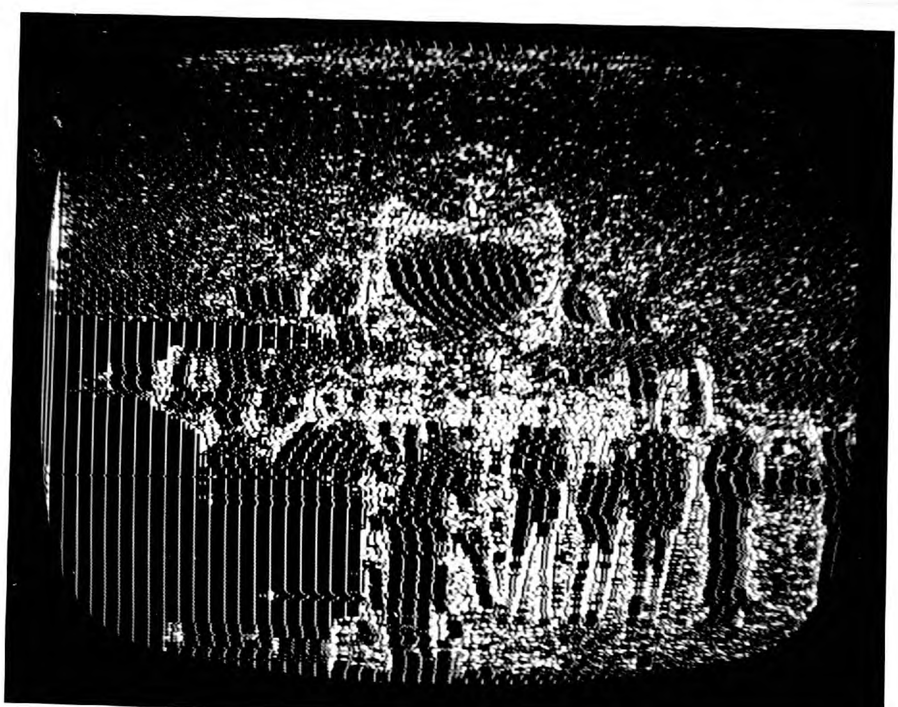


Figure 4.38. (d) Sampling pulses after run-length restriction (883 Kpps)

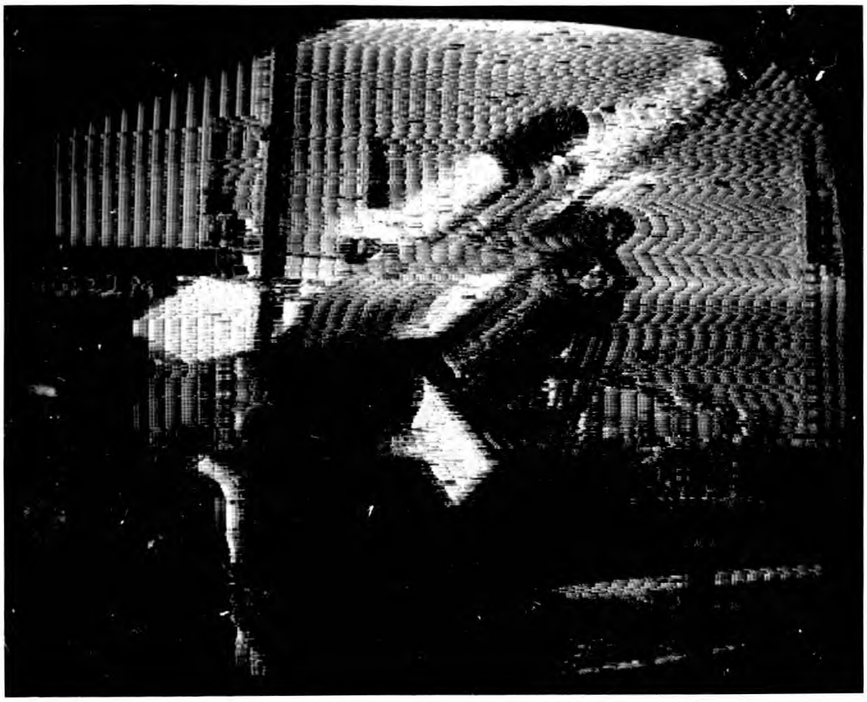


Figure 4.38. (e) "Rockets" Reconstructed at 760 Kpps

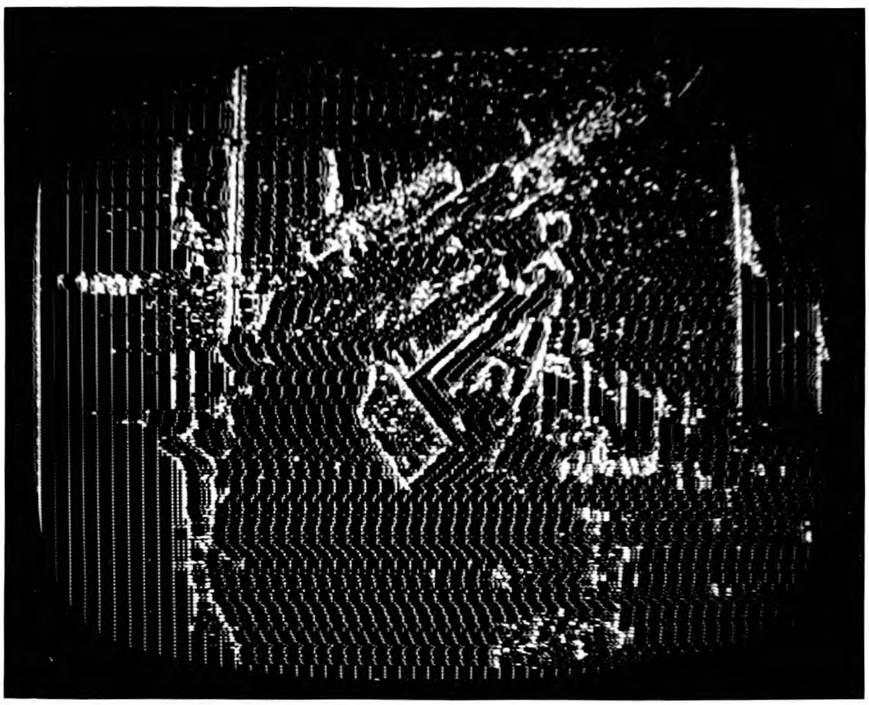


Figure 4.38. (f) Sampling pulses after restriction (760 Kpps)

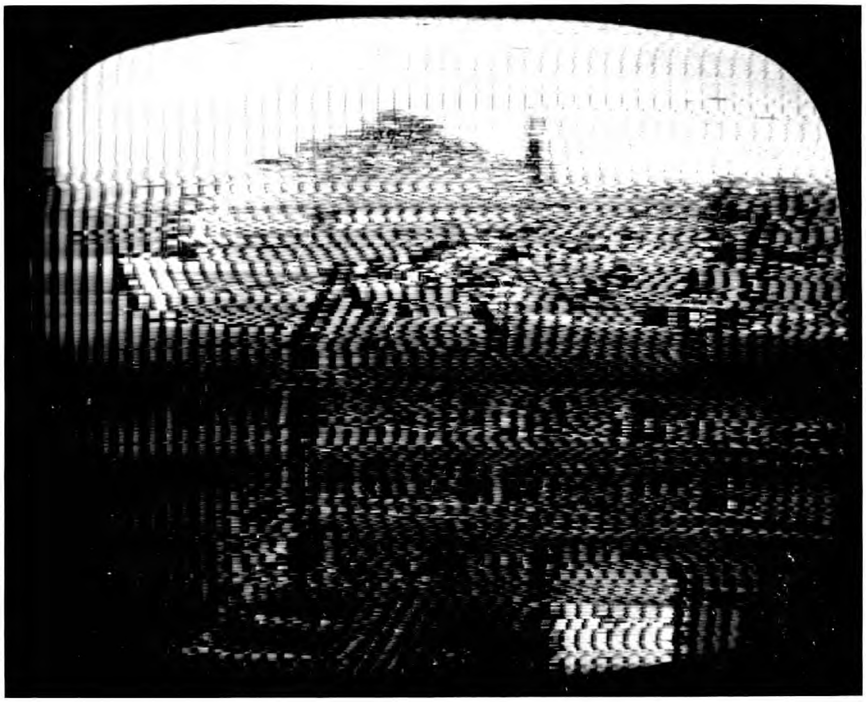


Figure 4.38. (g) "Albert Hall" Reconstructed at 770 Kpps.

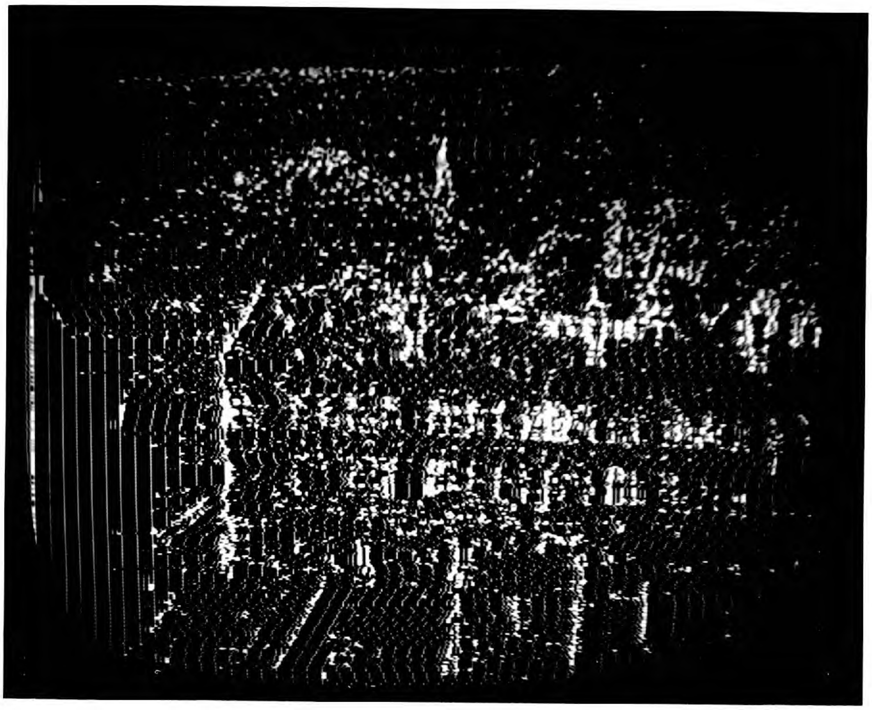


Figure 4.38. (h) Sampling pulses after run-length restriction (785 Kpps).

Figure 4.38. (i) "Jean" The original scene as displayed on a monitor screen.

Figure 4.38. (j) "Aircraft" The original scene as displayed on a monitor screen.

Figure 4.38. (k) "Rockets" The original scene as displayed on a monitor screen.

Figure 4.38. (l) "Albert Hall" The original scene as displayed on a monitor screen.

Alternative means of achieving serial readout for the resistor chain have already been mentioned (Section 4.7). Of the methods described, that using a number of different delay lines and one combining amplifier would probably prove least expensive and would certainly be less complex than a capacitor storage method. Any increase in complexity which can be tolerated should be directed towards better stability of amplifiers; this can be achieved by the provision of more forward gain and the use of considerable negative feedback.

An early section in this chapter dealt with theoretical consideration of reconstruction methods and showed the type and magnitude of errors which can be expected to result from the use of methods not matched to the run-end detection criterion. In the case of 1st Order Interpolation and Variable Velocity Scanning as alternative reconstruction methods when 1st Order run-end detection is used, the errors are usually small. Only in regions of high detail can the errors become large (i.e. an appreciable fraction of the total swing from black to white) and it is in this type of region that the eye is tolerant of amplitude errors. Whether the errors will be large enough to be significant can be shown only by subjective testing, however; if a full-scale elastic encoder is ever set up, together with its associated variable velocity receiver, this is one of the experiments which should be conducted.

5. SUMMARY OF CONCLUSIONS

Conclusions have been drawn at the end of each chapter of the thesis so far, but for convenience they will now be collected and summarised, overall conclusions will be presented, and suggestions for further work will be made.

