

STUDY OF CERTAIN FORMS OF
TERNARY PULSE CODE MODULATION SYSTEMS

A Thesis Submitted for the
Degree of Doctor of Philosophy

by

Thevalil Chathunny Sidhan

Electrical Engineering Department
Imperial College,
LONDON.

ABSTRACT

The contributions presented in this thesis can be classified under two categories:

A. Theoretical investigations leading to the development of analytical techniques for the analysis and synthesis of multiple-valued digital-data transmission systems, and

B. Experimental studies concerned with the implementation of practical systems for transmission of multiple-level signals, special emphasis being placed upon ternary pulse code modulation systems.

A. Theory

Certain new concepts are introduced that help to define an information theoretic measure for the coding efficiency of discrete-data systems. It is shown that certain existing statistical coding techniques can be further improved. The transmission efficiency of a communication system operating under noisy conditions is now defined, and its magnitude evaluated for different systems. The advantages of a ternary system over others for transmission of information are brought out. These investigations also led to a more general formulation of the concept of channel capacity.

Equivocation in binary and ternary systems is next studied.

Techniques of symbolic calculus are used to evaluate the signal to false-pulse noise ratio in a ternary PCM system. Methods for improving this ratio are suggested. The problem of optimum detection of ternary signals in the presence of noise is then formulated and solved.

The effect of quantization on system performance is studied in detail using certain well known results of Applied Statistics. Techniques are described by which the reduction in the entropy of the signal, caused by quantization, can be minimized.

A generalized class of discrete-data transmission systems, of which PCM and delta modulation are particular cases, is studied. The performance of these systems, for different types of statistical inputs, is evaluated. Studies in system stability led to some interesting investigations into the behaviour of non-linear sampled-data control systems.

B. Experimental Studies:

The development of a fast tristable circuit and the actual performance of a ternary feedback encoder incorporating tristable circuits, are described. The design of a simple ternary decoder is also given.

ACKNOWLEDGEMENT

It is with great pleasure that the writer acknowledges his indebtedness to Professor Colin Cherry, for his continued encouragement and help throughout the execution of the project.

Grateful acknowledgement is also made to the University Grant Commission whose equipment grant of £100 - made the experimental work mentioned in this thesis possible.

CONTENTS

	<u>Page</u>
Abstract	1
Acknowledgement	3
1. Introduction	8
1.1. Basic Operations in Pulse Code Modulation	9
1.2. Performance of PCM Systems	13
1.2.1. Quantization Distortion	13
1.2.2. Overloading	15
1.2.3. False Pulse Noise	15
1.2.4. Assessment of the Performance of PCM Systems	17
2. Coding Efficiency of an Encoding System with base b	20
2.1. Optimum Codes for Discrete Data	21
2.1.1. Determination of Optimum Base for Encoding.	21
2.1.2. Alternative Method for Determination of Optimum Base.	24
2.2. Significance of the Concept of Information Capacity per State of a Symbol	25
2.2.1. Significance of the Concept in the Study of Coding Systems	27
2.2.2. Assessment of Coding Efficiencies of Different Systems	29
2.3. Applications	31
2.3.1. Possibility of Compression of Binary Messages	31
2.3.2. Efficiency of Shannon-Fano Code	35
2.3.3. An Optimum Technique for Encoding Messages of Unequal Probability	36

3.	Transmission Efficiency of Pulse Code Modulation Systems	38
3.1.	Transmission Efficiency of a Communication System	40
3.2.	Transmission Efficiency of Binary PCM.....	41
3.2.1.	Capacity of Binary Systems... ..	48
3.3.	Transmission Efficiency of Ternary PCM ...	48
3.3.1.	Capacity of a Ternary System	58
3.3.2.	Comparison of Binary and Ternary Systems..	66
3.4.	Evaluation of Efficiency when Noise is not Gaussian	67
4.	Equivocation in Binary and Ternary PCM Systems.	70
5.	Signal to False-Pulse Noise Ratio in a Ternary System	73
5.1.	Signal to Noise Ratio for a Code Group consisting of 1 Digit... ..	73
1.1.	Anomalous Behaviour of S_0/N_0 for Low Values of S/N	78
5.2.	Signal to Noise Ratio for a Code Group consisting of n Digits	80
5.2.1.	Maximization of S_0/N_0 of an n-Digit Ternary System	86
5.3.	Detection of Ternary Signals in the Presence of Noise	89
5.3.1.	Graphical Representation of the Ternary Detection Problem	92
6.	Analysis of Quantization Distortion	96
6.1.	Evaluation of Quantization Distortion ...	96
6.1.1.	Quantization Distortion when l is large .	97
6.1.2.	Quantization Distortion when l is small .	100
* 6.2.	Minimization of Quantization Distortion .	111
* 6.1.3.	General expression for quantization noise.	107

6.3.	Effect of Quantization on Entropy of the Signal	113
6.3.1.	Choice of the Nonlinear Characteristic $N(A)$	119
6.3.2.	Choice of the Linear Shaping Network $L(j\omega)$	123
7.	Extension of the Philosophy of PCM	125
7.1.	Description of Suggested System	126
7.2.	Extrapolation of Sampled Signals	130
7.2.1.	Accuracy of Extrapolation	133
7.3.	Synthesis of the Extrapolator	134
7.4.	Representation of Stochastic Signals	146
7.4.1.	Least Square Estimation of Sample Magnitudes	146
7.4.1.1.	Least Square Estimate for Gaussian Signals	148
7.4.1.2.	Least Square Estimate for Non-Gaussian Signals	153
7.4.1.3.	Least Square Estimates for n^{th} order Markov Processes	155
7.5.	Exponential Extrapolation	161
7.6.	Improvement in System Performance due to the Use of the Extrapolator	165
7.6.1.	First Order Correlation	165
7.6.2.	Higher Order Correlations	167
7.7.	Effect of Quantization on System Performance	170
7.7.1.	Effect due to Quantization Noise	170
7.7.2.	Effect of Nonlinearity on Loop Gain of the System	173
7.7.3.	Statistical Transfer Function of Quantization Nonlinearity... ..	180
8.	Tristable Circuits	185

9.	Ternary Encoder....	...	...	...	...	...	194
9.1.	Encoder I....	...	...	...	...	...	196
9.2.	Encoder II ..	...	...	...	...	...	203
10.	Ternary Decoder	...	...	...	...	...	217
11.	Summary of Conclusions	...	...	...	...	...	225
12.	Appendices	...	...	...	...	...	229
13.	References	...	...	...	...	...	241

STUDY OF CERTAIN FORMS OF TERNARY PULSE CODE MODULATION SYSTEMS

1. INTRODUCTION

An important development in the history of communication has been the rapid growth in recent times, of long distance communication channels. The inherent limitations of conventional systems of modulation, such as amplitude modulation, for transmission of signals became more and more apparent as the communication channels became longer and the transmission requirements more stringent. An amplitude modulated signal, for instance, is affected by channel noise, inter-channel cross-modulation products and linear as well as non-linear distortion in the channel. If the transmission path is sufficiently long, the interference accumulated over the total path length may well be enough to make it impossible to retrieve the original signal. A useful method of reducing the magnitude of this external interference is to use several repeater sections of coaxial line or shielded cable pairs, in tandem, as the channel. This technique, however, does not help in discriminating against the internal noise, generated in the amplifiers themselves. As an average long distance transmission line may have as many as five hundred repeater sections or more, it is easy to see that the thermal noise from the repeaters alone would be sufficient to make high-fidelity transmission of signals over long distances, very nearly impossible.

A possible solution to this fundamental problem in long distance communication was brought to light when the noise-cleaning properties of frequency modulation were demonstrated

in 1936.⁽¹⁾ It was shown that by suitably processing the primary signal, which carries the message, it is possible to produce a secondary signal which can be "cleansed" of the noise, which it accumulates during transmission through the channel, as often as necessary. It was further demonstrated that it is possible in such a system to decrease the signal power required to achieve a given quality of transmission, in a channel of specified noise power, by suitably increasing the bandwidth of transmission. Also the noise-cleaning capacity of the system was shown to be related to the system's ability to trade bandwidth with the signal-to-noise ratio in the channel.

The interest aroused by this original concept led to two main forms of activity: one was concerned with the extension of earlier concepts of communication theory^(2,3) to include new factors such as communication in the presence of noise, multiple regeneration of signals, and the effect of statistical structure of signals and of noise on efficient communication.⁽⁴⁾ This theory made it meaningful to discuss the performance of an ideal communication system, and established qualitative criteria for comparing the performance of different systems. The other result was to stimulate further study of the performance of unorthodox types of modulation such as pulse position modulation⁽⁵⁾ and pulse duration modulation⁽⁶⁾ which had already been discovered, but had aroused, so far, little interest; and it also gave an increased impetus to the development of newer and more efficient systems of communication. It was during this period of sustained activity in the theory and development of modulation systems, that pulse-code modulation (or PCM), one of the most efficient and versatile communication systems in use today,⁽⁷⁻¹¹⁾ was revived.

1.1 Basic Operations in Pulse-Code Modulation

PCM incorporates certain concepts such as "quantizing" and "encoding" which make it radically different from other

modulation systems including other pulse modulation systems.⁽¹²⁾ A brief description of the principal operations of PCM is given below.

Sampling: The signal is applied to the input of an instantaneous sampler which periodically samples the signal at a rate slightly higher than twice the frequency of the highest frequency component present in the signal. According to the sampling theorem, these samples completely and unambiguously determine the original signal.^(13,14) These discrete samples represent data which can then be individually processed into a secondary signal possessing more suitable characteristics for transmission through a given channel.

Holding: The output of the sampler is passed on to a holding circuit, which maintains an output voltage level equal to that of a given sample during the interval until the next sample arrives. This operation facilitates the later operations of quantizing and encoding.

Quantization: The samples emerging from the holding circuit are next subjected to amplitude quantization. All samples falling within a given interval are specified by the same quantization level, usually chosen to be the midpoint of the interval. This operation somewhat distorts the signal as the amplitude of the quantized sample may differ from that of the original by as much as half the magnitude of the quantization interval.

Encoding: If the magnitude of the peak interference in the channel due to noise is less than half that of a quantum step, the received signal can be restored to its original value by regeneration at the receiver; an error can occur only if the peak noise is greater than this value. Encoding serves to transform the sampled and quantized primary signal, having a large number of levels of quantization and therefore very

susceptible to noise, into a secondary signal which has fewer and larger specified amplitude levels and consequently possesses more suitable characteristics for transmission through a noisy channel.

The transmission characteristics of the PCM system are determined by the choice of the code used for encoding. A code may be described as a scheme by which each member of a finite set of discrete magnitudes is represented by a particular arrangement of the members of a different finite set of discrete events. Each event in the latter set is called a code element. Each different possible arrangement of these code elements is called a code group, and represents the magnitude of a single member of the first set.

In PCM, all the possible different levels, say ℓ , of a quantized sample can be represented by the ordered set of integers $(0, 1, 2, \dots, (\ell - 1))$ and, therefore, by the symbols or values $(0, 1, 2, \dots, (\ell - 1))$ of a single digit belonging to a number system of base ℓ . The encoding process consists in setting up a one-to-one correspondence between this system, which is characterised by a single digit of base ℓ , and another system with a group of n digits of base b . The order of the code used is determined by the base b of the new system. For instance, if b is two, a binary code is obtained, each code element of which is restricted to either one of two symbols for example, the absence or presence of a pulse, or alternatively, a positive or a negative pulse. In a ternary code, for which b is three, a code element can have one of three distinct values: it can, for instance, be a pulse which can be positive, null or negative. An N 'ary code is obtained if b is equal to N , and is characterised by code elements which can have any one of N distinct values.

The number of discrete data required to specify a continuous signal of bandwidth W and duration T increases

from $2WT$ symbols before encoding, to $(n)(2WT)$ after, where n is the number of code-elements (or digits) per code group as given by the relation

$$\ell = b^n$$

If the transmission time remains constant, PCM therefore requires a bandwidth n times as great as that required for direct transmission. Since ℓ is, in general, very much greater than b , the increase in bandwidth is usually quite considerable.

Regenerative Repeaters: Repeaters are used at frequent intervals in a long distance communication channel to compensate for the attenuation as well as distortion and noise in the channel. In a PCM system, since the signal elements occur at distinct, predictable instants of time and are completely specified both in shape and magnitude, the accumulated distortion and noise can be almost completely eliminated by identifying the code element and generating a similar one at the repeater itself. The overall performance of a PCM system can therefore be made substantially independent of the total distance of transmission.

Decoding: At the receiver, each code group corresponding to a given quantized signal sample is separated from the rest. The contributions of each of the code elements constituting a given code group are then added together, after weighting each code element according to its own magnitude as well as its position in the code group. For instance, the ternary code group 0122 is made to yield the weighted sum

$$\begin{aligned} & [0 \times 3^0 + 1 \times 3^1 + 2 \times 3^2 + 2 \times 3^3] \\ &= 0 + 3 + 18 + 54 \\ &= 75 \end{aligned}$$

At the termination of each code group, a pulse equal in magnitude to the weighted sum of the code elements of the group, is generated. This pulse is applied to the input of a low pass

filter of cut-off frequency equal to that of the highest frequency component present in the primary signal before it was sampled. Under normal operating conditions, the output of the filter is an almost exact reproduction of the original signal except for the distortion introduced by the process of quantization at the transmitter.

1.2 Performance of PCM Systems

A signal transmitted by PCM is subject to three kinds of transmission impairment: a) quantization distortion, described above, b) overloading, due to the finiteness of the range of quantization and c) false pulse noise, which occurs when the transmitted code elements are incorrectly interpreted, either during regeneration at the repeaters or in the process of decoding at the receiver. These factors have been extensively studied in the literature.

1.2.1 Quantization Distortion

The error introduced by amplitude quantization varies from sample to sample and has a maximum value of plus or minus half a quantum step. Since the variation of the error with time is unpredictable, quantization distortion has the characteristic of noise rather than of distortion, and represents an irreversible loss of information.

Quantization error power is proportional to the square of a quantum step. Therefore, by reducing the size of a quantum interval by increasing the number of levels, the magnitude of this error power can be reduced. But, for a given system of coding, increasing the number of levels also increases the number of code elements required for sample specification, and consequently, the bandwidth for transmission. If the transmission bandwidth is fixed, the number of levels of quantization that is permissible is determined by the base ^b of the code used for encoding. The exact relation^{ship} between

the r.m.s. signal to quantization distortion ratio (S/N_Q) , the number (n) of code elements used and the base (b) of the code employed is given by the expression⁽¹⁵⁾

$$S/N_Q = [-10.8 + n (20 \log_{10} b)] \text{ db.}$$

so that the signal to quantization distortion ratio is proportional to the logarithm of the base b of the code. For instance, a unit increase in the number of digits per code group improves the signal to distortion ratio of a binary system by 6 db, of a ternary system by 9.54 db, of a quaternary system by 12 db and of a decimal system by 20 db.

The above formula is derived on the assumption that the input signal has a rectangular distribution of amplitude, that quantization levels are equally spaced and that their number is large. Under these conditions, quantization distortion has a rectangular amplitude distribution and is substantially uncorrelated with the input signal.⁽¹⁶⁾

Another important problem that has been investigated in detail is the possibility of reducing the magnitude of quantization distortion for signals having a non-rectangular amplitude distribution, such as gaussian or exponential, by logarithmically compressing the signal before subjecting it to quantization.⁽¹⁷⁻¹⁹⁾

The emphasis in all the work mentioned above is on the evaluation of the quantization error power. There has been very little study of the statistics of quantization noise, either of its own statistical structure or of the relationship between the signal and the quantization noise. This is especially true for the case when the number of levels of quantization is small. But, to evaluate the performance of

complex PCM systems (such as discussed in Sections 6 and 7) which make use of the probability constraints that exist between adjacent samples of the signal to obtain higher efficiency of communication, it is necessary to have a detailed knowledge of these statistics.

1.2.2. Overloading

The contribution to the distortion in the quantized signal output due to the finiteness of the total range of quantization, becomes vanishingly small and can be neglected for all practical purposes if the latter is greater than about four times the r.m.s. value of the input signal. It is easy to meet this condition when the number of quantization levels is large. If, however, the number of levels is small, the problem of minimization of the error power due to overloading becomes important. This problem, especially in the case of complex PCM systems employing feedback (Section 7), is, in general, not very simple to solve since it involves the simultaneous minimization of the quantization error power as well.

1.2.3. False-Pulse Noise

False-pulse noise is produced when a code element is wrongly interpreted⁽¹⁴⁾. Such misinterpretation can occur only if the channel noise is greater than half the separation between adjacent specified levels of the code element. If the interference is due to white noise, it is easy to show that if this amplitude separation is greater than a "threshold" value, the probability of error becomes so low that the transmission can be considered to be almost perfect.⁽¹⁴⁾ The threshold value is generally taken to be about ten times the r.m.s. value of the noise power in the channel.⁽¹⁴⁾

The characteristics of false-pulse noise have been studied in detail for binary systems. When the system is

operating above the threshold level, the statistics of false-pulse noise are very similar to those of impulse noise, and the quality of transmission is very little affected.⁽²⁰⁾ The introduction of even several hundred repeaters in the channel has been shown to have very little effect on the quality of the received signal, if the signal power in each repeater section is maintained at a slightly higher level than the threshold level for a communication channel consisting of a single repeater section.⁽²¹⁾

When a signal is encoded into a group of code digits, the misinterpretation of some code elements is more serious than the misinterpretation of others, since each code element is differently weighted. It has been shown that the false-pulse noise power can be minimised by weighting the power allotted to the various digits of the code group according to the contribution the un-weighted digit makes to the total false-pulse noise power.^(22,23)

Most results of this kind published relate to binary systems. There has been very little study of false-pulse noise in higher order systems such as ternary or quaternary PCM systems. The only significant work in this field was carried out by Oliver, Pierce and Shannon in their paper "Philosophy of PCM".⁽¹⁴⁾ Here they showed that the signal power required to ensure that the frequency of errors does not exceed a given limit, is approximately proportional to the square of the number characterizing the order of the system. But the usefulness of this result is rather limited. For instance, to mention only one of its limitations, it does not take into account the fact that the effect of an error of interpretation is not the same in the various systems, since the respective code elements have different weights in different systems.

1.2.4 Assessment of the Performance of PCM Systems

As mentioned earlier, it has been demonstrated that the signal power required to transmit a many-valued signal through a noisy channel with a specified frequency of errors is approximately proportional to the square of the number of levels of quantization of the signal.⁽¹⁴⁾ By encoding a signal having a large number of quantization levels into a group of code digits each of which is assigned only a small number of values, it was shown that it is possible to transmit the original signal with less signal power, but with an increase in the bandwidth of transmission. It is instructive to find out whether this trading between bandwidth and signal power has been economical, or more explicitly, whether the reduction, which was made possible by encoding, in the signal power required to maintain a given quality of transmission through a channel of specified noise power, really compensates for the increase in bandwidth of the new system.

This problem was studied by Oliver, Pierce and Shannon⁽¹⁴⁾, who compared the performance of a PCM system with that of an ideal system of communication.⁽⁴⁾ An ideal system can transmit information, with an infinitesimally small frequency of errors, at the rate of

$$W \log_2(1 + P/N) \text{ binary digits/sec}$$

where

W = bandwidth of transmission

P = average signal power, and

N = average noise power in the channel, noise being assumed to be "white".

Assuming that the PCM system is operating above the threshold level so that the frequency of errors is very small,

it was shown that a PCM system can transmit information at the rate of

$$C_{\text{pcm}} = W \log_2 \left(1 + \frac{12}{100} \frac{x S}{N} \right) \text{ binary digits per second}$$

where

S = average signal power required to keep the frequency of errors below the threshold level ≤ 1 part in 3 million

Comparing the "capacities" of the PCM and the ideal systems, it can be seen that in both systems

- (a) signal power and channel bandwidth are exchanged on a logarithmic basis and
- (b) information capacity is proportional to the bandwidth of the system.

On the other hand, in terms of the signal power required to achieve the same channel capacity as that of the ideal system, the efficiency of PCM is seen to be only 12%.

The expression quoted above for the rate of transmission of information in a PCM system, has often been used in the past as a relevant measure of the performance of a PCM system.^(12,13,24) This is rather unfortunate since it is hardly qualified to play such a role. Its main limitations may be summed up as follows:

A. The formula is not an expression for the information capacity of a PCM system, though it is often referred to by that designation in the literature. It is derived on the assumption that the PCM system is operating over the threshold, so that the frequency of errors is small. But, while the probability of an error is small, it is still finite. If the frequency of errors that can be tolerated in the system is made smaller and smaller, the "capacity" of the

system as defined by the formula would become less and less until, in the limit, it will vanish. The definition of the information capacity of a system precludes such behaviour.⁽⁴⁾

B. According to the formula, all PCM systems have the same capacity and therefore the same efficiency. This conclusion seems to be suspect for the following reasons:

i) The formula fails to take into account the fact that the effect of a misinterpretation of a code element is different in different systems. For instance, an error in interpretation of the symbol of a code element having 1000 distinct values produces only about 1 part in 500 as much error as a misinterpretation of a binary digit.

ii) The formula does not take into consideration the fact that the information carrying capacity per state of a symbol is different for different systems (Section 2). For instance, each ^{state} ~~symbol~~ of a binary digit carries more information than each ^{state} ~~symbol~~ of a decimal digit, but less information than each ^{state} ~~symbol~~ of a ternary digit.

2. Coding Efficiency of an Encoding System with Base b .

A quantized pulse sample that can possess any one of ℓ distinct magnitudes, is specified, after the encoding operation, by n separate digits each of which can have one of b distinct levels. It is often assumed^(12-14,24) that the efficiency of a pulse code modulation system in transmitting information is independent of the base b of the code used for encoding*. The validity of this assumption is discussed in the present and the following sections (Sections 2 and 3) where it is shown that coding to certain bases is more efficient than coding to others and that there exists an optimum base for encoding of quantized signals.

The choice of a code for the representation of discrete data is determined by considerations of whether

- 1) the data is to be used in a local static system such as an information storage unit, or whether
- 2) the data is to be transmitted to a distant place through a noisy channel. For the sake of clarity of presentation, the present section is devoted to a critical exposition of the basic ideas involved in the choice of a code to meet only the first requirement; the treatment of the more complex latter case is postponed to the next Section.

* It may, however, be mentioned that it has been shown that choice of e as the radix for the representation of a number minimises the total volume of equipment necessary for a computer of given capacity⁽²⁵⁾. But the assumptions made to derive this result are not well defined

The points of view presented in the following discussion are also rather different from those given in Reference 25.

2.1 Optimum Codes for Discrete Data

In the language of topology, coding maps the original signal from a given signal space on to a new signal space. An optimum code, for a given class of signals, may therefore be defined as that code for which the new signal space has least volume. (13)

2.1.1. Determination of Optimum Base for Encoding

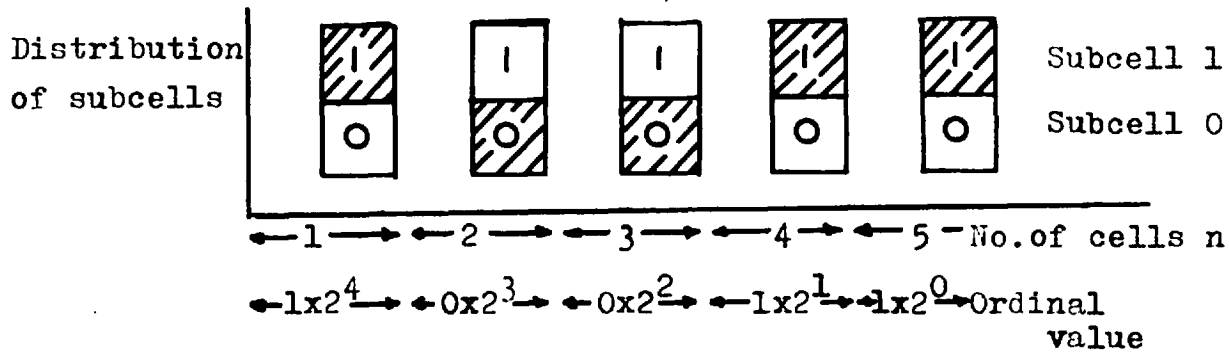
A quantized sample that can possess any one of (ℓ) levels can be represented by a linear array of (ℓ) cells, the k^{th} cell denoting a sample of magnitude (k) , where $k \leq \ell$. After encoding the (ℓ) levels may be considered to be represented by an array of (n) cells each of which has (b) sub-cells, as shown in Fig. 2.1. Any given magnitude (k) of a quantized sample is now unambiguously specified by the selection of (n) sub-cells, one sub-cell being chosen from each cell.

In the discussion which follows a sub-cell is also referred to as a "simple element" while a cell consisting of b sub-cells is called an "element of order b ". For instance a cell consisting of two sub-cells is called an element of order two or alternatively, a binary element.

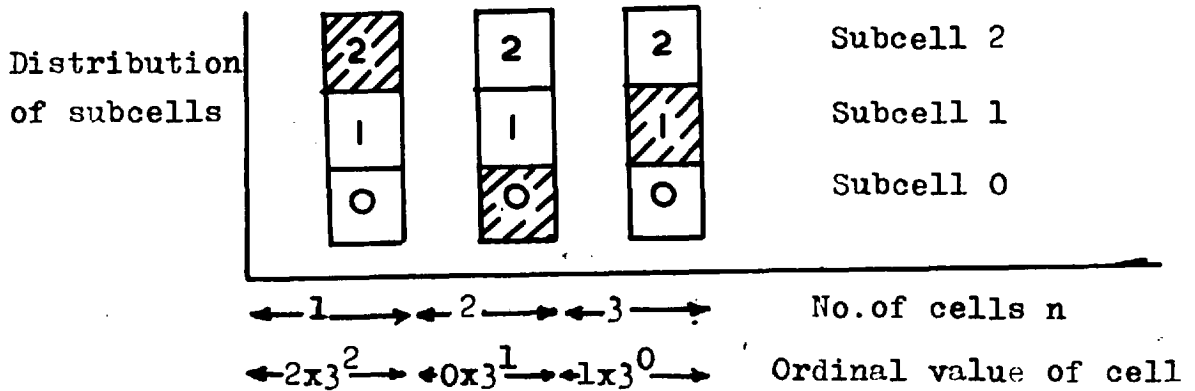
The total number of sub-cells necessary to specify any of the (ℓ) given levels, unambiguously, is equal to $(n b)$. If the number (b) of sub-cells per cell is less than (ℓ) , the total number (nb) of sub-cells will, in general, be less than (ℓ) since

$$\ell = b^n$$

The difference $(\ell - nb)$ may be considered to be a measure of the saving in signal space made possible by coding. The optimum code is obtained when (b) is chosen such that (nb) is



A. Representation by binary cells



B. Representation by ternary cells

Fig. 2.1 Representation of number "NINETEEN" by combination of n cells of order b.

(The term "Cell of Order b" may be used interchangeably with "Device with b Discrete States" or "Multivalued Digit of Radix b")

Occupied subcells marked by



Unoccupied subcells marked by



Total No. of subcells required using binary cells = $2 \times 5 = 10$

Total No. of subcells required using ternary cells = $3 \times 3 = 9$

a minimum for a given value of (l) .

The optimum value of (b) is easily obtained by the method of Lagrange's undetermined multipliers as shown below:

The function to be minimized is $(n b)$, subject to the constraint that (l) is constant.

But

$$l = b^n \quad 2.1.$$

that is $\log_k l = n \log_k b$ 2.2.

where k is an arbitrary constant.

Consider the function

$$U = nb + \lambda (n \log_k b) \quad 2.3.$$

where λ is a Lagrangian multiplier.

Partially differentiating U with respect to (n) and (b) , and equating to zero,

$$\frac{\partial U}{\partial n} = b + \lambda \log_k b = 0 \quad 2.4.$$

$$\frac{\partial U}{\partial b} = n + \lambda \frac{n \log_k e}{b} = 0 \quad 2.5.$$

Solving (2.5) gives

$$\lambda = -b / \log_k e$$

Substituting in (2.4),

$$b - \frac{b}{\log_k e} \cdot \log_k b = 0$$

Therefore, $b = e$.

Since (e) is an irrational number, it is impossible to divide a cell into (e) exact parts. However, since (e) is bounded by the two natural integers 2 and 3, and b^n is a continuous function, it is reasonable to assume that the optimum base is either 2 or 3.

Keeping (nb) constant and allowing l to vary, the

maximum number of elements (l_2) that can be unambiguously specified by coding to base 2 is $2^{(nb)/2}$ while the number of elements (l_3) that can be specified by coding to base 3 is $3^{(nb)/3}$.

Comparing $l_2 = 2^{(nb)/2} = (2\sqrt{2})^{nb}$ and

$$l_3 = 3^{(nb)/3} = (3\sqrt{3})^{nb}$$

it is seen that

$$l_3 > l_2$$

since

$$3\sqrt{3} > 2\sqrt{2}$$

It is thus seen that coding to base 3 provides the greatest economy in the total number of sub-cells or simple elements required to completely represent a given number of discrete data. [Also see Section 2.2]

2.1.2 An Alternative Method for the Determination of Optimum Base

If a sequence of digits

$$a_1 \ a_2 \ \dots \ a_{n-1} \ a_n$$

each of which can have b states is transmitted, the entropy (H) per digit is given by the expression⁽⁴⁾

$$H = - \sum_{i=1}^b p(i) \log p(i)$$

where $p(i)$ is the probability of occurrence of state i of the digit. H may also be written as

$$H = + \sum_{i=1}^b h(i)$$

where

$$h(i) = -p(i) \log p(i)$$

is defined as the entropy per state (i) of the digit.

For a given value of (b), the entropy H per symbol is a maximum when all the states are equally probable so that

$$\begin{aligned} p(i) &= \frac{1}{b} \\ &= \text{constant.} \end{aligned}$$

The maximum value of H is given by the expression

$$\begin{aligned} H_{\max} &= \max \sum_{i=1}^b h(i) \\ &= b \left[-\frac{1}{b} \log \frac{1}{b} \right] \\ &= b [h_b] \end{aligned}$$

The quantity h_b , inside the brackets, represents the average information capacity per state of a digit, for a given value of (b). [Also see Appendix 2.1] The optimum number of states per digit or, in other words, the optimum base for encoding, is given by that value of (b) for which the ~~average~~ information capacity per state of a symbol is maximum. It is shown in Appendix 2.1 that the optimum integral value of (b) is 3.

The magnitude of the ~~average~~ information capacity per state of a symbol (h_b) is plotted as a function of ($1/b$) in Fig. 2.2.

2.2. Significance of the Concept of Information Capacity per State of a Symbol.

The magnitude of the base (b) of the optimum encoding system has been shown above to be 3, using two independent

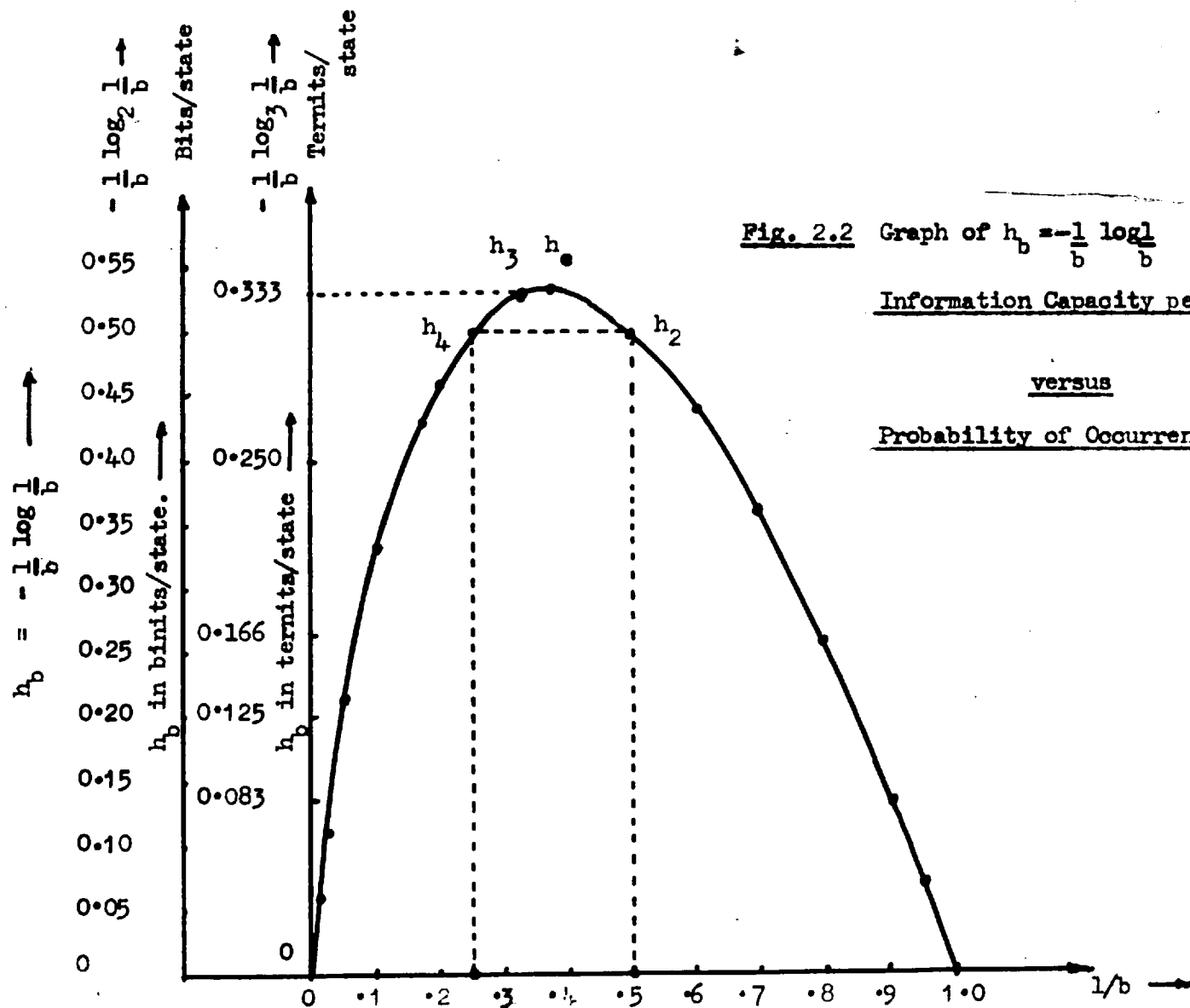


Fig. 2.2 Graph of $h_b = -\frac{1}{b} \log \frac{1}{b}$ versus $\frac{1}{b}$
Information Capacity per State of
a Symbol
versus
Probability of Occurrence of State

criteria. This is not surprising because these represent but two aspects of an efficient encoding system, namely

- 1) determination of the minimum amount of information that would completely specify the message and
- 2) optimal representation of this information.

The concept of ~~average~~ information capacity per state of a symbol provides an optimum method for measuring the information content of the message by specifying how the data should be grouped and each group then sub-grouped so that the total number of operations necessary to specify the information content of the message is a minimum. On the other hand, the concept of minimisation of the number of sub-cells required for the specification of the data provides a criterion for determining the optimum system for representation of the message. Now, if the system is to be used only for the storage of information or only for transmitting the information through a noiseless channel, the problem of guarding against the mutilation of information-carrying symbols by noise does not arise. Under these conditions, both the requirements of an efficient coding system mentioned *earlier* are equivalent: a one-to-one correspondence can be established between the arithmetical procedure by which the information content of the signal is evaluated and the means of representation of the signal, as is actually done, for instance, in the construction of the Shannon-Fano Code for non-equiprobable messages.^(4,26)

2.2.1 Significance of the Concept in the Study of Coding Systems

The concept of information capacity per state of a symbol is important because it has applications not only in the field of pulse code modulation systems, with respect to which

it is studied in later sections, but also in any coding problem involving the processing of discrete data. Studies in coding are generally concerned with only the statistical structure of the message.^(4,12,26) But an efficient encoding system has to take into account not only the statistical structure of the message but the properties of the physical mechanism as well, whereby the encoding is effected - in this case, the properties of ~~denumerably finite classes of~~ ^{The class of} numbers. This is made possible by introducing the concept of information capacity per state of a symbol.

An optimum coding system, such as the Shannon-Fano Code, operates by maximising the relative entropy of the encoded message.⁽¹²⁾ Now, relative entropy is the ratio of the entropy of the message source to the maximum value it could have, subject to the condition that in both cases the same set (or alphabet) of symbols is used.^(4,12) In other words, the concept of relative entropy applies only to the case where the original message and the encoded message make use of the same alphabet. It does not provide any useful criterion to judge the efficiency of encoding when the alphabet is changed. For instance, though the information content of a discrete message is invariant with respect to any reversible* transformation into sequences of binary, ternary, or, in general, N'ary digits, neither the minimum number of operations required for the measurement of the information content of the message, nor~~x~~ the minimum number of discrete elements necessary for its representation is, in general, invariant under such a transformation.**

* An alphabet A is said to be reversibly transformed into another alphabet B if the total number of possible messages in B is greater than or equal to the number of messages in A.

** The only exception to this rule is the transformation between binary and quaternary systems. Both of these systems are equally efficient. The reason for this anomaly is associated with the fact that two (2) is the only number for which the values obtained by adding the number to itself and by multiplying it by itself, are equal: $2 + 2 = 2 \times 2$.

2.2.2 Assessment of the Coding Efficiencies of Different Systems

The expression $h = \left[\frac{1}{b} \log \frac{1}{b} \right]$ for the average information capacity per state of a symbol, has the following properties:

- 1) h is continuous in b .
- 2) If the total number of elements (nb), defined as the number of states (b) per symbol times the number of symbols (n) is kept constant the total number of things that can be specified by these elements, is a monotonic function of h .
- 3) The function h attains a unique maximum for a fixed, finite value of the base b of the system, corresponding to the case of the optimum encoding system.

It, therefore, appears justifiable to define the coding efficiency of a system in terms of the average information capacity (h) per state of a symbol as follows:

The relative coding efficiency (η_b) of an encoding system with base b is the ratio of the average information capacity per state of a symbol of the given system to the average information capacity per state of a symbol of a hypothetical system with base $b = e$:

$$\eta_b = \frac{\left[\frac{1}{b} \log \frac{1}{b} \right]}{\left[\frac{1}{e} \log \frac{1}{e} \right]}$$

The percentage relative coding efficiency defined as $\angle 100 \times \eta_b \angle$ is plotted against the base b employed for coding in Fig. 2.3 .

It will be seen from Figure 3.3 that a ternary coding system has an efficiency of nearly 99.6% while coding to base (2) gives an efficiency of only about 94.4% . It may be

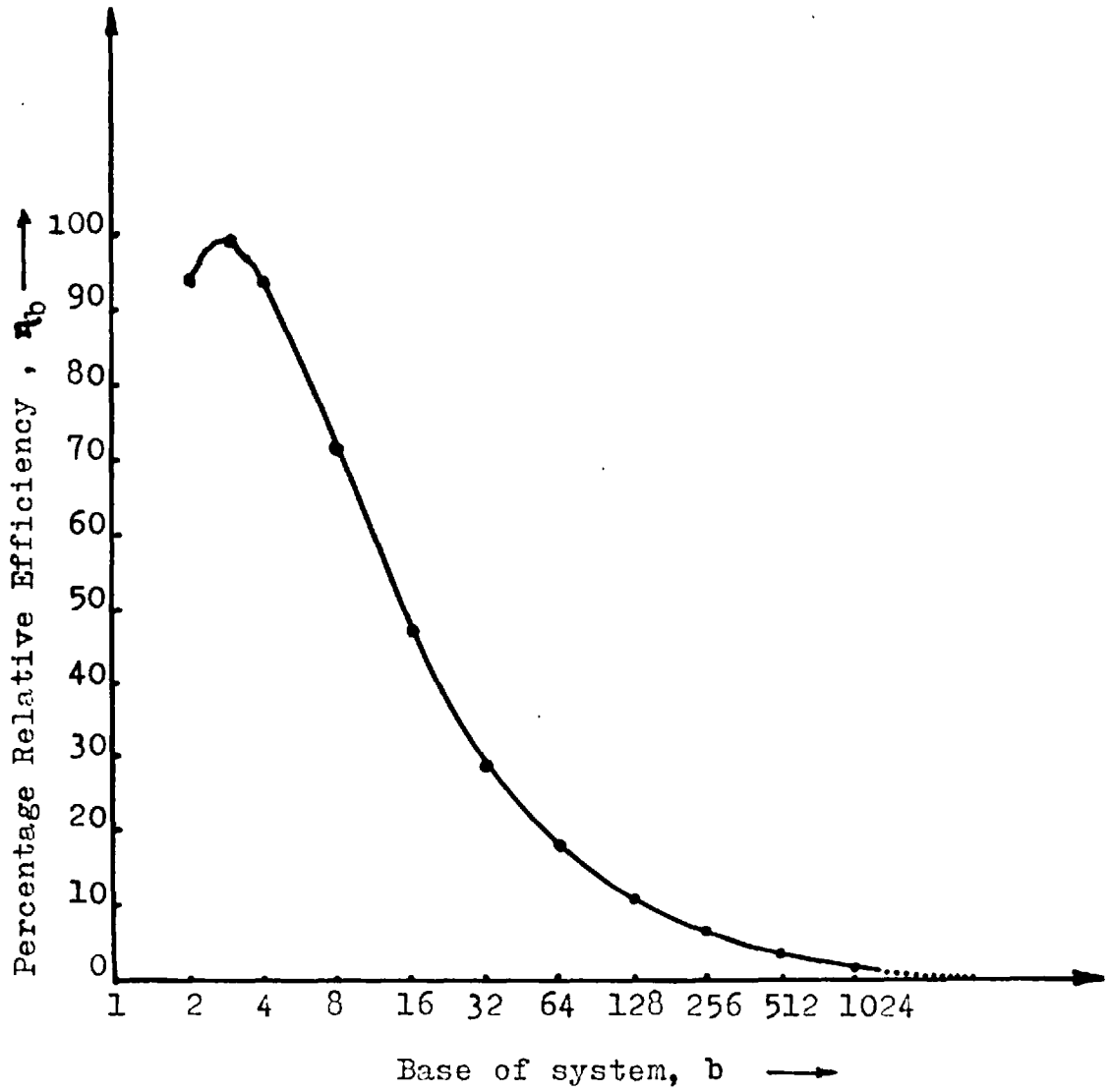


Fig. 2.3 Percentage Relative Coding Efficiency
versus
Base of Encoding System .

noted that coding to base (4) gives a system with the same relative efficiency as an encoding system with base (2). As the base b of the encoding system approaches infinity, the relative coding efficiency approaches zero.

2.3 Applications

The concepts developed above can be made use of to show that certain so-called optimum codes mentioned in the literature (12,13,26) are not really optimum. In the present Section it is shown that an important coding technique called the Shannon-Fano Code, which may be considered to be representative of this class of "optimum" codes, is not optimum; an optimum coding technique, which is a modified version of the Shannon-Fano Code, is then presented.

However, before proceeding to the discussion of these codes, it is instructive to analyse an assumption on which the codes mentioned above seem to be at least in part, based. The assumption is that "binary messages whose states are equiprobable cannot be condensed and that there is no more economical way to store them than to put each message separately into a binary storage unit." (27) It is shown below that this statement is true only in a very restricted class of cases.

2.3.1 Possibility of Compression of Binary Messages

A more precise formulation of the statement quoted above is as follows: "If every possible message has equal probability of being generated by the source, and if the number of possible messages is 2^m , the relative entropy is unity and it is impossible to improve upon the simple code which uses a block of m binary elements to represent each message." (12)

Assuming that there are $N = 2^m$ possible messages and that these are all equally probable, it is easily seen that the

minimum number of binary elements required to represent each possible message is m .

Since two separate simple elements are necessary to specify a binary element, the total number of simple elements necessary for the representation of the message is $(2 \times m)$.

But, any of the messages can be represented by a sequence of n ternary elements where

$$n = \log_3 l$$

Since a ternary element requires three simple elements, for its specification, the minimum number of simple elements that would be necessary to completely represent the message is

$$\begin{aligned} & 3 \log_3 2^m \\ &= 3m \log_3 2 \\ &= 3m / 1.585 \end{aligned}$$

which is less than $(2m)$.

From the above result, it would appear that coding to a ternary base would always require a lesser number of simple elements than in the case of binary coding. This, however, is not the case. For it is implicitly assumed in the above derivation (as well as the derivations in Section 2.1) that the total number of binary or ternary elements used need not be an integral multiple of unity - an assumption that is untenable from the point of view of physical realizability of the system. This would be evident from considering the number of simple elements necessary for encoding a group of, say, 15 messages in the binary and ternary systems: a binary system requires only 8 elements since

$$2^3 < 15 < 2^4$$

while a ternary system requires 9 elements since

$$3^2 < 15 < 3^3$$

It is thus seen that, for a given number of messages, coding to a ternary base can be less efficient than coding to a binary base, under certain conditions.

A more detailed investigation of the comparative coding efficiencies of binary and ternary systems for a given number of messages, yields the following results:

1) A ternary system is always more efficient than a binary system if the number of messages is greater than a minimum value ℓ_0 (equal to 262,144) corresponding to a sequence of 18 binary elements. [The value of ℓ_0 is derived in Appendix 2.2]

2) If the number of messages is less than or equal to ℓ , there are three possible situations:

i) the ternary system is more efficient than the binary system if the number (ℓ) of messages is specified by either of the relations

$$a) 2^{n/2} < \ell \leq 3^{n/3}$$

where

n is an even multiple of 3, or

$$b) 2^{\frac{n-1}{2}} < \ell \leq 3^{n/3}$$

where

n is an odd multiple of 3.

ii) the binary system is more efficient than the ternary system if

$$2^{n/2} \leq \ell < 3^{n/3}$$

where

n is even but not an integral multiple of 3.

iii) the two systems are equally efficient if

$$\left[2^{\frac{n}{2} - 1} + 1 \right] \leq \ell \leq 2^{n/2}$$

where

n is an even multiple of three.

It may be noted that the relationships given above refer only to comparative efficiencies of binary and ternary coding for a given number of messages. If, on the other hand, the maximum number (λ) of simple elements that may be used is fixed, the following simple results are obtained:

1) A ternary system is always more efficient than a binary system if the maximum total number of elements λ is equal to or greater than 33

2) If λ is less than 33, there are only two possible alternatives:

i) The ternary system is more efficient than the binary system if either

a) λ is an odd or even multiple of 3

or b) λ is odd, but not a multiple of 3, and also $\lambda \geq 17$
or $\lambda = 7$

ii) A binary system is more efficient than a ternary system if either

a) λ is even, but not a multiple of 3.

or b) λ is odd, but not a multiple of 3 and also $\lambda \leq 17$,
except when $\lambda = 7$

It is interesting to note that in contrast to the case of a given number of messages, when the maximum number of elements is fixed, there are only two possible alternatives: either the binary or the ternary system is more efficient, but they are never equally efficient for any non-zero value of λ [Appendix 2.3]

Now, any binary message source is bound to produce, within a given length of time τ , a message sequence consisting of at least 18 binary elements, if the duration τ is sufficiently large, say τ_0 . Therefore, if coding delays of the order of τ_0 or greater can be tolerated, a binary message can always be compressed (using fewer elements) by employing a

ternary encoding scheme.

2.3.2 Efficiency of the Shannon-Fano Code

A discrete message, in general, consists of a sequence of, say, N symbols each of which can have any of, say, l distinct values. The relative entropy of such a ^{source of} message is less than unity when the symbols are not independent of each other and/or the probabilities of occurrence of the different states of each symbol are not equal. It is possible to encode such a message into a new sequence of symbols ^{such that the} whose relative entropy ^{of the source} approximates very closely to unity. The Shannon-Fano code provides a simple means for effecting such a transformation.^{4,12,26} Employing a repeated-division process into binary groups of nearly equal probability, it encodes the original message into a sequence of binary elements ^{such that the} whose relative entropy ^{of the source} is nearly unity. (12)

It is easy to show that though this procedure maximises the relative entropy ^{of the source} of the coded sequence, it is, nevertheless, not truly optimum in the sense it requires more elements for the representation of the message than is strictly necessary. For, if the original message is encoded into the binary sequence

$$a_1 \ a_2 \ a_3 \ \dots\dots\dots a_{n-1} \ a_n$$

the sequence requires $(2n)$ simple elements or sub-cells for its specification. But it was shown in Section 2.3.1 that any binary sequence consisting of $(2n)$ simple elements can be uniquely represented by a sequence of ternary elements consisting of fewer than $(2n)$ simple elements if $(2n)$ is equal to or greater than 36 .

It may be noted that in a Shannon-Fano Code, the representation of the message is effected by identifying the process of grouping and sub-grouping necessary to isolate a

given message from the ensemble of messages, with the message itself. Therefore, if the representation of the message is not optimum, it also follows that the binary encoding technique employed to effect the representation is also not optimum in the sense that the total number of elementary operations required to encode the message is more than is strictly necessary. [The total number of elementary operations is here defined as the number of operations per individual decision (for instance, the number of sub-groups into which a group is divided, at each step of the encoding process) times the total number of decisions].

2.3.3 An Optimum Technique for Encoding Messages of Unequal Probability

The concepts developed in the present Section can be made use of to develop a technique for encoding ^{sources of} messages of unequal probability into a sequence of symbols ^{such that the} ~~having a~~ relative entropy of ^{the new source is} nearly unity, satisfying the following two conditions:

- 1) the number of elementary operations necessary to encode the message is a minimum and
- 2) the number of simple elements used to represent the message is a minimum.

The coding scheme described below satisfies all these requirements. It is a simple modification of the Shannon-Fano Code.

The messages are arranged in order of decreasing probability. This ordered series is then divided into three groups such that the total probability of each group is nearly the same. If the message is in the first group, the first ternary element of the sequence to be used to represent the message, is 0 ; if the message is in the second group, it is 1 ; if the message is in the third group, it is 2 . Each group is then sub-divided into three sub-groups of equal total probability, as before. Depending upon whether the message is in the first,

second or third sub-group, the second ternary element of the sequence is either 0, 1 or 2 respectively. The process of sub-division and allotment of succeeding ternary digits is repeated until each sub-group contains only one message. Now, every possible message is represented by a unique sequence of ternary elements, the high-probability messages being specified by short sequences and low-probability messages by long sequences.

3. Transmission Efficiency of Pulse Code Modulation Systems

We shall next consider whether a ternary system, which has been shown above to be ^{optimum} for representation of data, is also optimum for transmission of information through noisy channels.

A ternary system requires fewer "simple elements" to transmit a given amount of information; therefore, hypothetically, for a given average transmitter power, each element of a ternary digit can be assigned more power than the corresponding element of any other system including binary pulse code modulation, so that the probability of error in the detection of a ternary element due to interference by noise, is less than for any other system. But, on the other hand, the equivocation (E) in a ternary or higher order system is more complex than in a binary system. It now consists of two parts: the first part (E_1) is a measure of the uncertainty involved in estimating the position of the errors in the received sequence of digits, while the second (E_2) provides a measure of the uncertainty in estimating which of the various possible states of the wrongly received symbol was actually transmitted. The latter quantity (E_2) is zero for binary systems, but not for higher order systems including the ternary, as shown below:

Let the received sequence consist of N symbols. Let the probability of a given symbol being correctly received be q and let

$$p = 1 - q \quad \dots 3.1$$

Now, the probability of obtaining a sequence containing p errors and q correct symbols is

$$P_S = p^{Np} q^{Nq} \quad \dots 3.2$$

∴ Average entropy per symbol, (E_1), required to locate the errors:

$$\begin{aligned} E_1 &= \frac{1}{N} \ln P_S \text{ nepits}^{**} \\ &= - [p \ln p + q \ln q] \quad \dots 3.3 \end{aligned}$$

Once the position of the error is known, only $(b-1)$ alternatives have to be considered since the state of the symbol as indicated when received, is known to be wrong.

∴ Entropy per symbol (E_2) required to correct the errors:

$$\begin{aligned} E_2 &= p \ln (b-1) \text{ nepits} \quad \dots 3.4^* \\ &= 0, \text{ for binary PCM} \\ &= p \ln 2, \text{ for ternary PCM} \\ &= p \ln 3, \text{ for quaternary PCM} \end{aligned}$$

∴ Total equivocation E

$$\begin{aligned} E &= E_1 + E_2 \\ &= -[p \ln p + q \ln q] + p \ln(b-1) \dots 3.5 \end{aligned}$$

The extent to which the improvement in signal transmission made possible by the availability of more signal power per simple element in a ternary system, is offset by the increase in the total equivocation per symbol can be determined only by a more quantitative analysis of the problem.

Two possible approaches to the problem are outlined in the present Section and in Section 4 respectively.

* The expression 3.4 for E_2 is correct only if each of the $(b-1)$ possible alternatives is equally probable. This condition is not always satisfied. A more suitable expression for E_2 and for the total equivocation (E) is derived later.

** Defined in Ref: 24.

3.1 Transmission Efficiency of a Communication System

A quantitative measure of the performance of a communication system is provided by the expression for the transmission efficiency of the system, It may be defined as follows:

"The transmission efficiency (η) of a system of communication is the ratio of the rate of transmission of information (R) of the given system to the rate of transmission of information (C_I) of an ideal system as defined by the modified Hartley Law,^(4,28) the bandwidth (W), the signal power (P) and the noise power (N) being the same in both systems." (29,30,31)

$$\eta = R/C_I \quad \dots 3.6A$$

where

$$R = H(X) - H_Y(X) \quad \dots 3.6B$$

$$C_I = W \ln \frac{P+N}{N} \quad \dots 3.6C$$

$$\begin{aligned} H(X) &= \text{Entropy of the transmitted signal } \{x\} \\ &= -\sum_x p(x) \ln p(x) \quad \dots 3.6D \end{aligned}$$

$$\begin{aligned} H_Y(X) &= \text{Equivocation or conditioned entropy} \\ &\quad \text{of } \{x\} \text{ when the received signal } \{y\} \text{ is} \\ &\quad \text{known.} \\ &= - \sum_y \sum_x p(y) p_y(x) \ln p_y(x) \quad \dots 3.6E \end{aligned}$$

It may be noted that R represents the maximum rate of transmission of information that is possible if optimum

coding techniques are employed, rather than the actual rate of transmission achieved in simple practical systems.

The performance of different pulse code modulation systems are compared below on the basis of the transmission efficiency of the system, with special reference to binary and ternary systems. The discussion is confined to the study of balanced systems, that is, systems in which the extreme excursions of the transmitted symbols are symmetrically situated about zero. It has been shown that a balanced pulse code modulation system of any order b , requires less signal power than the corresponding unbalanced system having the same performance.⁽¹⁴⁾

In Sections 3.2 and 3.3 it is assumed that the interfering channel noise has a gaussian distribution of amplitude and a flat spectrum, while in Section 3.4 the noise is assumed to possess a rectangular amplitude distribution and a flat spectrum.

3.2 Transmission Efficiency of Binary PCM

In balanced binary pulse code modulation, the amplitude of the coded signal can assume either of two values: $(+V)$ or $(-V)$ units. It is convenient to express the signal amplitude V in terms of the root mean square value σ of the interfering noise, as follows:

$$V = \alpha \sigma \quad \dots 3.7$$

where

α is a dimensionless multiplicative constant

$\sigma = \sqrt{N}$, and

N is the noise power.

Assuming that the probability of occurrence $p(x)$ of the two states at the transmitter are equal, the entropy per

second $H(x)$ of the source is

$$H(x) = 2 \times W \sum_{x=1}^2 p(x) \log_2 p(x)$$

$$= [2W \log_2 2] \text{ népits} \quad \dots 3.8$$

To determine the equivocation $H_Y(X)$ it is necessary to determine the output probability distribution $p(y)$ and the conditional probability distribution $p_y(x)$ at the receiver. $p(y)$ is given by the relation

$$p(y) = \sum_x p(x) p_{x(y)} \quad \dots 3.9A$$

or, in matrix notation,

$$\{p(y)\} = \left\| \overset{\sim}{p_x(y)} \right\| \{p(x)\} \quad \dots 3.9B$$

that is

$$\begin{pmatrix} p'(1) \\ p'(2) \end{pmatrix} = \left\| \overset{\sim}{\begin{pmatrix} p_1(1) & p_1(2) \\ p_2(1) & p_2(2) \end{pmatrix}} \right\| \begin{pmatrix} p(1) \\ p(2) \end{pmatrix}$$

where

$$\left\| \overset{\sim}{p_x(y)} \right\| \text{ is the transpose of } \|p_x(y)\| \quad \dots \quad (?)$$

Now, white ^{thermal} noise has a symmetrical probability density ^{function} $f(x)$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{x^2}{2\sigma^2}} \quad \dots 3.10$$

Therefore,

$$\left\| \overset{\sim}{p_x(y)} \right\| = \|p_x(y)\| \quad \dots 3.11$$

since

$$\begin{aligned} p_1(1) &= p_2(2) = r \\ p_1(2) &= p_2(1) = s \end{aligned}$$

Also, since $p(x)$ is constant,

$$\begin{aligned} \begin{Bmatrix} p'(1) \\ p'(2) \end{Bmatrix} &= \begin{Bmatrix} r & s \\ s & r \end{Bmatrix} \begin{Bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{Bmatrix} \\ &= \begin{Bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{Bmatrix} \end{aligned} \quad \dots 3.12$$

since

$$r + s = 1$$

Now,

$$\begin{aligned} p_y(x) &\equiv \frac{p(x, y)}{p(y)} \\ &\equiv \frac{p(x) p_x(y)}{p(y)} \end{aligned}$$

But $p(x) = p(y)$, by equation 3.12

$$\therefore p_y(x) = p_x(y)$$

Consequently,

$$\begin{aligned} H_Y(X) &= \sum_y p(y) \sum_x p_y(x) \ln p_y(x) \\ &= \sum_x p(x) \sum_y p_x(y) \ln p_x(y) \\ &= H_X(Y) \end{aligned} \quad \dots 3.13$$

In a balanced binary system, the detection level is at

zero. (Fig. 3.1). Therefore, if the received pulse is positive it is detected as being in state (1), and if negative, as being in state (2). Hence,

$$\begin{aligned}
 p_1(1) &= p_2(2) \\
 &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\alpha\sigma}^{+\infty} e^{-\frac{x^2}{2\sigma^2}} dx \\
 &= \frac{1 + \Psi(\alpha)}{2} \quad \dots 3.14A
 \end{aligned}$$

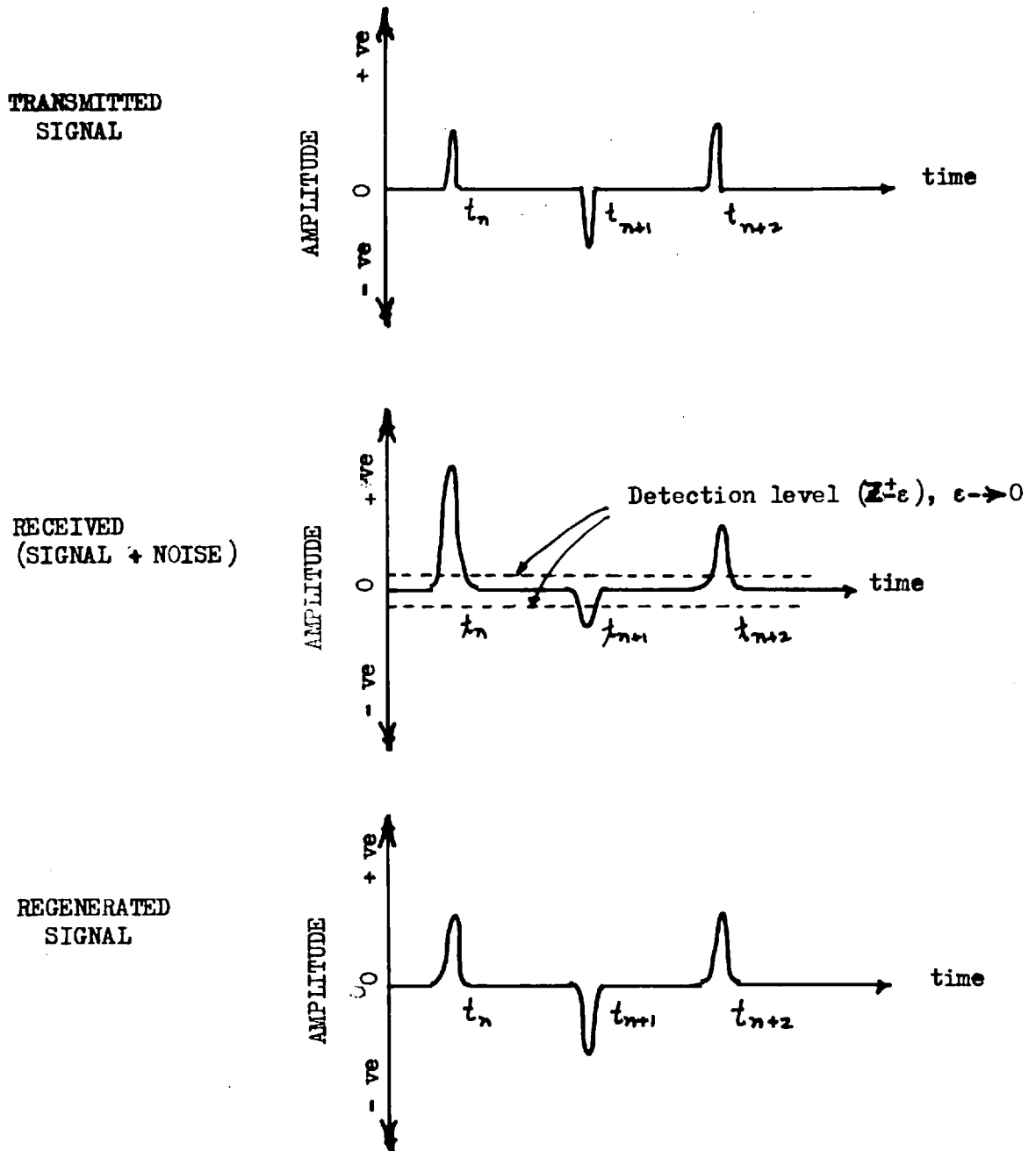
$$\begin{aligned}
 p_1(2) &= p_2(1) \\
 &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{-\alpha\sigma} e^{-\frac{x^2}{2\sigma^2}} dx \\
 &= \frac{1 - \Psi(\alpha)}{2} \quad \dots 3.14B
 \end{aligned}$$

where

$$\Psi(\alpha) = \int_{-\alpha}^{+\alpha} \psi(x) dx \quad \dots 3.15A$$

$$\psi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \quad \dots 3.15B$$

$$\begin{aligned}
 \therefore H_Y(X) &= -2W \left[\frac{1 + \Psi(\alpha)}{2} \log_2 \frac{1 + \Psi(\alpha)}{2} + \frac{1 - \Psi(\alpha)}{2} \log_2 \frac{1 - \Psi(\alpha)}{2} \right] \\
 &\quad \text{neps per second} \quad \dots 3.16
 \end{aligned}$$



DETECTION LEVEL, $Z = 0$.

FIG. 3.1. BALANCED BINARY PULSE COMMUNICATION.

∴ The rate of transmission R

$$= 2W \left[\ln 2 - \frac{1+\psi(\alpha)}{2} \ln \frac{1+\psi(\alpha)}{2} - \frac{1-\psi(\alpha)}{2} \ln \frac{1-\psi(\alpha)}{2} \right]$$

nepits per second ...3.17*

The signal power P of the binary system is

$$\begin{aligned} P &= \sum_{x=1}^2 x^2 \cdot p(x) \\ &= (\alpha \sigma)^2 \\ &= \alpha^2 N \end{aligned}$$

...3.18

The channel capacity of an ideal system using the same signal power as the binary system is

$$\begin{aligned} C_I &= W \ln (1 + P/N) \\ &= W \ln (1 + \alpha^2) \text{ nepits per} \\ &\quad \text{second} \end{aligned}$$

...3.19

The transmission efficiency $\eta = R/C_I$ is plotted as a function of the signal to noise ratio $(\frac{S}{N} = \alpha^2)$ in Fig. 3.2.

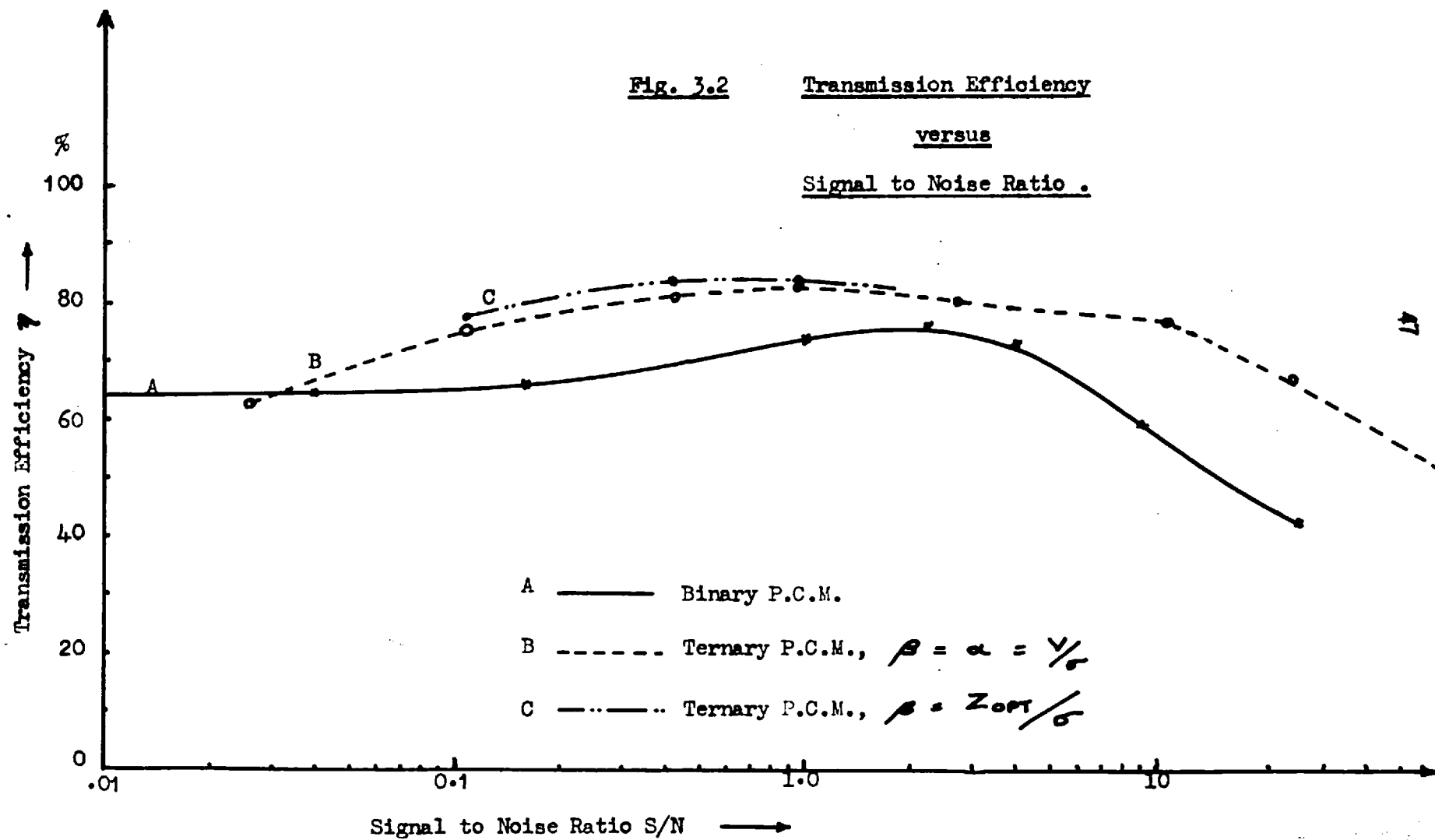
*

Equation 3.17 has been often mentioned in the literature. For instance, in references (29) and (3), it is partially derived without giving the full proof. It is, however, instructive to derive equation 3.17 in detail: this shows how the degenerate character of a (2×2) symmetric matrix makes binary pulse code modulation unique in certain respects. For example, it is the only pulse code modulation system for which the following relations hold:

$$H(X) = H(Y) \quad (\text{follows from eqn. 3.12})$$

and

$$H_Y(X) = H_X(Y) \quad \text{eqn. 3.13}$$



3.2.1 Capacity of Binary Systems

In the determination of the transmission efficiency, the expression for R was derived on the assumption that the input probability distribution $p(x)$ is the same for the two states of the binary symbol. It is easily shown that this, namely, the equiprobable, distribution is also the optimum input probability distribution for maximum rate of transmission of information (Appendix 3.1). In other words, the expression 3.17 for R also gives the capacity for a binary system.

It may be noted that this result is a direct consequence of the fact that a binary system which is symmetric, also possesses cyclic symmetry. In other words, if a binary channel is disturbed by noise having a symmetric probability distribution, the conditional probability matrix $[p_{x|y}]$ possesses both symmetry as well as cyclic symmetry. Under these conditions, the probability distribution $p(x)$ at the input must be an equiprobable one for maximum R . This point is discussed further in Section 3.3.

3.3 Transmission Efficiency of Ternary PCM

In balanced ternary systems, the transmitted signal can assume any of three values: $(+2V)$, 0 or $(-2V)$ units. Assuming that all three states are equally probable,

$$p(1) = p(2) = p(3) = \frac{1}{3}$$

the entropy $H(x)$ is

$$H(x) = 2W \cdot \log_2 3 \text{ nepits per second} \quad \dots 3.20$$

Now

$$R = H(Y) - H_X(Y) \quad \dots 3.21^*$$

*

It is easier to compute $H(Y) - H_X(Y)$ rather than $H(X) - H_Y(X)$, since, to determine $H_Y(X)$, it is necessary to evaluate not only $p(y)$ but also $p_y(x)$ from $p(x)$ and $p_x(y)$.

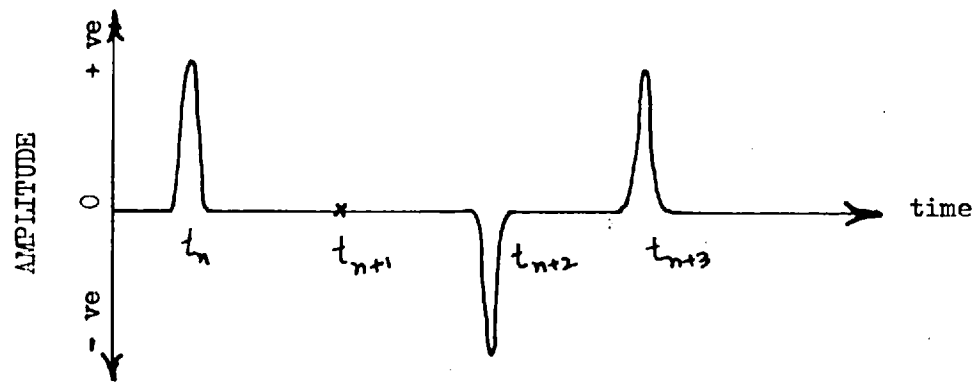
Here, $H(Y)$ and $H_X(Y)$ are functions of the conditional probabilities $p_X(y)$ which are, in turn, determined by the characteristics of the noise as well as the means of detection employed at the receiver. The receiver is assumed to be a simple detector which decides that the transmitted signal is

- (1) in state (1) if the received pulse is greater than some pre-assigned detection level, $Z_1 = +Z$
- (2) in state (2) if the received pulse lies between $(+Z)$ and $(-Z)$ and
- (3) in state (3) if the pulse is more negative than $Z_2 = -Z$. (Fig. 3.3.)

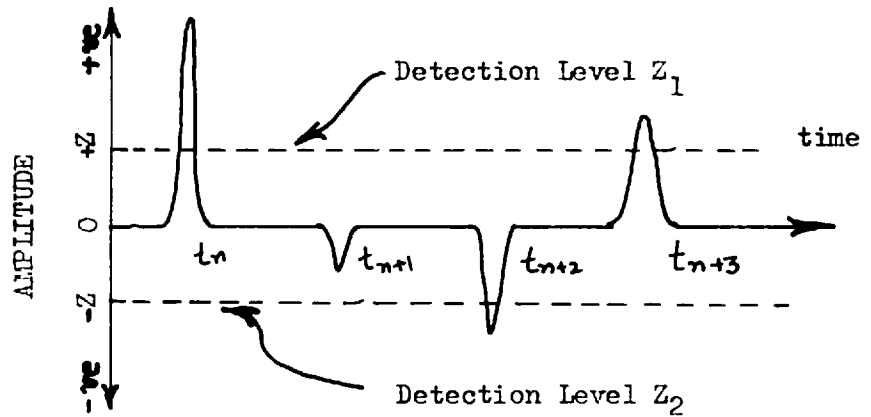
It may be noted that the absolute magnitudes of the two detection levels, positive and negative, are the same, being both equal to Z . This must be so for the optimum detector since the system is symmetrical, the positive pulses of state (1) and the negative pulses of state (3) being of equal magnitude and of equal probability of occurrence. But, in contrast to the case of binary systems, where the detection level is taken as the middle of the extreme excursions of the positive and negative pulses, (namely, the zero voltage level), the optimum slicing levels Z for ternary systems are not, in general, situated half way between the excursions for successive states, namely, $\pm V$ units with respect to the zero level. This is because, when $Z=V$,

- 1) the conditional probability $p_2(2)$ for the intermediate state is less than the corresponding conditional probabilities $p_1(1)$ and $p_3(3)$ for the extreme states (1) and (3), the difference becoming very pronounced for small values of signal to noise ratio (S/N)
- 2) the conditional probabilities $p_1(3)$ and $p_3(1)$ are larger than the corresponding conditional probabilities $p_1(2)$ and $p_3(2)$ for very low values of (S/N), and

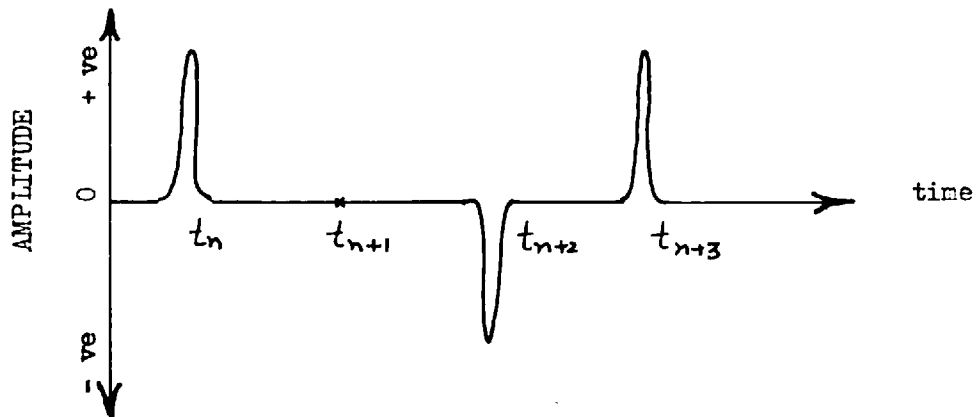
TRANSMITTED
SIGNAL



RECEIVED
(SIGNAL + NOISE)



REGENERATED
SIGNAL



DETECTION LEVELS

$$Z_1 = +Z$$

$$Z_2 = -Z$$

FIG. 3.3. BALANCED TERNARY PULSE COMMUNICATION

3) for small values of S/N , the conditional error probabilities $p_2(1)$ and $p_2(3)$ for state (2) are greater than the no-error probability $p_2(2)$. Due to the above reasons, the output probability $p(y=2) \approx p'(2)$ for state (2) tends to be less than those for the other two states. Consequently, $H(Y)$ tends to be less than what it would be if $p'(1)$, $p'(2)$ and $p'(3)$ were equal, the difference being very marked when (S/N) is small.

For certain values of (S/N) the effect of (1), (2) and (3) in lowering the magnitude of $H(Y)$ and hence of R , can be reduced by making Z greater than V . For certain other values of (S/N) , however, making Z greater than V causes $H_X(Y)$ to increase faster than $H(Y)$; under these conditions in general, R is maximum when Z is less than V .