The programme of work which has been carried out and reported in the thesis has covered two main aspects of data reduction in television signals - run-end detection and run-length restriction. The two aspects are interdependent, however: the investigation of one required the investigation of the other and the assessment of the effects of each required the equipment for carrying out the other to be operational.

As part of the work on detail detection, the relative merits of various models for television signals have been considered, the exponential being found to be the most useful of the three-parameter curves examined. In parallel with this, the definition of a run and a rational general procedure for detecting run-ends have been dealt with. Of considerable importance is the justification given in Section 2.8 of the least-squares fitting of curves to data. The exact nature of this process when applied to picture signals has not previously been critically examined.

These two lines of attack have naturally converged on the least squares fitting of exponential curves to data. It emerges that this will best be done by distorting the data using a logarithmic amplifier so that exponential runs become linear and linear fitting can be employed. Such distortion of the signal samples strictly requires that each distorted sample be given a weight proportional to the square of its amplitude before distortion. The effect of neglecting this requirement has been shown to be negligible in a number of examples using realistic signal-to-noise ratios, however.

The fact that first-order run-end detection will be of considerable importance led to a pilot experimental study of this operation. The run-end detector instrumented operated in good accordance with theory and the slope change it was required to detect was estimated as 0.21mV/μS, approximately the same fractional change in the run-defining parameter as in the zero-order case.

Run-length restriction is the other facet of data reduction which has been dealt with. The problem of devising and constructing suitable equipment to carry out this process has been satisfactorily solved. The equipment chosen employs electromagnetic delay cable, logic gates and monostable multi-vibrators to generate blanking signals for gating clock pulses. The resulting pulse train has standard pulse spacings only but contains sampling pulses whenever this is required by the detail detector. The availability of this equipment has permitted direct measurement of the optimum standard run-lengths to be undertaken for the first time, the first-order run-end detector being used to derive detail information for this purpose. The standard run-lengths of 1, 4 and 9 Nyquist intervals were found to give most economy over the range of unstandardised sampling rates tested (from almost zero to one quarter of the Nyquist rate), although the trend of the curves suggests that 1, 3, 8 may be more advantageous at very high average rates.

Completion of the study by work on 1st order interpolation included a comparison of this technique with variable velocity scanning as a means of reconstructing pictures. Consideration of a number of examples showed that the discrepancies between the reconstructions produced by the two methods could be expected to be small, particularly in regions of low detail. Where the detail content is high, the discrepancies can be expected to be appreciable but it is in such regions that the eye is known to be tolerant to considerable amplitude errors. It is therefore anticipated that the discrepancies will not be found objectionable in practice. During this discussion it was pointed out that 1st order interpolation is the reconstruction method which is matched to the run-end detection method and variable velocity decoding can be looked upon as the error generating process.

The by-passing of the encoding-decoding process by interpolation between non-uniformly spaced samples was initially conceived when it was necessary to use valves for all high speed circuits and the design and construction of an elastic encoder would have been virtually impossible. Since that time, transistors and integrated circuits have made vast advances and a full-scale encoder handling binary-coded signals is quite feasible. Despite this, the interpolator is a valuable tool for it has the potential of allowing the degradation due to detail detection to be separated from that due to elastic encoding and decoding.

The method of 1st order interpolation which was devised and instrumented is believed to be completely novel, a chain of identical resistors being employed to reconstruct samples intermediate between those in the restricted run-length train. Although the pictures reconstructed by this means left much to be desired, they demonstrated the feasibility of the resistor-chain interpolator and allowed the effects of 1st order run-end detection and run-length restriction to be assessed. Fair pictures could be reconstructed using about 800,000 samples per second, a data reduction of about 7.5:1. It is felt that this figure is valuable in that it represents an upper bound to the data reduction which can be expected using 1st order techniques. Improved equipment can be expected to lower the ratio, since at present, any improvements resulting from an increase in sampling rate are masked by degradations due to the shortcomings of the equipment.

The programme has inevitably left many questions unanswered and suggested further topics for research. The completion of the investigation of the exponential representation is one such topic. This should probably be preceded, however, by further work on 1st order run-end detection with more samples examined simultaneously. Complementary to this would be a programme aimed at improved 1st order interpolation.

Less closely associated with my own programme are problems such as the nature of detail and how best to sample a picture signal (the early sections of Chapter 2 have sought to make an initial assault on this problem but have barely scratched the surface), the amount of degradation which can be tolerated under various viewing conditions and how this is related to the type of scene, and the transmission of parameters other than brightness.

The continually increasing speeds of digital computers should facilitate all this work, eventually making possible easily alterable systems capable of real-time processing of television signals. When this comes about, the answers to the questions mentioned above will rapidly become available as will answers to questions which it is inconceivable to tackle by the construction of hardware. One such is the question of whether any useful result can obtain from the adoption of a reconstruction method which is not matched at all well to the method of detail detection.

6. APPENDICES

Appendix 1. Circuits

Both valves and transistors have been used in the equipment constructed although the two have never been mixed within a unit. At the beginning of the experimental programme valves were used exclusively as suitable fast switching transistors were not available or were excessively expensive. Later in the work the choice between valved and transistorised circuits depended mainly on the type of device used in adjacent equipment and the availability of power supplies. Where a transistorised unit had to be connected to a valved unit protection was incorporated so that the transistors were not damaged by surges during the switching on or off of the valved unit.

Although stabilised power supplies were used throughout the equipment, it was considered prudent, in most stages, to employ decoupling circuits to prevent undesired interactions. A series 100 ohm resistor together with a 0.01μF capacitor to earth frequently performed this function. Direct coupling between stages was used where possible; in cases where resistance-capacity coupling could not be avoided, l.f. compensation was used, the circuits being adjusted for best response to a 25 c/s square wave.

For valved equipment, voltages of +250 v., and +150 v. and -150 v. were provided. Transistorised units used +12 v., -6 v. and -12 v. In view of the ease with which the transistors employed could be damaged by reverse emitter-base voltages, the power supply units were fitted with a cut-out so that an overload on any supply caused all three to be switched off simultaneously.

Two types of valve were used almost exclusively. These were special quality, high mutual conductance devices. Where triodes were required the E88CC double triode found application, the two halves being independent (except for common heater connections). Requirements for pentodes were satisfied by the E180F. Valve type numbers are not given in the circuit diagrams unless types other than these were employed.

The circuit diagrams are self explanatory if examined in conjunction with the block diagrams given in the main body of the thesis.

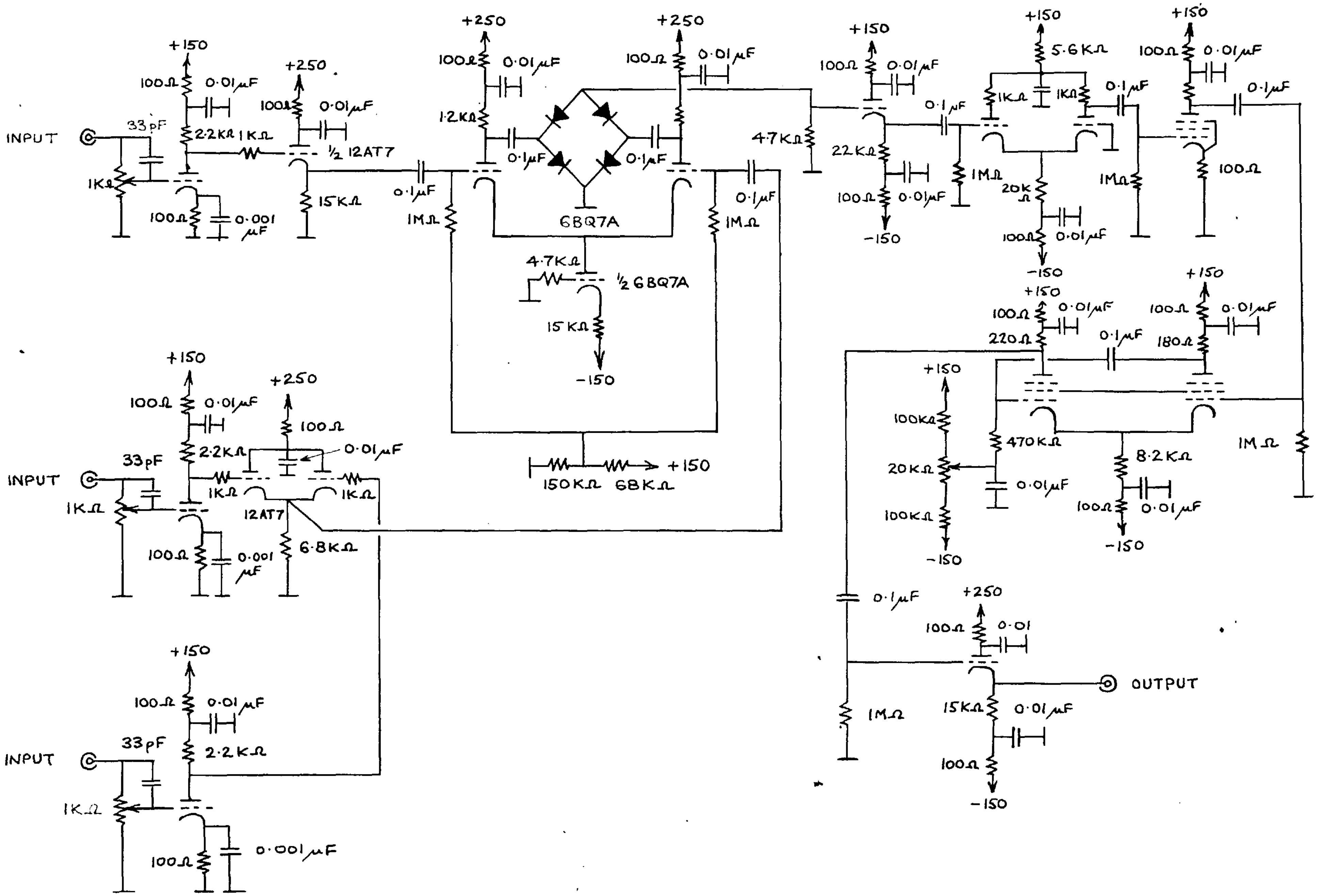


FIGURE 6.1. THE FIRST ORDER RUN-END DETECTOR.