It is instructive to find out the conditions under which Z is greater or less than V . This problem can be investigated by considering the function

$$R_Z = R_V + \Delta R \quad \dots 3.22A$$

where

R_Z = Rate of transmission when Z is adjusted at the optimum value

R_V = Rate of transmission when $Z=V$.

$$= \left[H(Y) - H_X(Y) \right]_{Z=V} \quad \dots 3.22B$$

ΔR = Increment in Rate of transmission

$$\doteq \Delta V \left[\frac{d}{dZ} H(Y) \right]_{Z=V} - \Delta V \left[\frac{d}{dZ} H_X(Y) \right]_{Z=V} \quad \dots 3.22C$$

$$\Delta V = Z - V \quad \dots 3.22D$$

Since ΔR has to be positive for maximum R , it is easily seen that ΔV is positive if

$$\left[\frac{d}{dz} H(Y) \right]_{z=V} - \left[\frac{d}{dz} H_X(Y) \right]_{z=V} > 0 \quad \dots 3.23A$$

ΔV is negative if

$$\left[\frac{d}{dz} H(Y) \right]_{z=V} - \left[\frac{d}{dz} H_X(Y) \right]_{z=V} < 0 \quad \dots 3.23B$$

If the detection levels are $+Z$ and $-Z$, the conditional probabilities $p_X(y)$ are

$$\begin{aligned} p_1(1) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-(2V-Z)}^{\infty} e^{-\frac{x^2}{2\sigma^2}} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-(2\alpha-\beta)}^{\infty} e^{-\frac{x^2}{2}} dx, \text{ putting } Z=\beta \\ &= \frac{1 + \bar{\psi}(2\alpha-\beta)}{2} \quad \dots 3.25A \end{aligned}$$

$$\begin{aligned} p_1(2) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-(2V+Z)}^{-(2V-Z)} e^{-\frac{x^2}{2\sigma^2}} dx \\ &= \frac{\bar{\psi}(2\alpha+\beta) - \bar{\psi}(2\alpha-\beta)}{2} \quad \dots 3.25B \end{aligned}$$

$$\begin{aligned}
 p_1(3) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{-(2V+Z)} e^{-\frac{x^2}{2\sigma^2}} dx \\
 &= \frac{1 - \Psi(2\alpha + \beta)}{2} \quad \dots 3.25C
 \end{aligned}$$

$$\begin{aligned}
 p_2(1) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{+Z}^{\infty} e^{-\frac{x^2}{2\sigma^2}} dx \\
 &= \frac{1 - \Psi(\beta)}{2} \quad \dots 3.25D
 \end{aligned}$$

$$\begin{aligned}
 p_2(2) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-Z}^{+Z} e^{-\frac{x^2}{2\sigma^2}} dx \\
 &= \Psi(\beta) \quad \dots 3.25E
 \end{aligned}$$

$$p_2(3) = p_2(1) \text{ , by symmetry} \quad \dots 3.25F$$

$$p_3(1) = p_1(3) \text{ , " " } \quad \dots 3.25G$$

$$p_3(2) = p_1(2) \text{ , " " } \quad \dots 3.25H$$

$$p_3(3) = p_1(1) \text{ , " " } \quad \dots 3.25I$$

The output probability distribution $p(y)$ is given by

$$\begin{aligned}
 \begin{Bmatrix} p'(1) \\ p'(2) \\ p'(3) \end{Bmatrix} &= \begin{bmatrix} \frac{1+\Psi(2\alpha-\beta)}{2} & \frac{\Psi(2\alpha+\beta)-\Psi(2\alpha-\beta)}{2} & \frac{1-\Psi(2\alpha+\beta)}{2} \\ \frac{1-\Psi(\beta)}{2} & \Psi(\beta) & \frac{1-\Psi(\beta)}{2} \\ \frac{1-\Psi(2\alpha+\beta)}{2} & \frac{\Psi(2\alpha+\beta)-\Psi(2\alpha-\beta)}{2} & \frac{1+\Psi(2\alpha-\beta)}{2} \end{bmatrix} \begin{Bmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{Bmatrix} \\
 &= \begin{Bmatrix} \frac{1}{2} - \frac{\Psi(2\alpha+\beta)+\Psi(\beta)-\Psi(2\alpha-\beta)}{6} \\ \frac{\Psi(2\alpha+\beta)+\Psi(\beta)-\Psi(2\alpha-\beta)}{3} \\ \frac{1}{2} - \frac{\Psi(2\alpha+\beta)+\Psi(\beta)-\Psi(2\alpha-\beta)}{6} \end{Bmatrix} \dots 3.26
 \end{aligned}$$

Substituting the above values for $p(y)$ and $p_x(y)$ in equation 3.21, we get, after simplification,

$$\begin{aligned}
 R=2W \left\{ -2 \left[\frac{1}{2} - \frac{\Psi(2\alpha+\beta)+\Psi(\beta)-\Psi(2\alpha-\beta)}{6} \right] \ln \left[\frac{1}{2} - \frac{\Psi(2\alpha+\beta)+\Psi(\beta)-\Psi(2\alpha-\beta)}{6} \right] \right. \\
 - \left[\frac{\Psi(2\alpha+\beta)+\Psi(\beta)-\Psi(2\alpha-\beta)}{3} \right] \ln \left[\frac{\Psi(2\alpha+\beta)+\Psi(\beta)-\Psi(2\alpha-\beta)}{3} \right] \\
 + \frac{2}{3} \left[\frac{1+\Psi(2\alpha-\beta)}{2} \right] \ln \left[\frac{1+\Psi(2\alpha-\beta)}{2} \right] \\
 + \frac{2}{3} \left[\frac{\Psi(2\alpha+\beta)-\Psi(2\alpha-\beta)}{2} \right] \ln \left[\frac{\Psi(2\alpha+\beta)-\Psi(2\alpha-\beta)}{2} \right] \\
 + \frac{2}{3} \left[\frac{1-\Psi(2\alpha+\beta)}{2} \right] \ln \left[\frac{1-\Psi(2\alpha+\beta)}{2} \right] \\
 + \frac{2}{3} \left[\frac{1-\Psi(\beta)}{2} \right] \ln \left[\frac{1-\Psi(\beta)}{2} \right] \\
 \left. + \frac{1}{3} [\Psi(\beta)] \ln [\Psi(\beta)] \right\} \text{ nepits per sec. } \dots 3.27
 \end{aligned}$$

R is maximum when Z is adjusted at the optimum value; this value for Z is determined by the relation

$$\frac{dR}{dZ} = 0 \quad \dots 3.28A$$

$$\begin{aligned} \text{i.e.} \quad & - \sum_y \left[\frac{1}{Z} + \ln p(y) \right] \frac{d}{dZ} p(y) \\ & + \frac{1}{3} \sum_x \sum_y \left[\frac{1}{Z} + \ln p_x(y) \right] \frac{d}{dZ} p_x(y) = 0 \quad \dots 3.28 \end{aligned}$$

Equation 3.28 is an integro-differential equation involving the determination of the limits of integration of the probability integral Ψ , as a function of Z. An approximate method of solution of the equation is presented in Appendix 3.2.. The results obtained are plotted in Fig. 3.4..

For a given value of signal-to-noise ratio S/N , equation 3.28 is satisfied by more than one value of $Z = Z_{\wedge}$, thereby indicating the existence of more than one local maxima $\{R_{\wedge}\}$. The true optimum value of $Z = Z_{\text{opt}}$ is obtained by evaluating R for the various values of $Z_{\wedge}$ and choosing that value of $Z_{\wedge}$ for which R attains a unique maximum numerical value, R_{max} .

Applying the above procedure to the determination of Z_{opt} and R_{max} , for $V = 0.4 \sigma$, it is seen that

$$\left. \begin{aligned} Z_{\text{opt}} &= .56 \sigma \text{ (approximately) ,} \\ R_{\text{max}} &= .2958 \text{ nepits per second} \end{aligned} \right\} \quad \dots 3.29A$$

Comparing the above results with the value of

$$\left. \begin{aligned} R &= .2872 \text{ nepits per second} \\ \text{corresponding to} \\ Z &= V = .4 \sigma \end{aligned} \right\} \quad \dots 3.29B$$

TRANSMISSION EFFICIENCY

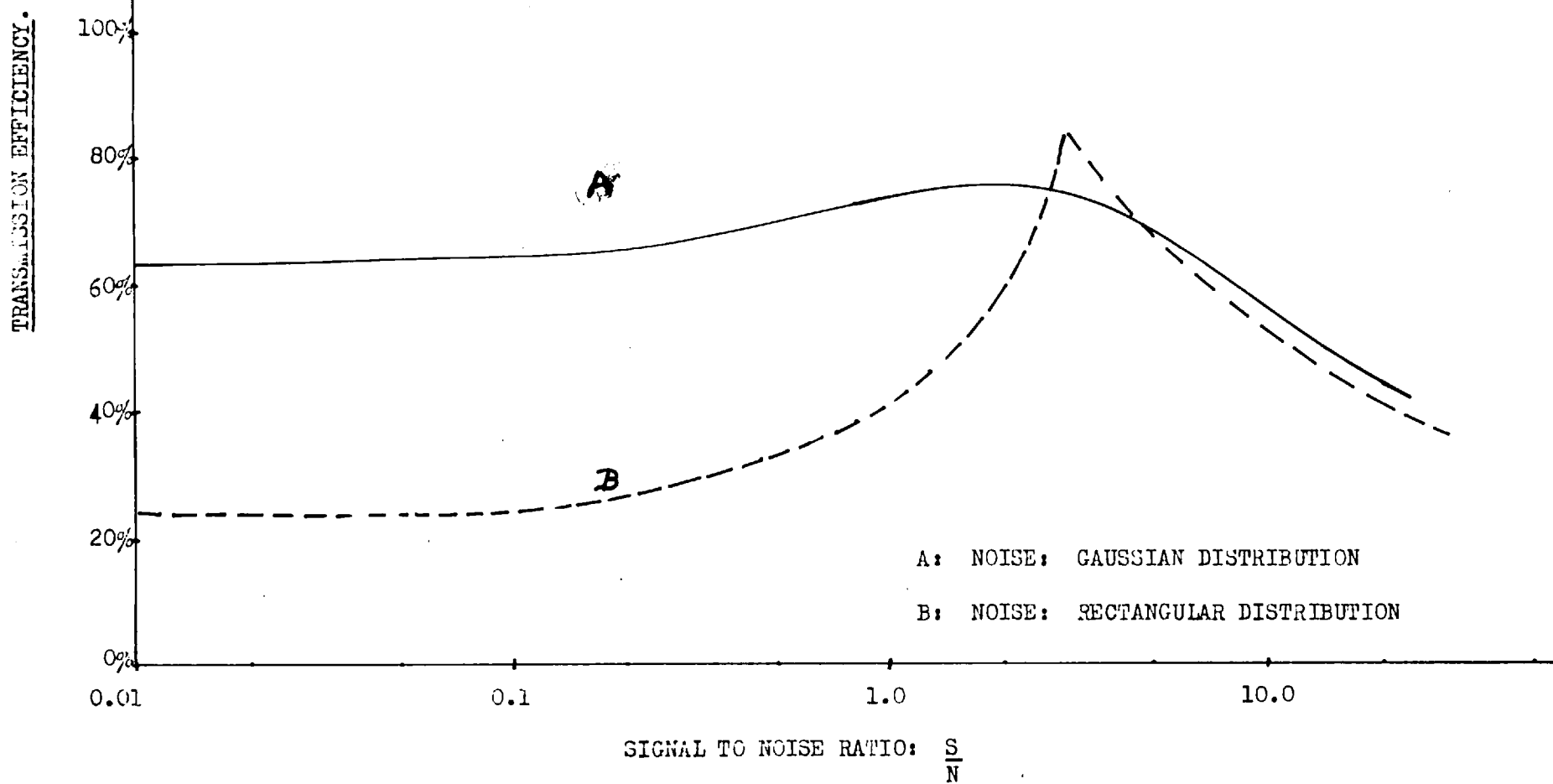
Vs

SIGNAL TO NOISE RATIO

FOR

BINARY PULSE CODE MODULATION

FIG: 3.4.



it is seen that proper choice of the detection level Z increases the rate of transmission R , in this instance, by approximately 3%.

The average signal power required to transmit a ternary symbol is

$$\begin{aligned}
 P &= \sum_{i=1}^3 x_i^2 p(x_i) \\
 &= \frac{1}{3} [4V^2 + 0 + 4V^2] \\
 &= (2.667\alpha^2) \sigma^2 \quad \dots 3.30
 \end{aligned}$$

The channel capacity of an ideal system using the same signal power is

$$\begin{aligned}
 C_I &= W \ln \left[1 + \frac{(2.667\alpha^2) \sigma^2}{\sigma^2} \right] \\
 &= W \ln [1 + 2.667\alpha^2] \text{ nepits} \\
 &\quad \text{per second} \quad \dots 3.31
 \end{aligned}$$

The transmission efficiency $\eta = R/C_I$ is plotted against signal to noise ratio S/N in Fig. 3.2..

If the detection levels are chosen as $\pm|Z| = \pm|V|$, the expressions for $p_x(y)$, $p(y)$ and R become considerably simpler. They are now given by

$$\left\| p_x(y) \right\| = \left\| \begin{array}{ccc} \frac{1+\Psi(\alpha)}{2} & \frac{\Psi(3\alpha)-\Psi(\alpha)}{2} & \frac{1-\Psi(3\alpha)}{2} \\ \frac{1-\Psi(\alpha)}{2} & \Psi(\alpha) & \frac{1-\Psi(\alpha)}{2} \\ \frac{1-\Psi(3\alpha)}{2} & \frac{\Psi(3\alpha)-\Psi(\alpha)}{2} & \frac{1+\Psi(\alpha)}{2} \end{array} \right\| \dots 3.32$$

$$\left\{ \begin{array}{l} p'(1) \\ p'(2) \\ p'(3) \end{array} \right\} = \left\{ \begin{array}{l} \frac{3 - \bar{\Psi}(3\alpha)}{6} \\ \frac{\bar{\Psi}(3\alpha)}{3} \\ \frac{3 - \Psi(3\alpha)}{6} \end{array} \right\} \quad \dots 3.33$$

and

$$\begin{aligned} R_{Z=V} = 2W \left\{ -2 \left[\frac{3 - \Psi(3\alpha)}{6} \right] \ell_n \left[\frac{3 - \Psi(3\alpha)}{6} \right] \right. \\ \left. - \left[\frac{\Psi(3\alpha)}{3} \right] \ell_n \left[\frac{\Psi(3\alpha)}{3} \right] \right. \\ + \frac{2}{3} \left[\frac{1 + \Psi(\alpha)}{2} \right] \ell_n \left[\frac{1 + \Psi(\alpha)}{2} \right] + \frac{2}{3} \left[\frac{\Psi(3\alpha) - \Psi(\alpha)}{2} \right] \ell_n \left[\frac{\Psi(3\alpha) - \Psi(\alpha)}{2} \right] \\ + \frac{2}{3} \left[\frac{1 - \Psi(3\alpha)}{2} \right] \ell_n \left[\frac{1 - \Psi(3\alpha)}{2} \right] + \frac{2}{3} \left[\frac{1 - \Psi(\alpha)}{2} \right] \ell_n \left[\frac{1 - \Psi(\alpha)}{2} \right] \\ \left. + \frac{1}{3} \left[\Psi(\alpha) \right] \ell_n \left[\Psi(\alpha) \right] \right\} \text{ nepits per second} \quad \dots 3.34^* \end{aligned}$$

The magnitude of $\eta_{Z=V} = \frac{R_{Z=V}}{C_I}$ is plotted against S/N in Fig. 3.2..

3.3.1 Capacity of a Ternary System

In contrast to the case of binary systems, the expression for the rate of transmission R for a ternary system as given by equation 3.27 does not represent the capacity of a ternary system. This is because the equiprobable distribution assumed at the input for the various states of the ternary symbol, does not maximise the rate of transmission R . An equiprobable distribution is the optimum distribution for the input symbols only if the

* A different, but numerically equivalent expression has been derived in a recent publication (Ref. 29), in a slightly different manner.

conditional probability matrix $[p_x(y)]$ possesses cyclic symmetry, that is, only if

$$p_i(j) = p_{i+1}(j+1) = \dots = p_{i+b-1}(j+b-1) \quad \dots 3.35$$

where

b is the period of the group so that

$$p_i(j) \equiv p_{i+b}(j+b) \quad \dots 3.36$$

$$p_{i+b-1}(j+b-1) \equiv p_{i-1}(j-1) \quad \dots 3.350$$

Under these conditions, it is easily seen that the expression for $H_X(Y)$

$$H_X(Y) = \sum_x p(x) \sum_y p_x(y) \ln p_x(y)$$

is completely determined by the characteristics of the noise as given by the conditional probabilities $p_x(y)$ and is not dependent on the probability distribution $p(x)$ of the input signal. Therefore, the channel capacity C is given by

$$\begin{aligned} C &= \text{Max } [H(Y) - H_X(Y)] \\ &= \text{MAX } [H(Y)] - H_X(Y) \end{aligned} \quad \dots 3.36$$

which is obtained when $p(y)$ is constant so that

$$p(y) = \left\{ \begin{array}{l} 1/b \\ 1/b \end{array} \right\} \quad \dots 3.37$$

But

$$\{p(y)\} = \| \tilde{p}_x(y) \| \{p(x)\}$$

Therefore

$$\{p(x)\} = \| \tilde{S}_{x,y} \| \{p(y)\}$$

$$\text{where } \| \tilde{S}_{x,y} \| = \| \tilde{p}_x(y) \|^{-1},$$

$$= \left\{ \begin{array}{l} 1/b \\ 1/b \end{array} \right\} \quad \text{if the inverse exists.} \quad \dots 3.38$$

because of equations 3.35 and 3.37, as can be easily seen by expanding $\|p_x(y)\|^{-1}$ using the Cayley-Hamilton theorem, and simplifying using equations 3.35 and 3.37.

But, in a ternary pulse code modulation channel, which is assumed to be balanced and is perturbed by white ^{thermal} noise which has a smooth, symmetric amplitude distribution curve, equation 3.35 can never be satisfied for finite values of S/N . For instance, for very low values of S/N , the conditional probability matrix is approximately given by

$$\|p_x(y)\| = \begin{vmatrix} \frac{1+k\epsilon}{2} & k\epsilon & \frac{1-3k\epsilon}{2} \\ \frac{1-k\epsilon}{2} & k\epsilon & \frac{1-k\epsilon}{2} \\ \frac{1-3k\epsilon}{2} & k\epsilon & \frac{1+k\epsilon}{2} \end{vmatrix} \quad \dots 3.39$$

where

$$k = \sqrt{\frac{2}{\pi}}$$

$$\epsilon = \frac{V}{\sigma}$$

When $\epsilon \ll 1$, it is easily shown that R can be increased by keeping the probability of transmission of the intermediate state (2) to a value much below that for the other two states. In the limiting case where $S/N \rightarrow 0$, so that $\epsilon \rightarrow 0$, the maximum rate of transmission is obtained by not transmitting state (2) at all.

Equation 3.35 is satisfied for the ternary system only when S/N tends to infinity. Under these conditions,

Limit

$$\frac{S}{N} \rightarrow \infty, p_i(j) \rightarrow \delta_{ij} \quad 3.40A$$

where

$$\delta_{ij} = \begin{cases} 0, & i \neq j \\ 1, & i = j \end{cases} \quad \dots 3.40B$$

so that

$$\| p_x(y) \| = \left\| \begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array} \right\| \quad \dots 3.40C$$

The problem of exact determination of the capacity of a ternary channel for finite values of S/N is complicated and difficult from the point of view of numerical evaluation, but very interesting from a theoretical standpoint. Two possible situations arise here: in the first case, the conditional probabilities $p_x(y)$ are known, while in the second, only certain relationships among the $p_x(y)$, and not the conditional probabilities themselves, are known. Both these situations are discussed below, but since the former has been studied in detail by Shannon it is only briefly mentioned here.

Case A : If $p_x(y)$ are known, C is obtained by maximising

$$R = H(Y) - H_X(Y) \quad \dots 3.41A$$

with respect to $p(x)$ subject to the constraint

$$\sum_{i=1}^3 p(x_i) = 1 \quad \dots 3.41B$$

This is equivalent to maximising

$$I = H(Y) - H_X(Y) + \lambda \sum_{i=1}^3 p(x_i) \quad \dots 3.41C$$

where λ is a Lagrangian Multiplier. This leads, by the method of Lagrange, to the system of equations

$$\frac{\partial}{\partial p(x_i)} \left[H(Y) - H_X(Y) + \lambda \sum_{i=1}^3 p(x_i) \right] = 0, \quad i = 1, 2, 3. \quad \dots 3.42$$

After a certain amount of manipulation, equations 3.42 yield the following transcendental equations for $p(x_i)$:

$$p(x_i) = \frac{\sum_{j=1}^3 s_{ij}}{\sum_{j=1}^3 s_{ij} e^{c \sum_{i=1}^3 s_{ij} p_i(j) \ln p_i(j)}} \quad i = 1, 2, 3 \quad \dots 3.43A^*$$

where

$$c = - \sum_{j=1}^3 p_i(j) \ln \left[\frac{p_i(j)}{\sum_{i=1}^3 p(i) p_i(j)} \right] \quad \dots 3.43B$$

$$\| s_{ij} \| = \| p_i(j) \|^{-1}, \quad \text{if the inverse exists} \quad \dots 3.43C$$

Case B : To appreciate the significance of the second case, it must be realized that channel capacity, as defined by Shannon, is a variational quantity that is maximised with respect to the input probability distribution $p(x)$, when $p_x(y)$ are known. But in any practical system, $p_x(y)$ are determined by the characteristics of the perturbing noise as well as the method of detection employed at the receiver. ~~However~~ as indicated in Section 3.3, the optimum detection levels (Z) at the receiver for maximum R , and therefore the conditional probabilities $p_x(y)$, are themselves functions of $p(x)$ - which is sought to be adjusted to maximise R -

* Equations 3.43 are directly adapted from Shannon's general result (Ref.4).

since R is a function of $p(x)$. It is thus seen that to maximise R to get C , it is not enough to optimise the probability distribution $p(x)$, since this would only optimise the system with respect to a receiving system which may (for example, a binary system) or may not (for instance, a ternary system) be optimum with respect to the newly adjusted values of $p(x)$. To obtain C , it is necessary to optimise the system with respect to not only $p(x)$ but also the variable or adjustable aspect of the receiver: for a ternary system, the receiver may be considered to be specified by a function, say, $f(Z)$ where Z represents the variable aspect, namely, the magnitude of the detection level and f stands for the invariant aspect of the receiver. Optimising the receiver is, however, equivalent to optimising the conditional probabilities $p_x(y)$ subject to certain constraints specified by the inter-relationships that exist among the $p_x(y)$ due to the invariance of the functional form f of the receiver $f(x)$ and of the channel noise $\mathbb{T}$.

It is therefore necessary to modify the functional equation 3.41C for the determination of the capacity of a ternary system. The function to be maximised now is

$$I = H(Y) - H_X(Y) + \lambda \sum_{x=1}^3 p(x) + \sum_{r=1}^3 \mu_r \sum_{y=1}^3 p_{x=r}(y) \quad \dots 3.44A$$

where

λ , μ_1 , μ_2 , μ_3 are Lagrangian multipliers and the equations of constraint* are

$$\sum_{x=1}^3 p(x) = 1 \quad \dots 3.44B$$

$$\sum_{y=1}^3 p_{x=r}(y) = 1, \quad r = 1, 2, 3 \quad \dots 3.44C$$

* Constraints due to the receiver are introduced later.

Differentiating I partially with respect to $p(x)$ and $p_x(y)$ and equating to zero, yields $m = 3+3^2 = 12$ equations:

$$\frac{\partial I}{\partial q_k} = Q_k = 0, \quad k = 1, 2, \dots, m \quad \dots 3.45$$

where

$$\begin{aligned} q_k &= p(x=k), \quad k \leq 3, \\ &= p_{x=1}(y=j), \quad 3 < k \leq 2 \times 3 \\ &\quad j = 1, 2, 3 \\ &= p_{x=2}(y=j), \quad 2 \times 3 < k \leq 3^2 \\ &\quad j = 1, 2, 3 \\ &= p_{x=3}(y=j), \quad 3^2 < k \leq 3^2 + 3 \\ &\quad j = 1, 2, 3 \end{aligned} \quad \dots 3.45B$$

These m equations 3.45 together with equations 3.44B,C comprise $(m+4)$ equations in the $(m+4)$ unknowns λ , μ_1, μ_2, μ_3 , $p(x)$ and $p_x(y)$. If $\lambda, \mu_1, \mu_2, \mu_3$ are eliminated $p(x)$ and $p_x(y)$ can be determined.

Since there are $m(=12)$ independent equations (3.45) and four equations of constraint (3.44B,C), this system of equations possess $(m-4)=8$ degrees of freedom. This does not correspond to the actual conditions of operation of the system, as there are additional constraints imposed upon the $p_x(y)$ by the characteristics of the receiver; in fact, the actual number of degrees of freedom among the $p_x(y)$, for a given channel, is the same as the number of detection levels (Z) to be determined at the receiver. Therefore, for a general ternary system, which may or may not be balanced, there are two degrees of freedom with respect to the $p(x)$ and an additional two degrees of freedom with respect to the $p_x(y)$. Therefore, to reduce the number of degrees of freedom of the

system of equations mentioned above from (8) to (4), it is necessary to introduce 4 additional equations of constraint. This can be done by formulating expressions connecting any two independent conditional probabilities, say $p_1(1)$ and $p_1(2)$, to the four other independent conditional probabilities $p_2(1)$, $p_2(2)$, $p_3(1)$ and $p_3(2)$. [It may be noted that the other conditional probabilities $p_1(3)$, $p_2(3)$ and $p_3(3)$ are not independent since they are related to the rest of the $p_x(y)$ through equations 3.44C.] In general, these equations can be formulated as follows:

$$\begin{aligned} \Phi_1(q_1, q_2, \dots, q_m) &= 0 \\ \vdots \\ \Phi_4(q_1, q_2, \dots, q_m) &= 0 \end{aligned} \quad \dots 3.46$$

Equation 3.44A has to be modified to take into account these additional constraints. This can be done as shown below:

The variational equations which follow from 3.46 are first formed:

$$\begin{aligned} \frac{\partial \Phi_1}{\partial q_1} \delta q_1 + \frac{\partial \Phi_1}{\partial q_2} \delta q_2 + \dots + \frac{\partial \Phi_1}{\partial q_m} \delta q_m &= 0 \\ \vdots \\ \frac{\partial \Phi_4}{\partial q_1} \delta q_1 + \frac{\partial \Phi_4}{\partial q_2} \delta q_2 + \dots + \frac{\partial \Phi_4}{\partial q_m} \delta q_m &= 0 \end{aligned} \quad \dots 3.47$$

where δq_k is the first variation of q_k . These equations are multiplied by the Lagrangian multipliers λ_1 , λ_2 , λ_3 and λ_4 respectively, integrated over the interval of variation and then added to equations 3.45.. m modified equations are now obtained:

$$\xi_1 = Q_1 + \lambda_1 \frac{\partial \Phi_1}{\partial q_1} + \lambda_2 \frac{\partial \Phi_2}{\partial q_1} + \dots + \lambda_4 \frac{\partial \Phi_4}{\partial q_1} = 0$$

$$\xi_k = Q_k + \lambda_1 \frac{\partial \Phi_1}{\partial q_k} + \lambda_2 \frac{\partial \Phi_2}{\partial q_k} + \dots + \lambda_4 \frac{\partial \Phi_4}{\partial q_k} = 0$$

$$\xi_m = Q_m + \lambda_1 \frac{\partial \Phi_1}{\partial q_m} + \lambda_2 \frac{\partial \Phi_2}{\partial q_m} + \dots + \lambda_4 \frac{\partial \Phi_4}{\partial q_m} = 0 \quad \dots 3.48$$

These m equations together with equations 3.44B,C and 3.46 comprise $(m+8)$ equations in the $(m+8)$ unknowns λ , $\mu_1 \rightarrow \mu_4$, $\lambda_1 \rightarrow \lambda_4$, $p(x)$ and $p_x(y)$. If λ , $\mu_1 \rightarrow \mu_4$, $\lambda_1 \rightarrow \lambda_4$ are eliminated, $p(x)$ and $p_x(y)$ can be determined and hence the channel capacity C .

The procedure given above for maximising R by adjusting $p(x)$ and $p_x(y)$ such that the transmitter, the channel and the receiver are all simultaneously matched to each other, can be easily extended to higher order systems such as the quaternary and others.

Summing up, it may be mentioned that the modification outlined above of the original definition of channel capacity is essential if the concept of capacity is to provide a valid criterion for the comparison of the performance of communication systems for which the conditional probabilities are not invariant, but are defined in terms of the channel noise and the adjustable parameters of the receiver.

3.3.2 Comparison of Binary and Ternary Systems

It can be seen from Fig. 3.3. that at signal-to-noise ratios greater than about -30 db, ternary pulse code modulation is always more efficient than the binary system.

Study of curve C, which corresponds to the use of optimum detection levels at the receiver, indicates that this lower limit of -30 db can be decreased further by using the optimum receiver₇. The maximum efficiency of the ternary system is approximately 83.4% while that of the binary system is only about 75% to 76%. For high signal-to-noise ratios of the order of 15 db or greater, the difference in the efficiencies of the two systems varies from about 15% to about 24%. On the other hand, when the signal-to-noise ratio is very low, say, less than -30 db, the binary system is more efficient. This can be accounted for by the fact that, under these conditions, R is maximum when the probability of occurrence $p(2)$ of the intermediate state (2) is less than that for the other two states - not equal to, as assumed in the derivation of R (Section 3.3.1). The extent to which the efficiency of the ternary system can be made to approach (or probably exceed) that of the binary system by adjusting the initial probabilities $p(x)$, was not investigated due to the complexity of the numerical computations involved.

3.4 Evaluation of Efficiency when Noise is not Gaussian

It is instructive to study the performance of pulse code modulation systems when the perturbing noise is not gaussian. In the present Section, the transmission efficiency of a binary system perturbed by peak-limited noise having a rectangular amplitude distribution $\phi(x)$ and a flat spectrum, is evaluated. It is assumed that the average signal power (P), the average noise power (N) and the bandwidth (W) of the system are the same as for that dealt with in Section 3.2..

The probability density function $\phi(x)$ of the noise

is given by

$$\phi(x) \, dx = \begin{cases} \frac{1}{2a} & , \, |x| \leq a \\ 0 & , \, |x| > a \end{cases} \quad \dots 3.49$$

where $|a|$ is the maximum amplitude of the noise.

The average noise power N_0

$$\begin{aligned} N_0 &= \int_{-a}^{+a} x^2 \phi(x) \, dx \\ &= \frac{a^2}{3} \text{ units} \end{aligned} \quad \dots 3.50A$$

is equal to the gaussian noise power N if

$$a = \sqrt{3} \, \sigma \quad \dots 3.50B$$

The rate of transmission R is obtained by substituting $\Phi(\alpha)$ for $\bar{\Gamma}(\alpha)$ in equation 3.17, where

$$\begin{aligned} \bar{\Phi}(\alpha) &= \int_{-\alpha\sigma}^{+\alpha\sigma} \phi(x) \, dx \\ &= \begin{cases} \frac{V}{a} & , \, V \leq a \\ 1 & , \, V > a \end{cases} \\ &= \begin{cases} \frac{\alpha}{3} & , \, \alpha \leq \sqrt{3} \\ 1 & , \, \alpha > \sqrt{3} \end{cases} \end{aligned} \quad \dots 3.51$$

The channel capacity of an ideal system using the same signal power P

$$P = (\alpha\sigma)^2$$

as the binary system is

$$C_I = W \log \left(1 + \frac{P}{N'} \right) \quad \dots 3.52A$$

where

N' is the entropy power of the noise

$$= \frac{1}{2\pi e} \cdot e^{2H'} \quad \dots 3.52B$$

where

H' = entropy of the noise

Since

$$\begin{aligned} H' &= \int_{-\infty}^{+\infty} \phi(x) \log_e \phi(x) dx \\ &= \frac{1}{2} \ln (12\sigma^2) \end{aligned} \quad \dots 3.53$$

$$\therefore N' = \frac{6}{\pi e} \sigma^2$$

Therefore

$$C_I = W \ln \left(1 + \frac{\pi e}{6} \alpha^2 \right) \quad \dots 3.54$$

The transmission efficiency $\eta = \frac{R}{C_I}$ is plotted against $\frac{S}{N}$ in Fig. 3.4.

It can be seen from Fig. 3.4 that there exists a sharply defined critical region where the transmission efficiency of the system is very high - higher than that for the corresponding binary system perturbed by gaussian noise. The S/N ratio corresponding to the peak of the curve is the same as that for which the equivocation falls to zero. The curve falls away rapidly for values of S/N ratio different from this value, the corresponding transmission efficiencies being much less than in the case of gaussian noise.

It is significant that the magnitude of the maximum efficiency is the same as that for the ternary system perturbed by gaussian noise.

4. Equivocation in Binary and Ternary Pulse Code Modulation Systems

The rate of transmission R through a noisy channel is always less than the source entropy $H(x)$. For a given source, the ratio $\frac{R}{H(x)}$, which may be called normalized rate of transmission, R_{norm} , provides a quantitative measure of the ability of the system to fight noise.

When $\frac{R}{H(x)}$ is very nearly unity, it may be useful to consider the complement of $\frac{R}{H(x)}$ namely

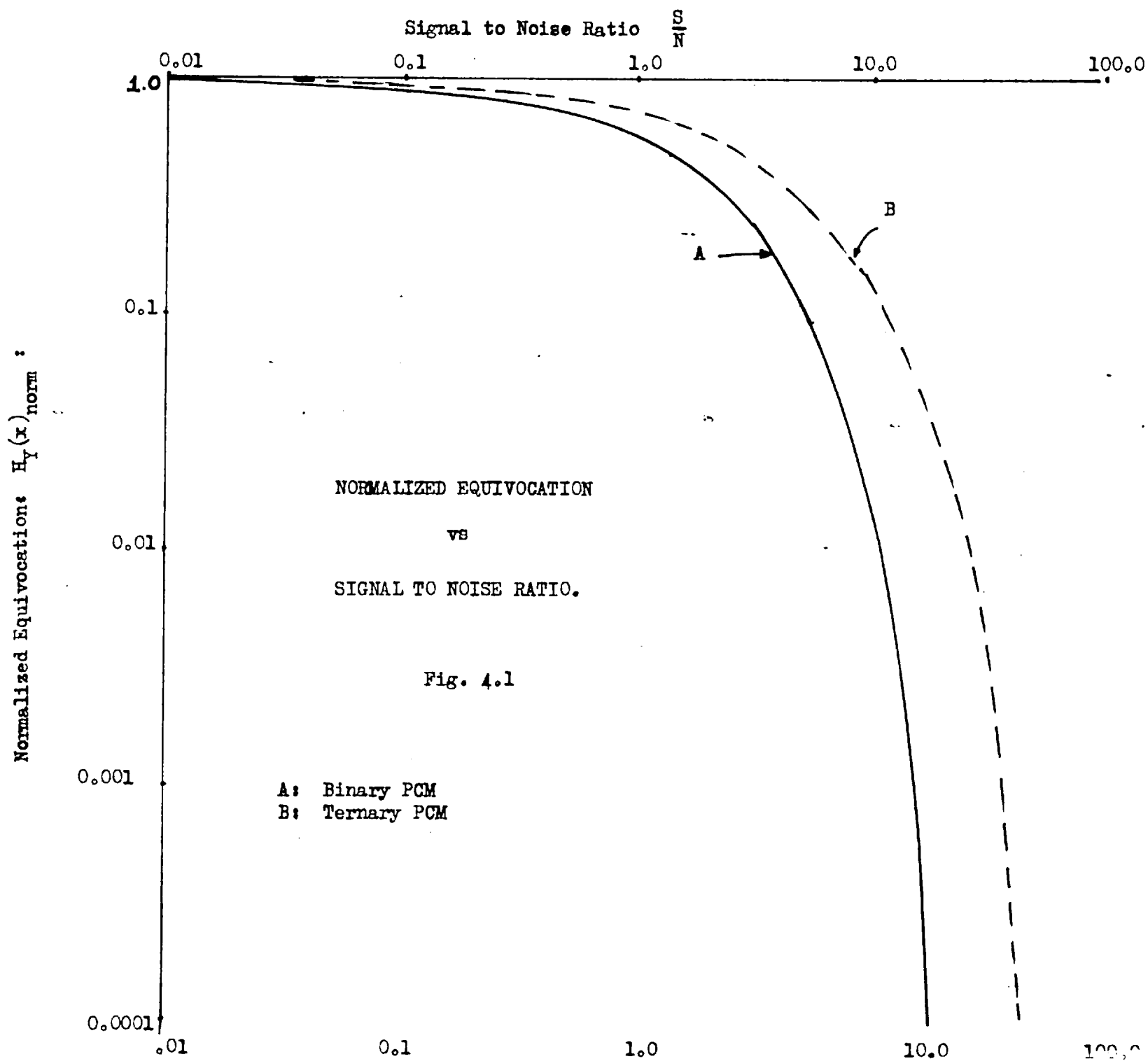
$$1 - \frac{R}{H(x)} = \frac{H_Y(X)}{H(X)} \quad \dots 4.1$$

$\frac{H_Y(X)}{H(X)}$ may be called the normalized equivocation, $H_Y(X)_{\text{norm}}$, of the system.

The normalized equivocation of binary as well as of ternary PCM is plotted against signal to noise ratio in Fig. 4.1.

It is evident from Fig. 4.1 that the normalized equivocation for binary systems is always less than that for ternary, for the same signal to noise ratio. The significance of this is not that binary PCM is better than ternary, but that pulse code modulation is such that it offers more immunity from noise to the transmitted symbols if these are coded to a binary rather than to a ternary base. The reason why a ternary system is more efficient than the binary despite this handicap is, of course, that a ternary symbol carries more information per state as shown in Section 2.

It may be noted that the concept of normalized equivocation is more limited in its application than the concept of transmission efficiency. While the latter



provides a criterion for the assessment of the relative performance of all communication systems, the former is most useful in assessing to what extent the system of modulation used (for instance, PCM) for communication is adapted to the type of symbols used (for example, binary, ternary or quaternary symbols) for transmission.

5. Signal to False-Pulse-Noise Ratio in a Ternary System

To achieve the rate of transmission R as evaluated in the previous Sections, not only very long - theoretically, infinitely long - processing delays will have to be tolerated both during coding, before transmission as well as during decoding, after reception, but it would also be necessary to use extremely complex instrumentation to perform the coding and decoding operations. It is obvious that a simple, practical system in which the coding delay is no more than is necessary to code a single sample of the primary signal, can never achieve this rate. While it is difficult to derive an exact expression for the rate of transmission of such a system, it is nevertheless possible to define and to evaluate an alternative measure of the performance of the system. In the present Section, the relevant measure is taken to be the signal to false-pulse-noise ratio of the system.

Techniques of symbolic (or Blissard) calculus⁽³²⁻³⁵⁾ are made use of in derivations given below, of the signal to noise ratio of the ternary channel.