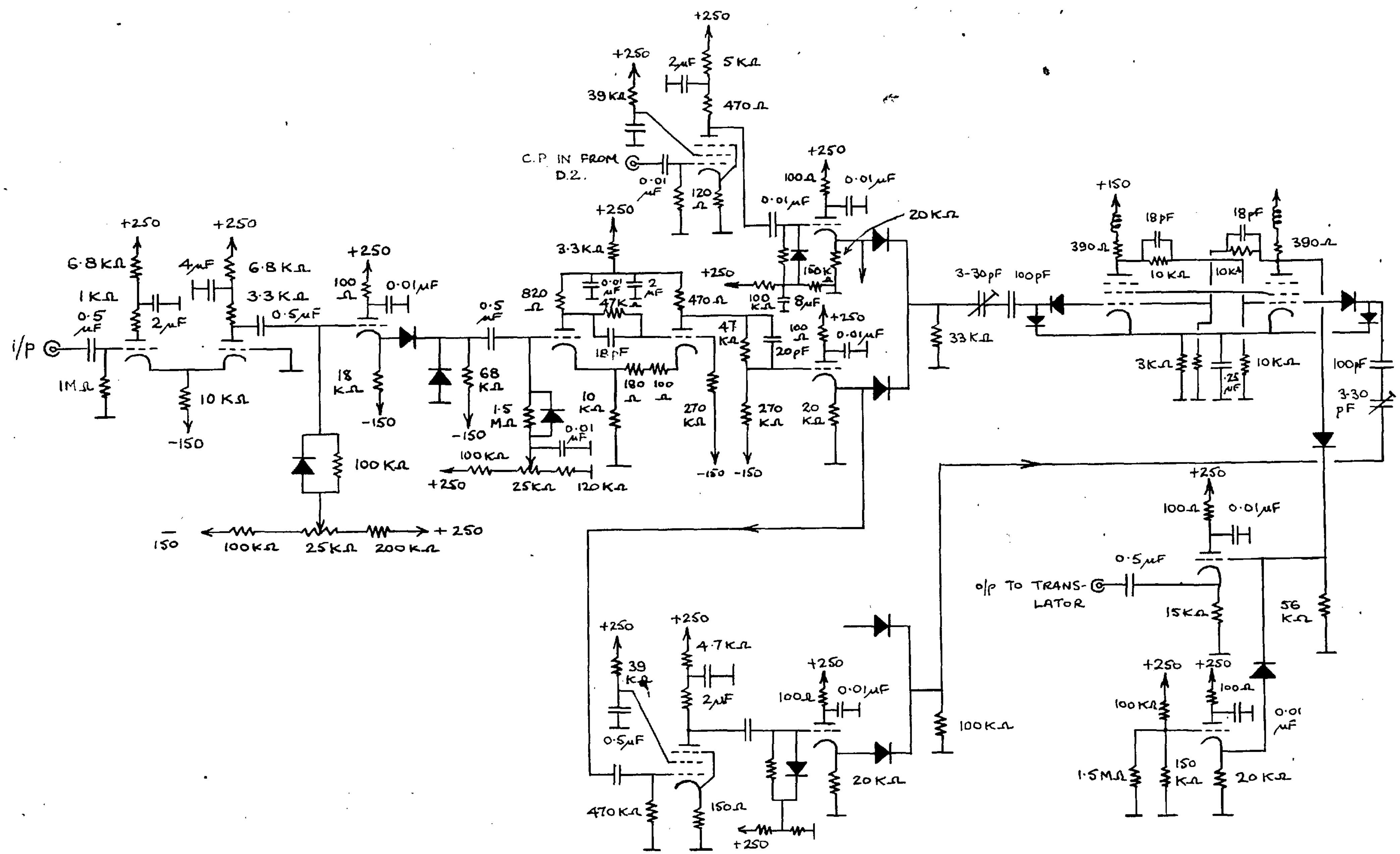


FIGURE 6.2. THE DETAIL FUNCTION SHAPER.

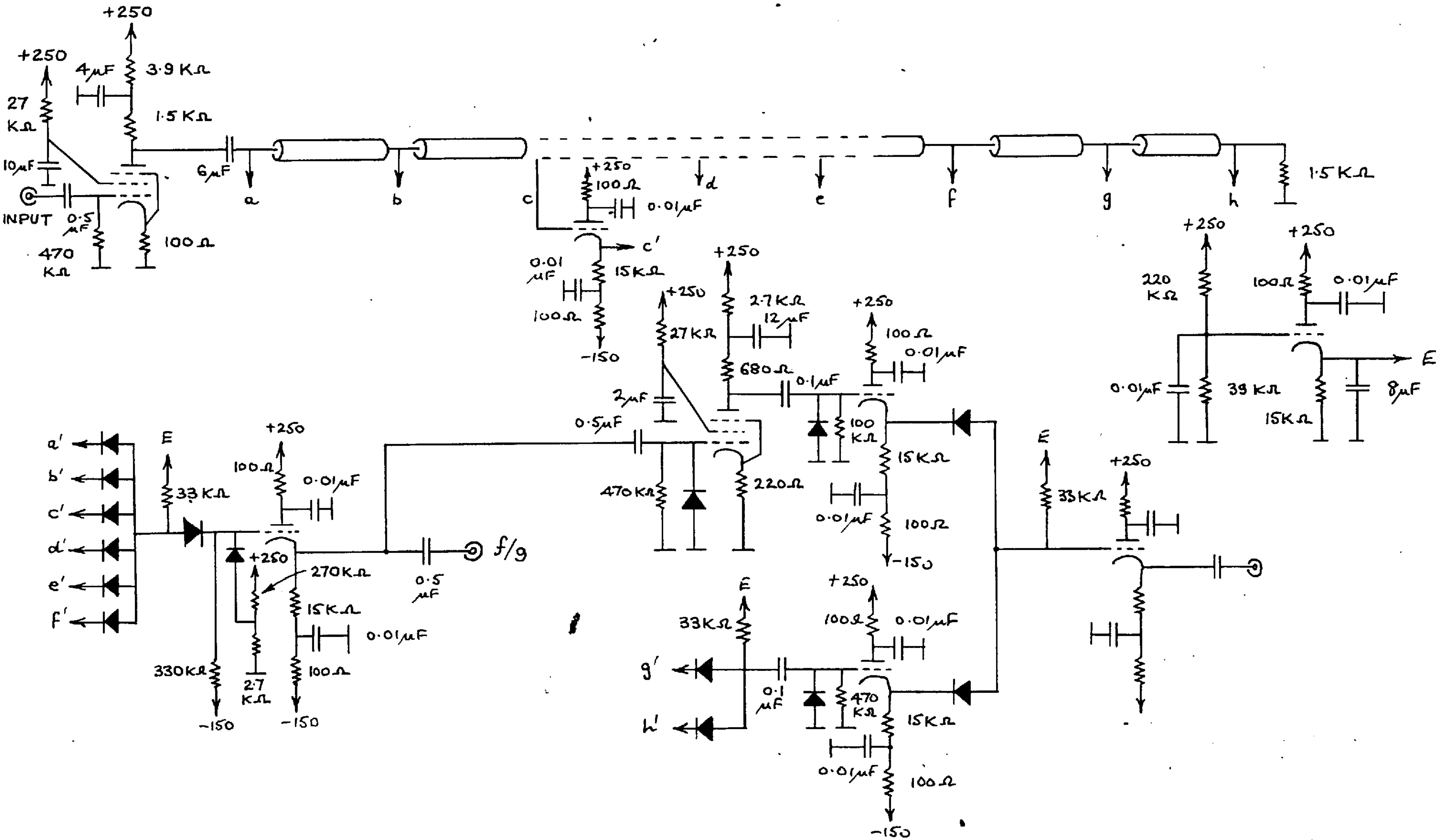


FIGURE 6.3. THE TRANSLATOR.

- 230 -

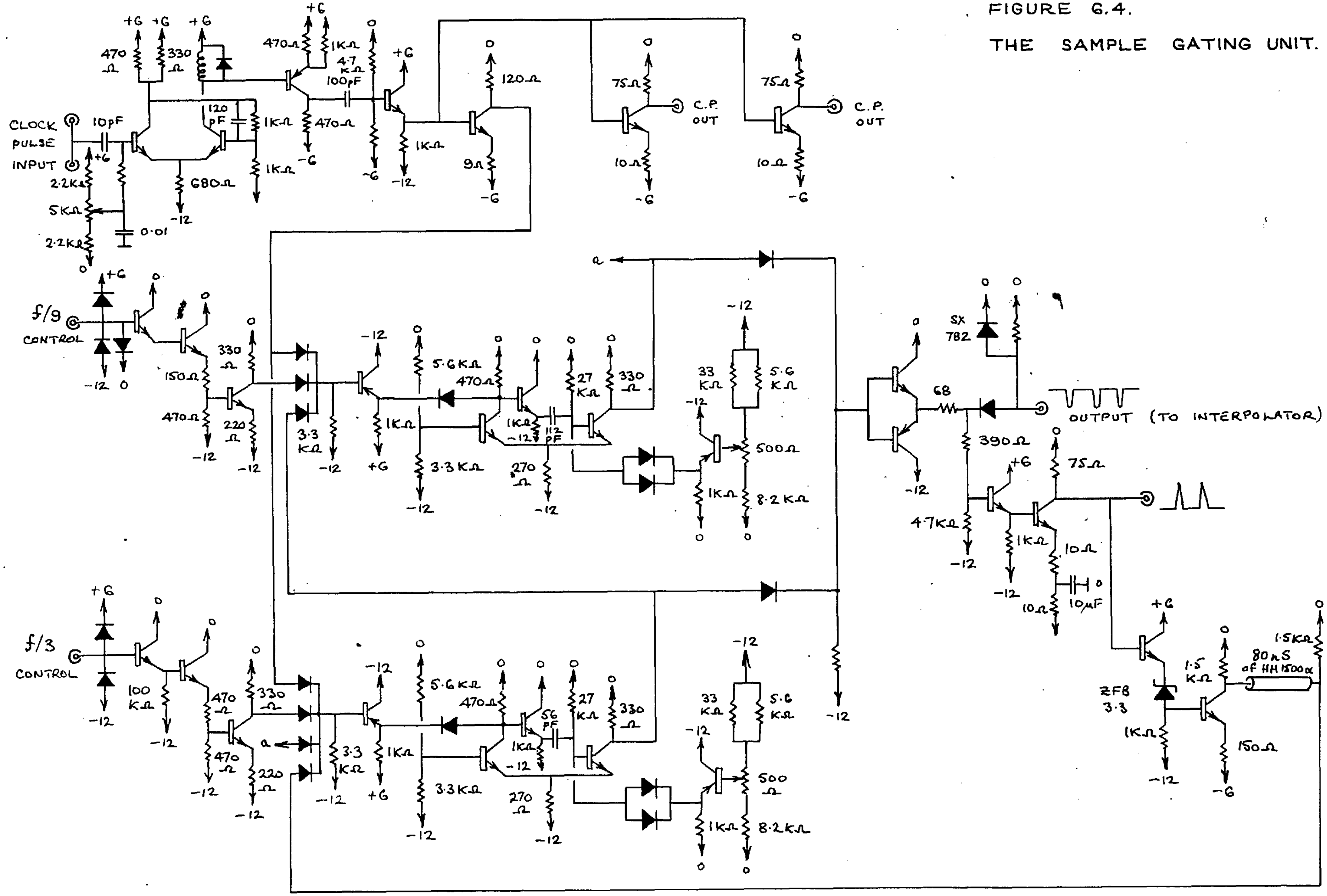


FIGURE 6.4.

THE SAMPLE GATING UNIT.

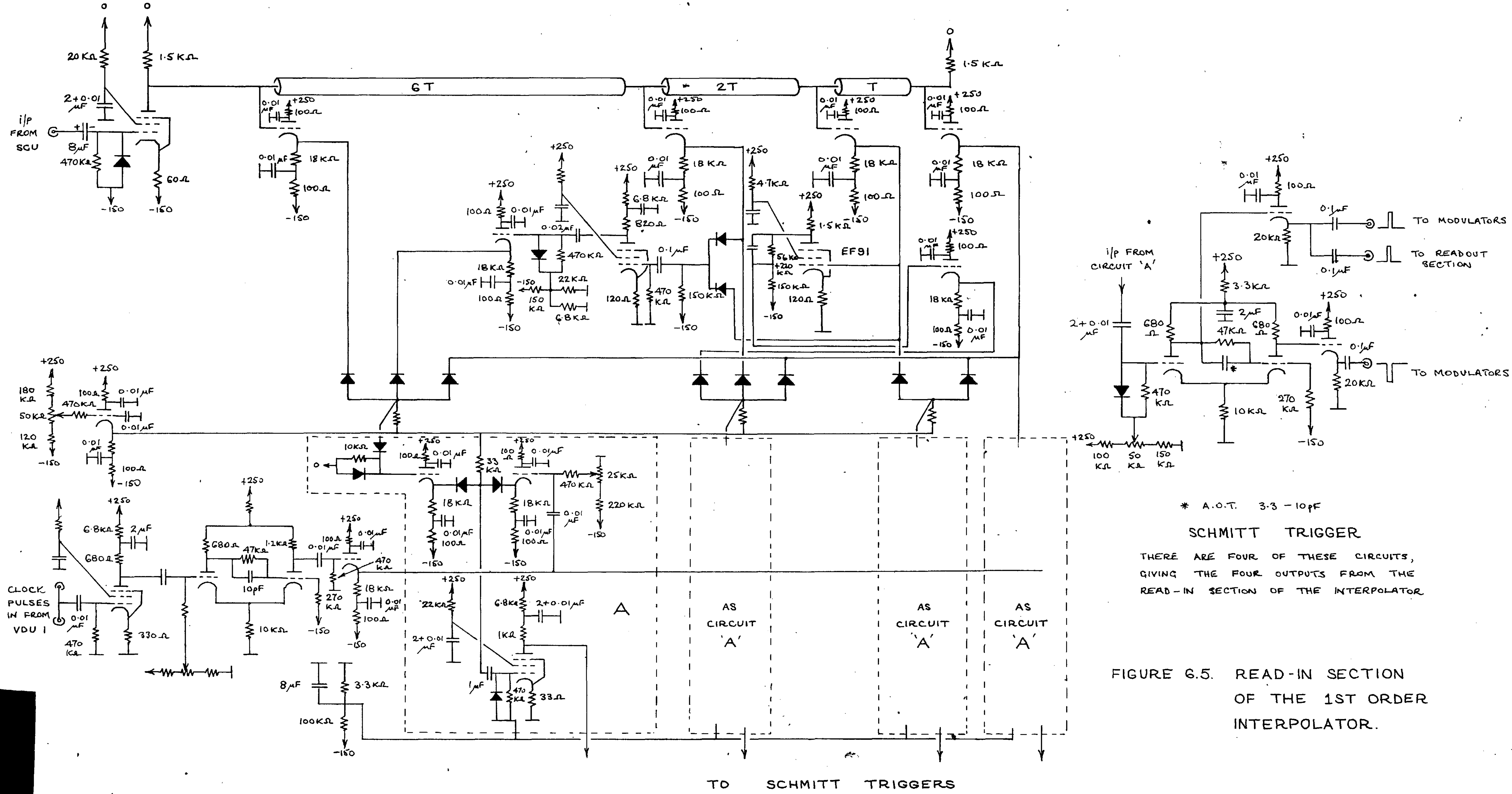


FIGURE 6.5. READ-IN SECTION OF THE 1ST ORDER INTERPOLATOR.

DOUBLE MODULATOR CIRCUITS (AS SHOWN CONNECTED TO P.7.) ARE ALSO AT P.1., P.9. & P.10.

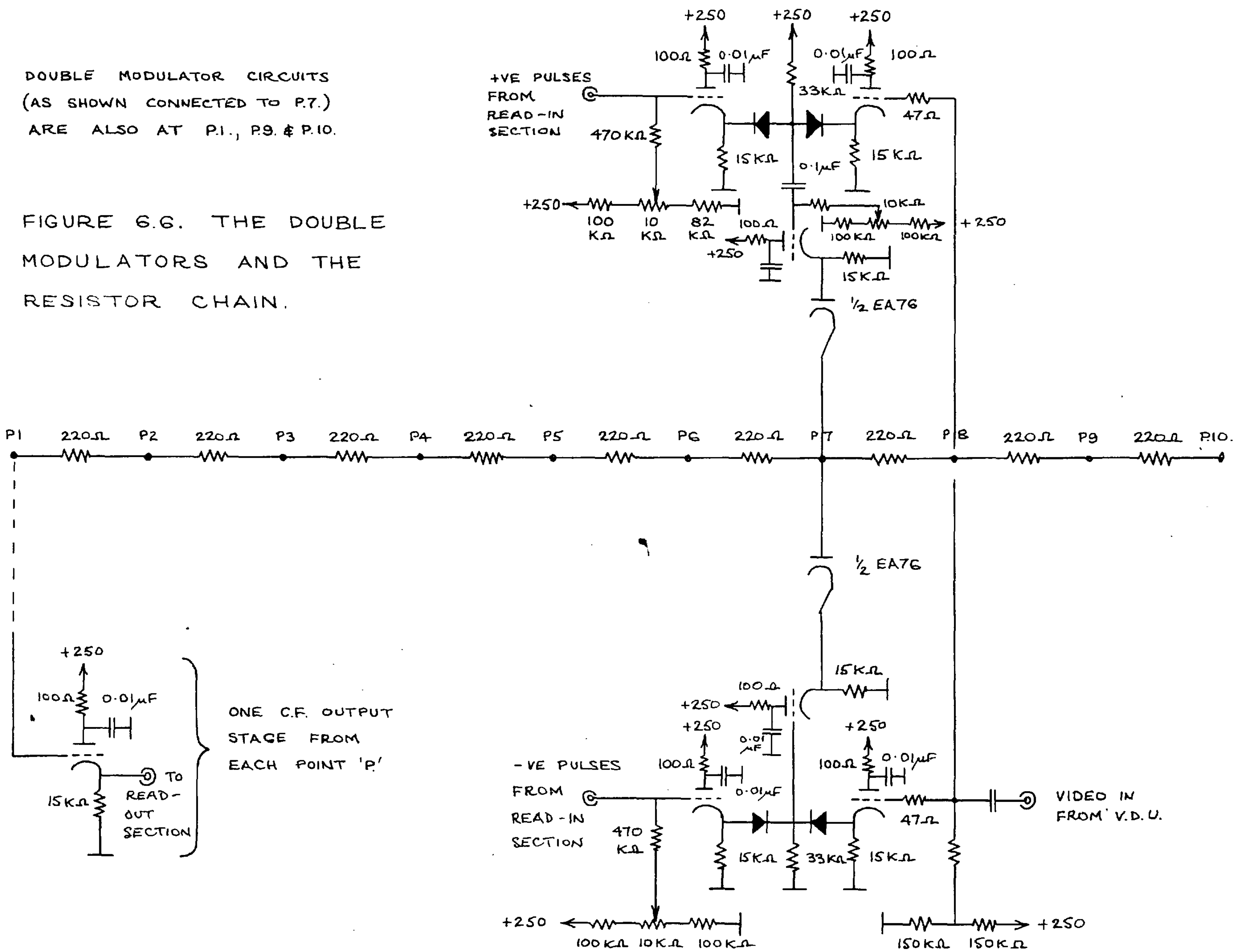


FIGURE 6.6. THE DOUBLE MODULATORS AND THE RESISTOR CHAIN.

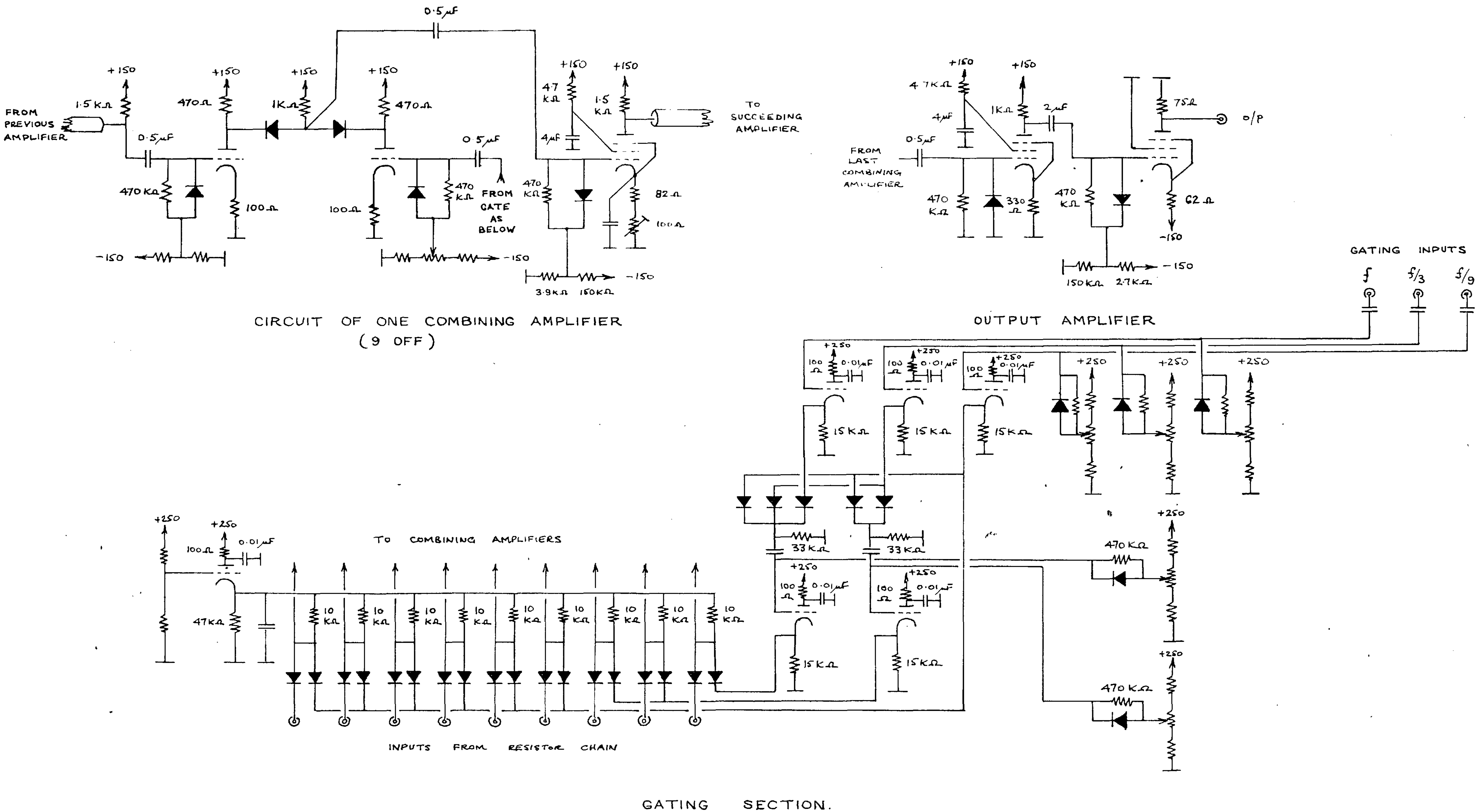


FIGURE 6.7. THE READOUT SECTION OF THE INTERPOLATOR.

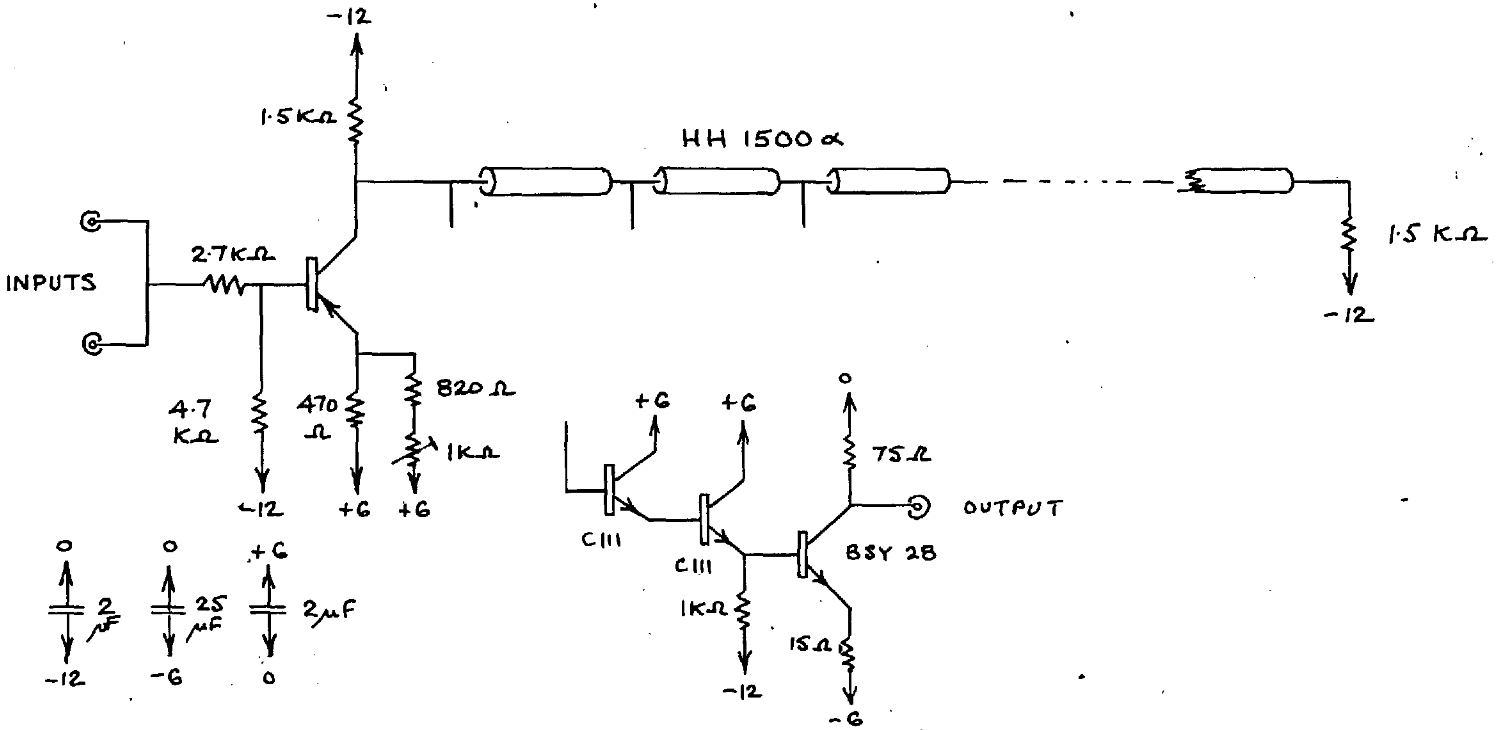


FIGURE 6.8. THE VIDEO DELAY UNIT.