5.1 Signal to Noise Ratio for a Code Group Consisting of one Digit

Let symbol $x(i)$ be transmitted and symbol $y(j)$ be received, so that the error ϵ is given by

$$\epsilon = x(i) - y(j) \quad \dots 5.1$$

Since $x(i)$ and $y(j)$ can each assume either of the three values $+h$, 0 , $-h$ units ϵ can assume any of the five values: 0 , $\pm h$, $\pm 2h$ units. Let the probabilities that ϵ is 0 , $|h|$ and $|2h|$ be

$$\Pr(\epsilon = 0) = \alpha$$

$$\begin{aligned}\Pr(\epsilon = |h|) &= \Pr(\epsilon = -h) + \Pr(\epsilon = +h) \\ &= \beta\end{aligned}$$

$$\begin{aligned}\Pr(\epsilon = |2h|) &= \Pr(\epsilon = -2h) + \Pr(\epsilon = +2h) \\ &= \gamma\end{aligned}$$

Let the events y_0, y_1, y_2 correspond to the situations $\epsilon = 0$, $\epsilon = |h|$, and $\epsilon = |2h|$ respectively.

In a statistical sense, the amplitude of the received signal may be considered to be undiminished and of magnitude unity, if $\epsilon = 0$. This is equivalent to the statement that the event y_0 makes a contribution of magnitude plus unity to the amplitude of the received signal. Similarly, the events y_1 and y_2 contribute (-1) unit and (-2) units respectively to the signal amplitude.

The probability generating function G_P of the amplitude of the received signal is now given by

$$\begin{aligned}G_P &= [\alpha t^{y_0} + \beta t^{y_1} + \gamma t^{y_2}] \\ &= [\alpha t^{+1} + \beta t^{-1} + \gamma t^{-2}] \quad \dots 5.3\end{aligned}$$

putting $y_0 = 1$, $y_1 = -1$ and $y_2 = -2$

The received signal power S_0 and noise power N_0 can be calculated from the first and second moments of the probability density function of the amplitude. But, it is easier to obtain these moments by making use of the moment generating function G_M which can be derived from G_P by substituting e^a for t .

$$G_M = [\alpha e^a + \beta e^{-a} + \gamma e^{-2a}] \quad \dots 5.4$$

Expanding G_M in powers of a ,

$$\begin{aligned}
G_M &= \alpha \left[1 + a + \frac{a^2}{2!} + \frac{a^3}{3!} + \dots \right] \\
&+ \beta \left[1 - a + \frac{a^2}{2!} - \frac{a^3}{3!} + \dots \right] \\
&+ \gamma \left[1 - 2a + \frac{4a^2}{2!} - \frac{8a^3}{3!} + \dots \right] \\
&= m_0 + m_1 \cdot \frac{a}{1!} + m_2 \cdot \frac{a^2}{2!} + \dots + m_n \cdot \frac{a^n}{n!} + \dots
\end{aligned}
\tag{5.5}$$

where $m_0, m_1, m_2, \dots, m_n, \dots$ are the moments of the distribution of order 0, 1, 2, ..., n, ..., respectively.

Equating coefficients,

$$m_0 = \alpha + \beta + \gamma = 1 \tag{5.6A}$$

$$\begin{aligned}
m_1 &= (\alpha - \beta - 2\gamma) \\
&= (1 - 2\beta - 3\gamma)
\end{aligned}
\tag{5.6B}$$

$$\begin{aligned}
m_2 &= (\alpha + \beta + 4\gamma) \\
&= (1 + 3\gamma)
\end{aligned}
\tag{5.6C}$$

m_1 gives the mean value of the amplitude of the received signal while m_2 gives the total received power. The average noise power N_0 is equal to the mean square value of the fluctuations of the received signal about the mean value m_1 and is therefore given by the second central moment μ_2 :

$$N_0 = \mu_2 \tag{5.7}$$

$$\begin{aligned}
\text{where } \mu_2 &= m_2 - m_1^2 \\
&= (1 + 3\gamma) - (1 - 2\beta - 3\gamma)^2 \\
&= 4\beta(1 - \beta) + 9\gamma(1 - \gamma) - 12\beta\gamma
\end{aligned}
\tag{5.7A}$$

Since the total received power is m_2 , the received signal power is

$$\begin{aligned}
S_0 &= m_2 - N_0 \\
&= m_1^2
\end{aligned}
\tag{5.8}$$

The signal to false-pulse-noise ratio S_0/N_0 of a pulse code modulation system using a single ternary digit per sample of signal is therefore

$$\begin{aligned}\frac{S_0}{N_0} &= \frac{m_1^2}{\mu_2} \\ &= \frac{(1-2\beta-3\gamma)^2}{4\beta(1-\beta)+9\gamma(1-\gamma)-12\beta\gamma} \quad \dots 5.9\end{aligned}$$

If the channel is perturbed by gaussian white noise, and the two detection levels are chosen midway between zero and the extreme positive and negative excursions respectively of the transmitted signal, as discussed in Section 3.3, then

$$\begin{aligned}\alpha &= \frac{1}{3} [p_1(1) + p_2(2) + p_3(3)] \\ &= \frac{1}{3} [1 + 2\Psi(V)] \\ \beta &= \frac{2}{3} [p_1(2) + p_2(3)] \\ &= \frac{1}{3} [1 + \Psi(3V) - 2\Psi(V)] \\ \gamma &= \frac{2}{3} [p_1(3)] \\ &= \frac{1}{3} [1 - \Psi(3V)] \quad \dots 5.10.\end{aligned}$$

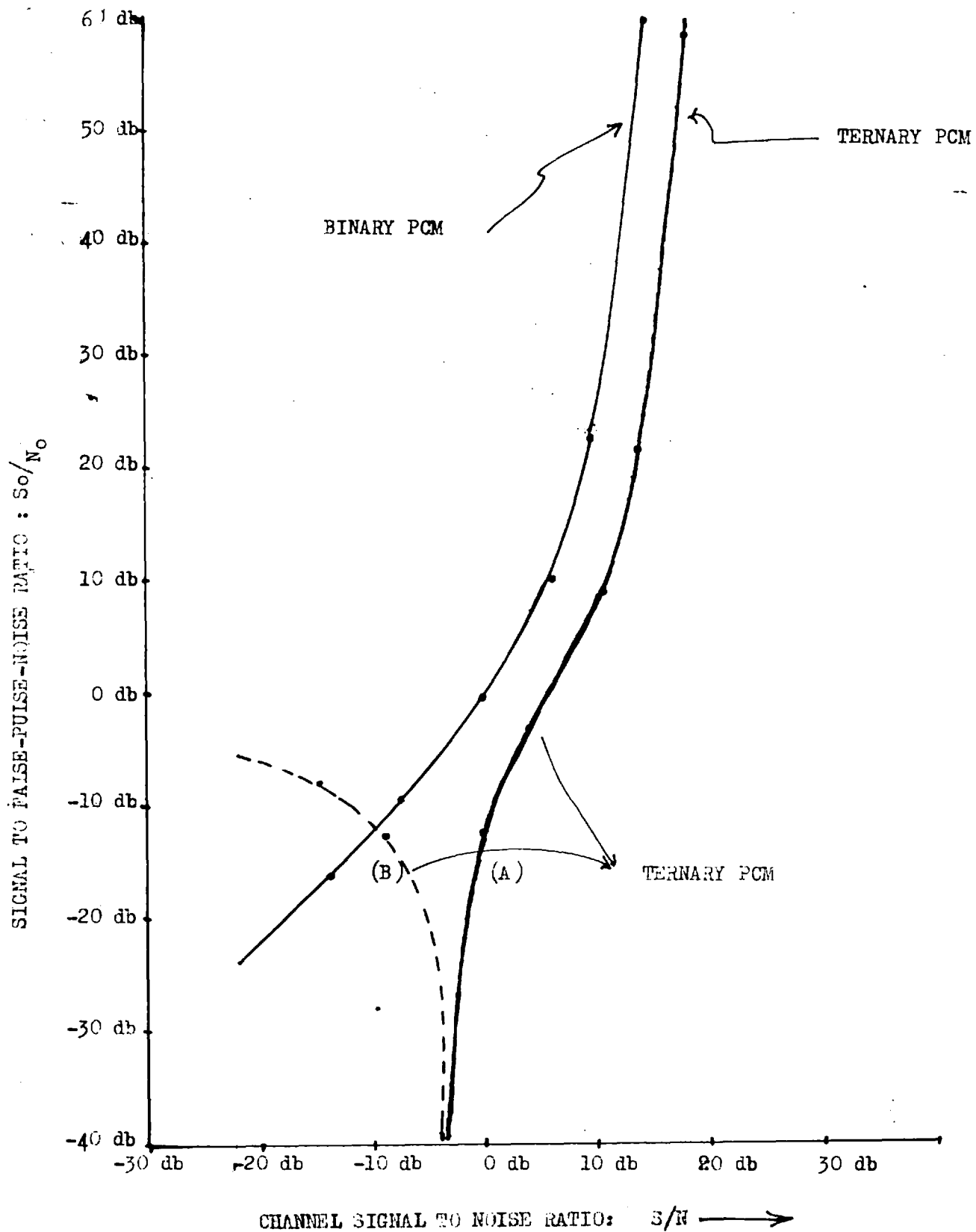
where

$$\Psi(V) = \frac{1}{\sqrt{2\pi}} \int_{-V}^{+V} e^{-\frac{x^2}{2}} dx$$

The signal to false-pulse noise ratio S_0/N_0 of a ternary channel operating under these conditions, is plotted as a function of the signal to channel noise ratio S/N , in Fig. 5.1. For purposes of comparative study of the two

SIGNAL TO FALSE PULSE-NOISE RATIO
vs
CHANNEL SIGNAL TO NOISE RATIO

FIG: 5.1.



Footnote to Fig. 5.1

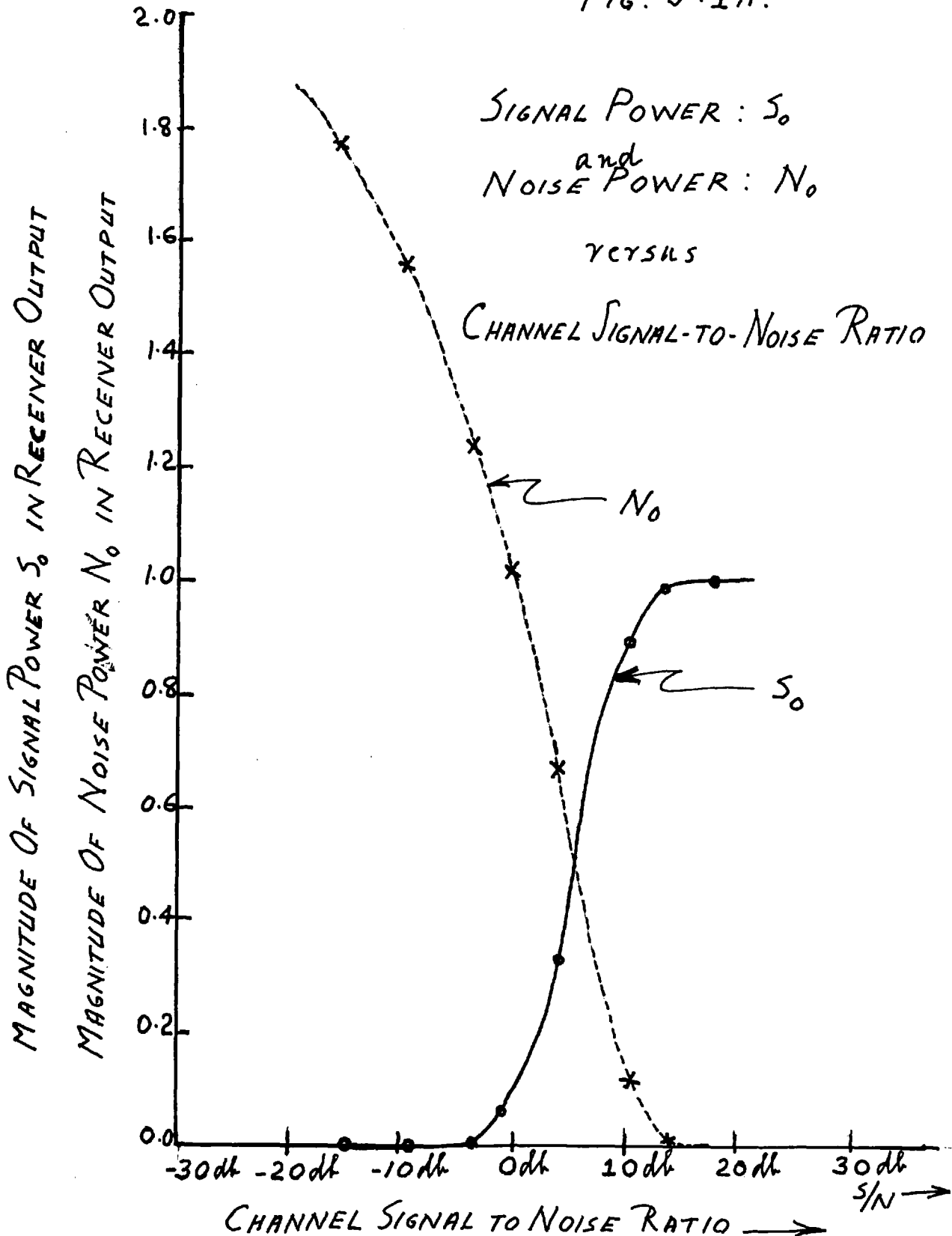
As the channel signal to noise ratio S/N falls from infinity to the value 0.42, the mean value (m_1) of the amplitude of the signal in the receiver output falls from unity to zero. If the channel signal to noise falls further, below this critical value, the mean value m_1 of the received signal amplitude becomes negative. In the statistical sense in which m_1 is used in the present context, a negative value of m_1 can only be interpreted as indicating the absence of any signal at all in the receiver output. In other words, for values of channel signal to noise ratio S/N less than 0.42, the signal to false-pulse-noise ratio of the system may be considered to be equal to zero. Under these conditions, the system may be considered to be acting as a source of pure noise with output power given by

$$N = m_1^2 + \mu_2 = m_2^2 = 1/3 \gamma$$

It is, therefore, obvious that the part designated B_2 of the signal to false-pulse-noise ratio (S_0/N_D) versus the channel signal to noise ratio (S/N) curve (Fig. 5.1) does not correctly represent the true signal to noise ratio of the system.

The output signal power S_0 and the false-pulse-noise power N_0 of the ternary system are plotted separately in fig. 5.1A.

FIG: 5.1A.



systems, the signal to false-pulse ratio $(S_o/N_o)_{\text{Binary}}$ of a binary channel operating under the same conditions, using the same signal power S as the ternary system, is also plotted in Fig. 5.1 using the formula^(31,36).

$$(S_o/N_o)_{\text{Binary}} = \frac{1}{4rs} - 1$$

where

$$r = \frac{1}{2} [p_1(1) + p_2(2)]$$

$$s = \frac{1}{2} [p_1(2) + p_2(1)]$$

the threshold effect common to all PCM systems,

It may be noted that while both curves exhibit the threshold effect for the ternary channel occurs at a higher signal to channel noise ratio S/N than for the corresponding binary. Also, for all values of S/N , the output signal to noise ratio S_o/N_o is higher for the binary system.

5.1.1 Anomalous Behaviour of S_o/N_o for Low Values of S/N

It is evident from equation 5.9 that S_o/N_o becomes zero when the mean signal amplitude m_1 at the receiver is zero. The important point to note here is that m_1 becomes zero before the signal to channel noise ratio S/N approaches zero. For instance, S_o/N_o becomes identically equal to zero when

$$\alpha = 0.54, \beta = 0.38, \gamma = 0.08$$

corresponding to the non-zero, finite value of $S/N = 0.42$.

[It may be noted that

$$\alpha = \beta = \gamma = \frac{1}{3} \text{ when } S/N = 0]$$

The reason for this anomalous behaviour is that a γ error (that is, an error corresponding to event y_2) contributes four times as much power at the output as the events y_0 or y_1 .

A closer study of equation 5.9 or, alternatively, of Fig. 5.1, further reveals that S_0/N_0 increases from zero to a finite value (0.29) as S/N decreases further to zero. For values of S/N in this range (0 to 0.42) it is obvious that S_0/N_0 , as given by equation 5.9, does not represent the true signal to noise ratio of the system. m_1 is negative in this range and, in the statistical sense in which m_1 is used in the present context, a negative value of m_1 can possibly be interpreted only as indicating the absence of any signal in the receiver output. A ternary system operating in this range of values of S/N may be considered to be acting as a pure noise generator of output power

$$m_2 = 1 + 3\gamma$$

The ternary system under discussion (Section 5.1) can therefore be profitably used to transmit a signal only if S/N is greater than a fixed, finite, value. It is instructive to find out whether it is possible to extend the range of values of S/N for which m_1 is finite and positive. Since it is the presence of γ errors which causes m_1 to take on negative values, it is immediately obvious that any scheme for extending the working range of a ternary system would involve techniques for decreasing the probability of occurrence of γ errors. Two methods present themselves:

1. Completely suppressing the transmission of the intermediate state (state 2) so that γ errors cannot occur. $m_1 = (1-2\beta)$ is now non-negative since β is never greater than 0.5. The system is, however, no longer ternary, it is binary!

2. Reducing the probability of occurrence, $p(1)=p(3)=p$,

of the extreme states, namely, state 1 and state 3. If the amplitude ($2V$) of the transmitted signal is made sufficiently large, the conditional probabilities $p_i(j)$, $i \neq j$, will be low enough to ensure that

$$m_1 = (1 - 2\beta - 3\gamma) > 0$$

For a given value of S/N , or, in other words, for a given transmitter power S , the signal amplitude ($2V$) can always be chosen as large as desired by making $p(1)$ correspondingly small according to the relation

$$p(1) \leq \frac{S}{8V^2} \quad \dots 5.11$$

If the signal amplitude is chosen to be too large, however, $p(1)$ and $p(3)$ become much smaller than $p(2)$ so that the source entropy $H(x)$ and therefore the rate of transmission of information, become very low. Therefore V must not be chosen to be any larger than is necessary to ensure that m_1 is positive and (m_1^2 / μ_2) is equal to the desired signal to false-pulse noise ratio S_0/N_0 .

The magnitude of incidence of γ errors can be reduced by adopting the above technique at the transmitting end of the channel. The signal to false-pulse noise ratio S_0/N_0 obtained by this means can be further improved by using a more refined detection procedure at the receiver, than the simple one discussed in the present Section. Such an optimum detection procedure for ternary signals is given in Section 5.3.

5.2 Signal to Noise Ratio for a Code Group Consisting of n Digits

Let $2A_0$ be the peak to peak amplitude of the unquantized signal. Assuming that the quantization is uniform and that n ternary digits $a_1 a_2 \dots a_n$ are used

to represent the original signal, the maximum amplitude A of the quantized signal in either direction, positive or negative, is

$$\begin{aligned} A &= 2A_0 \left(\frac{1}{3} + \frac{1}{3^2} + \text{-----} + \frac{1}{3^n} \right) \\ &= 2A_0 \left(\frac{1}{2} - \frac{1}{2 \cdot 3^n} \right) \\ &= A_0 \left(1 - \frac{1}{3^n} \right) \end{aligned} \quad \dots 5.12$$

The first digit is the most significant, the relative contribution of the k^{th} digit being inversely proportional to 3^k .

Let the average signal power P_k used for the transmission of the k^{th} digit be such that the probability of occurrence of the events $(y_0)_k$, $(y_1)_k$ and $(y_2)_k$ for the k^{th} digit are α_k , β_k and γ_k respectively. The relative contributions of $(y_0)_k$, $(y_1)_k$ and $(y_2)_k$ to the amplitude of the signal reconstructed at the receiver are $2A_0/3^k$, $-2A_0/3^k$ and $-4A_0/3^k$ respectively. Therefore, the probability generating function $(G_F)_k$ of the contribution due to the k^{th} digit is

$$(G_P)_k = \left(\alpha_k \cdot t^{\frac{2A_0}{3^k}} + \beta_k \cdot t^{-\frac{2A_0}{3^k}} + \gamma_k \cdot t^{-\frac{4A_0}{3^k}} \right) \quad \dots 5.13$$

The probability generating function of the received signal amplitude as given by the weighted sum of the contributions due to the various digits, is, therefore,

$$G_P(t) = \prod_{k=1}^n \left\{ \alpha_k \cdot t^{\frac{2A_0}{3^k}} + \beta_k \cdot t^{-\frac{2A_0}{3^k}} + \gamma_k \cdot t^{-\frac{4A_0}{3^k}} \right\} \quad \dots 5.14$$

the coefficient of t^y in the continued product giving the

probability of the amplitude of the received signal being equal to y . It is possible, though rather tedious, to obtain the moments of this compound distribution and, from these, the relative magnitudes of the signal and the noise power in the receiver output, using the moment generating function $G_M(a)$

$$G_M(a) = \prod_{k=1}^n \left\{ \alpha_k \cdot e^{a \frac{2A_0}{3^k}} + \beta_k \cdot e^{-a \frac{2A_0}{3^k}} + \gamma_k \cdot e^{-a \frac{4A_0}{3^k}} \right\} \dots 5.15$$

in a manner essentially similar to that used in Section 5.1. A more elegant method, however, both from the point of view of ease of computation as well as of obtaining a physical picture of the processes involved, is to make use of the cumulant generating function of the distribution as shown below.

The cumulant generating function $G_C(a)$

$$G_C(a) = C_1 a + C_2 \frac{a^2}{2!} + C_3 \frac{a^3}{3!} + \dots + C_n \frac{a^n}{n!} \dots 5.16$$

where $C_1, C_2, \dots, C_n$ are cumulants of order 1, 2, ..., n respectively, of the amplitude of the received signal is given by the relation⁽³²⁾

$$G_C(a) = \ln G_M(a) \dots 5.17$$

The moments $m_1, m_2, \dots, m_n$ and the corresponding cumulants $c_1, c_2, \dots, c_n$ are related as follows:

$$m_n = A_n(1; c_1, c_2, \dots, c_n) \dots 5.18A$$

$$c_n = A_n(f; m_1, m_2, \dots, m_n) \dots 5.18B$$

where $A_n(f; g_1, g_2, \dots, g_n)$ are

Bell polynomials⁽³⁵⁾ derived from the generating function

$$\begin{aligned} A(t) &\equiv f \left[g(t) \right] \\ &\equiv e^{ag(t)} \end{aligned} \quad \dots 5.19$$

so that

$$\begin{aligned} A_1(f; g_1) &= f_1 g_1 \\ A_2(f; g_1, g_2) &= f_1 g_2 + f_2 g_1^2 \\ A_3(f; g_1, g_2, g_3) &= f_1 g_2 + 3f_2 g_1 g_2 + f_3 g_1^2 \\ &\vdots \\ A_n(f; g_1, g_2, \dots, g_n) &= \left[e^{-ag(t)} \right] \left[\frac{d^n}{dt^n} e^{ag(t)} \right] \end{aligned} \quad \dots 5.20$$

where

$$\begin{aligned} f_1 &= \frac{d}{dg} f \left[g(t) \right] \\ &\vdots \\ f_n &= \frac{d^n}{dg^n} f \left[g(t) \right] \\ &= (-1)^{n-1} (n-1)! \end{aligned} \quad \dots 5.21$$

$$\begin{aligned} g_1 &= \frac{d}{dt} g(t) \\ &\vdots \\ g_n &= \frac{d^n}{dt^n} g(t) \end{aligned} \quad \dots 5.22$$

Now, cumulants possess the interesting property that when

independent systems are compounded, the r^{th} cumulant C_r of the compound system is given by the sum $\{(C_r)_1 + (C_r)_2 + \dots + (C_r)_n\}$ of the r^{th} cumulants of the separate systems. Since the latter can be easily calculated from the corresponding moments, the cumulants of the compound system are easily determined.

Now, the moments m_1 and m_2 of the distribution of the received signal amplitude due to the k^{th} digit, are

$$(m_1)_k = \frac{2A_0}{3^k} (1 - 2\beta_k - 3\gamma_k) \quad \dots 5.23$$

$$(m_2)_k = \frac{4A_0^2}{3^{2k}} (1 + 3\gamma_k) \quad \dots 5.24$$

from equations 5.4, 5.5 and 5.15. The corresponding cumulants $(c_1)_k$ and $(c_2)_k$ are, therefore, given by

$$\begin{aligned} (c_1)_k &= (m_1)_k \\ &= \frac{2A_0}{3^k} (1 - 2\beta_k - 3\gamma_k) \quad \dots 5.25 \end{aligned}$$

$$\begin{aligned} (c_2)_k &= (m_2)_k - (m_1)_k^2 \\ &= \frac{4A_0^2}{3^{2k}} [-4\beta_k(1-\beta_k) + 9\gamma_k(1-\gamma_k) - 12\beta_k\gamma_k] \quad \dots 5.26 \end{aligned}$$

from equations 5.18B and 5.20

The cumulants C_1 and C_2 of the compound system (that is, of the probability distribution of the total amplitude of the received signal) are now given by

$$C_1 = \sum_{k=1}^n \frac{2A_0}{3^k} (1 - 2\beta_k - 3\gamma_k) \quad \dots 5.27$$

$$C_2 = \sum_{k=1}^n \frac{4A_0^2}{3^{2k}} [-4\beta_k(1-\beta_k) + 9\gamma_k(1-\gamma_k) - 12\beta_k\gamma_k] \quad \dots 5.28$$

If it is assumed that α , β and γ are the same for all the n digits, then

$$\begin{aligned}
 C_1 &= \sum_{k=1}^n (c_1)_k \\
 &= (1-2\beta-3\gamma) \sum_{k=1}^n \frac{2A_0}{3^k} \\
 &= A_0 \left(1 - \frac{1}{3^n}\right) (1 - 2\beta - 3\gamma) \quad \dots 5.29
 \end{aligned}$$

$$\begin{aligned}
 C_2 &= \sum_{k=1}^n (c_2)_k \\
 &= \left[-4\beta(1-\beta) + 9\gamma(1-\gamma) - 12\beta\gamma \right] \sum_{k=1}^n \frac{4A_0^2}{3^{2k}} \\
 &= \frac{1}{2} A_0^2 \left(1 - \frac{1}{3^{2n}}\right) \left[-4\beta(1-\beta) + 9\gamma(1-\gamma) - 12\beta\gamma \right] \dots 5.30
 \end{aligned}$$

From equations 5.18A, 5.20 and 5.22, it is easily seen that the moments M_1 and M_2 of the compound distribution are

$$\begin{aligned}
 M_1 &= C_1 \\
 M_2 &= C_1^2 + C_2 \quad \dots 5.31
 \end{aligned}$$

so that the central moment μ_2 is

$$\begin{aligned}
 \mu_2 &= M_2 - M_1^2 \\
 &= C_2 \quad \dots 5.31A
 \end{aligned}$$

Therefore, the output signal to false-pulse noise ratio (S_0/N_0) of the compound ternary system is

$$\begin{aligned}
\frac{S_0}{N_0} &= \frac{M_1^2}{\mu_2} \\
&= \frac{A_0^2 (1 - \frac{1}{3^n})^2 (1 - 2\beta - 3\gamma)^2}{\frac{A_0^2}{2} (1 - \frac{1}{3^{2n}}) [4\beta(1-\beta) + 9\gamma(1-\gamma) - 12\beta\gamma]} \\
&= \frac{2(1 - 1/3^n)(1 - 2\beta - 3\gamma)^2}{(1 + \frac{1}{3^n}) [4\beta(1-\beta) + 9\gamma(1-\gamma) - 12\beta\gamma]} \quad \dots 5.32
\end{aligned}$$

5.2.1 Maximization of S_0/N_0 of an n-digit Ternary System

The relative contribution of the k^{th} digit to the total noise power N_0 at the receiver is given by the ratio

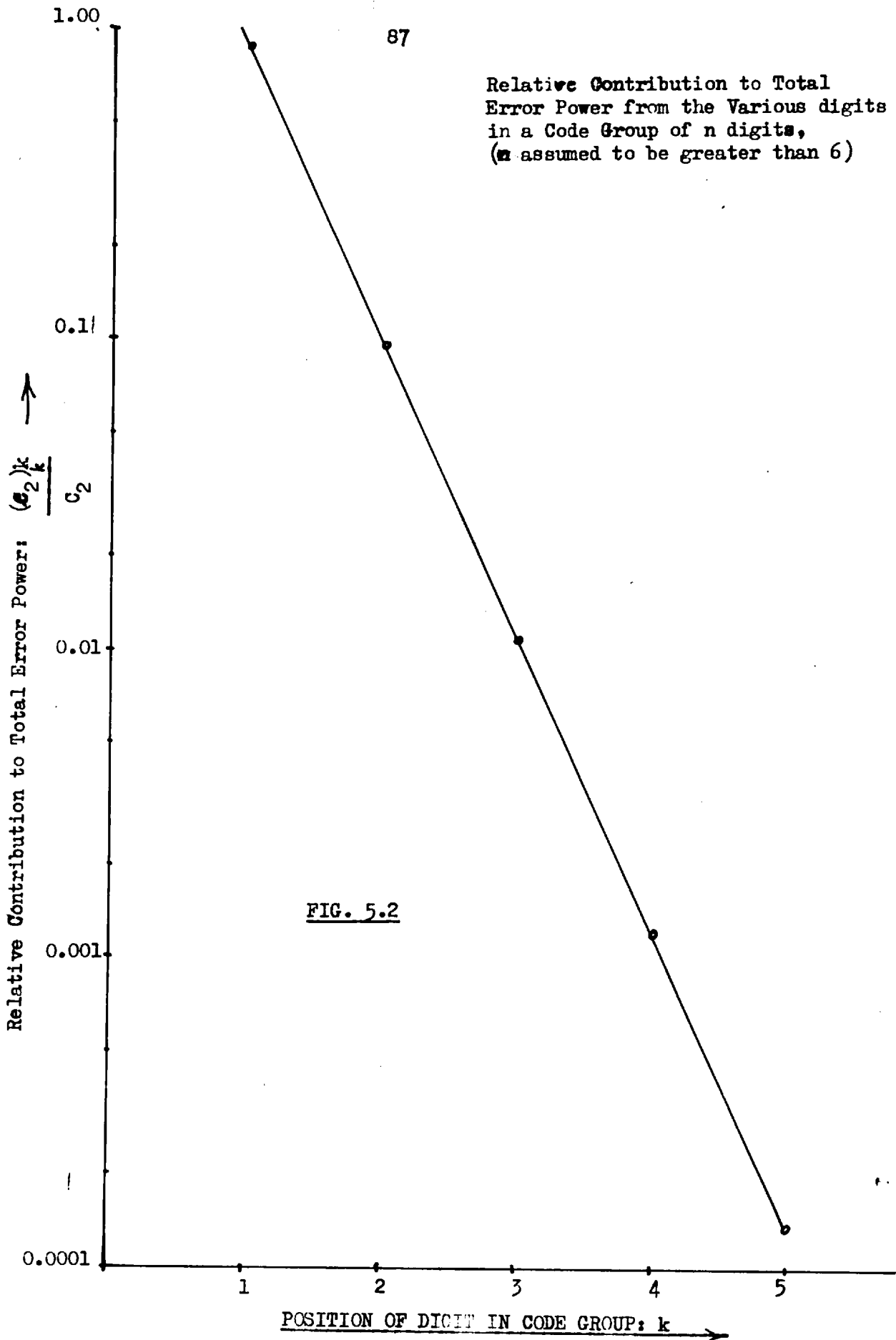
$$\begin{aligned}
\frac{(N_0)_k}{N_0} &= \frac{(c_2)_k}{C_2} \\
&= \frac{4A_0^2}{3^{2k}} \bigg/ \left[\frac{A_0^2}{2} \left(1 - \frac{1}{3^{2n}} \right) \right] \\
&= \frac{8}{3^{2k} (1 - \frac{1}{3^{2n}})} \quad \dots 5.33
\end{aligned}$$

assuming that the probability coefficients α_k , β_k and γ_k are invariant with respect to k .

The ratio of $(c_2)_k / C_2$ is plotted in Fig. 5.2 as a function of k .

It may be seen from Fig. 5.2 that approximately 98% of the total noise power is caused by wrong interpretation of the first two digits. It was, however, shown in Section 5.1

Relative Contribution to Total
Error Power from the Various digits
in a Code Group of n digits,
(n assumed to be greater than 6)



that when the system is operating over the threshold, a very slight increase in the transmitted signal power S causes the output signal to noise ratio S_0/N_0 to increase exponentially with S (Fig. 5.1). Therefore, it is reasonable to expect an increase in S_0/N_0 for the same value of channel signal to noise ratio S/N , if the amounts of signal power P_k allotted to the various digits are re-distributed such that the more significant digits, such as the first and the second, are allotted proportionately more power than the less significant. Assuming that the amplitude of the k^{th} digit is $(2V_k)$ and that all the three states of the digit are equally probable, the output signal to noise ratio S_0/N_0 is maximum if V_k and, therefore, the corresponding α_k , β_k and γ_k are chosen such that the function I , defined below, is maximum:

$$I = \frac{\left[\sum_{k=1}^n \frac{2A_0}{3^k} (1-2\beta_k-3\gamma_k) \right]^2}{\sum_{k=1}^n \frac{4A_0^2}{3^{2k}} \left[4\beta_k(1-\beta_k)+9\gamma_k(1-\gamma_k)-12\beta_k\gamma_k \right]} + \lambda S \quad \dots 5.34$$

where

$$\begin{aligned} \lambda & \text{ is a Lagrangian multiplier,} \\ S &= \sum_{k=1}^n P_k \\ &= \sum_{k=1}^n \sum_{i=1}^3 p(i) (2V_k)_i^2 \\ &= \sum_{k=1}^n \frac{8}{3} V_k^2, \text{ since } p(i) = \frac{1}{3} \end{aligned}$$

I is maximum if

$$\frac{\partial I}{\partial V_k} = 0, \quad k = 1, 2, \dots, n \quad \dots 5.35$$

β_k and γ_k being functions of V_k .

5.3 Detection of Ternary Signals in the Presence of Noise

Let the probability of transmission $p(i)$ of state i ($i = 1, 2, 3$) of the ternary symbol be assumed to be known. If the amplitude x_i of the transmitted symbol and also the characteristics of the interfering noise are known, the conditional probability $p_i(y)$ that the amplitude of the received signal is y if state i is transmitted, can be determined. The corresponding a posteriori probability $p_y(i)$ that state i is transmitted if the received signal amplitude is y , is given by

$$p_y(i) = \frac{p(i) p_i(y)}{\sum_{i=1}^3 p(i) p_i(y)}, \quad i = 1, 2, 3 \quad \dots 5.36$$

The problem of detection now resolves itself into dividing the sample space Y consisting of the ensemble of all received signals, into three regions Y_j ($j = 1, 2, 3$) such that if the received signal falls into class Y_j then decision D_j ($j = 1, 2, 3$) is optimum according to some criterion. (Decision D_j is said to be taken if it is decided at the receiver that the transmitted signal was in state j). For instance, the simple detection process assumed in Sections 5.1 and 5.2 makes its decisions on the assumption that decision D_j is optimum if it makes the corresponding a posteriori probability $p_y(i=j)$ maximum. On this basis, it can be easily shown that if the channel noise has a symmetrical amplitude distribution with zero mean and if the amplitude of the transmitted signal is $2V_k$, then the optimum division

Y_j of the sample space Y is

$$(y > V_k) \subset Y_1$$

$$(-V_k \leq y \leq +V_k) \subset Y_2$$

$$(y < -V_k) \subset Y_3$$

...5.37*

Y_1, Y_2, Y_3 corresponding to decisions D_1, D_2, D_3 respectively.**

The procedure outlined above gives an equal weight to all types of errors of decision. Therefore, from the point of view of minimisation of false-noise power this method is not optimum since, as shown in Section 5.1, certain types of error distort the reconstructed signal more than others - a γ error contributes four times as much error power as a β error. An optimum detector which minimises false-pulse error power can however be designed using an extension of the principles of statistical decision theory^(37,38,39) as shown below.

If state i of a symbol is transmitted but the decision at the receiver is D_j , let the noise power introduced into the reconstructed signal be denoted by N_{ij} . For a ternary system, the noise power matrix is

$$\|N_{ij}\| = \begin{vmatrix} 0 & 1 & 4 \\ 1 & 0 & 1 \\ 4 & 1 & 0 \end{vmatrix} \quad \dots 5.38$$

* $(y > V_k) \subset Y_1$ implies that a sample of amplitude greater than V_k forms a subset of sample space Y_1 .

** It maybe noted that Y_2 may have boundary points in common with Y_1 and Y_3 . These common points, however, form a null class of probability zero, and may be neglected.

The average value or expectation of noise power in the reconstructed signal if the decision is D_j , is now given by

$$\begin{aligned} E(N|D_j) &= \sum_{i=1}^3 \frac{p(i) p_i(y)}{\sum_{i=1}^3 p(i) p_i(y)} N_{ij} \\ &= \sum_{i=1}^3 p_y(i) N_{ij} \end{aligned} \quad \dots 5.39$$

where $p_y(i)$ is the a posteriori probability of transmission of state i . If the optimum detector is defined as one that minimises false ^{pulse} noise error power, the decision problem resolves itself into:

- 1) computing $p_y(i)$ for $i = 1, 2, 3$
- 2) computing $E(N|D_j)$ for different values of j and
- 3) choosing that value of j for which $E(N|D_j)$ is minimum.

In other words, the transmitted symbol is detected as being in state 1 if

$$E(N|D_1) \leq E(N|D_2) ; E(N|D_1) \leq E(N|D_3) \quad \dots 5.40A$$

as being in state 2 if

$$E(N|D_2) \leq E(N|D_1) ; E(N|D_2) \leq E(N|D_3) \quad \dots 5.40B$$

as being in state 3 if

$$E(N|D_3) \leq E(N|D_1) ; E(N|D_3) \leq E(N|D_2) \quad \dots 5.40C$$

Substituting actual numerical values for N_{ij} , we have:

decision D_1 is optimum if

$$p_y(2) + 4p_y(3) \leq p_y(1) + p_y(3) \quad \dots 5.41A$$

and

$$" \leq 4p_y(1) + p_y(2)$$

decision D_2 is optimum if
 $p_y(1) + p_y(3) \leq p_y(2) + 4p_y(3)$...5.41B

and " $\leq 4p_y(1) + p_y(2)$

decision D_3 is optimum if
 $4p_y(1) + p_y(2) \leq p_y(2) + 4p_y(3)$...5.41C

and " $\leq p_y(1) + p_y(3)$

5.3.1 Graphical Representation of the Ternary Detection Problem

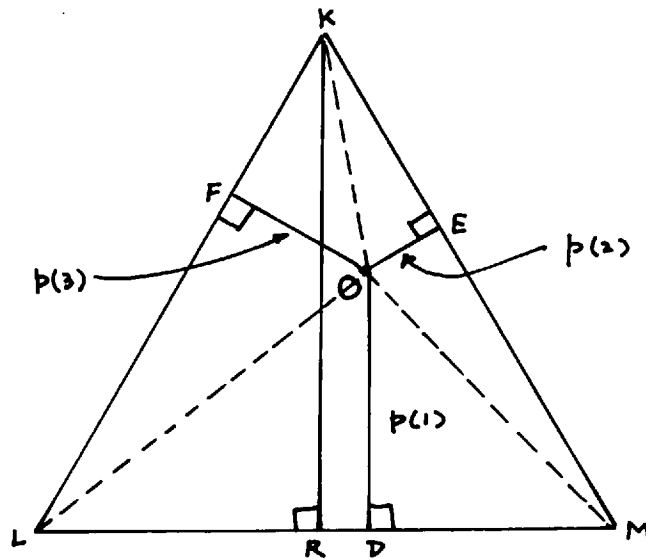
It can be easily shown that the sum of the perpendicular distances to the three sides of an equilateral triangle from a point inscribed inside it, is equal to the height of the triangle. An event specified by the ternary system of probabilities $p(1), p(2), p(3)$

$$\sum_i p(i) = 1$$

can therefore be represented by a point inscribed in an equilateral triangle of altitude unity, if the coordinates of the point as specified by its distances from the three sides are equal to $p(i), i = 1, 2, 3$. (Fig. 5.3).