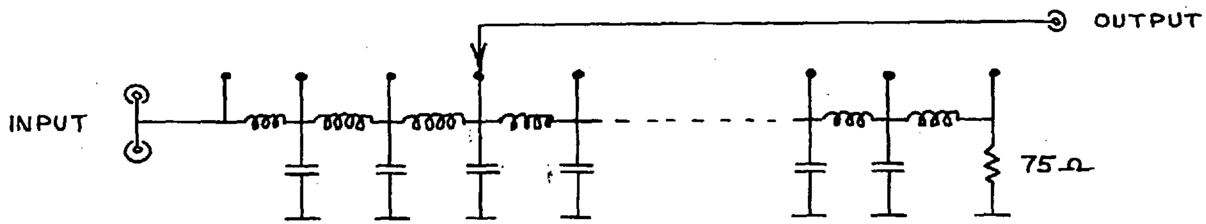


FIGURE 6.9. VARIABLE DELAY UNIT NO. 1.

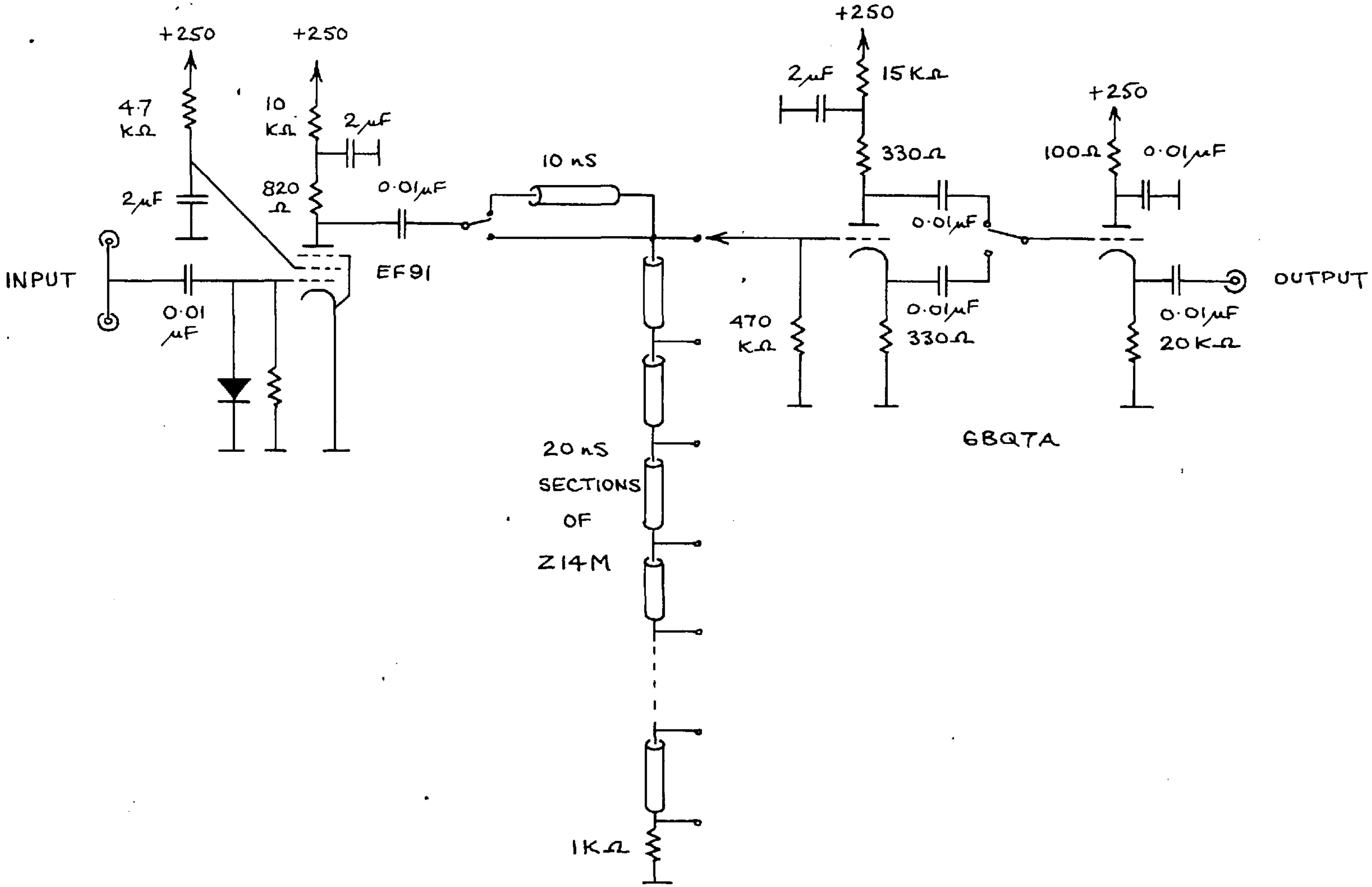


FIGURE 6.10. VARIABLE DELAY UNIT NO. 2.

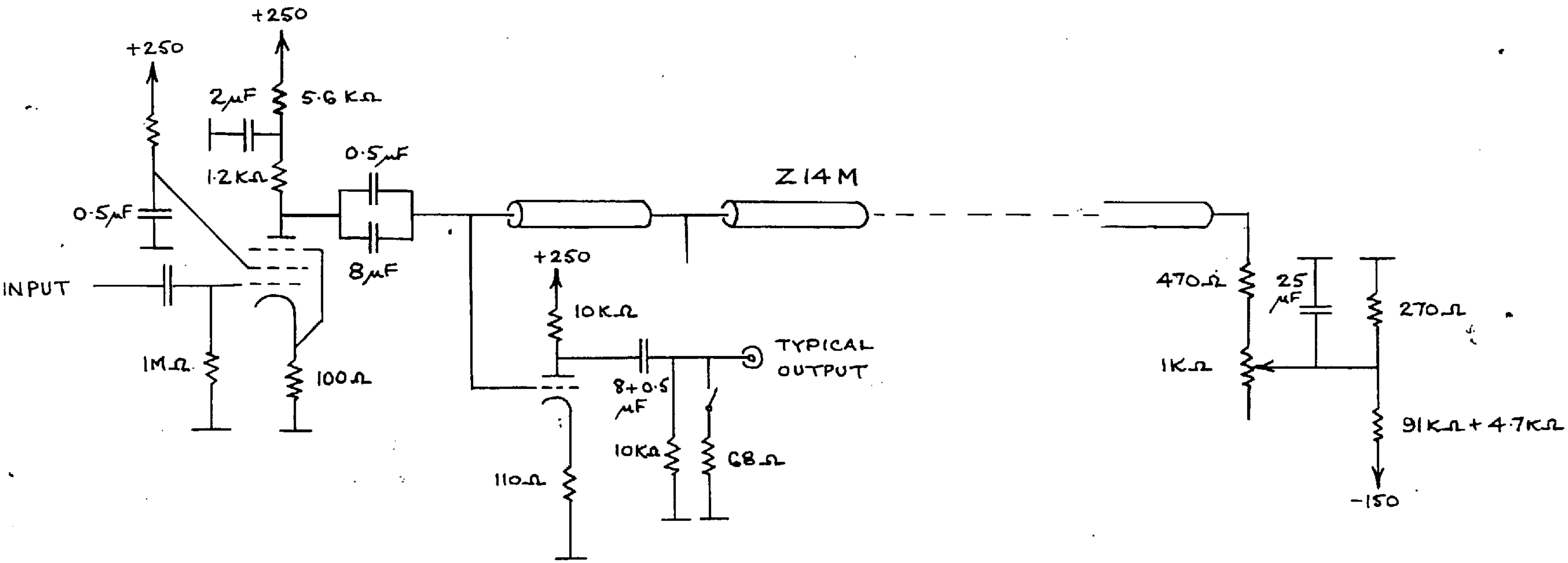


FIGURE 6.11. THE VALVED VIDEO DELAY UNIT.

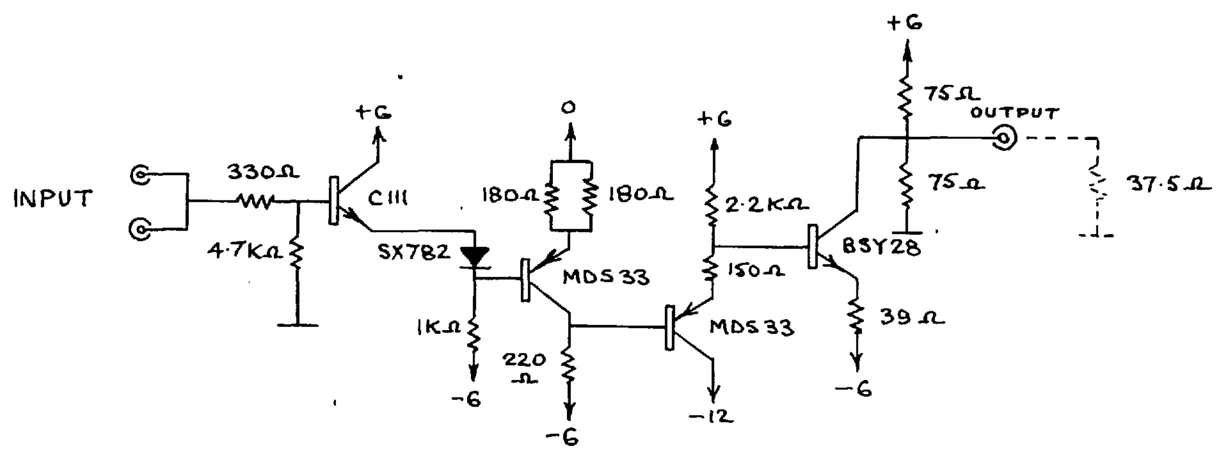


FIGURE 6.12. THE VIDEO AMPLIFIER USED TO DRIVE THE 605 C2 DELAY LINE.

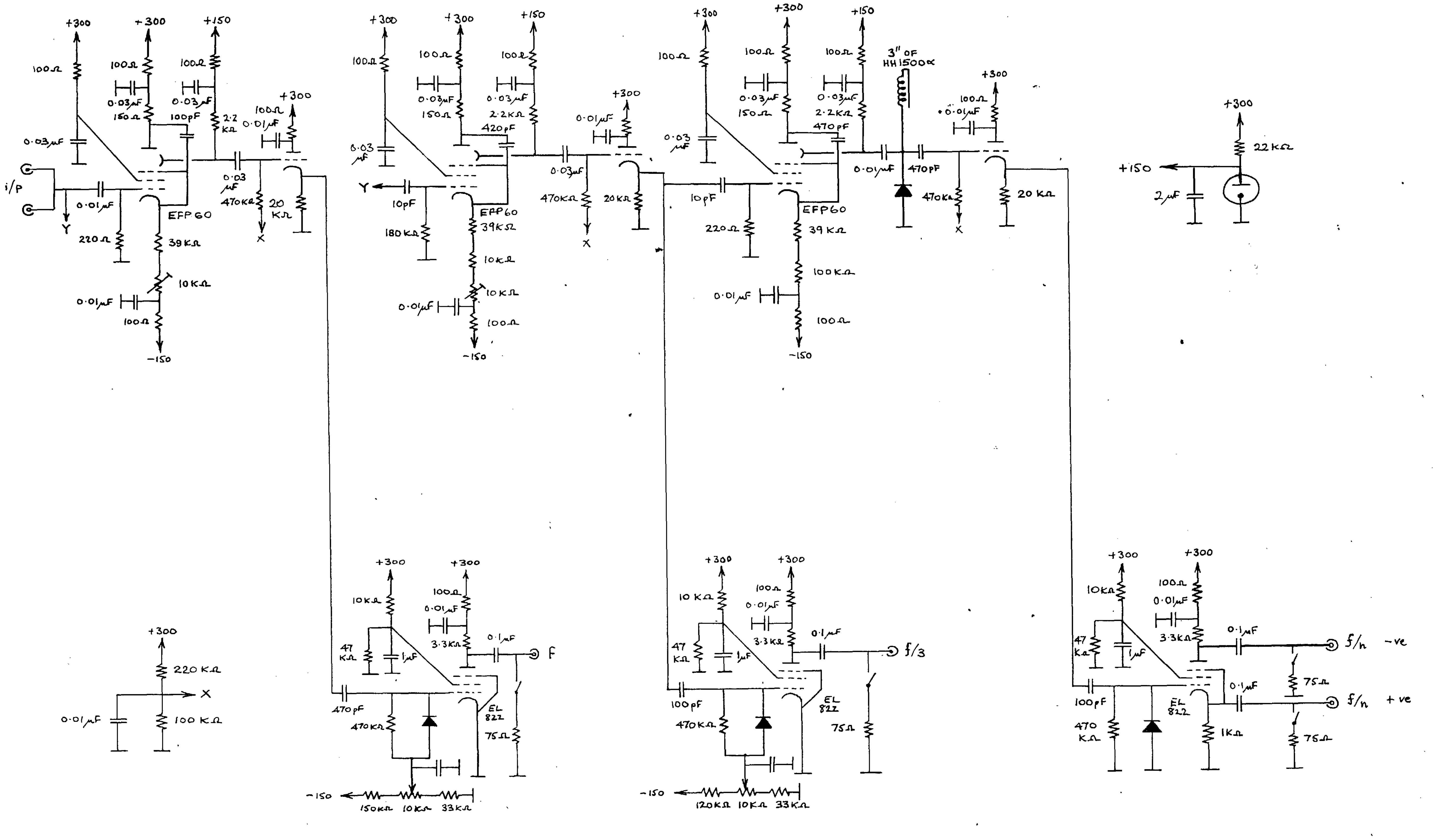


FIGURE 6.13. THE NEW CLOCK-PULSE GENERATOR.

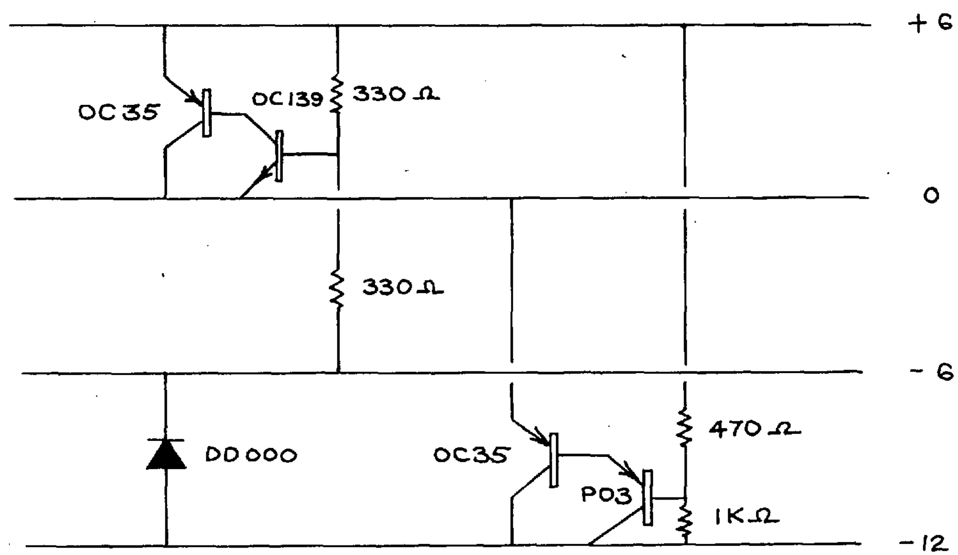


FIGURE 6.14. TRANSISTOR POWER SUPPLY PROTECTION CIRCUITS.

Appendix 2. The Flying-Spot Scanner and Associated Equipment

In order to allow slight simplifications in television receivers, television signals radiated by the B.B.C. and I.T.A. are locked in synchronism with the 50 c/s a.c. mains. Previous students in the Communications Section at Imperial College found that this severely limited the usefulness of such signals as all equipment locked to the television signal would have been required to remain in synchronism over a frequency range of 8%, the mains frequency varying between 48 c/s and 52 c/s. In particular, a frequency multiplier designed to multiply 50 c/s up a sampling frequency of about 6 Mc/s was found to be exceedingly unreliable in operation as the input frequency was varied slightly. Thus for work previously reported[33,44] and for the work described in the main body of this thesis, all the equipment was synchronised to a 6 Mc/s crystal oscillator, signals for synchronisation of monitors, etc. being derived by frequency division.

A flying-spot scanner was used to obtain television signals Flying-spot techniques are little used in either studio or industrial applications nowadays. However, the system is capable of generating a signal with better signal-to-noise ratio than that from any other type of pick-up device. It is thus the natural choice as a signal source when run-end detection is to be investigated.

The scanner used was originally supplied by Marconi's Wireless Telegraph Co. Extensive modifications had been carried out by previous users, however. The principles of its operation will be apparent from the following description. Figure 6.15 is a simplified block diagram of the Flying-Spot scanner, showing also the crystal oscillator and the frequency-division equipment.

A special cathode-ray tube having a fine-grain phosphor was employed. The phosphor had a very short decay time constant (about 50 nS) so that the transmission of only one point on the

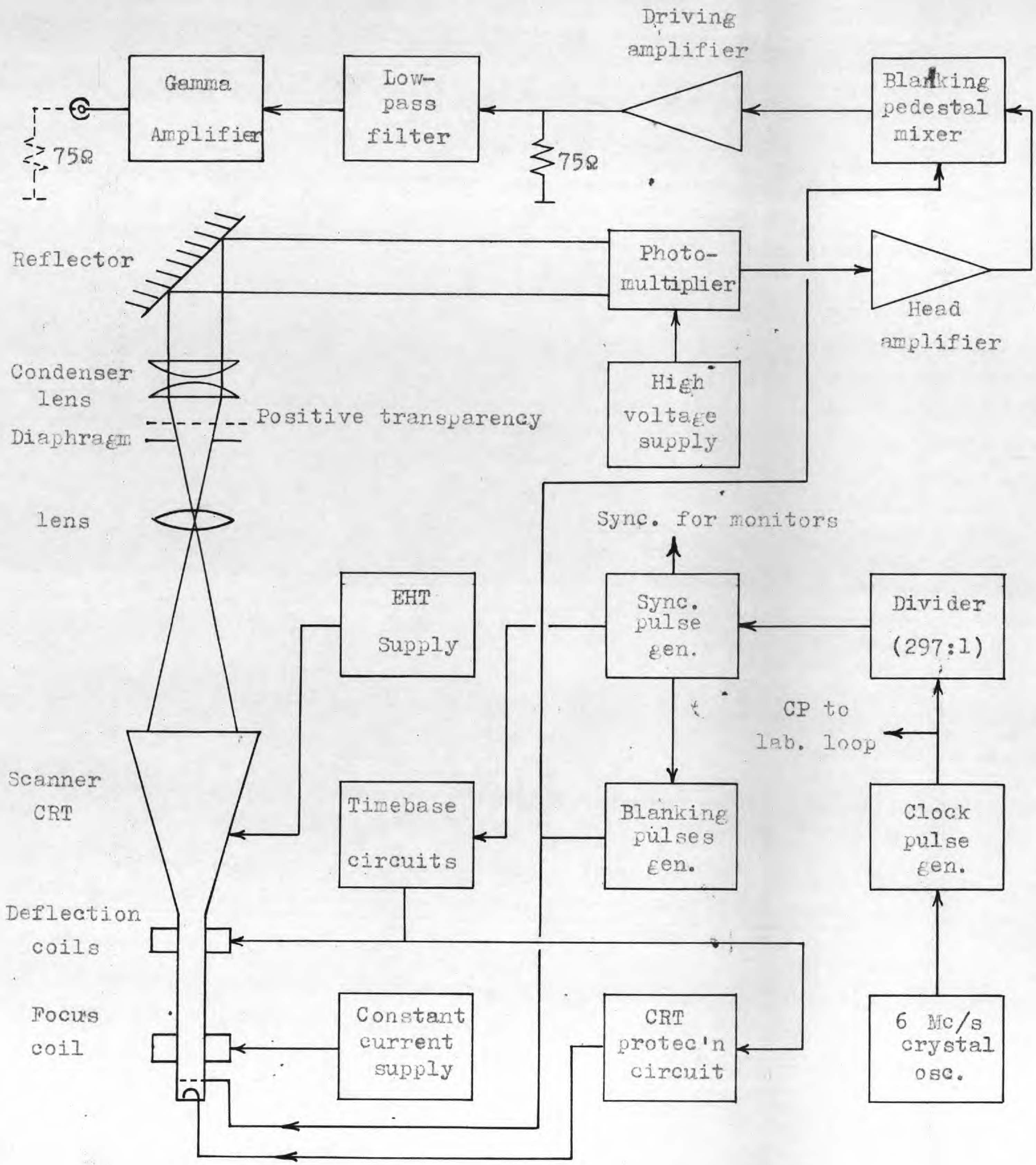


Figure 6.15. Simplified block diagram of the Flying Spot Scanner.

diapositive transparency was measured by the photomultiplier. The spot was caused to scan a raster on the face of the cathode-ray tube by means of a time base generator and deflection coils. The spot was focused on the transparency by a wide-aperture lens (f/3), the light transmitted being collected by a condenser lens before falling on the cathode of the photomultiplier. This device was chosen to have a spectral response well matched to that of the C.R.T. phosphor. The achievement of a large signal level was also aided by the use of a high final-anode voltage on the C.R.T., a very bright spot being thereby produced. As this would quickly have damaged the phosphor if allowed to remain stationary, a protection circuit was provided and this would have cut off beam current had scanning signals failed.

A head amplifier, close to the photomultiplier provided a low-impedance output suitable for driving the cable leading to the blanking pedestal mixer. Here a pedestal was added to the signal in order to assist the action of the d.c. restorers employed in monitors and other equipment. In order that no signal arrived at the mixer during the flyback of the spot, and also to establish a true black level, a signal was also sent to the C.R.T. by the blanking pulse generator.

To cut out high frequency noise and to define clearly the bandwidth of the video signal before processing, a 3 Mc/s low-pass filter was used. To obtain a convenient signal amplitude (1 v.) it was necessary to use a totem-pole driving amplifier as the filter required to be terminated in 75 ohm at both sides.

The current generated by a photomultiplier is linearly proportional to the light energy falling upon the cathode. If correctly exposed and developed the transparency will have a density which is directly proportional to the brightness of the scene photographed. Thus the signal voltage developed by passing the photomultiplier current through a load resistor is linearly proportional to the scene brightness. At the reproducing end however, the beam current of the C.R.T. is proportional to the grid-cathode voltage above cut-off raised to the power 2.2.

$$\text{i.e.} \quad I_b = K_1 E_{gk}^{2.2} \qquad (6.1)$$

Since the reproduced luminance is directly proportional to beam current, an amplifier with the characteristic

$$v_0 = K_2 \; v_{in}^{1/2.2} \qquad (6.2)$$

is required in order to give the system a linear overall response The factor 1/2.2 is frequently given the symbol γ so that the amplifier having this response is known as a "gamma corrector". Such an amplifier was employed in conjunction with the flying-spot scanner as shown in the block diagram.