Since each element y of the sample space Y consisting of the ensemble of all received signals, is completely specified by the ternary system of a posteriori probabilities $p_y(i)$ associated with y , the sample space Y can be set in one-to-one correspondence with the total area A of the equilateral triangle.

Optimum decision is now effected by the following means:



Equating triangles OLM, OMK and OKL to the equilateral triangle KLM,

$$(1/2) OD \cdot LM + (1/2) OE \cdot MK + (1/2) OF \cdot KL = (1/2) KR \cdot LM$$

Therefore,

$$OD + OE + OF = KR,$$

Or,

$$p(1) + p(2) + p(3) = 1,$$

if altitude KR of KLM is equal to unity.

FIG. 5.3. REPRESENTATION OF A TERNARY SYSTEM OF PROBABILITIES.

- 1) Making use of the inequalities 5.41A,B,C which specify the optimum division of the sample space Y into Y_j , divide the total area A of the ^{triangle} into three regions A_j , ($j = 1, 2, 3$) such that A_j corresponds to Y_j .
- 2) Compute the relative a posteriori probabilities $p_y(i) \equiv \phi_i$ corresponding to the given received signal y .
- 3) Locate the point corresponding to this system of probabilities inside the triangle.
- 4) If the point falls in region A_j , make decision D_j .

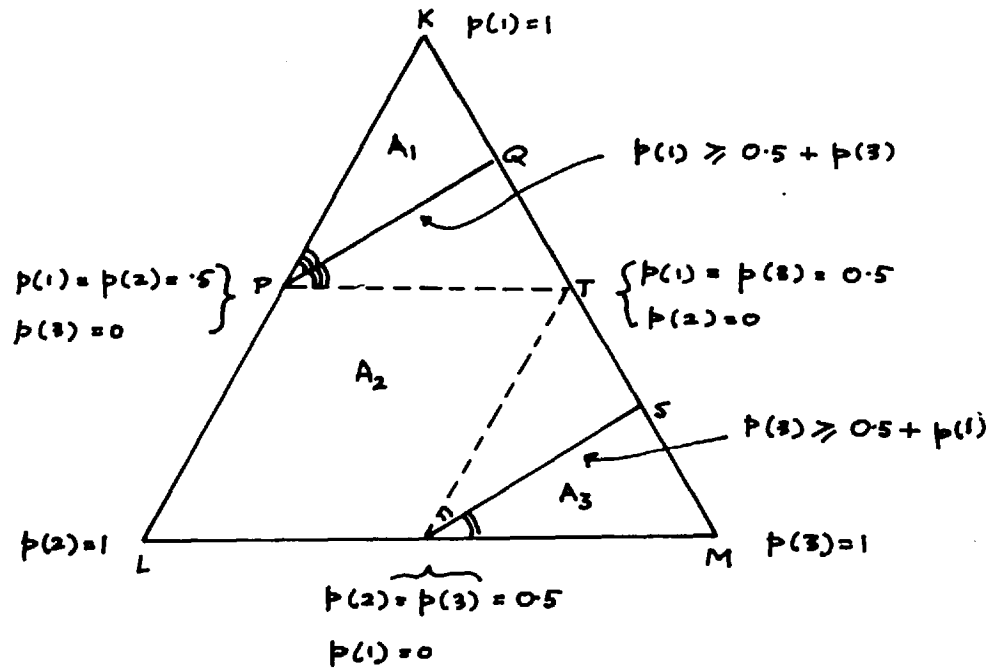
Equations 5.41A,B,C are not in a form suitable for simple geometrical construction. By simple algebraic manipulation it is however possible to derive from them the following equations to specify the loci of points which specify the boundaries of A_j :

$$\begin{aligned} A_1 : \quad p_1 &\geq 0.5 + p_3 \\ \text{and} \quad p_1 &\geq p_3 \end{aligned} \qquad \dots 5.42A$$

$$\begin{aligned} A_2 : \quad p_1 &\leq 0.5 + p_3 \\ \text{and} \quad p_3 &\leq 0.5 + p_1 \end{aligned} \qquad \dots 5.42B$$

$$\begin{aligned} A_3 : \quad p_3 &\geq p_1 \\ \text{and} \quad p_3 &\geq 0.5 + p_1 \end{aligned} \qquad \dots 5.42C$$

The results are plotted in Fig. 5.4



A_1, A_2, A_3 : DECISION REGIONS

FIG. 5.4. GRAPHICAL REPRESENTATION OF OPTIMUM TERNARY DETECTION PROBLEM.

6. Analysis of Quantization Distortion

Apart from false-pulse noise, the performance of a PCM system is subject to two further types of transmission impairment. These are 1) overloading and 2) quantization distortion. The error power N_{OL} due to overloading is easily computed from the relation

$$N_{OL} = 2 \int_{A_0}^{\infty} p(x) \cdot x^2 dx \quad \dots 6.1$$

where $p(x)$ is the probability density function of the signal amplitude and A_0 is half the quantization range. For most types of random inputs of practical interest, such as the gaussian, this can be neglected if A_0 is chosen to be greater than four times the r.m.s. value of the input signal.⁽¹⁶⁾ The present Section is therefore devoted to a discussion of only quantization distortion. The effect of quantization, *on the fidelity of transmission* as given by the signal-to-noise ratio, at the output of the quantizer, as well as on the rate of transmission of information, is discussed.

6.1 Evaluation of Quantization Distortion

If the peak to peak amplitude range of the signal is finite and equal to $2A_0$, the input signal power S is given by

$$S = \int_{-A_0}^{+A_0} x^2 p(x) dx \quad \dots 6.2$$

Let quantization be assumed to be uniform. If the number of quantization levels is ℓ and the magnitude of a quantization interval is h , then the probability density p_i associated with the i^{th} quantization level is

$$p_i = \int_{-h/2}^{+h/2} p(x_i + u) du \quad \dots 6.3$$

so that the total power of the output of the quantizer is

$$\begin{aligned} S_0 &= \sum_{i=-(\ell-1)/2}^{+(\ell-1)/2} p_i x_i^2 \\ &= \sum_{i=-(\ell-1)/2}^{+(\ell-1)/2} x_i^2 \int_{-h/2}^{+h/2} p(x_i + u) du \quad \dots 6.4 \end{aligned}$$

As ℓ increases from low to sufficiently high values, there is not only a quantitative change in the magnitude of S_0 but also a qualitative change in its structure as shown below.

6.1.1 Quantization Distortion when ℓ is large.

According to the Euler-Maclaurin summation theorem, we have ⁽⁴⁰⁾

$$\begin{aligned} \int_{\alpha}^{\alpha+\ell h} f(x) dx &= h \left\{ \frac{1}{2}f(\alpha) + f(\alpha+h) + \dots + f(\alpha+\ell-1h) + \frac{1}{2}f(\alpha+\ell h) \right. \\ &\quad \left. - \sum_{j=2}^{m-1} B_j \frac{h^j}{j!} \left\{ f^{(j-1)}(x) \right\}_{\alpha}^{\alpha+\ell h} + h R_m \right\} \quad \dots 6.5 \end{aligned}$$

where

B_j is the Bernoulli number of order j , defined as the coefficient of $(t^j/j!)$ in the expansion of $t/(e^t-1)$

and

$$R_m = B_m \frac{\ell h^m}{m!} f^{(m)}(\alpha + \theta \ell h) \quad 0 \leq \theta \leq 1 \quad \dots 6.5A$$

Let

$$f(x) = x^2 \int_{-h/2}^{+h/2} p(x+u) du \quad \dots 6.6$$

If $p(x)$ and its first m derivations $p'(x)$, $p''(x)$, ----, $p^{(m)}(x)$ - which are assumed to exist - vanish at $x = -A_0$ and $x = +A_0$, then $f(x)$ and its first $(m+1)$ derivatives will not only be continuous in the range $(-A_0, +A_0)$ but also vanish at $x = -A_0$ and $+A_0$. Consequently, equation 6.5 reduces to

$$\int_{\alpha}^{\alpha + \ell h} f(x) dx = h \sum_{i=1}^{2a} f(\alpha + ih) + h R_{m+1} \quad \dots 6.7$$

where

$$a = (\ell h - 1)/2 \quad \dots 6.7A$$

Making use of equations 6.4, 6.6 and 6.7 we have

$$\begin{aligned} \int_{-A_0}^{+A_0} x^2 dx &= \int_{-h/2}^{+h/2} p(x+u) du \\ &= h \sum_{i=-a}^{+a} x_i^2 \int_{-h/2}^{+h/2} p(x_i+u) du + h R_{m+1} \\ &= h S_0 + h R_{m+1} \quad \dots 6.8 \end{aligned}$$

Also

$$\begin{aligned} \int_{-A_0}^{+A_0} x^2 dx \int_{-h/2}^{+h/2} p(x+u) du &= \int_{-A_0}^{+A_0} \int_{-h/2}^{+h/2} x^2 p(x+u) du dx \\ &= \int_{-A_0}^{+A_0} \int_{-h/2}^{+h/2} (x-u)^2 p(x) du dx \end{aligned}$$

substituting x for $(x+u)$ and
integrating with respect to u *

$$\begin{aligned} &= \int_{-A_0}^{+A_0} p(x) \frac{(x+h/2)^3 - (x-h/2)^3}{3} dx \\ &= h \int_{-A_0}^{+A_0} p(x) x^2 dx + \frac{h^3}{12} \int_{-A_0}^{+A_0} p(x) dx \\ &= h(S + h^2/12) \quad \dots 6.9 \end{aligned}$$

$$\therefore S_0 + R_{m+1} = S + h^2/12 \quad \dots 6.10$$

If the tails of the probability density function $p(x)$ approach the variate axis sufficiently close such that $p(x)$ and at least its first two** derivatives tend to zero at the extremities of the distribution and if, in

* This is permissible if the double integral is absolutely convergent.

** It is not necessary to take account of the third derivative since $B_{2k+1} \equiv 0$
for all integral values of $k \geq 1$

addition, λ is large so that h is small, then it can be shown^{40,41} that R_{m+1} can be neglected in comparison with $h^2/12$. It is thus seen that the quantizing operation causes the output power to be greater than the input signal power by an amount N_Q :

$$\begin{aligned} N_Q &= S_o - S \\ &= h^2/12 \end{aligned} \quad \dots 6.11$$

6.1.2 Quantization Distortion when λ is Small

When the number of quantization levels is small, it is difficult to compute the error power N_Q for the following reasons:

- a) If the quantization distortion is to be kept within bounds, it is necessary to reduce the dynamic range of quantization. Under these conditions, the error power due to overloading cannot be neglected.
- b) Reduction of the dynamic range of quantization is equivalent to truncating the input probability distribution at points where the probability density function $p(x)$ is not small. Since the new truncated distribution $p_T(x)$ and its derivatives $p_T'(x)$, $p_T''(x)$, ---- do not vanish at the extremities of the distribution the term

$$\sum_{j=2}^{m-1} B_j \frac{h^j}{j!} \left\{ f^{(j-1)}(x) \right\}_{\alpha}^{\alpha + \ell h}$$

in equation 6.5. cannot be neglected.

- c) Since the quantization intervals are large, the remainder term R_m in equation 6.5 is not small.

Evaluation of R_m is however complicated by the fact that the magnitude of Θ is not known.

The maximum bounds to the error due to quantization can however be computed in an approximate manner as shown below.

The quantized signal power S_0 may be written as

$$S_0 = \sum_i z^2 \int_{z-h/2}^{z+h/2} p(x) dx \quad \dots 6.12$$

where

$$z = (i + \frac{\theta}{2\pi}) h \quad \dots 6.12A$$

Fisher has shown that equation 6.12 is equivalent to⁴²

$$S_0 = C_0 + C_1 \sin \theta + C_2 \sin 2\theta + \dots \\ + D_1 \cos \theta + D_2 \cos 2\theta + \dots \quad \dots 6.13$$

where

$$C_0 = \frac{1}{2\pi} \sum_i \int_0^{2\pi} d\theta \int_{z-h/2}^{z+h/2} z^2 p(x) dx \quad \dots 6.13A$$

$$C_r = \frac{1}{\pi} \sum_i \int_0^{2\pi} \sin(r\theta) d\theta \int_{z-h/2}^{z+h/2} z^2 p(x) dx \quad \dots 6.13B$$

$$D_r = \frac{1}{\pi} \sum_i \int_0^{2\pi} \cos(r\theta) d\theta \int_{z-h/2}^{z+h/2} z^2 p(x) dx \quad \dots 6.13C$$

It can be shown that

$$C_0 = \frac{1}{h} \int_{-\infty}^{+\infty} p(x) dx \int_{x-h/2}^{x+h/2} z^2 dz \quad \dots \text{Ref.42}$$

$$= \frac{1}{h} \int_{-\infty}^{+\infty} \int_{-h/2}^{+h/2} x^2 p(x+u) dx du \quad \dots 6.14$$

But the right hand side of equation 6.14 is the same as the left hand side of equation 6.8 multiplied by h . Therefore,

$$C_0 = S + h^2/12 \quad \dots 6.15.$$

Assuming the input signal distribution to be gaussian with zero mean and variance $\sigma^2 = S$, it can be further shown that⁽⁴²⁾

$$C_r = 0 \quad r = 1, 2, \dots$$

$$D_r = (-1)^r 4 \left(\sigma^2 + \frac{h^2}{4\pi^2 r^2} \right) e^{-\frac{2r^2 \pi^2 \sigma^2}{h^2}} \quad \dots 6.16$$

Therefore,

$$\begin{aligned} S_0 = S + \frac{h^2}{12} &- 4 \left(S + \frac{h^2}{4\pi^2} \right) e^{-\frac{2\pi^2}{h^2} S} \cos \theta \\ &+ 4 \left(S + \frac{h^2}{16\pi^2} \right) e^{-\frac{8\pi^2}{h^2} S} \cos 2\theta \\ &- 4 \left(S + \frac{h^2}{36\pi^2} \right) e^{-\frac{18\pi^2}{h^2} S} \cos 3\theta \\ &+ \dots \end{aligned}$$

The error power ($S_0 - S$) therefore consists of a steady " "

term $N_0 = h^2/12$ as well as an infinite series of periodic terms. Since the coefficients of the periodic terms form an asymptotic series, the maximum possible contribution due to the periodic component is

$$\delta_\theta = -4(S + h^2/4\pi^2)e^{-2\pi^2 S/h^2} \quad \dots 6.18$$

The relative magnitudes of this component as well as of the steady term N_0 are given below for various values of h .

$$\begin{aligned} h = \sigma : \\ \delta_\theta &= 4(S + S/4\pi^2)e^{-2\pi^2} \\ &= 0.8 \times 10^{-8} S \\ N_0 &= h^2/12 \\ &= 0.085 \end{aligned}$$

$$\begin{aligned} h = 1.5\sigma : \\ \delta_\theta &= 0.0006 S \\ N_0 &= 0.187 S \end{aligned}$$

$$\begin{aligned} h = 2\sigma \\ \delta_\theta &= 0.032 S \\ N_0 &= 0.33 S \end{aligned}$$

$$\begin{aligned} h = 3\sigma : \\ \delta_\theta &= 0.545 S \\ N_0 &= 0.75 S \end{aligned}$$

It may be noted that the magnitude of the periodic component is quite negligible even when the quantization interval (h) is as large as one standard deviation, . Therefore, if $h \leq \sigma$, it is possible to obtain the true second moment (S) of the original signal distribution by subtracting the correction term N_0 which is completely

defined in terms of h , from the second moment of the quantized signal distribution, namely the output power S_0 . Proceeding in a manner essentially similar to that used in deriving S_0 , it can be shown that the correction terms to be applied to the other moments, namely 1st, 3rd, 4th, . . . , of the output distribution to get the true moments of the original distribution is a function of only the quantization interval, h , if $h \leq \sigma$. Since the useful dynamic range of a signal is usually confined to about 8σ , recovery of the original signal statistics is possible if $\ell \geq 8$, ℓ being the number of levels.

There is a sudden increase in the magnitude of δ_θ when h exceeds σ . For instance, δ_θ attains a value of approximately 10% of the steady component when $h = 2\sigma$. Now, the magnitude of the periodic component $\delta_\theta \cos \theta$ is a function of θ and is different for different locations of the quantization levels along the variate (signal input) axis.* Since the magnitude of θ cannot be determined from a study of the output of the quantizer, a correction term involving the periodic component $\delta_\theta \cos \theta$ cannot be formulated even though δ_θ is known. Therefore, for values of

* This follows from the fact that even though the quantization interval h is fixed, if the levels of quantization are shifted, as a whole, along the variate axis, the groupings corresponding to the various levels change so that the value of S_0 also changes. It is easily seen, however, that S_0 is periodic of period h , if the number of levels is sufficiently large. This is the basis of the formulation of equation 6.12. If the tails of the distribution at the overloading limits $\pm\Delta_0$ of the quantizer, approach the variate axis very closely the limitation that ℓ must be large can be relaxed to a certain extent.

h greater than σ , the statistics of the original signal cannot be reconstructed with good accuracy as the periodic component now begins to assume relatively large values.

Even though the actual magnitude of the periodic component $\delta_\theta \cos \theta$ in any particular case is unknown, the statistics of the error introduced by this component on an average can be evaluated by studying the ensemble consisting of all quantizers. Assuming that θ has a rectangular probability density function $g(\theta)$

$$g(\theta) = \begin{cases} 0, & 0 > \theta > 2\pi \\ \frac{1}{2\pi}, & 0 \leq \theta \leq 2\pi \end{cases} \quad \dots 6.19$$

the probability density function $h(x)$ for

$$x = K \cos \theta \quad \dots 6.20$$

where

$$\begin{aligned} K &\equiv \delta_\theta \\ &= -4(S + h^2/4\pi^2)e^{-2\pi^2 S/h^2} \end{aligned}$$

can be determined as follows.

Inverting equation 6.20, we have

$$\begin{aligned} \theta &= f(x) \\ &= \arccos(x/K) \end{aligned} \quad \dots 6.21$$

Direct substitution of 6.21 in 6.19 gives

$$\begin{aligned} g(\theta) d\theta &= g[f(x)] \left| J(\theta/x) \right| dx \\ &= h(x) dx \end{aligned} \quad \dots 6.22$$

where $J(\theta/x)$ is the Jacobian of θ with respect to x .

Now, for each value of x in the range $-K$ to $+K$, there exist two values of θ .

Therefore,

$$J(\theta/x) = 2 \frac{d}{dx} f(x) \quad \dots 6.23$$

whence

$$h(x) = \begin{cases} 0, & -K < x < +K \\ 1/\pi (K^2 - x^2)^{1/2}, & -K \leq x \leq K \end{cases} \quad \dots 6.24$$

The expectation or average value of the periodic component is given by

$$\begin{aligned} E(x) &= \int_{-K}^{+K} x h(x) dx \\ &= \int_{-K}^{+K} \frac{x dx}{\pi (K^2 - x^2)^{1/2}} \\ &= 0 \end{aligned} \quad \dots 6.25$$

while the mean square value is

$$\begin{aligned} E(x^2) &= \int_{-K}^{+K} \frac{x^2 dx}{\pi (K^2 - x^2)^{1/2}} \\ &= \frac{1}{\pi} \left[-\frac{x}{2} (K^2 - x^2)^{1/2} + \frac{K^2}{2} \arcsin(x/K) \right]_{-K}^{+K} \\ &= \frac{1}{2} K^2 \end{aligned} \quad \dots 6.26$$

$K^2/2$ is also the variance of the distribution of the periodic component since $E(x)$ is zero.

A reasonable compromise between low overloading error power and high quantization noise on one hand and large overloading and low quantization noise on the other, can be achieved by keeping the dynamic range of quantization ($2A_0$) equal to approximately 6 when the permissible number of quantization levels is small. It may be noted that under these conditions, the use of a 3-level quantizer, which corresponds to $h = 2$, makes it possible to recover the statistics of the original signal without a great loss in accuracy, the error being approximately 10% in the estimation of quantization noise corresponding to an error of about 3% in the estimation of the variance S of the signal distribution. On the other hand, the error introduced by the use of a 2-level quantizer is 73% in the estimation of quantization noise and approximately 55% in the estimation of variance.

The above estimates of error are only approximately correct since the effect of the abrupt termination of the tails of the distribution caused by limiting the overloading range to 6σ is not taken into account.

6.1.3 General Expression for Quantization Noise

If $p(x)$ and its derivatives do not vanish at the extremities of the distribution then the terms

$$\sum_{j=2}^{m-1} \frac{h^j}{j!} B_j \left[f^{(j-1)}(x) \right]_{\alpha}^{\alpha+h} \quad \text{and} \quad R_m$$

cannot be neglected. For instance, if the probability

distribution is rectangular

$$p(x) = \begin{cases} \frac{1}{2A_0} , & -A_0 \leq x \leq +A_0 \\ 0 , & A_0 > |x| \end{cases} \quad \dots 6.27$$

and if the quantization levels are equally spaced, it can be shown that

$$S = \ell^2 h^2 / 12 \quad \dots 6.28A$$

$$\begin{aligned} S_0 &= \sum_{i=-a}^{i=+a} \frac{1}{\ell} x_i^2 \\ &= \frac{2}{\ell} h^2 \left[0 + 1^2 + 2^2 + \dots + \left(\frac{\ell-1}{2}\right)^2 \right] \\ &= \frac{\ell^2 - 1}{12} h^2 \quad \dots 6.28B \end{aligned}$$

The important point to note here is that in contrast to the previous case (equation 6.11) where the quantized signal power S_0 was greater than the input signal power S , in the present instance S_0 is less than S , though in both cases*

$$S_0 - S = h^2 / 12$$

Since the probability density function $p(x)$ of the amplitude of the signal is continuous in x in both these cases, it is reasonable to assume that there exists a particular probability distribution for which the input

*

If the input signal were characterized not by a rectangular distribution, but by a unimodal, U-shaped probability density function, not only would the quantized signal power S_0 be less than S , but $(S - S_0)$ would be greater than $h^2/12$ as well.

signal power and the quantized signal power are the same. Such a distribution would not have as abrupt terminations at its extremities as the rectangular distribution nor as smooth contact with the variate axis, for large values of the variate, as the one discussed in Section 6.1.1. Though $(S - S_0)$ vanishes for this distribution, it is obvious that quantization noise which is given by

$$N = \sum_{i=-a}^{+a} \int_{-h/2}^{+h/2} u^2 p(x_i + u) du \quad \dots 6.29$$

cannot be zero since h is finite. In other words, even if the number of quantization levels is large so that there is no correlation between the quantized signal and the error due to quantization, it can still be misleading to use the expression $(S - S_0)$ to assess the quantization noise power N_Q . It is, however, possible to assess N_Q in terms of S and S_0 if the cross-correlation between input and output is taken into account, as shown below.

If x represents samples of the input signal, y the output of the quantizer and n the number of samples, then the quantization noise power per sample is

$$N_Q = \frac{1}{n} \sum (x - y)^2 \quad \dots 6.30$$

the summation extending over ^{all} sample values.

But

$$\sum (x - y)^2 \equiv \sum x^2 + \sum y^2 - 2 \sum xy \quad \dots 6.31$$

Now, the covariance between the input and the output is

$$\text{cov}(x,y) = \frac{1}{n} \sum (x - \mu_x)(y - \mu_y)$$

where μ_x and μ_y are the mean values of x and y respectively

$$\begin{aligned} &= \frac{1}{n} \left(\sum xy - \mu_x \sum y - \mu_y \sum x + n \mu_x \mu_y \right) \\ &= \frac{1}{n} \left(\sum xy - n \mu_x \mu_y \right) \end{aligned} \quad \dots 6.32$$

since

$$\mu_x \sum y = \mu_y \sum x = n \mu_x \mu_y$$

Also, by definition,

$$\text{cov}(x,y) = \rho \sigma_x \sigma_y \quad \dots 6.33$$

where

ρ = correlation coefficient of x and y

σ_x^2 = variance of x and

σ_y^2 = variance of y

Therefore,

$$\begin{aligned} N_Q &= \frac{1}{n} \left(\sum x^2 + \sum y^2 - 2 \sum xy \right) \\ &= \frac{1}{n} (n \sigma_x^2 + n \mu_x^2 + n \sigma_y^2 + n \mu_y^2 - 2n \text{cov}(x,y) - 2n \mu_x \mu_y) \\ &= \sigma_x^2 + \sigma_y^2 - 2 \rho \sigma_x \sigma_y + (\mu_x - \mu_y)^2 \end{aligned} \quad \dots 6.34$$

If $\mu_x = \mu_y = 0$, then

$$N_Q = S + S_0 - 2\rho \sqrt{S \cdot S_0} \quad \dots 6.35$$

The usefulness of equation 6.35 is not confined to providing a means for assessing the magnitude of N_Q .

It also indicates the existence of a simple experimental procedure by which the cross-correlation between the input and the output of a quantizer or a similar non-linear sampled-data system can be determined. Since S , S_o and N_Q can be evaluated experimentally using long time-constant wattmeters and a unity-gain difference amplifier, the correlation is easily determined from the relation

$$\rho = \frac{S + S_o - N_Q}{2(S \cdot S_o)^{1/2}} \quad \dots 6.36$$

This formula is especially useful under conditions when it is difficult to compute ρ from purely theoretical considerations, due to either difficulties of mathematical computation or lack of knowledge of the detailed statistics of the original signal.

The significance of the correlation coefficient in the statistical analysis and design of communication or process control systems using quantizers is brought out in Section 7.

6.2 Minimization of Quantization Distortion

The magnitude of quantization distortion N_Q can be reduced by proper choice of the levels of quantization, x_i . Since the extreme quantization levels are already fixed by considerations of maximum permissible overloading error power, the optimum values of x_i , for minimum N_Q , can be obtained by solving the $(l-2)$ simultaneous equations:

$$\frac{\partial}{\partial x_i} \sum_{i=-a}^{+a} \int_{-h_i/2}^{+h_i/2} u^2 p(x_i + u) du = 0$$

$$i = (-a+1), (-a+2), \dots, -1, 0, +1, \dots, (a-2), (a-1)$$

...6.37

There are two sets of unknown quantities to be determined here: h_i and x_i . It is reasonable to assume, as a first approximation, that h_i is inversely proportional to the probability p_i associated with the i^{th} level, x_i . If the number of levels is large, p_i is directly proportional to $p(x_i)$. Therefore, the quantity h_i in equation 6.37 may be replaced by

$$h_i = \alpha h/p(x_i) \quad \dots 6.38$$

where α is a proportionality constant determined by the relation

$$\sum_i \frac{\alpha h}{p(x_i)} = 2A_0 \quad \dots 6.39$$

and

$$h = 2A_0/\ell \quad \dots 6.39A$$

Simultaneous solution of equations 6.37 and 6.39 gives, approximately, the desired optimum values of x_i . If L is large and the probability density function $p(x)$ is well behaved, the method of successive approximations can be used to obtain increased accuracy.

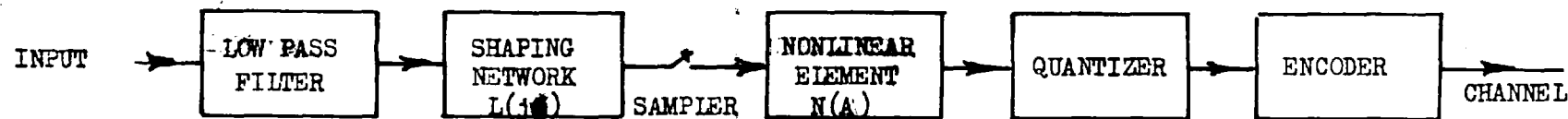
Though equation 6.37 can be solved, in any particular case, using the methods of numerical analyses the labour involved is usually very great. In practice, if the amplitude distribution has a pronounced "hump" or kurtosis⁽⁴³⁾ it has been shown that it is possible to achieve a reduction in quantization noise by using logarithmically-tapered quantum steps.^(16-19,44) The reduction in quantization distortion obtained by this means is, in general, less than what is possible with optimum spacing of levels, but the problems of design and of instrumentation of a

logarithmic quantizer are very much simpler. Since, however, these aspects have been studied in considerable detail in the literature, they will not be discussed here further.

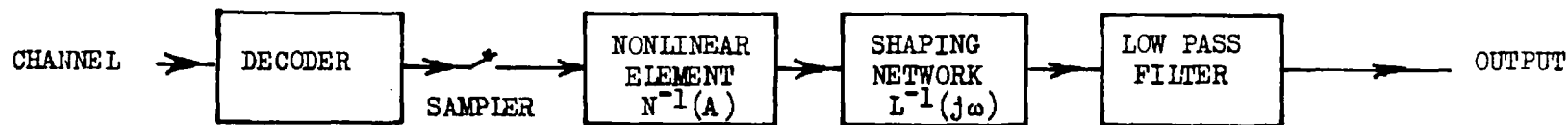
6.3 Effect of Quantization on Entropy of the Signal

If it is assumed that samples of the original signal have not been corrupted by external noise before they are quantized, then the effect of quantization is to reduce the entropy per sample from an infinite to a finite value. It can be easily shown that the magnitude of the final entropy of the quantized signal is a maximum if the probability of occurrence of the various levels at the output is the same for all levels and, further, if successive quantized samples are independent. Assuming that quantization is uniform, this can be achieved only if the input to the quantizer has a rectangular amplitude distribution and a flat power spectrum. But many types of naturally occurring signals do not possess these characteristics, an important instance being the class of speech signals which have an approximately gaussian amplitude distribution and a falling power spectrum.⁽⁴⁵⁻⁴⁷⁾ In such circumstances, the entropy of the output of the quantizer can be maximized by subjecting the signal to the following sequence of operations (Fig. 6.1)

- 1) A modified signal having a bandwidth-limited, flat spectrum is obtained by passing the original, unsampled signal through a spectrum-shaping network having a transfer function $L(j\omega)$
- 2) This modified signal is sampled at the corresponding Nyquist rate.⁽¹²⁾
- 3) The sampled signal is passed through a suitable non-linear network $N(A)$ which is such that the



TRANSMITTER



RECEIVER

FIG. 6.1. OPTIMIZATION OF CONVENTIONAL PCM SYSTEMS.

probability density function of amplitude of the output is rectangular.

4) The processed signal is now uniformly quantized.

The corresponding operations at the receiver to retrieve the original signal are:

- 1) The decoder output is sampled and applied to a non-linear network $N^{-1}(A)$ having an input-output relationship that is the inverse of that used at the transmitter.
- 2) This signal is passed through a shaping network having a transfer function $L^{-1}(j\omega)$ to compensate for the spectral shaping done at the transmitter.
- 3) This is now applied to a low pass filter having the same bandwidth as the original signal.

It may be noted that the shaping network $L(j\omega)$ and the sampler precede the non-linear element $N(A)$. The reason for this is that a continuous band-limited signal when passed through a non-linear network, is so distorted due to the introduction of high order harmonics of the components of the input signal that it requires, in general, a very much higher sampling frequency for exact specification. This difficulty is eliminated by subjecting the signal to any non-linear operation only after it is sampled, all the necessary linear operations on the signal being performed before sampling is effected.

An important result in information theory, on which the above technique for optimizing system performance by spectral shaping is based, is derived and discussed below.

Even if the transmission medium between the transmitter and the receiver is noiseless, the maximum rate of transmission of information of a PCM system is only $R_0 = (2W \log_2 \ell)$ ternits/second. From the information theory

point of view, the problem of optimum design of the system is therefore equivalent to that of reducing the entropy of the original signal from infinity to a value that is as nearly as possible equal to R_0 (but not greater than R_0), subject to the constraint that the performance of the system as specified by a suitably chosen "criterion of fidelity",⁽⁴⁾ is an optimum. On this basis, the PCM transmitter may be split up into two parts: the first, consisting of the optimizing linear and non-linear networks and the quantizer, which reduces the source entropy from infinity to R_0 relative to some fidelity criterion, and the second part consisting of the encoder, which codes the quantized signal into a form suitable for transmission through a channel having a capacity not greater than R_0 . The quantizer may now be considered to be a noisy channel perturbed by noise n_Q of magnitude N_Q . If x and y represent the input and the output respectively of the quantizer, then the entropy of the joint event (x, y) is

$$\begin{aligned} H(x, y) &= H(x) + H_x(y) \\ &= H(x) + H_x(x+n_Q), \text{ assuming that } x \text{ and } \\ &\quad n_Q \text{ are completely additive} \\ &= H(x) + H(n_Q), \end{aligned} \quad \dots 6.40$$

since quantization noise is independent of the input signal. Hence, the rate of transmission is

$$\begin{aligned} R &= H(x) + H(y) - H(x, y) \\ &= H(y) - H(n_Q) \end{aligned} \quad \dots 6.41$$

and attains a maximum value $R_{\max}$ when $H(y)$ is maximized with respect to the input probability distribution $p(x)$, $H(n_Q)$ being independent of x .

The problem of design now is that of finding the power spectra $P(f)$ of the input signal that would maximize the performance of the system on the basis of some relevant

criterion of fidelity, while the rate of transmission is held constant at $R_{\max}$, $R_{\max}$ being less than or equal to R_0 . A clue to the choice of a suitable criterion of fidelity is provided by the fact that the higher the signal-to-noise ratio S_0/N_Q at the output in any given band of the spectrum, the better the performance of the system in that particular band. Let $P(f)$ and $N_Q(f)$ represent the power spectra of the input signal and quantization noise respectively. If the total bandwidth of transmission, W , is now divided into an infinite number of narrow bands df , it would be reasonable to assume that the spectra of both signal and noise are white inside each of these bands. Let $P_1(f) df$ and $N_1(f) df$ denote the entropy power of signal and noise respectively in one of these narrow bands. Since it is assumed that the amplitude distribution of the input to the quantizer is rectangular and that quantization is uniform, it follows that

$$\frac{P_1(f) df}{N_1(f) df} = \frac{P(f) df}{N_Q(f) df} \quad \dots 6.42$$

The rate of transmission in a typical infinitesimal band is now given by

$$\begin{aligned} dR &= H(y) - H(n_Q) \\ &= df \log_e 2\pi e \sqrt{P_1(f) df + N_1(f) df} \\ &\quad - df \log_e 2\pi e \sqrt{N_1(f) df} \\ &= \log_e \left(1 + \frac{P_1(f)}{N_1(f)} \right) df \quad \dots 6.43 \end{aligned}$$

while the signal to noise ratio in that band is

$$\frac{S_0}{N_Q} = \frac{N_1(f) + P_1(f)}{N_1(f)} \quad \dots 6.44$$

The signal-to-noise ratio of the output will be a maximum

for a given value of $R_{\max}$ if

$$I = \int_0^W \left[\frac{N_1(f) + P_1(f)}{N_1(f)} + \lambda \log_e \frac{N_1(f) + P_1(f)}{N_1(f)} \right] df \quad \dots 6.45$$

where λ is a Lagrangian multiplier, is a maximum

Equating

$$\frac{dI}{dP_1(f)} = 0 \quad \dots 6.46$$

gives

$$\frac{N_1(f)}{P_1(f) + N_1(f)} = -\lambda, \text{ a constant} \quad \dots 6.47$$

Therefore

$$\frac{P_1(f)}{N_1(f)} = \frac{P(f)}{N_Q(f)} \text{ must be a constant}$$

for optimum performance. Since the spectrum of quantization noise is flat within the frequency range $0 - W$,⁽¹⁶⁾ it follows that the spectrum of the signal must also be flat.

It is important to note that the above technique of system optimization by adjustment of input signal spectrum is based upon the assumption that the statistics of quantization noise is completely independent of that of the input signal. This assumption is quite valid in almost all cases of practical interest, the correlation function of quantization noise being quite negligible unless the correlation function of the input is greater than about 0.99997, which is an extremely high figure, very unlikely to be met in practice.

It may be noted that the spectral reshaping performed at the receiver to reconstruct the original signal does not affect the fidelity of reproduction since signal and noise

are both affected to the same extent.

In the transmission of a typical signal, such as speech, the improvement in communication efficiency achieved by spectral reshaping is quite considerable. The power spectrum of speech⁽⁴⁵⁾ (Fig. 6.2) remains approximately constant at 40 db between 100 cps and 500 cps, then falls to 0 db at about 8000 cps, the "gain" varying approximately linearly with frequency. Now, if a signal confined to the bandwidth $0 - W$, having an entropy power P , is passed through a filter having the characteristic $Y(f)$, then the entropy power of the output ensemble is⁽⁴⁾

$$P_e = \frac{1}{W} \int_0^W \log_e |Y(f)|^2 df$$

Using this result, it can be shown that if the signal spectrum is confined to the range $0 - 8000$ cps, the entropy power of the signal is approximately

$$P_e = 2 \cdot \frac{8000 - 500}{8000} = \frac{15}{8} \quad \text{or} \quad P_e$$

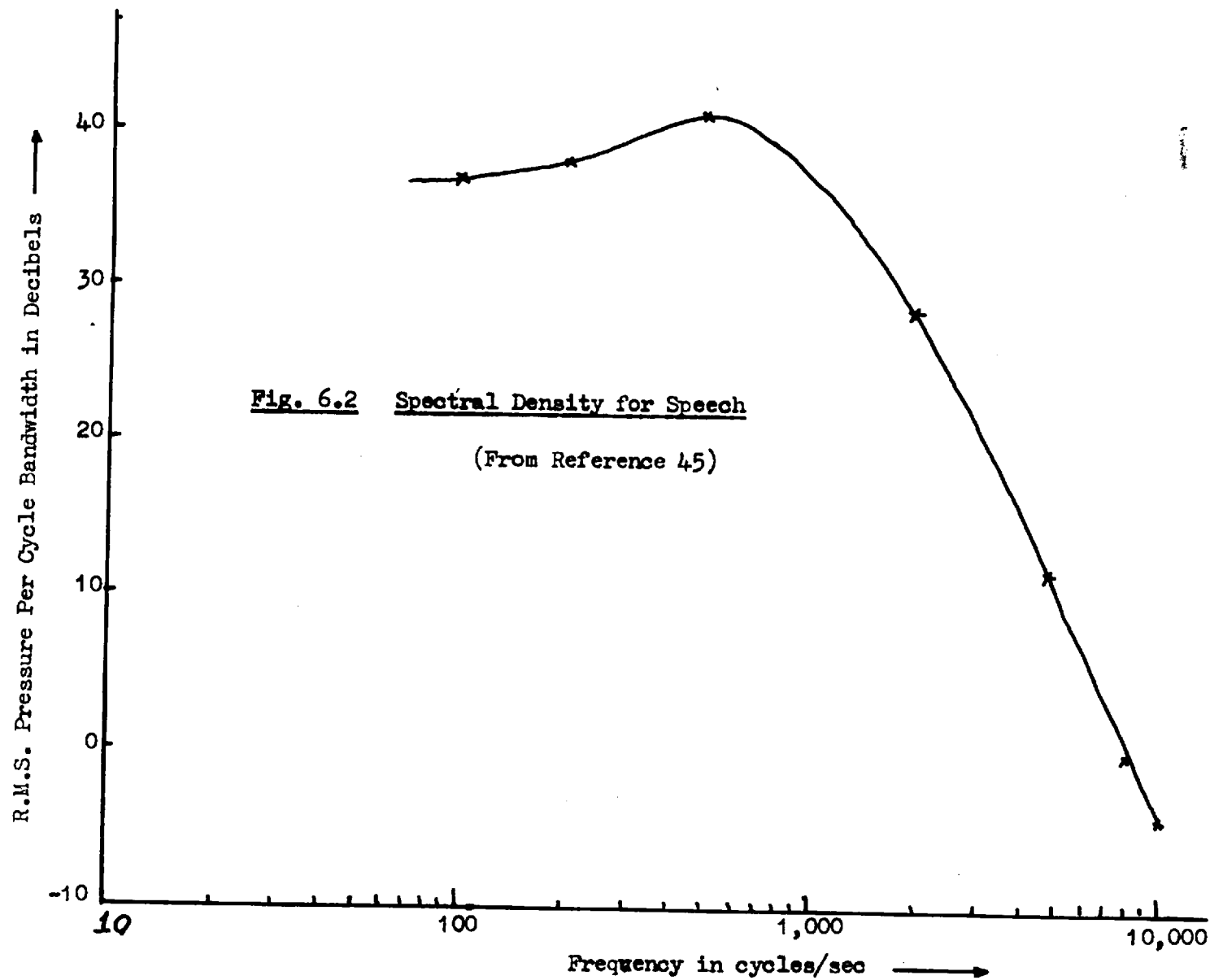
that of white noise confined to the same bandwidth. In other words, transmitting the speech signal without modifying its spectrum causes a deterioration in the performance of the system equivalent to a reduction of approximately (8.15) db in the received signal-to-noise ratio.

6.3.1 Choice of the Non-linear Characteristic $N(A)$

Any probability density function $p(x)$ that is defined within the range $(\alpha, \alpha + \beta)$ and is continuous in x can be transformed into the rectangular probability density function $q(y)$

$$q(y) = \begin{cases} \frac{1}{\beta}, & \alpha \leq y \leq \alpha + \beta \\ 0, & \alpha > y > \alpha + \beta \end{cases} \quad \dots 6.48$$

as shown below.



Equating the elements of probability density $p(x) dx$ and $q(y) dy$, we get

$$\begin{aligned} p(x) dx &= q(y) dy \\ &= q \cdot \left[\frac{dN(x)}{dx} \right] dy \end{aligned} \quad \dots 6.49$$

where

$y = N(x)$ is the desired functional relationship between the original and transformed variates.

Therefore,

$$dy = \frac{p(x) dx}{q \left[\frac{dN(x)}{dx} \right]}$$

Integrating,

$$\begin{aligned} y &= \int_{-\infty}^x \frac{p(x) dx}{q \left[\frac{dN(x)}{dx} \right]} \\ &= \beta \cdot P(x) \end{aligned} \quad \dots 6.50$$

since $q \left[\frac{dN(x)}{dx} \right]$ is independent of x within $(\alpha, \alpha + \beta)$.

where

$P(x)$ is the cumulative distribution function

$$P(x) = \int_{-\infty}^x p(x) dx \quad \dots 6.50A$$

= It is therefore possible to convert any ensemble having an arbitrary, continuous probability distribution into one having a rectangular distribution by passing the ensemble through an instantaneous non-linear element having the same characteristic as the cumulative probability distribution function of the input.

In any practical communication system such as PCM, where the peak power handling capacity of the system is

necessarily limited, there is considerable advantage in dealing with signals that have a rectangular distribution of amplitude rather than gaussian. Assuming, for instance, that the unprocessed signal has a gaussian amplitude distribution of variance σ^2 , and that the quantizer has a peak to peak range $(2A_0)$ of 8σ , the entropy power P'_G of the truncated gaussian signal is

$$P'_G = \frac{1}{2\pi e} e^{2H_G}$$

where H_G is the entropy of the signal.

Neglecting the reduction in entropy caused by limiting the range to 8σ , we get

$$P'_G = \frac{1}{2\pi e} e^{2\log_e(2\pi e \sigma^2)^{1/2}}$$

The entropy power of an ensemble having a rectangular amplitude distribution confined to the same range $(-4\sigma, +4\sigma)$ is

$$P'_R = \frac{1}{2\pi e} e^{2H_R}$$

where H_R is the entropy of the rectangular distribution

$$\begin{aligned} P'_R &= \frac{1}{2\pi e} e^{2 \int_{-4\sigma}^{+4\sigma} \frac{1}{8\sigma} \log_e\left(\frac{1}{8\sigma}\right) dx} \\ &= \frac{1}{2\pi e} e^{2\log_e(8\sigma)} \\ &= 3.75 \sigma^2 \end{aligned}$$

Transforming the input distribution from truncated gaussian to rectangular, therefore, makes it possible, in this case,

to increase the entropy power of the transmitted signal by approximately 5.74 db . Performance of the system is therefore improved since quantization noise ^{power} is constant, spacing of levels being the same in both cases. The improvement achieved is greater or less than 5.74 db depending upon whether the quantization range is greater or less than 8σ .

6.3.2 Choice of the Linear Shaping Network $L(j\omega)$

Let $G_{ii}(\omega)$ be the power spectrum of the input signal. If $L(j\omega)$ is the transfer function of the shaping network, the power spectrum of the output is

$$G_{oo}(\omega) = G_{ii}(\omega) |L(j\omega)|^2 \quad \dots 6.51$$

Since it is desired that the output spectrum be white,

$$|L(j\omega)|^2 = \frac{1}{G_{ii}(\omega)} \quad \dots 6.52$$

Now $G_{ii}(\omega)$ can be approximated to any required degree of accuracy by a rational function $G(\omega)$. It therefore follows that the input power spectrum can be expressed as the ratio of two polynomials, $M(\omega)$ and $N(\omega)$, in ω . Since the power spectrum function is a real-valued, even, non-negative function of ω for real values of ω , $M(\omega)$ and $N(\omega)$ are polynomials in ω^2 with real coefficients. Therefore

$$\begin{aligned} G_{ii}(\omega) &\simeq G(\omega) \\ &= \frac{M(\omega)}{N(\omega)} \\ &= \frac{m_0 + m_1(\omega^2) + \dots + m_a(\omega^{2a})}{n_0 + n_1(\omega^2) + \dots + n_b(\omega^{2b})} \quad \dots 6.53 \end{aligned}$$

$M(\omega)$ and $N(\omega)$ being of degree $2a$ and $2b$ respectively.

Since the coefficients $m_0, \dots, m_a; n_0, \dots, n_b$ are real, the poles and zeros of $M(\omega)$ and $N(\omega)$ are either real or occur in complex conjugate pairs. Hence, $M(\omega)$ and $N(\omega)$ can each be expressed as the product of two conjugate factors

$$\begin{aligned} M(\omega) &= M_L(j\omega) \cdot M_L^*(j\omega) \\ N(\omega) &= N_L(j\omega) \cdot N_L^*(j\omega) \end{aligned}$$

such that $M_L(j\omega)$ and $N_L(j\omega)$ have no poles or zeros in the right half of the complex ω plane.

Therefore

$$\left| L(j\omega) \right|^2 = \left| \frac{N_L(j\omega)}{M_L(j\omega)} \right|^2$$

that is

$$\left| L(j\omega) \right| = \frac{N_L(j\omega)}{M_L(j\omega)}$$

Though the output of $L(j\omega)$ would yield a signal having a flat spectrum, this does not necessarily mean that the spectrum of the input to the quantizer would be white since the non-linear element $N(A)$ also distorts the signal spectrum to a certain extent. For instance, uniform quantization applied to a signal that has been passed through a logarithmic element introduces discrimination against high frequency components of the signal. In such instances, the effect of the non-linear element on signal spectra should be taken into account in the determination of $L(j\omega)$.

7. Extension of the Philosophy of PCM

Though the performance of PCM systems can be improved by spectral shaping, this method suffers from the following disadvantages:

(a) It is assumed that an ideal low pass filter is available for reconstruction of the original signal. But, such a filter is not physically realizable. In practice, interpolation between successive samples has to be achieved by some other means such as passing the samples through a holding circuit and then through another network that partially compensates for the 'aperture effect' introduced by the holding circuit⁽⁴⁸⁾. This is equivalent to the use of a low pass filter with a finite transition region between the transmission and the attenuation bands, the attenuation in the attenuation band being also finite. Since the attenuation presented by the system to unwanted frequencies is finite, the fidelity of transmission of the system is also limited,⁽⁴⁹⁾.

(b) It is, in general, necessary to keep the transmission delay through the system, caused by the use of low pass filters at the transmitter (before sampling) and at the receiver (after decoding), within reasonable proportions. [This is especially important in telemetry and control systems applications.] This makes it necessary to use a sampling frequency considerably in excess of twice the bandwidth of the sampled signal⁽⁴⁹⁾. Spectral shaping techniques cannot take advantage of the appreciable correlation that may exist between successive samples under these conditions.

(c) Spectral shaping cannot make use of second- and higher-order correlations that may be present in the signal.

In the analysis of Section 6, the techniques employed to derive the relevant stochastic model for the actual signal were based on the principle of "approximation in the vicinity of a point"; that is, it is assumed that the characteristics of a curve inside a given interval can be specified as a function of the derivatives of the curve at a point within the interval. There exists another method of approximation, that is more relevant to sampled-data systems, in which the given function is represented by a polynomial whose values coincide with those of the given function at a sufficient number of specified points in time. It is possible not only to derive a mathematical model for a stochastic process on this assumption but also to devise a practical communication system based upon this model, as shown below.

7.1. Description of Suggested System

Consider the system shown in Fig. 7.1.. The input signal $p(t)$ is applied to one terminal of a difference amplifier, while the predicted value $r(t)$ of the signal as computed by the extrapolator J , is applied to the other terminal. The difference signal $\epsilon(t)$ is sampled and then quantized by the quantizer Q before it is applied to the input of the extrapolator J . The sampled and quantized signal $q(t)$ is simultaneously applied to the encoder which then transmits the coded equivalent of $q(t)$ to the receiver through the given noisy channel. At the receiver, the coded signal is decoded before it is sampled and then applied to the input of a network J' that is identical to the extrapolator J at the transmitter. It may be noted (see also Fig. 7.2) that the receiver output $r'(t)$ is effectively the same as the output $r(t)$ of the extrapolator J , apart from the transmission delay caused by the channel and the

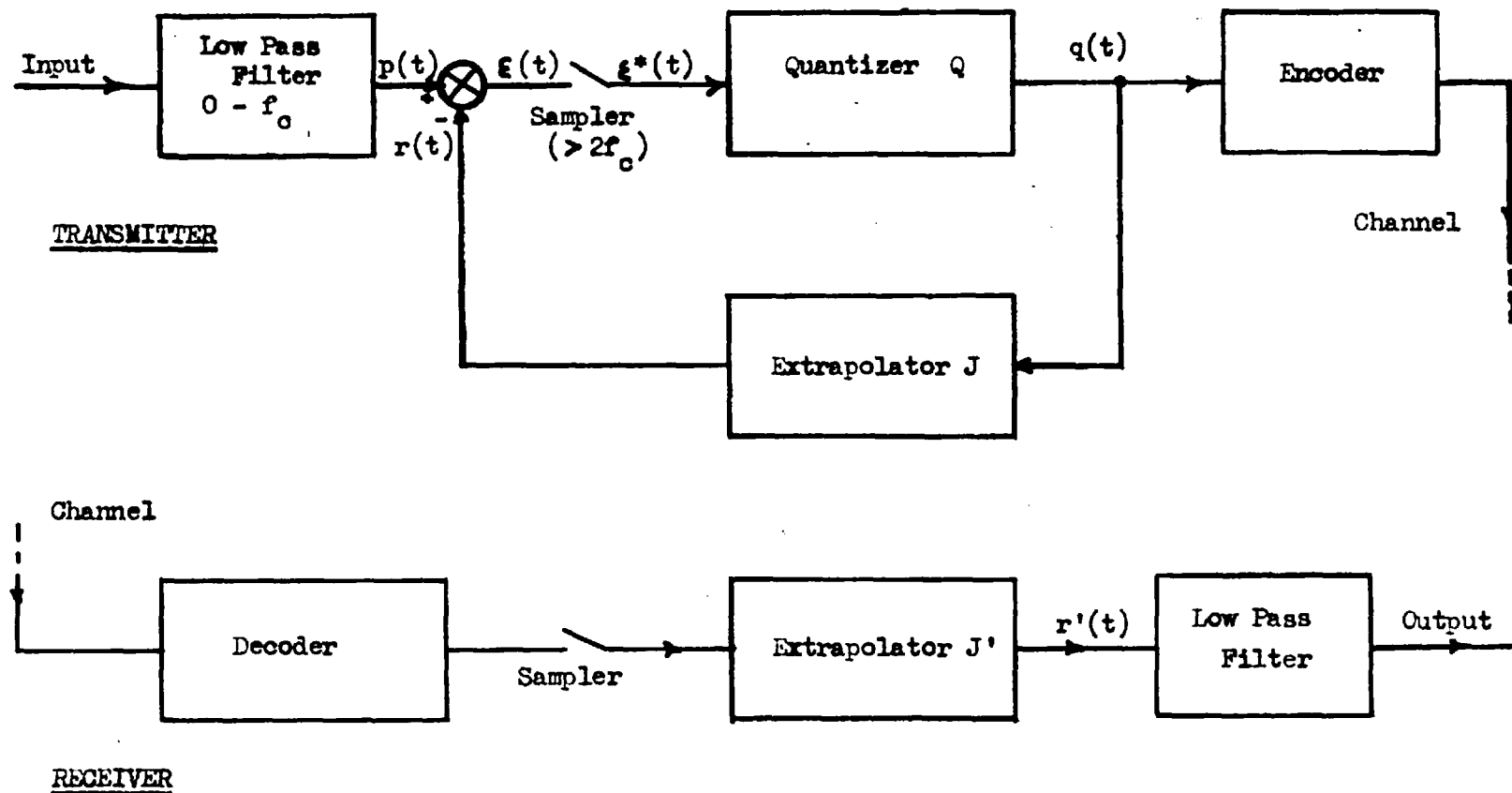


Fig. 7.1 Modified P.C.M. System

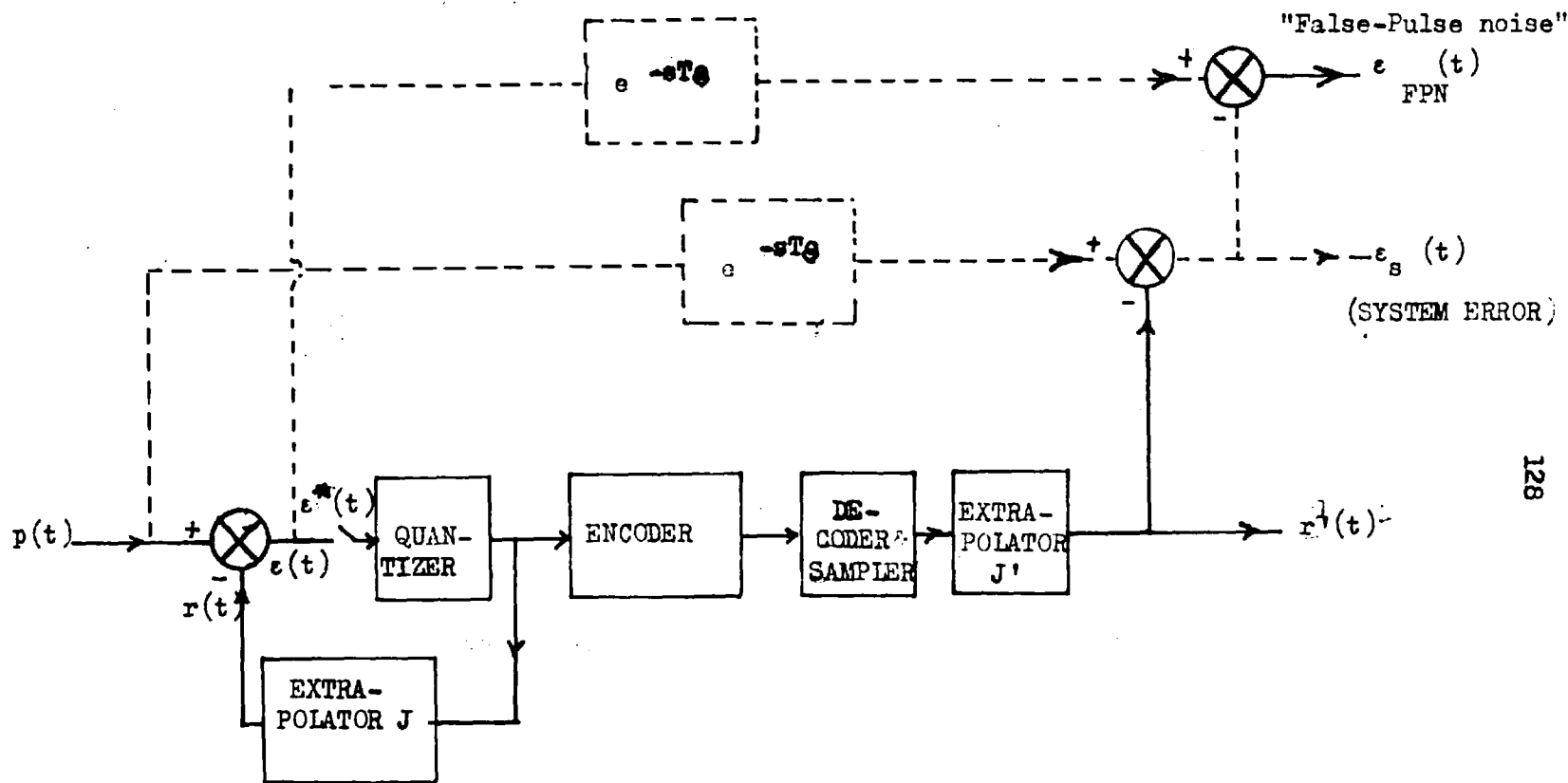


FIG. 7.2. ERROR DIAGRAM FOR FIG. 7. 1

decoding circuits, if it can be assumed that no decoding errors have been caused due to the presence of noise in the transmission channel*. Further, if the criterion of fidelity for the system is chosen to be minimization of the mean-square system error $\mathcal{E}_s$, given by the difference between the output signal $r'(t)$ and the input signal $p(t)$ delayed by the dead-time lag (T_o) of the system

$$\mathcal{E}_s = |r'(t) - p(t+T_o)| \quad \dots 7.1.$$

then, it can be seen that system optimization is equivalent to minimization of the mean square of the error signal $\mathcal{E}(t)$

$$\mathcal{E}(t) = p(t) - r(t) \quad \dots 7.2.$$

Now, the greater the accuracy of extrapolation, the smaller the error signal $\mathcal{E}(t)$. If the number of quantization steps (ℓ) is fixed, it would therefore be possible to specify the signal to a greater degree of precision than would be possible if the input signal $p(t)$ itself were quantized. [See Section 7.4]. A high fidelity communication system can therefore be designed if an accurate extrapolator can be instrumented. Problems associated with physical realizability of extrapolators and their instrumentation, as well as the optimization of extrapolators with respect to the statistical structure of the input-signal, are discussed in the following Sections.

* It is easy to see that the signal deterioration caused in this case by channel-noise interference (that is, false pulse noise, $\mathcal{E}_{FPN}$) is of a different pattern from that produced in ordinary PCM systems. Since now the effect of false-pulse noise is cumulative, it will be assumed in further discussions that there is provision for correction of errors in the transmitted signal, either through use of error-correcting codes for transmission of the coded signal, or alternatively, through use of an additional channel between the transmitter and the receiver for interrogating the transmitter either periodically for routine checks or on detection of an error in the received signal. *group, assuming* error-detecting but not error-correcting codes are used. (ref: 50, 51).

7.2. Extrapolation of Sampled Signals

The simplest method of estimating the future value of a function $f(t)$ at time $t=(nT+\tau)$, given only the values of $f(t)$ at the discrete instants of time

$$t = (n-k)T, \quad 0 < k < n$$

where T is the sampling period, is to assume that $f(t)$ remains constant at $f(nT)$ in the interval (nT) to $(nT+\tau)$. If, however, the signal is slowly varying, a better estimate of the signal at time $(nT+\tau)$ can be obtained by employing linear extrapolation in which knowledge of the magnitude of the immediately previous two samples

$$f_n = f(nT)$$

$$\text{and} \quad f_{n-1} = f(\overline{n-1} T)$$

are made use of to compute the average rate of change of the signal

$$\frac{f(nT) - f(\overline{n-1} T)}{T}$$

in the interval $(n-1)T$ to (nT) , it being then assumed that the rate of change remains unchanged in the interval (nT) to $(nT+\tau)$. The predicted value of $f(nT+\tau)$ is now given by

$$\begin{aligned} f_P(t) &= f_P(nT+\tau) \\ &= f(nT) + \frac{f(nT) - f(\overline{n-1} T)}{T} (t-nT) \\ &= f_n + \frac{f_n - f_{n-1}}{T} \cdot \tau \quad \dots 7.3 \end{aligned}$$

An even better estimate of $f(nT+\tau)$ could have been obtained

if allowance had been made for possible curvature of the function $f(t)$. Knowledge of $f(nT)$ and $f(\overline{n-1} T)$ alone is insufficient to specify whether $f(t)$ is convex or concave in the neighbourhood of $f(nT)$. If, however, $f_{n-2}=f(\overline{n-2} T)$ is also known, then it becomes possible to take the curvature of the signal into account using the quadratic function in t

$$\begin{aligned} f_P(t) &= f_P(nT+\tau) \\ &= f_n + \frac{f_n - f_{n-1}}{T} (t-nT) \\ &\quad + \alpha (t-\overline{n-1} T)(t-nT) \quad \dots 7.4 \end{aligned}$$

as the basis of the extrapolation scheme. The structure of the quadratic term of eqn. 7.4. is such that it vanishes when $t=(n-1)T$ and $t=nT$. This ensures that the extrapolated curve has the correct magnitude $f(\overline{n-1} T)$ and $f(nT)$ at $t=(n-1)T$ and $t=nT$ respectively. The constant α is now chosen to make the function $f_P(t)$ have the correct magnitude f_{n-2} at $t=(n-2)T$. Putting $t=(n-2)T$, we get

$$f_{n-2} = f_n + \frac{f_n - f_{n-1}}{T} (-2T) + \alpha (-T)(-2T) \quad \dots 7.5$$

giving

$$\alpha = \frac{f_n - 2 f_{n-1} + f_{n-2}}{2T^2} \quad \dots 7.6$$

Therefore,

$$\begin{aligned} f_P(nT+\tau) &= f_n + \frac{f_n - f_{n-1}}{T} \tau + \frac{f_n - 2 f_{n-1} + f_{n-2}}{T^2} \underbrace{(T+\tau)\tau}_2 \\ &\quad \dots 7.7. \end{aligned}$$

Further increase in the accuracy of extrapolation can be obtained by adding a cubic term

$$\beta (t - \overline{n-2} T) (t - \overline{n-1} T) (t - nT)$$

to eqn. 7.4. and then solving for β by making the extrapolated curve $f_p(nT+\tau)$ have the correct magnitude f_{n-3} , f_{n-2} , f_{n-1} and f_n at the four sampling instants $t = (n-3)T$, $(n-2)T$, $(n-1)T$ and nT , respectively. The technique can be indefinitely repeated to give the general extrapolation formula (sometimes called "Newton's Formula for Backward Interpolation" (52):

$$\begin{aligned}
 f_P(nT+\tau) = & f_n + \frac{\delta_1 f_n}{T} \tau \\
 & + \frac{\delta_2 f_n}{T^2} \frac{(T+\tau)\tau}{\underline{12}} \\
 & + \text{---} \\
 & + \frac{\delta_k f_n}{T^k} \frac{(\overline{k-1} T+\tau)(\overline{k-2} T+\tau)\text{---}(T+\tau)\tau}{\underline{k}} \\
 & + \text{---} \dots 7.8
 \end{aligned}$$

where

$$\delta_1 f_n = f_n - f_{n-1}$$

$$\begin{aligned} \delta_2 f_n &= \delta_1 f_n - \delta_1 f_{n-1} \\ &= f_n - 2f_{n-1} + f_{n-2} \end{aligned}$$

$$\begin{aligned} \delta_3 f_n &= \delta_2 f_n - \delta_2 f_{n-1} \\ &= f_n - 3f_{n-1} + 3f_{n-2} - f_{n-3} \end{aligned}$$

$$\delta_{k f_n} = \delta_{k-1} f_n - \delta_{k-1} f_{n-1} \quad \dots 7.8A$$

 $\delta_k f_n$ being the k^{th} order back-difference of $f(nT)$

If $\tau = T$,

then

$$f_{p(n+1)} = f_n + \delta_1 f_n + \delta_2 f_n + \dots + \delta_k f_n + \dots \quad \dots 7.9$$

The number of past samples that have to be taken into account in extrapolating the given function one interval into the future is therefore greater by one than the order of the highest-order back - difference occurring in the expression defining the given function.

By analogy with the terminology of numerical mathematical analysis, a system that is capable of extrapolating making use of back-difference of order k , would be called hereafter an extrapolator of k^{th} order.

7.2.1. Accuracy of Extrapolation

Putting $nT=t$ in equation 7.8., we get

$$\begin{aligned} f_p(t+\tau) = f(t) &+ \frac{f(t) - f(t-T)}{T} \tau \\ &+ \frac{f(t)-2f(t-T)+f(t-2T)}{T^2} \cdot \frac{(T+\tau)\tau}{2} \\ &+ \dots \text{ to } \infty \end{aligned} \quad \dots 7.10.$$

Taking the Laplace transform of both sides we get

$$\begin{aligned} \mathcal{L} f_p(t+\tau) &= F_p(s) \\ &= f(s) + \frac{f(s)-f(s)e^{-sT}}{T} \tau \\ &\quad + \frac{f(s)-2f(s)e^{-sT}+f(s)e^{-2sT}}{T^2} \frac{(T+\tau)\tau}{2} \\ &\quad + \dots \text{ to } \infty \\ &= f(s) \left[1 + \frac{1-e^{-sT}}{T} \tau + \frac{(1-e^{-sT})^2}{T^2} \frac{(T+\tau)\tau}{2} + \dots \text{ to } \infty \right] \\ &= f(s) \left[1 - (1 - e^{-sT}) \right]^{-\tau/T} \\ &= \left[f(s) \right] \left[e^{s\tau} \right] \\ &= \mathcal{L} f(t+\tau) \end{aligned} \quad \dots 7.11$$

Since $(e^{s\tau})$ is the transfer function of an ideal predictor with prediction time of τ units, it follows that the extrapolated function $f_p(t+\tau)$, as given by equation 7.8, provides exact prediction, if an infinite number of back differences (if they exist) are taken into account.

7.3. Synthesis of the Extrapolator

A k^{th} order extrapolator can accurately predict the future values of any function whose back-differences of order greater than k , are zero, after an initial settling period. For instance, since a polynomial of order p cannot *have* back-differences of order greater than p , its future values can be accurately predicted by an extrapolator of order greater than (or equal to) p , after an initial settling time of p periods. If, however, the order of the extrapolator is less than that of the polynomial, the settling time of the system becomes infinite and the difference $\epsilon(t)$ between the actual signal $p(t)$ and the predicted signal $r(t)$ is always finite, except possibly at a finite number of discrete instants in time, specified by a null class of probability zero. For instance, if a zero-order extrapolator which can reproduce only a constant or a step input perfectly, or a first-order extrapolator which can exactly reproduce only signals of no greater complexity than a ramp function

$$y = a + bt, \quad \dots 7.12$$

is used to extrapolate the future values of a parabolic function

$$z = c + dt + e t^2 \quad \dots 7.13$$

it is easy to see that the error term would be "almost never" equal to zero for any value of t . While it would not be

possible to exactly reproduce the original input using an extrapolator of lower order than the input signal, it is possible to reduce the magnitude of the error term $e(t)$ by making use of a feedback configuration similar to that suggested by Porter and Stoneman⁽⁵³⁾ for use in sampled servomechanisms, in which the magnitude of the error term itself is used to control the output of the extrapolator. The actual implementation of such a system is discussed below.

Let the extrapolator be designed such that its output $r(t)$ is equal to $P_n, P_{n-1}, \dots, P_{n-k}$ at the sampling instants $(nT), (n-1T), \dots, (n-kT)$ respectively, where P_{n-k} is the true amplitude of the input signal $p(t)$ at time $(n-k)T$. The output of the extrapolator can be expressed as k^{th} order polynomial:

$$r(t) = r_0 + r_1(t-nT) + r_2(t-nT)^2 + \dots + r_k(t-nT)^k \quad \dots 7.14$$

If either the order of the extrapolator is less than that of the signal or if the signal is not purely deterministic, at the instant $t=(n+1)T$ there would be an extrapolation error $e(n+1T)$ between the actual signal input $p(t)$ and the predicted signal $r(t)$

$$e(n+1T) = p(n+1T) - r(n+1T) \quad \dots 7.15$$

For the present, it will be assumed that the optimum mode of operation of the extrapolator is such that when the error term $e(n+1T)$ is detected, the extrapolator generates a new polynomial $s(t)$

$$s(t) = s_0 + s_1(t - \overline{n+1T}) + s_2(t - \overline{n+1T})^2 + \dots + s_k(t - \overline{n+1T})^k \quad \dots 7.16$$

that passes through the most recent $(k+1)$ values of the

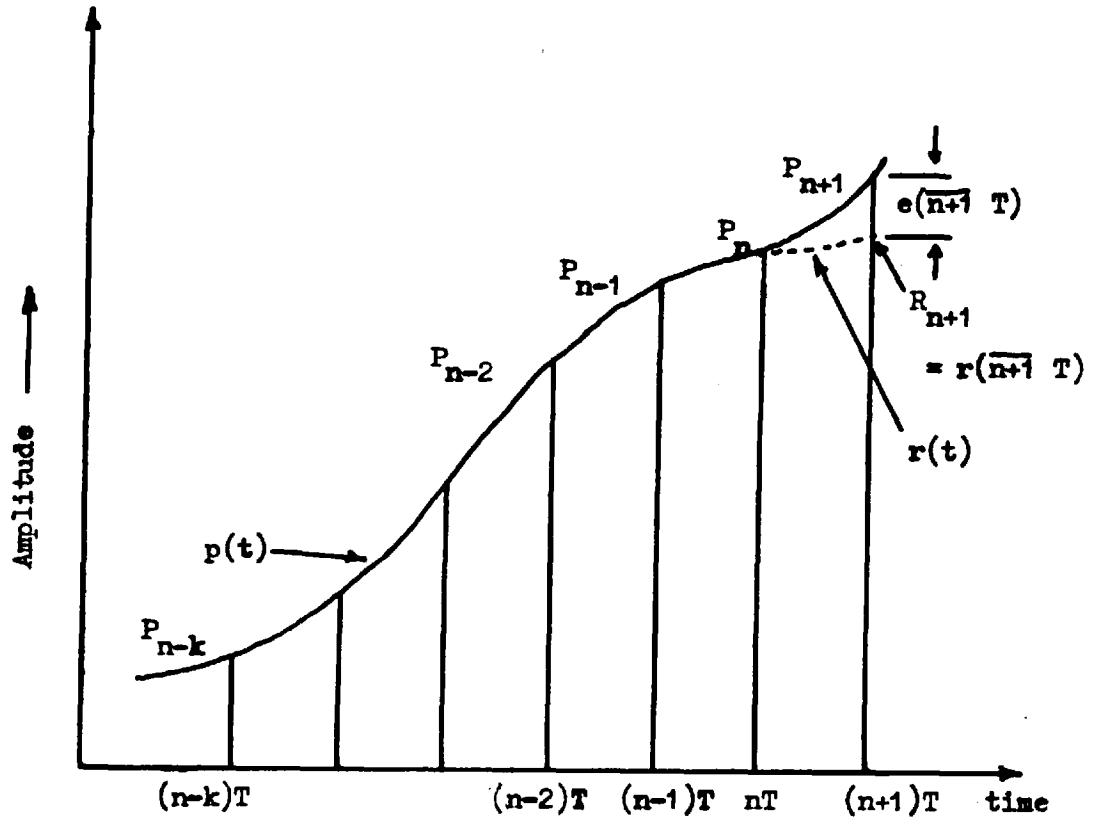


Fig. 7.3 Extrapolation of Sampled Signals.

input $p(t)$ namely

$$p(\overline{n+1}T), p(nT), p(\overline{n-1}T), \dots p(\overline{n+1-k}T),$$

k being the order of the extrapolator. In other words, the extrapolator takes account of the most recent discrepancy between the actual value of the signal and the predicted value, to make both magnitudes equal as well as to modify the extrapolation constants (namely, the coefficients of the generated polynomial) such that further extrapolation is performed on the basis of the most recent data available regarding the input signal. This can be done by using the error term $e(\overline{n+1}T)$ to generate a correction function $c(t)$

$$\begin{aligned} c(t) = & c_0 + c_1(t - \overline{n+1}T) + c_2(t - \overline{n+1}T)^2 \\ & + \dots + c_k(t - \overline{n+1}T)^k \end{aligned} \quad \dots 7.17$$

which is then added to $r(t)$ such that the new output is equal to $p(t)$ at $t = (n+1)T$. Since, now,

$$s(t) = r(t) + c(t) \quad \dots 7.18$$

it follows that

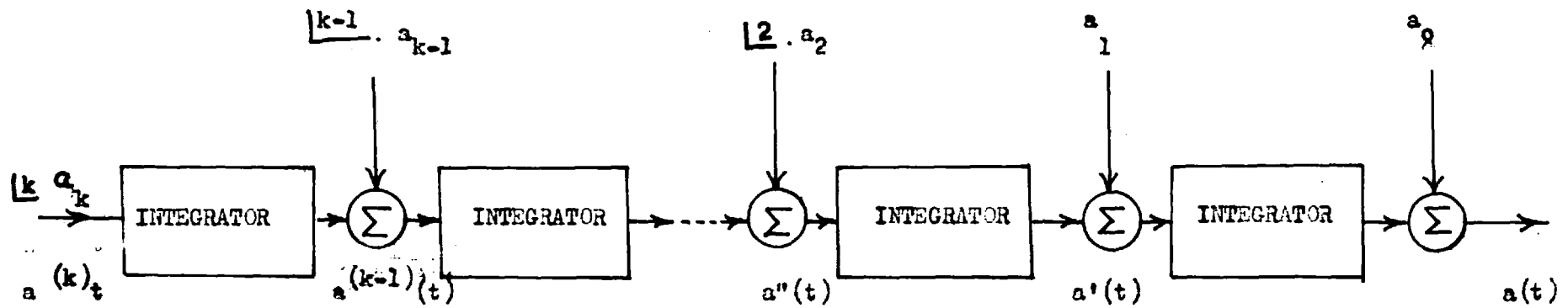
$$\left. \begin{aligned} c(t) &= e(\overline{n+1}T), \text{ at } t = \overline{n+1}T \\ &= 0, \text{ at } t = mT, \end{aligned} \right\} \quad \dots 7.19$$

where

$$(n - k) < m \leq n \quad \dots 7.19A$$

Equation 7.17 can therefore be expressed in the form

$$c(t) = e(\overline{n+1}T) \left[\left(1 + \frac{t - \overline{n+1}T}{T}\right) \left(1 + \frac{t - \overline{n+1}T}{2T}\right) \dots \left(1 + \frac{t - \overline{n+1}T}{kT}\right) \right] \quad \dots 7.20$$



$$a(t) = a_0 + a_1 t + a_2 t^2 + \dots + a_k t^k$$

FIG. 7.4: GENERATION OF POLYNOMIALS.

$$\begin{aligned}
s(T+\tau) &= r(T) + c(T+\tau) \\
&= \left[c(\tau) \right]_{\tau=T} + e(T+\tau) \\
&= e(0) (a_0 + a_1 T + \dots + a_k T^k) \\
&\quad + e(T) (a_0 + a_1 \tau + \dots + a_k \tau^k) \quad \dots 7.23B \\
\hline
s(nT+\tau) &= e(0)(a_0 + a_1 T + \dots + a_k T^k) \\
&\quad + e(T)(a_0 + a_1 T + \dots + a_k T^k) \\
&\quad + \dots \\
&\quad + e(n-1T)(a_0 + a_1 T + \dots + a_k T^k) \\
&\quad + e(nT) (a_0 + a_1 \tau + \dots + a_k \tau^k) \quad \dots 7.23C
\end{aligned}$$

Therefore

$$r(nT) = (a_0 + a_1 T + \dots + a_k T^k) \sum_{m=0}^{n-1} e(mT) \quad \dots 7.24$$

It is obvious from equation 7.24 that if the extrapolator J is to generate $r(nT)$ at time $t=nT$, the input to the polynomial generator $(a_0 + a_1 T + \dots + a_k T^k)$ should be the commulative sum of all error inputs $e(mT)$ previous to $t= nT$. The extrapolator proper, therefore, consists of two parts ; namely, the error integrator I and the polynomial generator R , as shown in Fig. 7.5.