Line and frame synchronising pulses for the timebases and blanking pulse generator of the flying-spot scanner were supplied by a Synchronising Pulse Generator (S.P.G.) built by a previous student[33]. This unit also provided a composite synchronising signal which was circulated around the laboratory in order to synchronise the several picture monitors used. Owing to the use of 2:1 interlace in television systems, successive field synchronising pulses differ in their phase relationship to the line synchronising pulses. (Figure 6.16.) The S.P.G. thus needed a supply of pulses at twice the line frequency. As the clock pulse repetition frequency was 6 Mc/s, a divider having a ratio of 297 was necessary to give a line frequency as close as possible to the standard line frequency of the 405-line system (10.125 lines/sec.). With this ratio, the line frequency was 10,101 lines/sec., the output of the divider being at 20,202 pulses/sec. In addition to the conventional line- and frame-synchronising pulses, the S.P.G. provided "Single Frame Trigger", a pulse occurring once per frame which could be used for

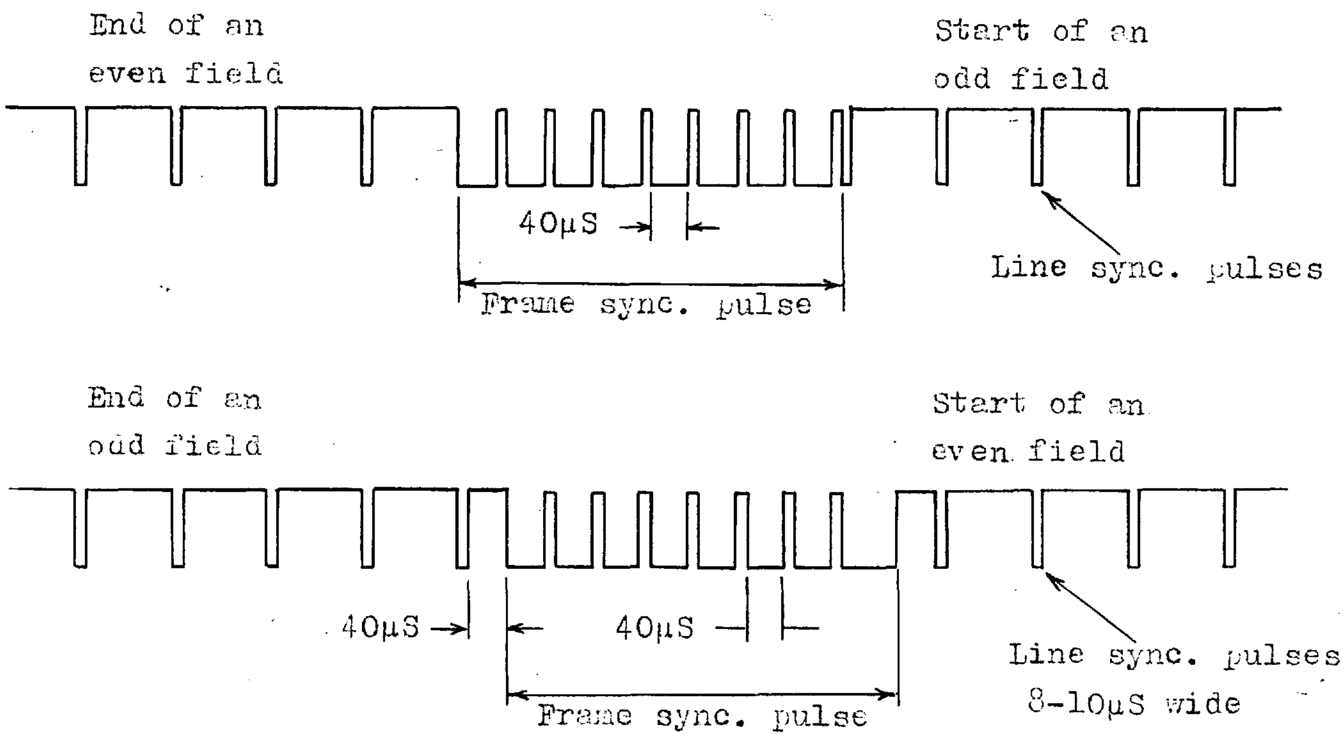


Figure 6.16. Field and line synchronising pulse relationships

synchronising oscilloscopes or, in conjunction with a variable delay, could be used to initiate a gating pulse or a marker for operations on a specific portion of the video signal.

Clock pulses at 6 Mc/s, to trigger the divider and for circulation around the laboratory, were produced by a unit using secondary emission valves. Although fast transistors were available at the time this unit was built, suitable valves and power supplies were immediately to hand and were thus employed. The clock pulse generator was driven by a crystal oscillator giving a sinusoidal output at 6 Mc/s

Circuit diagrams of the units for which I have been responsible are included in Appendix 1.

Appendix 3. Logic

In computers and many other items of electronic equipment, circuits which perform logical functions are employed. An assembly of such circuits is often known as "logic". Many electrical devices can be most conveniently arranged to operate as two-state elements i.e. the device is either "ON" or "OFF". Thus the signals in such systems have two levels only, the levels usually being designed "0" and "1" in accordance with the familiar binary number system. If the logical "1" level is more positive than the "0" level the logic polarity is said to be positive whilst if the converse applies, the logic polarity is negative. Both positive and negative logic polarities have been used in the equipment constructed.

Logic elements include "AND" and "OR" gates. A gate is, in general, a circuit having several inputs but only one output. In the case of the "AND" gate, a "1" is produced at the output only when all the inputs are at the "1" level. The "OR" gate, on the other hand, requires only one input at the "1" level to produce a "1" at the output. It is convenient, when assigning logic systems, to use conventional signs for the elements, rather than the actual circuits. In the case of "AND" and "OR" gates, the number of inputs required to be at the "1" level in order to produce a "1" at the output is written in a circle. Thus Figure 6.17(a) shows a 3-input "AND" gate while a 4-input "OR" gate is shown in Figure 6.17(b). As an alternative to this convention, the type of the gate and its number in the system can be written in the circle e.g. AND 1, AND 2, etc.

Another logic element frequently encountered is the inverter or "NOT" element. This is a single input-single output element, the output being at the "1" level when the input is "0" and vice versa. The symbol used is shown in Figure 6.17(c). The "NOT" element is frequently combined with an "AND" element to produce a gate which gives an output of "1" only when certain inputs are at the "1" level while others are at "0". (Figure 6.17(d)).

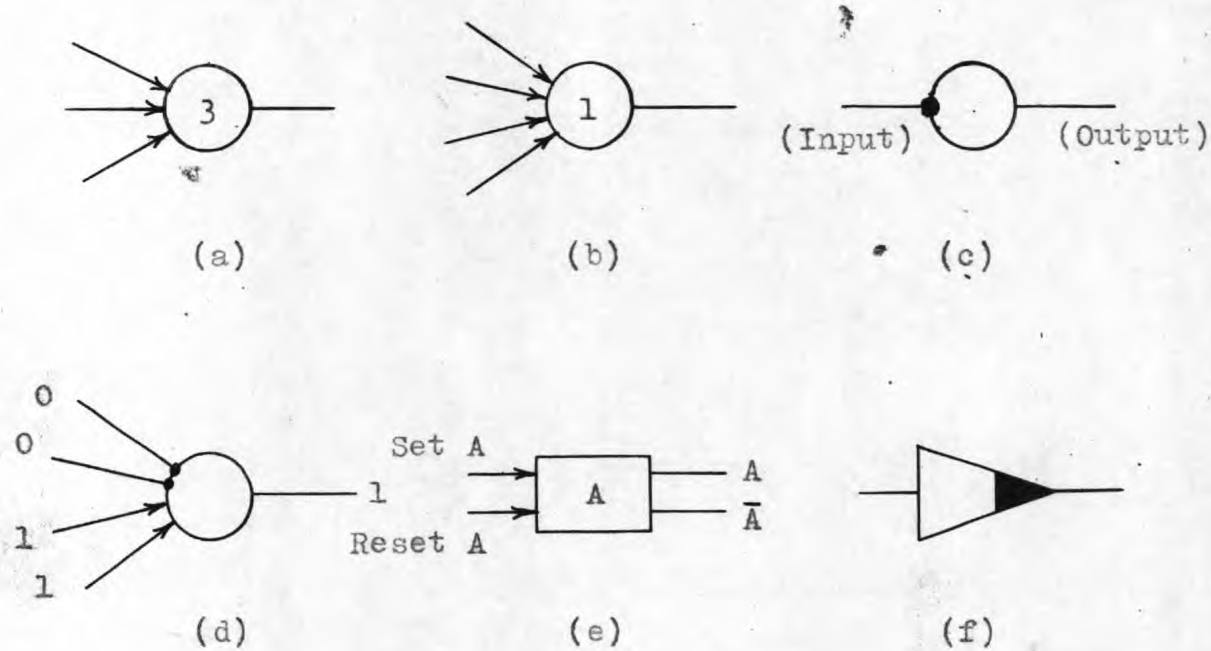


Figure 6.17. (a) a 3-input AND gate, (b) a 4-input OR gate, (c) a NOT element or inverter, (d) a NOT AND gate, (e) a bistable or toggle, (f) my own symbol for a cathode- or emitter-follower.

The remaining element found in pure logic systems is the bistable or toggle. This is a memory device, the output remaining at the "0" level until a "1" at the "SET" input switches the element into its other state. The output is now "1" until such time as a "1" is applied to the "RESET" input. A complementary output is often available also (Figure 6.17(e)).

Diode-resistor gates have been used extensively in my equipment and one point concerning them which could cause confusion should be mentioned here. The difficulty arises because an "OR" gate in positive logic is identical with an "AND" gate in negative logic and vice versa. Figure 6.18 shows a typical circuit. With positive excursions the circuit acts as an "OR" gate since the output must follow the most positive of the inputs. With negative signals, the output cannot drop below the quiescent voltage level Q until <u>all</u> the inputs make a negative excursion.

Emitter- or cathode-followers, while not strictly logic elements, are frequently used in logic systems as impedance transformers. A convenient symbol for this circuit has therefore been coined, the output end of the familiar amplifier symbol being filled in to suggest the "solid" or low impedance output characteristic. (Figure 6.17(f)).

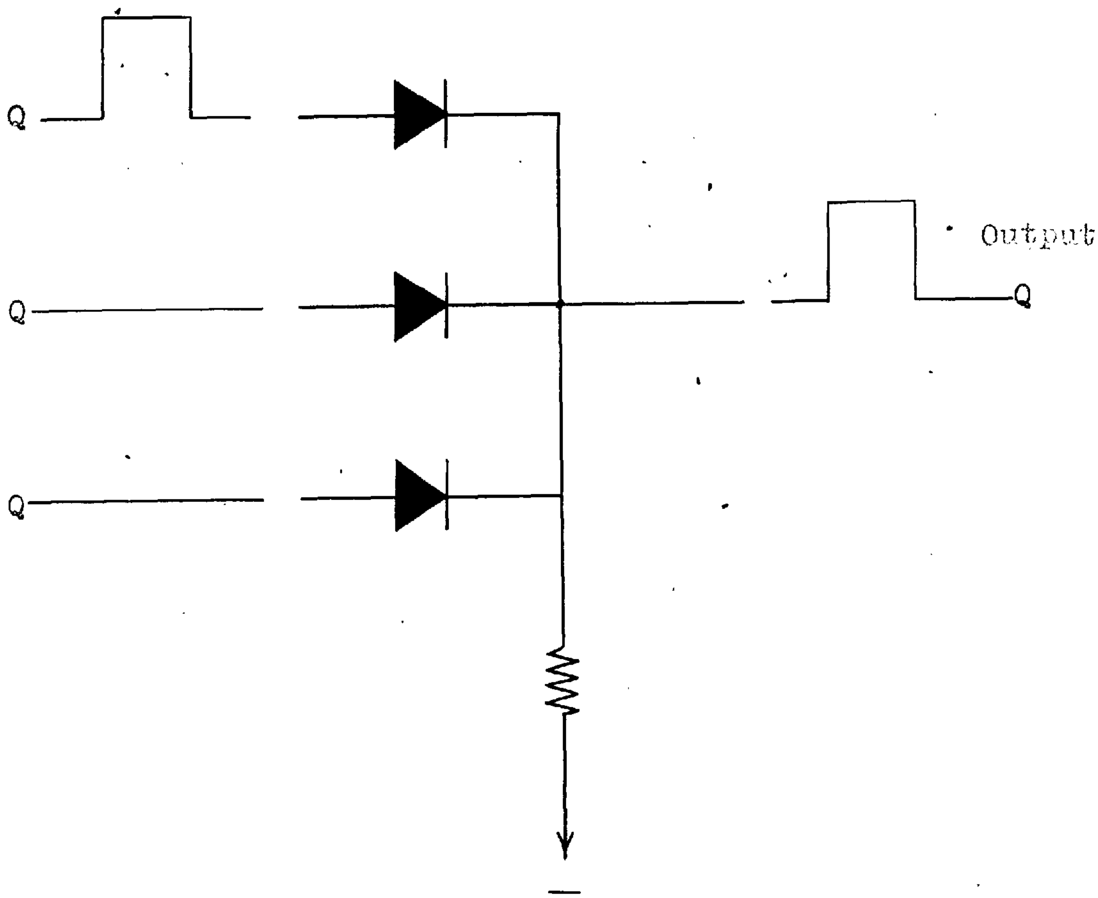


Figure 6.18. Circuit which is a 3-input OR-gate for positive-going signals but a 3-input AND-gate for negative-going signals

Appendix 4

A Number of possible scenes reproducible

Of the 405 lines forming each frame in the British television system, only 377 carry video information, the remainder being blanked out so as to be invisible during frame flyback. If it can be assumed that there should be as many pictureselements per unit length of line as there are lines per unit of height, then the number of elements per line is 377 x 4/3 since 4/3 is the standard aspect ratio. (This neglects the Kell factor[30] which allows for the reduction of resolution in the vertical direction due to line structure and is usually given a value of about 0.7.) Thus the total number of elements presented to the viewer in a single frame is

$$(377)^2 \times 4/3 = 189,700$$

If each picture element can have any one of 128 brightness levels (as suggested by Goodall[21]), and all elements are independent, then the number of possible scenes is

$$128^{189,700}$$

$$\text{i.e. } \underline{\underline{10^{339,000}}}$$

B Bandwidth required for a 405-line picture

In order to create an illusion of motion, at least 25 frames must be transmitted every second. This rate results in annoying flicker, however, so alternate lines are transmitted in one field (taking 1/50 sec.), the remaining lines forming another field which is transmitted in another 1/50 sec. This process is known as interlacing; whether it is adopted or not, it is still necessary to transmit 405 lines in 1/25 sec. making the line period 98.8μS. Of this, about 17.5μS is blanked out to allow for line flyback; thus 377 x 4/3 elements are transmitted in 81.3μS and the time per element is 0.162μS.

When elements are alternately black and white, the signal generated is ideally a square wave but in practice the finite size of the scanning spot and the limited resolution of the camera mosaic result in the fundamental component only being produced. The frequency of this component is given by

$$f_c = \frac{1}{2T}$$

where T is the time per picture element.

Thus $$f_c = \frac{1}{2 \times 0.162} = \underline{\underline{3.09 \text{ Mc/s}}}$$

At the other end of the frequency scale, response down to the frame frequency (25 c/s) must be provided. In practice a bandwidth requirement of 3 Mc/s is generally assumed.

7. REFERENCES

1. ANDO, H. - "Quantising System of Higher Components of T.V. Signals". J. Inst. Elec. Comm. Eng. Japan 42, 5, May 1959.

2. BECKER, F.K., HEFELE, J.R. and WINTRINGHAM, W.T. - "An Experimental Visual Communication System". B.S.T.J. 38 January 1959.

3. BEDDOES, M.P. - "Investigations into Transmission Systems Using Variable-Velocity Scanning for Bandwidth Reduction of Facsimile and Television Systems". D.I.C. Thesis, Imperial College, London, 1956.

4. BEDFORD, L.H. and PUCKLE, O.S. - "A Velocity Modulation Television System". J.I.E.E. 75 July 1934. pp.63-82.

5. BELL, D.A. and HOWSON, E.A. - "Two to one television bandwidth reduction scheme". Paper presented at Brit. I.R.E. Television Convention, Cambridge, July 1959.

6. BLUMLEIN, A.D., BROWNE, C.O., DAVIS, N.E. and GREEN, E. - "The Marconi-E.M.I. Television System". Proc. I.E.E. (Wireless Section) 14 No. 40, March 1939.

7. CHERRY, E.C. - "On Human Communication". Wiley and Chapman and Hall, 1957.

8. CHERRY, E.C. and BEDDOES, M.P. - "Television Compression System". British Patent Application No. 35465/56 November 1956.

9. CHERRY, E.C. and GOURIET, G.G. - "Some Possibilities for the Compression of Television Signals by Recoding". Proc. I.E.E. 100 Part 3, 1953.

10. CHERRY, E.C., KUBBA, M.H., PEARSON, D.E. and BARTON, M.P. - "Experimental Study of the Possible Bandwidth Compression of Visual Image Signals". Proc I.E.E.E. 51 No. 11 November 1963.

11. CHERRY, E.C., PRASADA, B. and HOLLOWAY, D.G. - "Improvements in or relating to Signal Transmission Systems". British Patent Application No. 32,376/58.

12. CHERRY, E.C., PEARSON, D.E., KUBBA, M.H., BARTON, M.P. and VOELCKER, H.B. Jr. - "Improvements in or relating to television and like data transmission systems". British Patent Application No. 33125/62 August 1962.

13. CUNNINGHAM, J.E. - "Picture Processing". Quarterly Progress Report No. 54. M.I.T. Research Lab. of Electronics, July 1959. P.138.

14. DEUTSCH, S. - "Narrow-band T.V. uses Pseudo-Random Scan". Electronics. 27th April, 1962.

15. DEUTSCH, S. - "The possibilities of reduced television bandwidth". I.R.E. Trans. on Broadcast and Television Receivers 1956.

16. ELIAS, P. - "A note on autocorrelation and entropy". Proc. I.R.E. 39 No. 7, July 1951 P.839 (Correspondence).

17. FANO, R.M. - "The transmission of information". M.I.T. Research Lab. of Electronics. Technical Report No. 65, 1949.

18. FIELLER, E.C., LEWIS, T., and PEARSON, E.S. - "Correlated Random Normal Deviates" Tracts for Computers No. XXVI, Cambridge University Press, 1955.

19. FINK, D.G. - "Television Engineering Handbook" McGraw-Hill, 1957.

20. GABOR, D. - "Theory of Communication". J.I.E.E. 93 Part III 1946 P.429.

21. GOODALL, W.M. - "Television by P.C.M." B.S.T.J. 30 1951, P.33.

22. GOURIET, G.G. - "The Compression of a Television Signal". Proc. I.E.E. 104 B No. 15, May 1957.

23. GRAHAM, R.E. - "Communication Theory applied to Television Coding". Acta Electronica 2 Nos. 1 and 2, 1958.

24. GRAHAM, R.E. - "Predictive Quantising of Television Signals". I.R.E. Wescon Convention Record. Part 4. 1958.

25. HARRISON, C.W. - "Experiments with Linear Prediction in Television". B.S.T.J. 31 July 1952. P.764.

26. HARTLEY, R.V.L. - "The Transmission of Information" B.S.T.J. 7. July 1928 P.535.

27. HILL, P.C.J. - "Television Band Compression by Contour Interpolation". Ph.D. Thesis, Faculty of Engineering, University of London, March 1960.

28. JESTY, L.C. - "Television as a Communication Problem". Proc. I.E.E. 99 Part IIIA No. 20 1952. P.761.

29. JULESZ, B. - "A Method of coding television signals based on edge detection" B.S.T.J. 38 No. 4 July 1959.

30. KELL, R.D., BEDFORD, A.V. and TRAINER, M.A. - "An experimental television system". Proc. I.R.E. 22 November 1934.

31. KRETZMER, E.R. - "Reduced alphabet representation of television signals". I.R.E. Convention Record. Part 4. 1956.

32. KRETZMER, E.R. - "Statistics of Television Signals". B.S.T.J. 31 July 1952.

33. KUBBA, M.H. - "Methods of Measuring picture detail in relation to television signal bandwidth compression." Ph.D. Thesis. Faculty of Engineering. University of London. April 1962.

34. KUBBA, M.H. - "Open Loop Television Bandwidth Compression System". Progress Report No. 2 October 1959.

35 MACNAMARA, T.C. and BIRKINSHAW, D.C. - "The London Television Service". Proc. I.E.E. (Wireless Section) 14 No. 40, March, 1939.

36. McCULLY, G.R. - "Experiments upon a method of Signal Compression for Electrical Communication" M.Sc. Thesis, University of London, 1954.

37. MICHEL, W.S., FLECKENSTEIN, W.O. and KRETZMER, E.R. - "A coded facsimile system". Bell Labs. 1957.

38. MILLMAN, J. and TAUB, H. - "Pulse and Digital Circuits". McGraw Hill. 1956.

39. MOOD, A.M. - "Introduction to the Theory of Statistics". 2nd Edition. McGraw Hill 1963.

40. N.R.D.C. - "Improvements in Television and like data transmission systems". British Patent Application No. 24053/62. June 1962.

41. NYQUIST, H. - "Certain Factors Affecting Telegraph Speed". B.S.T.J. 3 April 1924.

42. OLIVER, B.M. - "Efficient Coding". B.S.T.J. 31 July 1952.

43. OLIVER, B.M., PIERCE, J.R. and SHANNON, C.E. - "The Philosophy of PCM". Proc. I.R.E. 36. November 1948.

44. PINE, C.S.F. - "Studies in Coding for Bandwidth Reduction of Television Signals". Ph.D. Thesis, Faculty of Engineering, University of London, 1960.

45. PRASADA, B. - "Studies of Picture Signal Bandwidth Compression". Ph.D. Thesis, Faculty of Engineering, University of London 1960.

46. SASULY, M. - "Trend Analysis of Statistics". The Brookings Institution, Washington, 1934.

47. SCHREIBER, W.F. - "The measurement of third-order probability distributions of television signals". Research Department, Technicolour Corporation, Burbank, California.

48 SCHREIBER, W.F. and KNAPP, C.F. - "Television Bandwidth reduction by Digital Coding". I.R.E. Convention Record. 1958, Part 4, P.88.

49. SCHREIBER, W.F., KNAPP, C.F. and KAY, N.D. - "Synthetic Highs - Experimental Bandwidth reduction system". 84th S.M.P.T.E. Convention, October 1958.

50. SCHRÖTER, F. - "Speicherempfang und Differenzbild Fernshen". Archiv der Elektrischen Ubertragung, 7 1953, P.63.

51. SEKEY, A. - "Digital Detection in Television Signals". Proc. I.E.E.E. 53 No. 1. January 1965, P.75 (Correspondence).

52. SEYLER, A.J. - "Bandwidth reduction in television relaying". Proc. Inst. Rad. Eng. Australia, 16, July and August 1955.

53. SHANNON, C.E. - "The Mathematical Theory of Communication". B.S.T.J. 27, 1948, Pp. 379 and 623. Also in book form with WEAVER, W. University of Illinois Press, 1949.

54. SWIFT, J. - "Adventure in Vision". John Lehmann, London, 1950.

55. SZIKLAI, G.C. - "Some studies in the speed of visual perception". Trans. I.R.E. on Info. Theory. IT2, No. 3, September 1956.

56. TEER, K. - "Investigations into Redundancy and Possible Bandwidth Compression in Television Transmission". Philips Research Reports 14 and 15, December 1959 and February 1960.

57. TREUHAFT, M.A. - "Description of a system for transmission of line drawings with bandwidth-time compression." Final Report R339-53. PIB 274. Microwave Research Inst. Polytechnic Inst. Brooklyn 1, N.Y.

58. TSUKKERMAN, I.I. - "Transmission of Co-ordinates of Television Picture Elements". Radiotecknika 13, 4 Pp77-9, 1958 (In Russian).

59. VOELCKER, H.B. - "Notes on a class of Compression Schemes." Unpublished report, Imperial College, March 1959.

60. WOODWARD, P.M. - "Probability and Information Theory with applications to radar" Pergamon Press, London, 1957.

61. YOUNGBLOOD, W.A. - "Estimation of Channel Capacity required for Picture Transmission". Sc.D. Thesis, Department of Electrical Engineering, M.I.T. May 1958.