The complete system configurations, at the transmitter, of the modified PCM systems using zero, first, second and third order extrapolators are shown in Figures 7.6, 7.7, 7.8 and 7.9 respectively,

The effect of introducing the quantizer within the loop is considered in Sections 7.4 and 7.7.



IS EQUIVALENT TO

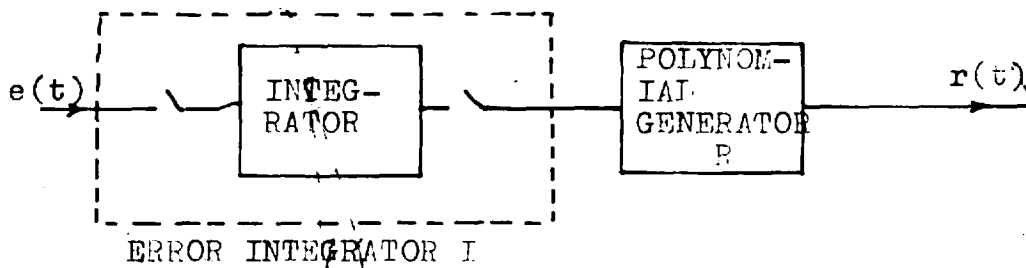


FIG: 7.5: SYNTHESIS OF THE EXTRAPOLATOR.

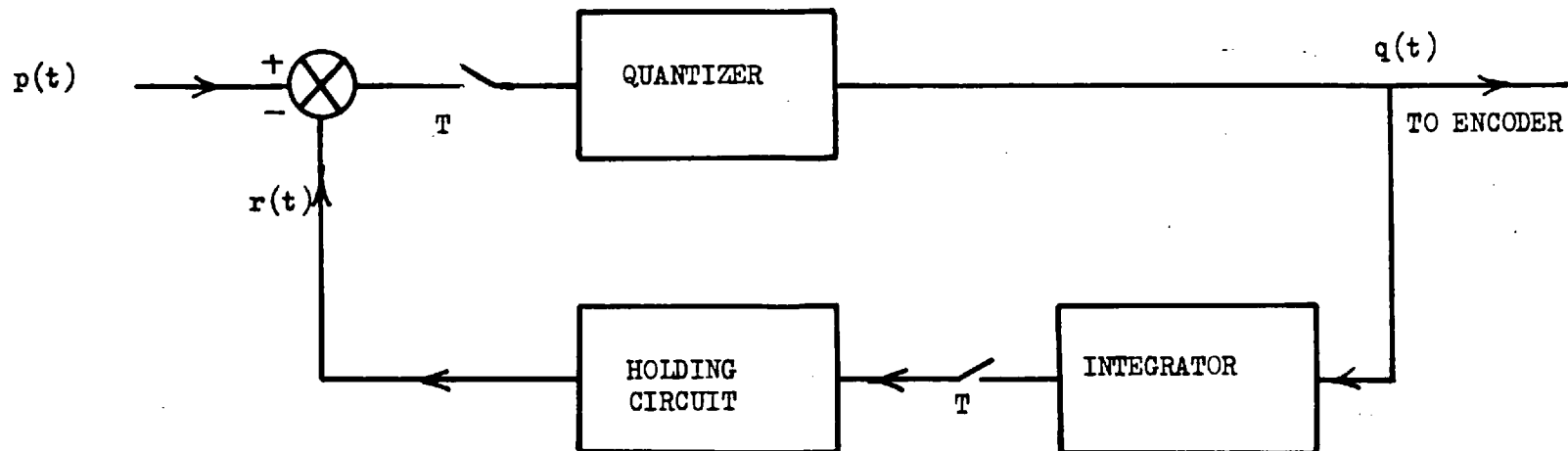


FIG. 7.6 MODIFIED PCM SYSTEM:- EXTRAPOLATION: ZERO ORDER

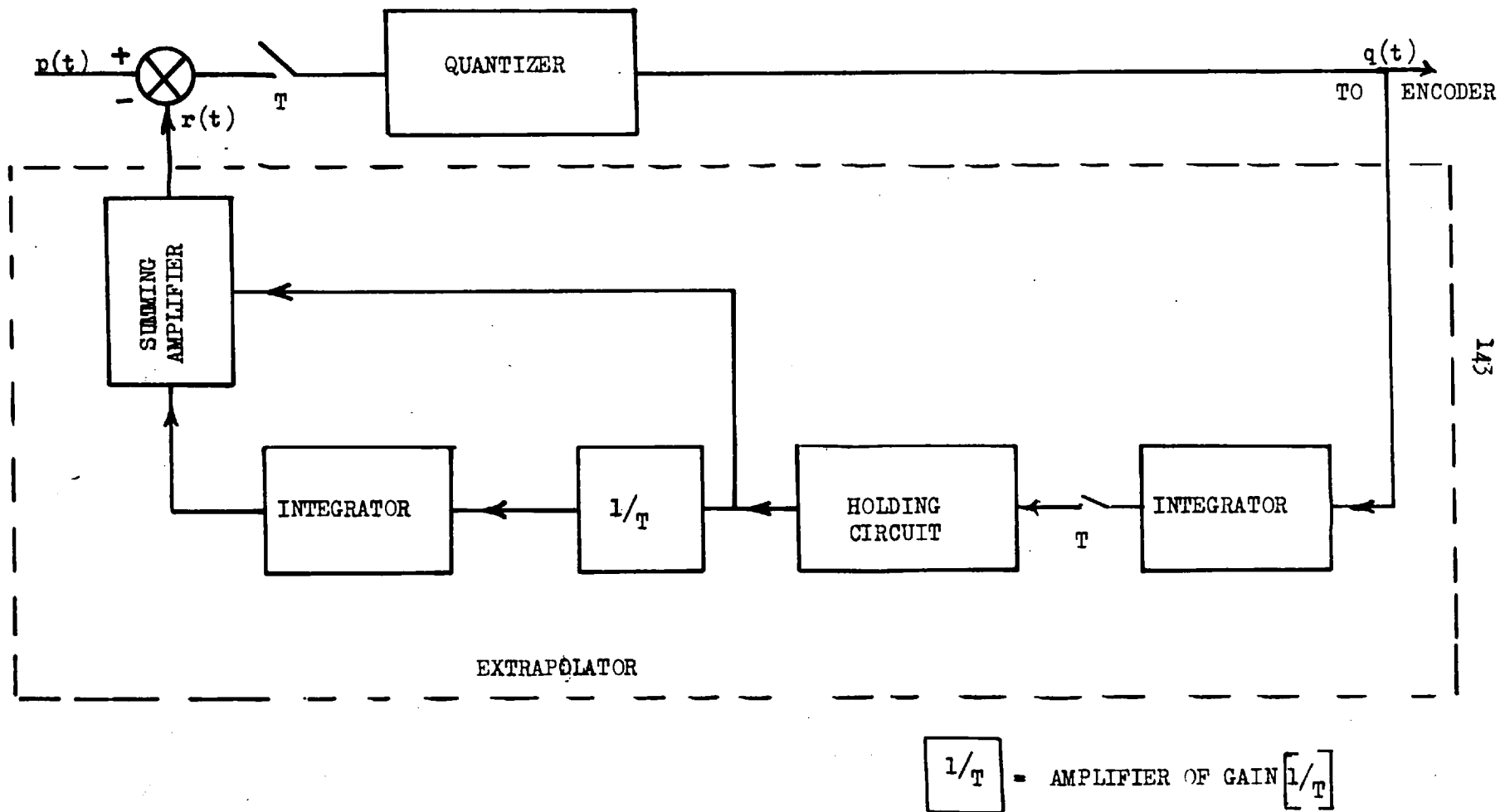
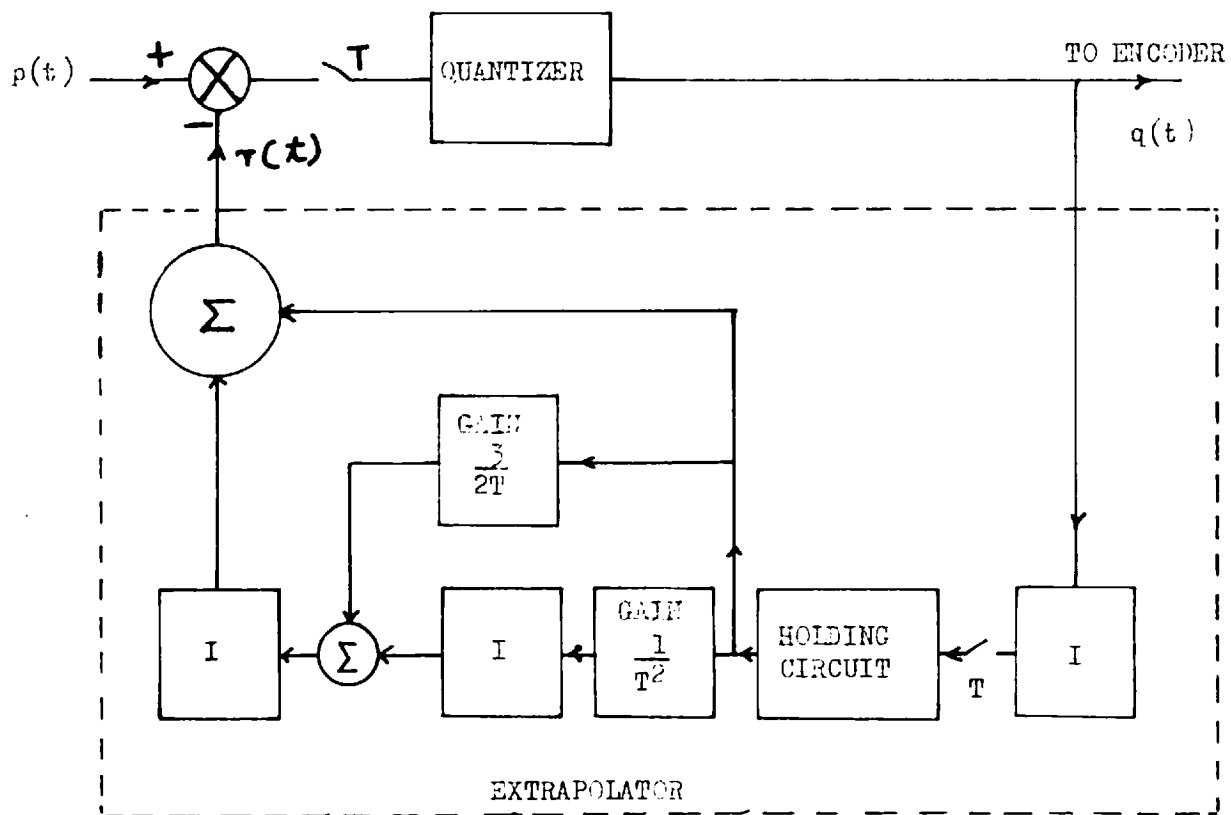
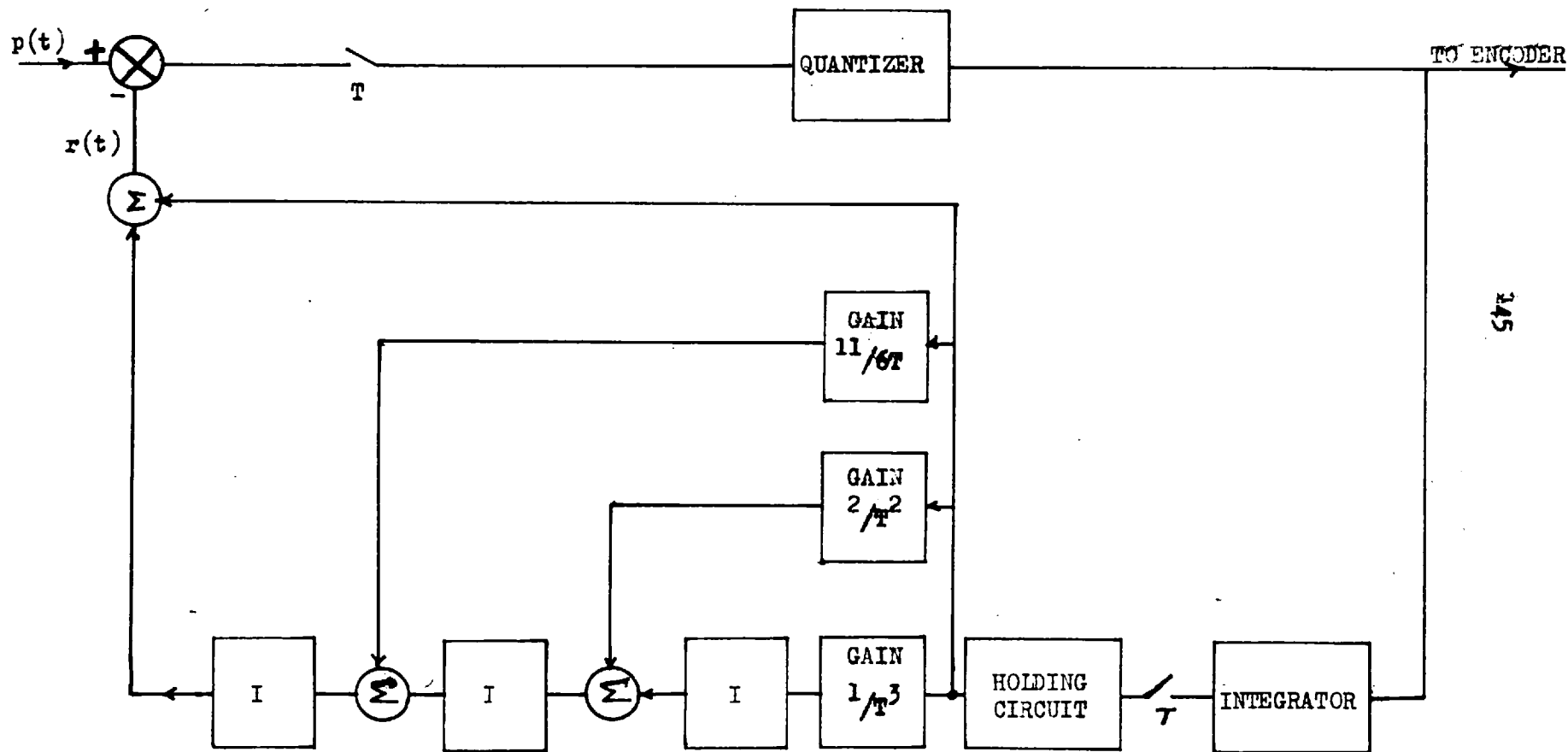


FIG. 7.7. MODIFIED PCM SYSTEM:—EXTRAPOLATION: FIRST ORDER.



$\boxed{I}$ = INTEGRATOR

FIG. 7.8. MODIFIED PCM SYSTEM:—EXTRAPOLATION: SECOND ORDER.



245

I = INTEGRATOR

FIG. 7.9. MODIFIED PCM SYSTEM-EXTRAPOLATION: THIRD ORDER.

7.4. Representation of Stochastic Signals

The conditions of validity of the assumption made in Section 7.3 that it is permissible to represent a signal at least locally (that is, over time intervals of restricted duration) by a polynomial of finite order, are discussed in the present Section.

7.4.1. Least Square Estimation of Sample Magnitudes

Let the joint probability density function ^{of} the variates x_1 and x_2 be

$$\begin{aligned} x_1 &= x(t) \\ x_2 &= x(t-\tau) \end{aligned} \quad \dots 7.25$$

be $p(x_1, x_2)$. Let the best estimate in the least square sense, of x_1 for a given value of x_2 be

$$x_1 = h(x_2) \quad \dots 7.26$$

Therefore, by definition,

$$\begin{aligned} &\text{Expectation of } [x_1 - h(x_2)]^2 \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} [x_1 - h(x_2)]^2 p(x_1, x_2) dx_1 dx_2 \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} [x_1 - h(x_2)]^2 p(x_2) p(x_1 | x_2) dx_1 dx_2 \\ &= \int_{-\infty}^{+\infty} p(x_2) dx_2 \int_{-\infty}^{+\infty} [x_1 - h(x_2)]^2 p(x_1 | x_2) dx_1 \end{aligned} \quad \dots 7.27$$

where

$$p(x_2) = \int_{-\infty}^{+\infty} p(x_1, x_2) dx_1 \quad \dots 7.27A$$

is a minimum.

But,

$$\begin{aligned} & \int_{-\infty}^{+\infty} [x_1 - h(x_2)]^2 p(x_1 | x_2) dx_1 \\ &= \int_{-\infty}^{+\infty} [x_1 - m_2(x_1) + m_2(x_1) - h(x_2)]^2 p(x_1 | x_2) dx_1 \end{aligned}$$

where $m_2(x_1)$ is the conditional mean value of x_1 for a given value of x_2 .

$$= \int_{-\infty}^{+\infty} [x_1 - m_2(x_1)]^2 p(x_1 | x_2) dx_1 + [m_2(x_1) - h(x_2)]^2 \quad \dots 7.28$$

since $m_2(x_1)$ is a function of x_2 only⁽⁵⁴⁾ is a minimum when

$$h(x_2) = m_2(x_1)$$

where

$$m_2(x_1) = \frac{\int_{-\infty}^{+\infty} x_1 p(x_1, x_2) dx_1}{\int_{-\infty}^{+\infty} p(x_1, x_2) dx_1} \quad \dots 7.29$$

Therefore, the best estimate of x_1 in terms of x_2 is the conditional mean value $m_2(x_1)$ of x_1 .

7.4.1.1. Least Square Estimate for
Gaussian Signal

If the signal is gaussian,

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{x_1^2}{\sigma_1^2} - \frac{2\rho x_1 x_2}{\sigma_1\sigma_2} + \frac{x_2^2}{\sigma_2^2}\right)\right]$$

where ρ is the correlation coefficient ...7.30

the best estimate of x_1 is given by

$$\begin{aligned} m_2(x_1) &= \frac{\int_{-\infty}^{+\infty} x_1 p(x_1, x_2) dx_1}{\int_{-\infty}^{+\infty} p(x_1, x_2) dx_1} \\ &= \frac{\int_{-\infty}^{+\infty} x_1 p(x_1, x_2) dx_1}{\frac{1}{\sigma_2\sqrt{2\pi}} \exp\left(-\frac{x_2^2}{2\sigma_2^2}\right)} \\ &= \int_{-\infty}^{+\infty} \frac{x_1 dx_1}{\sigma_1\sqrt{2\pi(1-\rho^2)}} \exp\left[-\frac{1}{2\sigma_1^2(1-\rho^2)}\left(x_1 - \frac{\rho\sigma_1}{\sigma_2}x_2\right)^2\right] \\ &= \rho \frac{\sigma_1}{\sigma_2} x_2 \\ &= \rho x_2, \end{aligned} \quad \dots 7.31$$

since $\sigma_1 = \sigma_2$,

x_1 and x_2 being both samples from the same ensemble.

The structure of the signal during the interval $(t-\tau)$ to (t) can be obtained by solving equation 7.31 which may be rewritten as

$$x(t) = \rho(\tau) x(t-\tau)$$

Putting

$$t = nT,$$

we get
$$x_n = \rho x_{n-1} \quad \dots 7.31A$$

Multiplying both sides by the integrating factor ρ^{-n-1} , we get⁽⁴⁰⁾

$$\rho^{-(n+1)} x_{n+1} - \rho^{-n} x_n = 0 \quad \dots 7.32$$

Therefore, the first difference

$$\delta_1(\rho^{-n} x_n) = 0$$

giving

$$\rho^{-n} x_n = k, \text{ an arbitrary constant.}$$

Substituting back the values for t , and neglecting the constant k , we get

$$x(t) = \rho^t \quad \dots 7.33$$

If the sampling frequency of the signal is considerably greater than the Nyquist rate or if the signal is characterized by a falling power spectrum, ρ will not be much less than unity. For instance, ordinary speech $\{x(t)\}$ is characterized by a falling power spectrum that can be approximately represented by the expression

$$G_{xx}(f) = \frac{2\pi A}{1+(f/f_2)^2} \quad \dots 7.34$$

where f_2 is the frequency at which the spectral density becomes half the maximum value attained by $G_{xx}(f)$. The corresponding autocorrelation function $\phi_{xx}(\tau)$ is given by⁽²⁰⁾

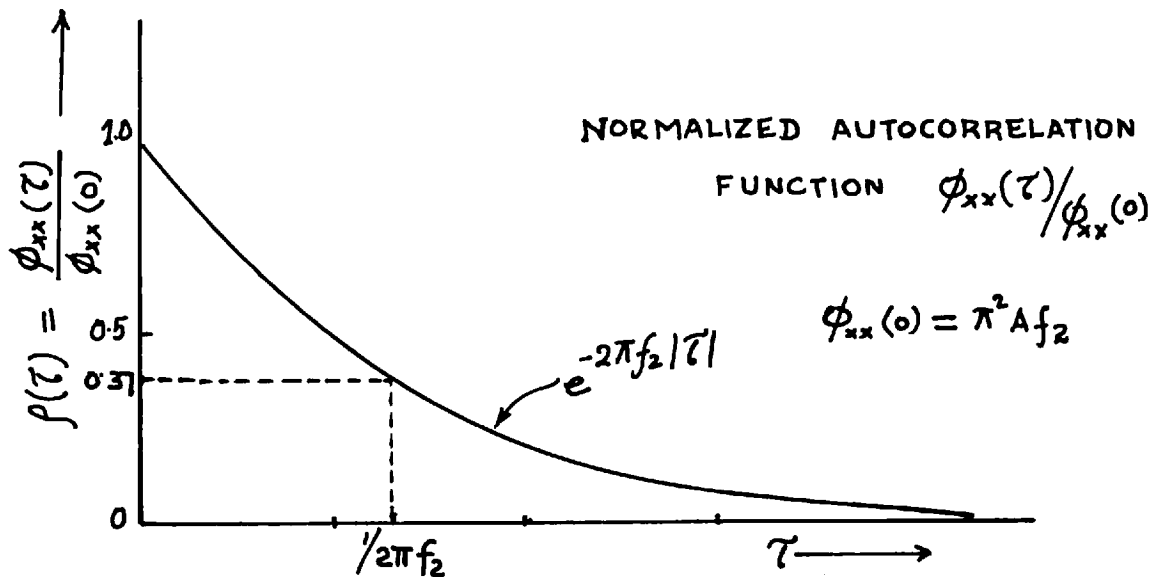
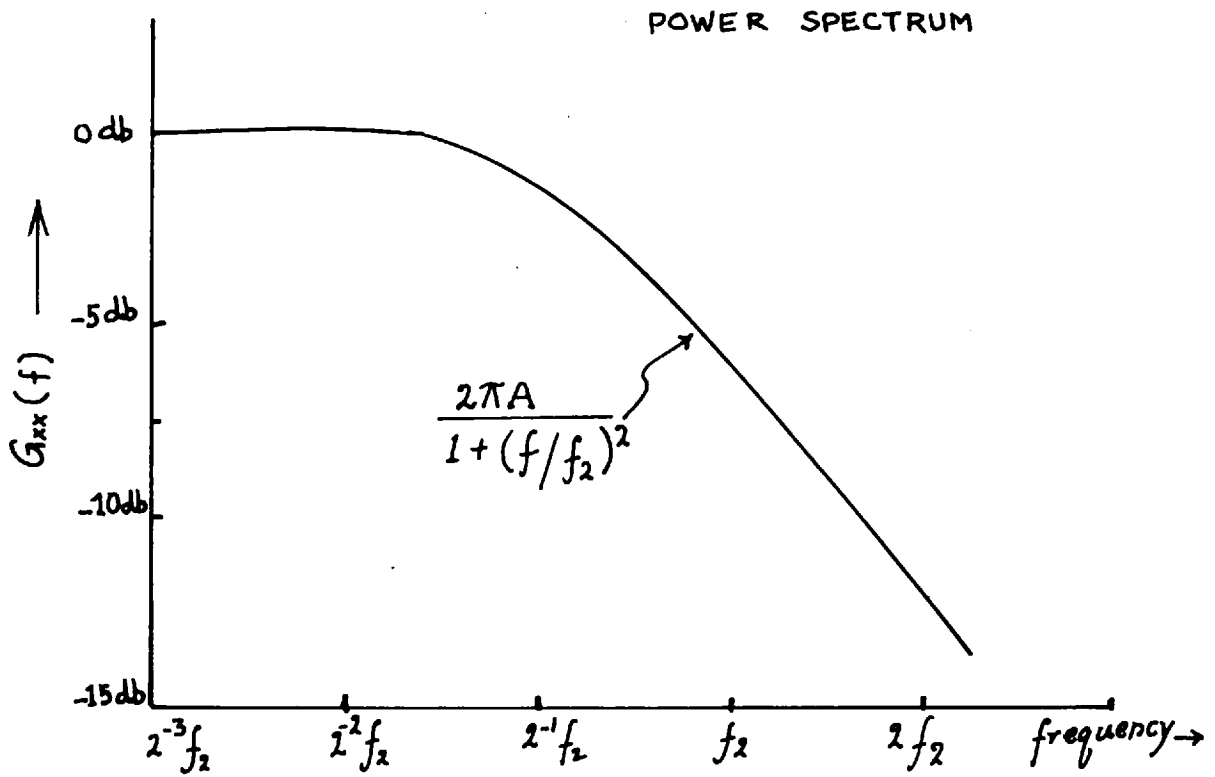


FIG. 7.10. POWER SPECTRUM & AUTOCORRELATION FUNCTION OF INPUT SIGNAL.

$$\begin{aligned}\phi_{xx}(\tau) &= \int_0^{151} G_{xx}(f) \cos 2\pi f \tau df \\ &= \phi_{xx}(0) e^{-2\pi f_2 |\tau|} \quad \dots 7.35\end{aligned}$$

where

$$\phi_{xx}(0) = \pi^2 A f_2^2 \quad \dots 7.35A$$

The normalized correlation coefficient $\rho(\tau)$ is now given by

$$\begin{aligned}\rho(\tau) &= \frac{\phi_{xx}(\tau)}{\phi_{xx}(0)} \\ &= e^{-2\pi f_2 |\tau|} \quad \dots 7.36\end{aligned}$$

For ordinary speech, f_2 is approximately 800 cps. If it is desired to pass all frequencies within the pass band 0-W cps, it is necessary to sample at a frequency f_s equal to or greater than $2W$, the corresponding value of τ being given by $1/f_s$. The calculated values of $\rho(\tau)$ for various values of W are given in Table 7.1. The sampling frequency f_s is assumed to be approximately 20% in excess of the theoretical Nyquist rate.

Table 7.1

W	f_s	f_s	$\rho(\tau)$
3.2. kc/s	8,000	cps	0.53
8.2. kc/s	20,000	cps	0.78
13 kc/s	30,000	cps	0.85
-----	50,000	cps	0.90
-----	100,000	cps	0.95

The values of $\rho(\tau)$ for the high sampling frequencies of 50 kc/s and 100 kc/s are given in the Table merely to indicate the high values of correlation existing between

successive samples in communication systems such as delta-modulation which employ pulse repetition frequencies of this order.⁽⁵⁵⁾

If f_s is very much greater than the half-power frequency f_2 , then $\rho(\tau)$ is close to unity. Under these conditions, expression 7.33 can be replaced by a polynomial of finite order to a high degree of accuracy:

$$\begin{aligned} x(t) &= \rho^t \\ &= 1 + t \ln \rho + t^2 \frac{(\ln \rho)^2}{2} + t^3 \frac{(\ln \rho)^3}{6} \\ &\quad + \dots + t^n \frac{(\ln \rho)^n}{n!} \dots 7.37 \end{aligned}$$

Since modulus of $\ln \rho$ tends to zero as ρ tends to unity, only a few terms of the expansion would have to be considered to obtain adequate accuracy if $(t \ln \rho)$ is less than unity. It, therefore, follows that a polynomial extrapolator of finite order is adequate to handle stochastic signals generated by gaussian processes.

It may be noted here that if the signal $x(t)$ has a constant or "direct-current" component m , then using the same statistical techniques as before, the best least square estimate of $x(t)$ can be shown to be given by

$$\begin{aligned} x(t) &= m + \rho(\tau) [x(t-\tau) - m] \\ &= m [1 - \rho(\tau)] + \rho(\tau)x(t-\tau) \\ &= \text{constant} + \rho(\tau) x(t-\tau) \dots 7.32A \end{aligned}$$

7.4.1.2. Least Square Estimate for Non-Gaussian Signals

For non-gaussian signals, the least square estimate of $x(t)$ as given by the conditional mean value of $x(t)$ in terms of $x(t-\tau)$ is not, in general, linear in $x(t-\tau)$. If, however, it is only desired to represent $x_1 = x(t)$ as a linear function of $x_2 = x(t-\tau)$

$$\begin{aligned} x_1 &= h(x_2) \\ &= \alpha + \beta x_2 \end{aligned} \quad \dots 7.38$$

thus restricting in advance the class of functions $h(x_2)$ which may be used to minimize the integral given in eqn. 7.27, then the optimum values of the constants α and β for which

$$\iint [x_1 - h(x_2)]^2 p(x_1, x_2) dx_1 dx_2$$

is a minimum, can be found as follows:

Let

m_1 and m_2 be the mean values of x_1 and x_2 respectively,

σ_1^2 and σ_2^2 be the variances of x_1 and x_2 respectively, *and*

λ_{12} be the covariance of x_1 and x_2 .

Then

$$\begin{aligned} I &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} [x_1 - h(x_2)]^2 p(x_1, x_2) dx_1 dx_2 \\ &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} [x_1 - \alpha - \beta(x_2)]^2 p(x_1, x_2) dx_1 dx_2 \end{aligned}$$

$$\begin{aligned}
&= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \left[x_1 - m_1 - \beta(x_2 - m_2) + m_1 - \alpha - \beta m_2 \right]^2 p(x_1, x_2) dx_1 dx_2 \\
&= \sigma_1^2 - 2\beta\lambda_{12} + \beta^2 \sigma_2^2 + (m_1 - \alpha - \beta m_2)^2 \quad \dots 7.39
\end{aligned}$$

I is a minimum if

$$\left. \begin{aligned} \frac{\partial I}{\partial \alpha} &= 0 \\ \frac{\partial I}{\partial \beta} &= 0 \end{aligned} \right\} \quad \dots 7.40$$

Solving equations 7.40 gives

$$\alpha = m_1 - \beta m_2 \quad \dots 7.41$$

$$\begin{aligned}
\beta &= \frac{\lambda_{12}}{\sigma_2^2} \\
&= \rho(\tau) \frac{\sigma_1}{\sigma_2} \quad \dots 7.42
\end{aligned}$$

where $\rho(\tau)$ is the normalized correlation coefficient

$$\frac{\lambda_{12}}{\sigma_1 \sigma_2}$$

$$\left. \begin{aligned} \text{Since } \sigma_1 &= \sigma_2 = \sigma \\ m_1 &= m_2 = m \end{aligned} \right\} \quad \dots 7.43$$

$$\text{therefore } x_1 = m [1 - \rho(\tau)] + \rho(\tau) x_2 \quad \dots 7.44$$

which is essentially the same as expression 7.32A for the case of gaussian signals. It is therefore obvious that if it is justifiable a-priori (see Section 7.4.1.3.) to use an estimate for x_1 that is linear in x_2 , then a polynomial extrapolator can be efficiently used for the processing of such signals, even if they are non-gaussian.

It may be noted that the least square estimate $h(x_2)$ is the same for both gaussian and non-gaussian signals. This is a direct result of the use of the least square error criterion for the determination of $h(x_2)$, the solutions of the linear least square problem and the corresponding general least square problem being identical if the process is normal⁽⁵⁵⁾.

7.4.1.3. Least Square Estimates for n^{th} Order Markov Processes.

So far it was assumed that the statistical structure of the signal $\{x(t)\}$ is such that the magnitude of the signal at time t is a function of the magnitude of the immediately previous sample $x(t-\tau)$ at time $(t-\tau)$ and only of it. This is, of course, equivalent to assuming the signal to be generated by a Markov process of first order⁽⁵⁷⁾. It will now be assumed that the intersymbol influence extends over n sampling intervals, that is, the magnitude of $x(t)$ depends upon $x(t-\tau)$, $x(t-2\tau)$, ----- $x(t-n\tau)$.

Let

$$\left. \begin{array}{l} x(t) = x_1 \\ x(t-\tau) = x_2 \\ \text{-----} \\ \text{-----} \\ x(t-n\tau) = x_{n+1} \end{array} \right\} \dots 7.45$$

If it is now desired to express x_1 as a function $h(x_2, x_3, \dots, x_{n+1})$ of $x_2, x_3, \dots, x_{n+1}$ such that the expectation of

$$\left[x_1 - h(x_2, x_3, \dots, x_{n+1}) \right]^2$$

is a minimum, then proceeding as in the previous Section, it can be shown that $h(x_2, \dots, x_{n+1})$ is given by the conditional mean value of x_1 when the magnitudes of $x_2, \dots, x_{n+1}$ are

known, that is,

$$h(x_2 \dots x_{n+1}) = \frac{\int_{-\infty}^{+\infty} x_1 p(x_1 x_2 \dots x_{n+1}) dx_1}{\int_{-\infty}^{+\infty} p(x_1 x_2 \dots x_{n+1}) dx_1} \quad \dots 7.46$$

This expression is, however, very unwieldy to use and difficult to solve. Also, it is, in general, quite difficult to obtain an accurate representation for the $(n+1)$ dimensional probability density function $p(x_1 x_2 \dots x_{n+1})$. For these reasons it is more sensible to use a simpler, though theoretically less accurate*, model for $h(x_2 \dots x_{n+1})$ that can be derived using experimentally obtainable data such as the moments and first-order product-moments of the variables $x_1 x_2 \dots x_{n+1}$. Therefore, it will hereafter be assumed that the form of the possible functions $h(x_2 \dots x_{n+1})$ will be restricted to functions belonging to the class of linear manifolds of $x_2 x_3 \dots x_{n+1}$. Therefore, the best estimate of x_1 is now given by

$$h(x_2 \dots x_{n+1}) = \beta_{12} x_2 + \beta_{13} x_3 + \dots + \beta_{1n} x_n + \beta_{1(n+1)} x_{n+1} \quad \dots 7.47$$

The coefficients $\beta_{12}, \dots, \beta_{1n}, \beta_{1(n+1)}$ can now be found by minimizing the integral

$$I = \text{Expectation of } \left[x_1 - \beta_{12} x_2 - \beta_{13} x_3 - \dots - \beta_{1(n+1)} x_{n+1} \right]^2 \quad \dots 7.48$$

* The temporal and spatial stability of high order joint probability distributions is, in general, poor. This is especially true of the conditions normally met with in communication systems. The greater the order of the distribution, the greater is the fluctuation in the magnitudes of the parameters of the distribution.

$$= \sum \left[x_1 - \beta_{12}x_2 - \beta_{13}x_3 - \dots - \beta_{1(n+1)}x_{n+1} \right]^2 \quad \dots 7.49$$

the summation extending over all sample values.

Differentiating I with respect to β_{1k} and equating to zero, we get

$$\begin{aligned} \frac{\partial I}{\partial \beta_{1k}} &= \sum (x_1 x_k) - \beta_{12} \sum (x_2 x_k) - \dots - \beta_{1k} \sum (x_k)^2 \\ &\quad - \dots - \beta_{1(n+1)} \sum (x_{n+1} x_k) \\ &= 0 \end{aligned} \quad \dots 7.50$$

where k takes on the values $2, 3, \dots, (n+1)$.

Therefore

$$\begin{aligned} \beta_{12} \sum x_2^2 + \beta_{13} \sum x_2 x_3 + \dots + \beta_{1(n+1)} \sum x_2 x_{n+1} &= \sum x_2 x_1 \\ \beta_{12} \sum x_3 x_2 + \beta_{13} \sum x_3^2 + \dots + \beta_{1(n+1)} \sum x_3 x_{n+1} &= \sum x_3 x_1 \\ \dots &\dots \\ \beta_{12} \sum x_{n+1} x_2 + \beta_{13} \sum x_{n+1} x_3 + \dots + \beta_{1(n+1)} \sum x_{n+1}^2 &= \sum x_{n+1} x_1 \end{aligned} \quad \dots 7.51$$

Using the notation λ_{ik} to denote the covariance between the variates x_i and x_k we get

$$\begin{aligned} \beta_{12} \lambda_{22} + \beta_{13} \lambda_{23} + \dots + \beta_{1(n+1)} \lambda_{2(n+1)} &= \lambda_{21} \\ \beta_{12} \lambda_{32} + \beta_{13} \lambda_{33} + \dots + \beta_{1(n+1)} \lambda_{3(n+1)} &= \lambda_{31} \\ \dots &\dots \\ \beta_{12} \lambda_{(n+1)2} + \beta_{13} \lambda_{(n+1)3} + \dots + \beta_{1(n+1)} \lambda_{(n+1)(n+1)} &= \lambda_{(n+1)1} \end{aligned} \quad \dots 7.52$$

Solving the system of equations 7.52 gives

$$\beta_{1k} = - \frac{\Delta_{1k}}{\Delta_{11}} \quad \dots 7.53$$

where Δ_{ik} is the cofactor of λ_{ik} in the determinant Δ

$$\Delta = \begin{vmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1n} \\ \lambda_{21} & \lambda_{22} & \dots & \lambda_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{n1} & \lambda_{n2} & \dots & \lambda_{nn} \end{vmatrix} \quad i, k = 1, 2, \dots, (n+1) \quad \dots 7.54$$

Consider now a second order process specified by the relationship

$$x(t) + a_1 x(t-\tau) + a_2 x(t-2\tau) = \varepsilon(t) \quad \dots 7.55$$

If a_1 and a_2 are chosen to minimize the mean square of the residual term $\varepsilon(t)$, then

$$\begin{aligned} a_1 &= -\beta_{12} \\ &= + \frac{\Delta_{12}}{\Delta_{11}} \\ &= - \frac{\begin{vmatrix} \lambda_{12} & \lambda_{23} \\ \lambda_{13} & \lambda_{33} \end{vmatrix}}{\begin{vmatrix} \lambda_{22} & \lambda_{23} \\ \lambda_{32} & \lambda_{33} \end{vmatrix}} \end{aligned} \quad \dots 7.56$$

But $\lambda_{ik} = \rho_{ik} \sigma_i \sigma_k \quad \dots 7.57$

where ρ_{ik} is the corresponding correlation coefficient

Further,

$$\sigma_i = \sigma_k, \quad \dots 7.58$$

since x_i and x_k are samples from the same ensemble, and

$$\rho_{ik} = \rho(\tau | k-i |) \quad \dots 7.59$$

since the sampling interval is constant.

Also,

$$\begin{aligned} \rho_{ii} &= \rho(0) \\ &= 1 \end{aligned} \quad \dots 7.60$$

$$\begin{aligned} \therefore a_1 &= - \frac{\begin{vmatrix} \rho(\tau) & \rho(\tau) \\ \rho(2\tau) & \rho(0) \end{vmatrix}}{\begin{vmatrix} \rho(0) & \rho(\tau) \\ \rho(\tau) & \rho(0) \end{vmatrix}} \\ &= - \frac{\rho(\tau) [1 - \rho(2\tau)]}{1 - \rho^2(\tau)} \end{aligned} \quad \dots 7.61$$

Similarly,

$$\begin{aligned} a_2 &= + \frac{\Delta_{13}}{\Delta_{11}} \\ &= \frac{\begin{vmatrix} \lambda_{12} & \lambda_{22} \\ \lambda_{13} & \lambda_{32} \end{vmatrix}}{\begin{vmatrix} \lambda_{22} & \lambda_{23} \\ \lambda_{32} & \lambda_{33} \end{vmatrix}} \\ &= \frac{\begin{vmatrix} \rho(\tau) & \rho(0) \\ \rho(2\tau) & \rho(\tau) \end{vmatrix}}{\begin{vmatrix} \rho(0) & \rho(\tau) \\ \rho(\tau) & \rho(0) \end{vmatrix}} \\ &= - \frac{\rho(2\tau) - \rho^2(\tau)}{1 - \rho^2(\tau)} \end{aligned} \quad \dots 7.62$$

It may be noted that if the autocorrelation function of the signal is of the exponential type given by eqn. 7.36, then

$$\begin{aligned}\rho(2\tau) &= e^{-2\tau/T_0} \\ &= \left[e^{-\tau/T_0} \right]^2 \\ &= \rho^2(\tau)\end{aligned}\quad \dots 7.63$$

Therefore, $\omega_2 = 0$

according to eqn. 7.62, so that the least square estimate of $x(t)$ is now given by

$$x(t) = a_1 x(t-\tau) \quad \dots 7.64$$

In other words, if the signal has a falling power spectrum similar to that possessed by white noise passed through a single stage RC integrating network, no greater accuracy in the representation of the signal is achieved by considering second or higher order processes as the statistical model for the signal.

The statistical estimation equation

$$x(t) + a_1 x(t-\tau) + a_2 x(t-2\tau) = 0 \quad \dots 7.65$$

can be solved⁽⁵⁸⁾ to give

$$x(t) = u^t (A \cos \omega t + B \sin \omega t) \quad \dots 7.66$$

where $u = + a_2^{1/2} \quad \dots 7.67$

$$\omega = \arccos \left[- \frac{a_1}{2 a_2^{1/2}} \right] \quad \dots 7.68$$

Since a_2 and therefore u are less than unity (according

to equation 7.62), $x(t)$ is characterized by a damped sinusoidal oscillation, that rapidly decays to zero. The order of the polynomial required to adequately represent $x(t)$ is determined by the relative magnitudes of u and ω .

The more general estimation equation

$$x(t) + a_1 x(t-\tau) + \dots + a_n x(t-n\tau) = 0 \quad \dots 7.69$$

can be solved to give

$$x(t) = H_1 h_1^t + H_2 h_2^t + \dots + H_n h_n^t + K \quad \dots 7.70$$

where $h_1, h_2, \dots, h_n$ are the roots of the characteristic equation

$$h^n + a_1 h^{n-1} + a_2 h^{n-2} + \dots + a_n = 0 \quad \dots 7.71$$

It is rather difficult to study the detailed behaviour of such a function. In any case, if the order of the equation is large, the order of the extrapolator required for adequate representation of the function becomes too large for consideration in practical cases.

7.5. Exponential Extrapolation

It was shown in Section 7.4 that for gaussian signals, the best estimate of $x(t)$ is given by

$$x(t) = \rho(\tau) x(t-\tau) \quad \dots 7.72$$

which can be solved to give

$$x(t) = \rho^t(\tau) \quad \dots 7.73$$

Rather than designing a high order polynomial extrapolator to generate a polynomial approximation to the exponential input signal, it is more logical to design the extrapolator to generate an exponential waveform. This can be done by

replacing the polynomial generator part only of the extrapolator by an exponential generator consisting of an RC network as shown in Fig. 7.11. The input to the exponential generator is, as in the polynomial case, the sampled output of the error integrator so that the actual input to the RC network is (except for the error introduced by quantization) an impulse of magnitude equal to that of the most recent sample $p(nT)$ of the signal. Now, the impulse response of the RC network is given by

$$\frac{1}{\gamma} \cdot e^{-t/\gamma}$$

where γ is equal to the time constant RC of the network. If R and C are chosen such that

$$\gamma = \frac{T}{\ln \frac{1}{\rho(\tau)}} \quad \dots 7.74$$

where T is the sampling interval, then it can be seen that the output of the exponential generator would decay to exactly $p(nT)\rho(\tau)$ at the next sampling instant $(n+1)T$.

An amplifier of gain γ is introduced between the output of the RC network and the difference amplifier to compensate for the attenuation caused by the former.

It may be noted that ordinary PCM and single-integration delta modulation systems^(55,60,61) are particular cases of the present system. The former can be derived by breaking the feedback section of the loop while single-integration delta-modulation systems (with n coded digits per sample) can be obtained by replacing the exponential extrapolator with a unity-gain holding circuit or alternatively, by increasing the time constant of the network to a value that is very large compared to the sampling interval.

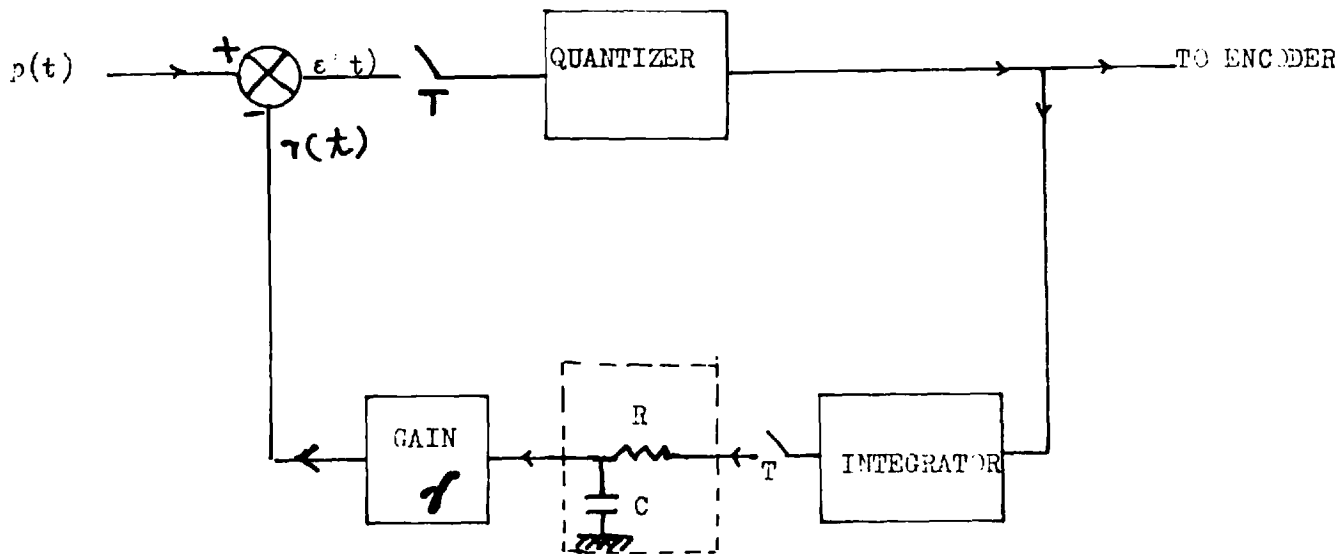


FIG. 7.11. EXPONENTIAL EXTRAPOLATION.

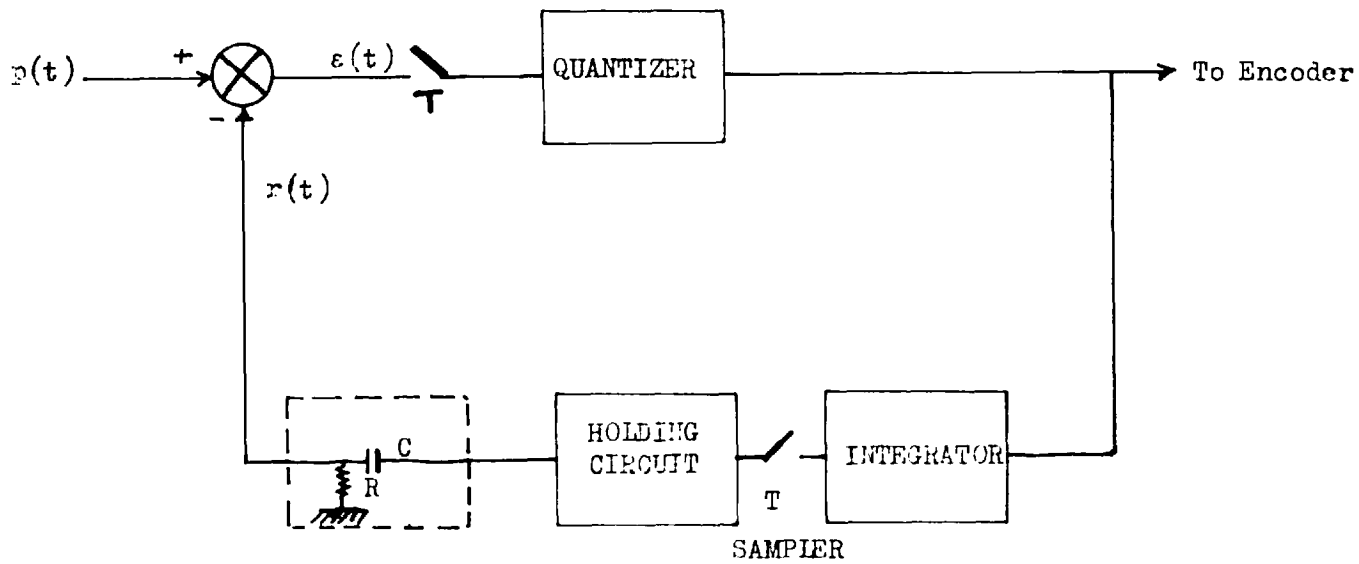


FIG. 7.11A. EXPONENTIAL EXTRAPOLATION USING PHYSICALLY REALIZABLE ELEMENTS.

7.6. Improvement in System Performance due to Use of Extrapolator.

It has been shown in previous sections that the existence of correlation between successive samples of the same signal can be considered to be due to the presence of a non-random component in the signal. If this systematic component is eliminated using extrapolation techniques, it is now necessary to transmit only the purely random component to the receiver for it to regenerate the original signal. The variance and, therefore, in general, the peak to peak variation of the error signal $e(t)$ given by the difference between the actual input signal $p(t)$ and the extrapolator output $r(t)$, are less than those of the input signal itself so that for a given, fixed number of levels of quantization, it is possible to represent the output of the quantizer to a greater degree of precision. The ratio of the variances of the actual system input $p(t)$ and of the input $e(t)$ to the quantizer may therefore be considered to be a suitable criterion to judge the improvement in system performance obtained by using the extrapolator.

Methods for evaluating the ratio are outlined below:

7.6.1. First Order Correlation.

Let $e(t)$ denote the residual component left after the best estimate of $x(t)$, namely $p(\tau) x(t - \tau)$, has been subtracted from $x(t)$

$$x_1 - p x_2 = e_1 \quad \text{..... 7.75.}$$

where

$$\begin{aligned} x_1 &= x(t) \\ x_2 &= x(t - \tau) \\ e_1 &= e(t) \end{aligned}$$

The variance of $e(t)$ is given by

$$\begin{aligned}
\sigma_{\varepsilon}^2 &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x_1 - \rho x_2)^2 p(x_1, x_2) dx_1 dx_2 \\
&= \sigma_1^2 - 2\rho \lambda_{12} + \rho^2 \sigma_2^2, \text{ from eq. 7.39.} \\
&= \sigma^2 (1 - \rho^2) \quad \dots 7.76.
\end{aligned}$$

since

$$\begin{aligned}
\lambda_{12} &= \rho \sigma_1 \sigma_2 \\
\sigma_1 &= \sigma_2 = \sigma
\end{aligned}$$

It is thus seen that the variance of the input to the quantizer is reduced by a factor of $(1 - \rho^2)$ by using the extrapolator.

The quantization noise introduced into the transmitted signal is approximately given by

$$\begin{aligned}
N_Q &= \frac{k \sigma_{\varepsilon}^2}{l^2} \\
&= \frac{k \sigma^2 (1 - \rho^2)}{b^{2n}} \quad \dots 7.77
\end{aligned}$$

where

l = number of quantization levels
 b = base of the coding system used
 n = number of coded digits per sample
 k = a constant of proportionality.

If it is decided to reduce the number of coded digits per sample from n_1 to n_2 but increase the sampling frequency f_s from $(1/T_1)$ to $(1/T_2)$ such that the product of the sampling frequency and the number of digits per sample is constant, that is

$$T_2 = T_1 \frac{n_2}{n_1} \quad \dots 7.78$$

then a reduction in quantization noise can be expected if

$$\frac{\sigma^2 (1 - \rho_2^2)}{b^{2n_2}} < \frac{\sigma^2 (1 - \rho_1^2)}{b^{2n_1}} \quad \dots \quad 7.79.$$

where

$$\left. \begin{aligned} \rho_1 &= \rho(T_1) \\ \rho_2 &= \rho(T_2) \end{aligned} \right\} \quad \dots \quad 7.80$$

Equation 7.79 yields the relation

$$b^{2(n_1 - n_2)} < \frac{1 - \rho_1^2}{1 - \rho_2^2} \quad \dots \quad 7.82$$

Since a very large increase in ρ_2 over ρ_1 is not, in general, obtainable even with a doubling of the sampling frequency it is obvious that equation 7.82 would normally be satisfied only when n_1 is small, say equal to 2 or 3. Also, it is more difficult to satisfy equation 7.82 when a ternary or higher order system rather than ^{the} conventional binary ^{system} is used.

7.6.2. Higher Order Correlations.

If high order correlations exist in the original signal, then $x(t)$ can be represented in the least square sense by a linear function of $x(t - \tau)$, $x(t - 2\tau)$, ... $x(t - n\tau)$, as in equation 7.47. The residual, random element $\varepsilon(t)$ obtained after subtracting the best estimate of x_1 , from x_1 is therefore given by

$$\begin{aligned} \varepsilon'(t) &= \varepsilon_1 = x_1 - \beta_{12}x_2 - \beta_{13}x_3 \dots - \beta_{1(n+1)}x_{n+1} \\ &= \frac{1}{\bigwedge_{11}} \sum_{k=1}^{n+1} \bigwedge_{1k} x_k \quad \dots 7.83. \end{aligned}$$

correlation that exists between x_1 and x_3 after the effect of any influence on x_1 and x_3 due to x_2 has been removed.

Thus, for the case mentioned in section 7.4. where appreciable correlation is known to exist over two sample intervals, the variance of the actual input to the quantizer, after the elimination of the systematic component in the signal, is given by

$$\begin{aligned}
 \sigma_e^2 &= \sigma^2 (1 - \rho_{12}^2) (1 - \rho_{13.2}^2) \\
 &= \sigma^2 (1 - \rho_{12}^2) \left[1 - \left\{ \frac{\hat{\rho}_{13}}{\sqrt{\hat{\rho}_{11} \hat{\rho}_{33}}} \right\}^2 \right] \\
 &= \sigma^2 (1 - \rho_{12}^2) \left[1 - \left\{ \frac{\rho_{13} - \rho_{12} \rho_{23}}{\sqrt{(1 - \rho_{12}^2)(1 - \rho_{23}^2)}} \right\}^2 \right] \\
 &= \sigma^2 \left[1 - \rho^2(\tau) - \frac{\{\rho(2\tau) - \rho^2(\tau)\}^2}{1 - \rho^2(\tau)} \right] \dots 7.87.
 \end{aligned}$$

It may be noted that if, in the above expression,

$$\rho(2\tau) = \rho^2(\tau) \dots 7.88.$$

or, in other words, if the autocorrelation function $\rho_{xx}(\tau)$ of the signal is exponential, then the variance of $s(t)$ is given by

$$\sigma_e^2 = \sigma^2 (1 - \rho^2(\tau)) \dots 7.89.$$

which is the same as the expression obtained for the case of extrapolation making use of only first order correlation. This result thus further confirms the deduction made earlier in section 7.24 (equations 7.62 to 7.64) that, for signals characterized by such autocorrelation functions, no advantage is gained by designing the extrapolator to make use of second- and higher order correlations in the sampled signal.

7.7. Effect of Quantization on System Performance.

Two main effects are produced by putting the quantizer in the forward section of the feedback loop of the system:

- A. The quantization noise thus introduced affects the proper operation of the extrapolator, and
- B. The quantizer makes the loop gain of the system amplitude dependent. Under certain conditions, this can lead to system instability.

These aspects are discussed in detail below.

7.7.1. Effect due to Quantization Noise.

If the number of levels of quantization is sufficiently large and if the sampling frequency is not very much greater than the Nyquist rate, then it is justifiable, as shown in section 6, to replace (Ref.62) the quantizer by a source of independent, random, additive, white noise $g(t)$ with amplitude distribution $p_1(g)$ given by

$$p_1(g) = \begin{cases} \frac{1}{h} & , \quad -\frac{h}{2} \leq g \leq +\frac{h}{2} \\ 0 & , \quad |g| > \frac{h}{2} \end{cases} \quad \dots 7.90$$

and variance

$$\sigma_g^2 = \frac{1}{12} h^2 \quad \dots 7.91$$

where h is the magnitude of a quantization interval (Fig.7.12A). Assuming the probability distribution of the amplitude of the error signal $e(t)$ to the quantizer to be gaussian, as given by the relation

$$p_2(e) = \frac{1}{\sqrt{2\pi} \sigma_e} \cdot e^{-e^2/2\sigma_e^2} \quad \dots 7.92.$$

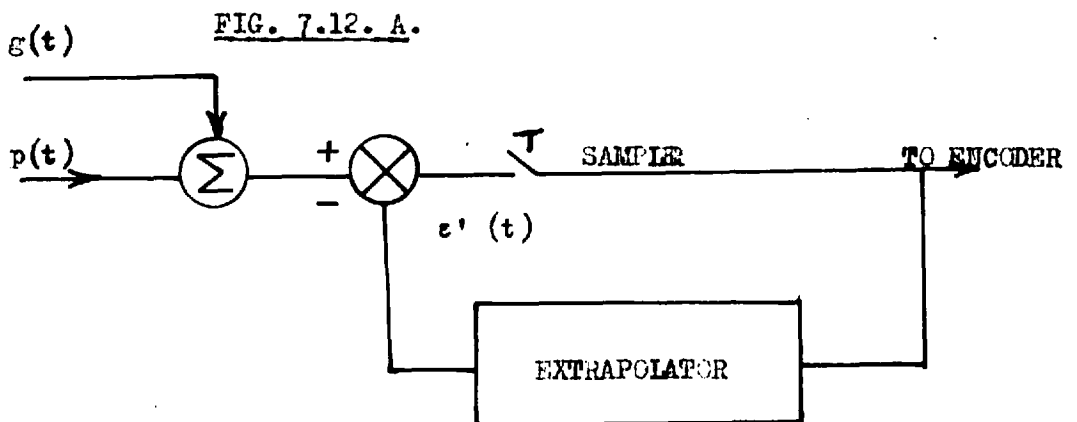
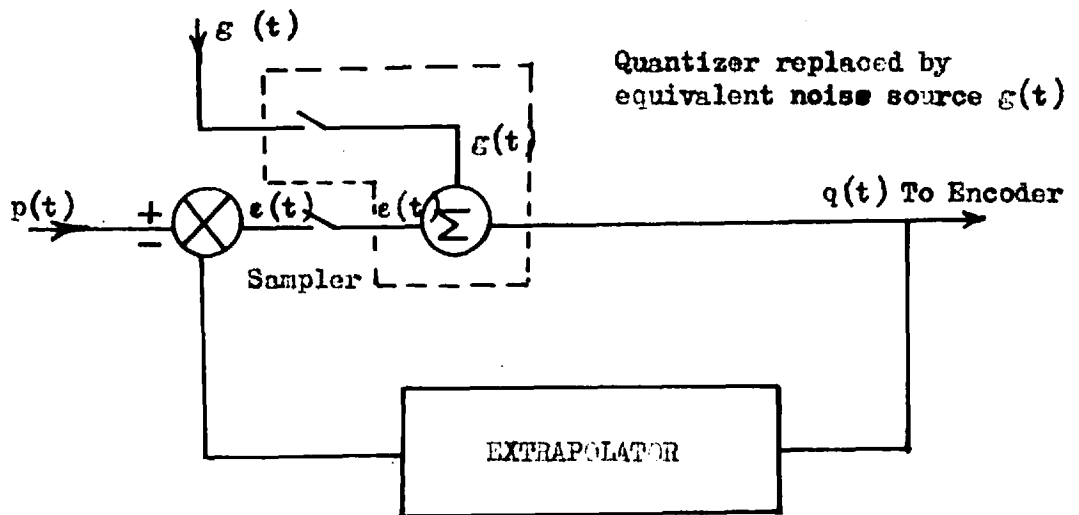


FIG. 7.12B. EQUIVALENT CIRCUIT FOR THE QUANTIZER.

the compound distribution $p_3(q)$ of the amplitude of the output

$$q(t) = g(t) + e(t) \quad \dots 7.93$$

of the quantizer is given by (Ref:63)

$$\begin{aligned} p_3(q) &= \int_{-\infty}^{+\infty} p_1(q-\varepsilon) p_2(\varepsilon) d\varepsilon \\ &= \frac{\sqrt{\pi/2}}{h} \left[I\left(\frac{q}{\sigma_\varepsilon} + \frac{h}{2\sigma_\varepsilon}\right) - I\left(\frac{q}{\sigma_\varepsilon} - \frac{h}{2\sigma_\varepsilon}\right) \right] \end{aligned} \quad \dots 7.94$$

where

$$I(\alpha) = \frac{1}{\sqrt{2\pi}} \int_{-\alpha}^{+\alpha} e^{-t^2/2} dt \quad \dots 7.95$$

If, however, the quantization noise power is small compared to σ_ε^2 , then it may be assumed that the amplitude distribution of $q(t)$ remains gaussian, with the difference that the output power σ_q^2 is greater than σ_ε^2 by an amount equal to σ_g^2 . Under these conditions, since $g(t)$ and $e(t)$ are uncorrelated, it is easily shown that the configuration of fig.7.12A can be replaced by the system shown in fig.7.12B in which the additive noise source within the loop is replaced by an equivalent input signal outside the loop.

The introduction of the quantizer is thus seen to be equivalent to merely increasing the variance of the random component of the input signal $p(t)$ by σ_g^2 .

One immediate conclusion which can be drawn from a study of fig.7.12A is that even if the input signal is completely systematic, there is no significant advantage gained by increasing the order of the extrapolator beyond the point where the k^{th} -order back difference of the signal is approximately equal to or less than the root mean square value of the quantization noise. For instance, for the gaussian signal mentioned in section 7.4, optimum utilization of the extrapolator may presumably be considered to be obtained if the k^{th} back difference

$$\delta_k x_n = (\rho - 1)^k x_n \quad \dots 7.96.$$

of the signal $x(t)$ at $t = nT$, is approximately equal to σ_g , since beyond this value there is very little reduction in the magnitude of the residual component to be obtained even if k is made much larger.

7.7.2. Effect of Nonlinearity on Loop Gain of System.

Since the quantizer possesses odd symmetry, only odd harmonics of the input signal are generated. It may therefore be assumed that the linear elements following the quantizer, being predominantly of the low pass type, (for instance, the RC integrator network used for exponential extrapolation), severely attenuate all the harmonics generated by the non-linearity, the order of the lowest-order harmonic present being three. Under these conditions, it is permissible to investigate the stability of the system using describing function techniques. (Ref: 64, 65).

The describing function $N(A)$ of a nonlinearity is defined as the ratio of the fundamental component $A_1 \sin(\omega t + \theta)$ of the output signal, to the input signal, when the latter is a sinusoidal function, $A \sin \omega t$

$$N(A) = \frac{A_1 \sin(\omega t + \theta)}{A \sin \omega t} \quad \dots 7.97$$

It is easily shown (Ref. 64) that for a steady state oscillation to be present, the loop gain of the system for the fundamental component must be unity - in other words,

$$W(j\omega) * N(A) = 1 \quad \dots 7.98$$

where $w(j\omega)$ is the open loop frequency response function of the feedback loop obtained by substituting $(j\omega)$ for (s) in the open loop transfer function $w(s)$, s being the complex frequency function

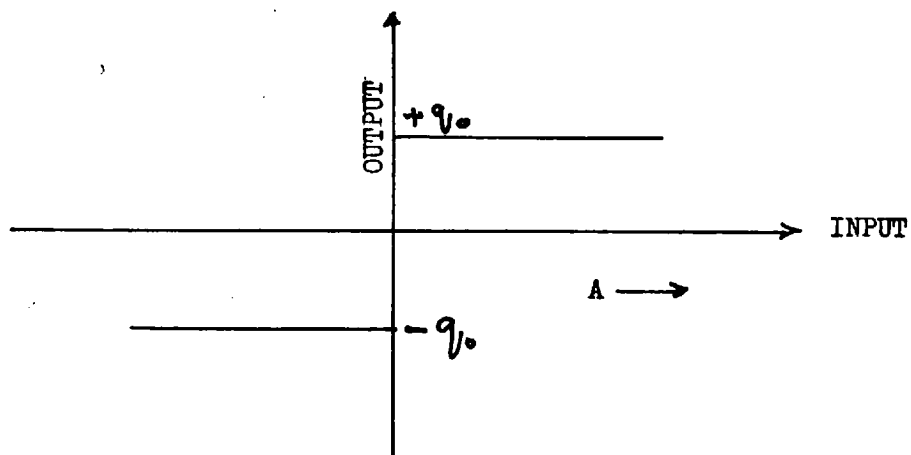


FIG. 7.13A. QUANTIZER NONLINEARITY: 2-LEVEL

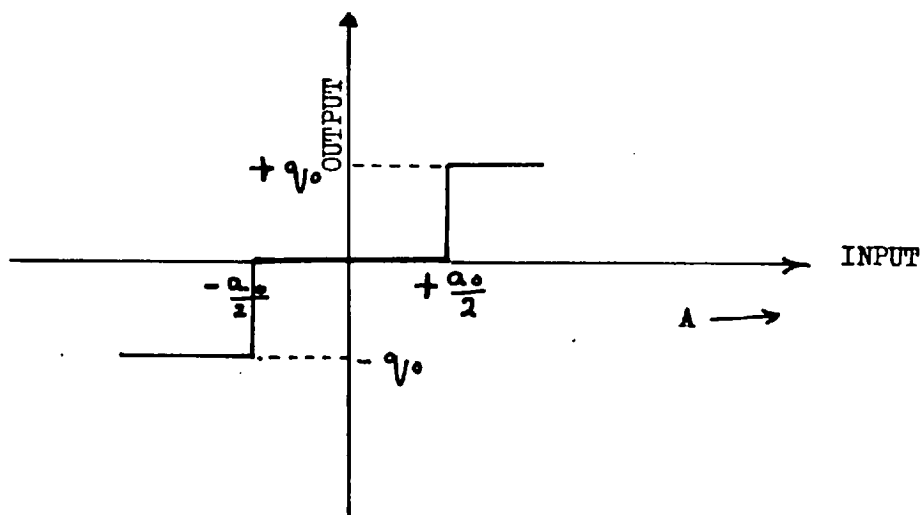


FIG. 7.13B QUANTIZER NONLINEARITY: 3-LEVEL.

$$s = \alpha + j \omega .$$

The amplitude and frequency of possible states of oscillation of the system are given by the roots of equation 7.98 which can be solved by plotting $w(j \omega)$ and $1/N(A)$ to real and imaginary coordinates in the w plane

$$w = \mu + j v ,$$

and noting the points of intersection of the two loci.

In a sampled data system, the delays introduced (Ref.66) by sampling and the use of holding circuits, cause the frequency response function to have the form of a spiral (Ref.67) as shown in fig.7.14. Further, the describing function of an ON-OFF nonlinearity, corresponding to the use of a 1-digit binary code, is given by (Ref.64)

$$N_2(A) = \frac{4q_0}{\pi A} \quad \dots 7.99$$

where $2 q_0$ is the peak to peak amplitude of the output of the nonlinearity. $1/N(A)$ is also plotted in fig.7.14. [It may be noted that the gain of the element is purely real - a consequence of the assumption that the nonlinear element has no hysteresis]. It is obvious that the system is unstable since the gain becomes infinite when the amplitude A of the input is zero. Though there is more than one point of intersection between the two loci possible, it can be shown (Ref.67) that the only stable mode of oscillation is that corresponding to the point w_0 at the extreme right of all the points of intersection. This point defines the existence of a limit cycle (Ref.68, 69) of frequency ω_0 and amplitude A_0 , oscillations at the lower frequency ω_0 suppressing those at ω_1 and ω_2 . It is possible to suppress these oscillations if, by some means, the gain of the nonlinear element can be prevented from increasing to infinity for small values of A . This can be accomplished by using a three-level nonlinearity,

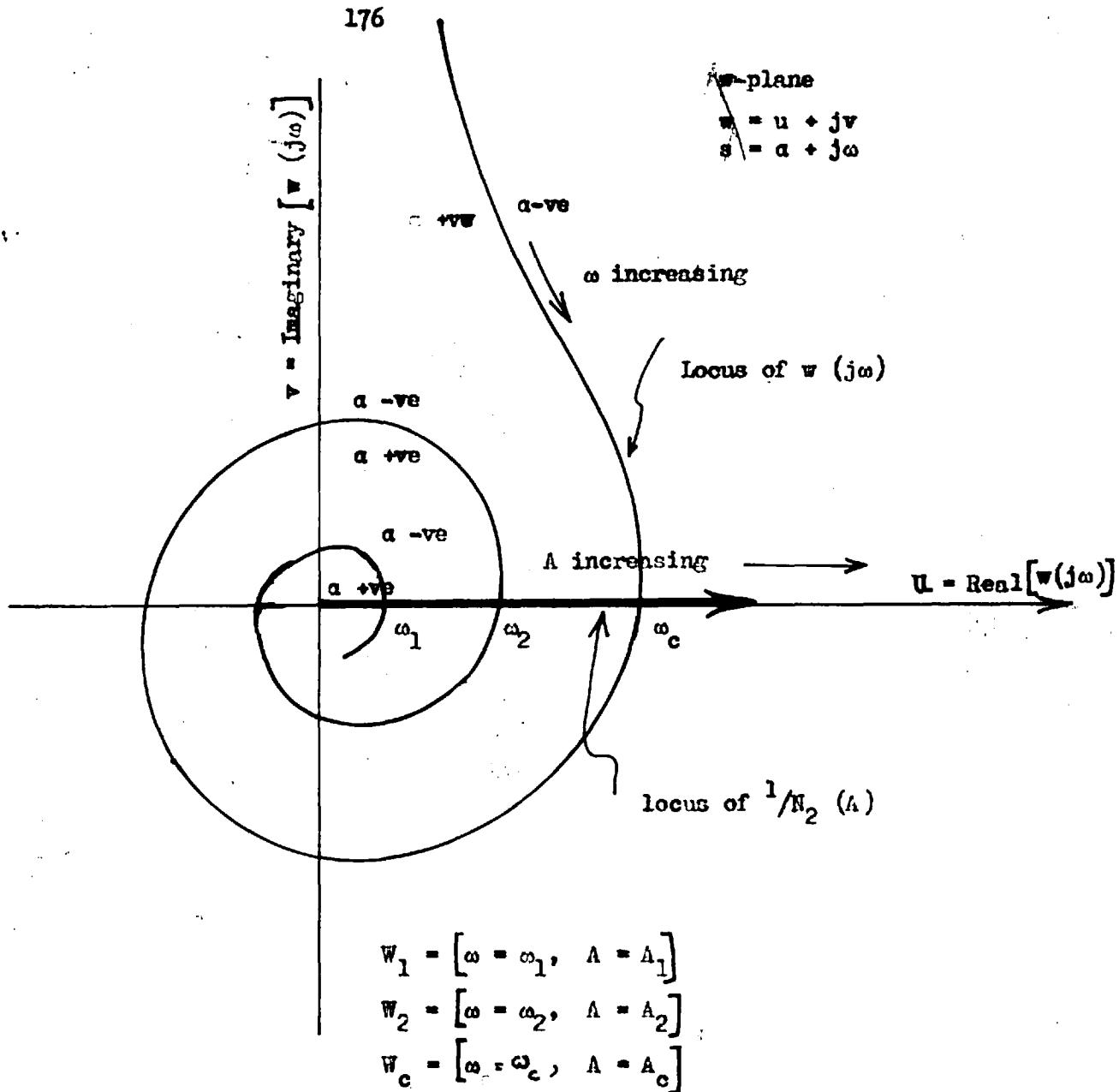


FIG. 7.14. STABILITY DIAGRAM FOR BINARY SYSTEM.

corresponding to the use of a 1-digit ternary code, as shown in fig. 7.13B. The describing function of such a nonlinearity is given by

$$N_3(A) = \begin{cases} \frac{4a_0}{A} \left[1 - \left(\frac{a_0}{2A} \right)^2 \right]^{\frac{1}{2}} & \text{or } 2A \geq a_0 \\ 0 & \text{or } 2A < a_0 \end{cases} \quad \dots 7.100$$

where

a_0 is the "dead-zone" of the nonlinearity

$1/N_3(A)$ is plotted in fig. 7.15. It may be noted that for certain values of amplitude, $1/N_3(A)$ is double-valued, its magnitude falling from infinity to a minimum and then again increasing, as A increases from zero to infinity. The magnitude of the input amplitude A corresponding to the point of closest approach of the function $1/N_3(A)$ to the origin can be determined by solving

$$\frac{d}{dA} N_3(A) = 0 \quad \dots 7.101$$

Solving equation 7.101 gives

$$\frac{a_0^2}{4A^4} = \frac{1}{A^2} \left(1 - \frac{a_0^2}{4A^2} \right)$$

giving

$$a_0 = \sqrt{2} A \quad \dots 7.102$$

For this value of a_0 , magnitude of $N_3(A)$ attains a maximum given by

$$N_3^{\wedge}(A) = \frac{4a_0}{\pi a_0} \quad \dots 7.103$$

If this value of $N_3(A)$ is less than the magnitude of $[N_2(A)]$ for $A = A_c$, corresponding to the limit cycle condition, that is, if

$$\begin{aligned} \frac{4a_0}{\pi a_0} &\leq N_2(A_c) \\ &\leq (4a_0)/\pi A_c \end{aligned}$$

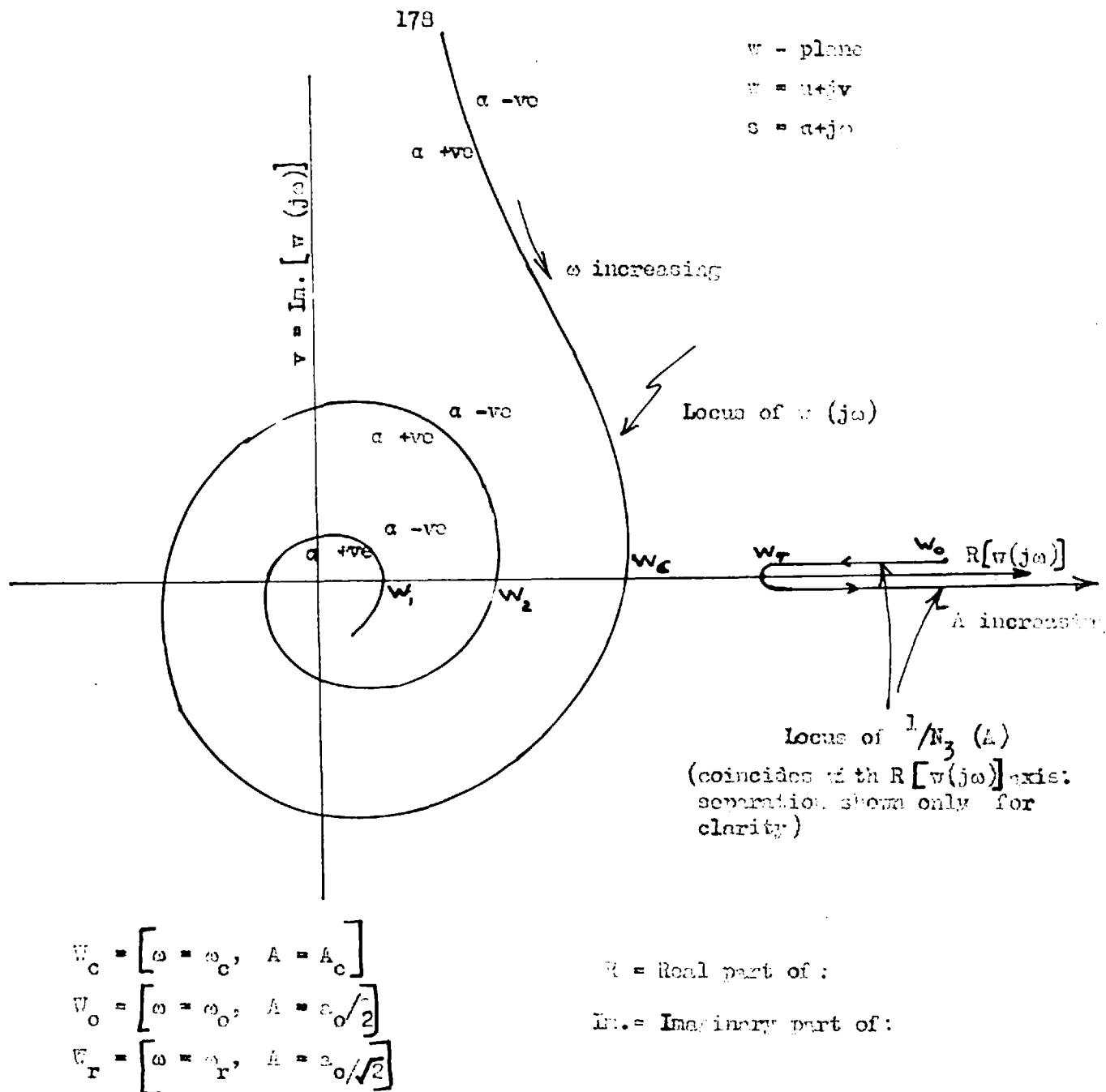
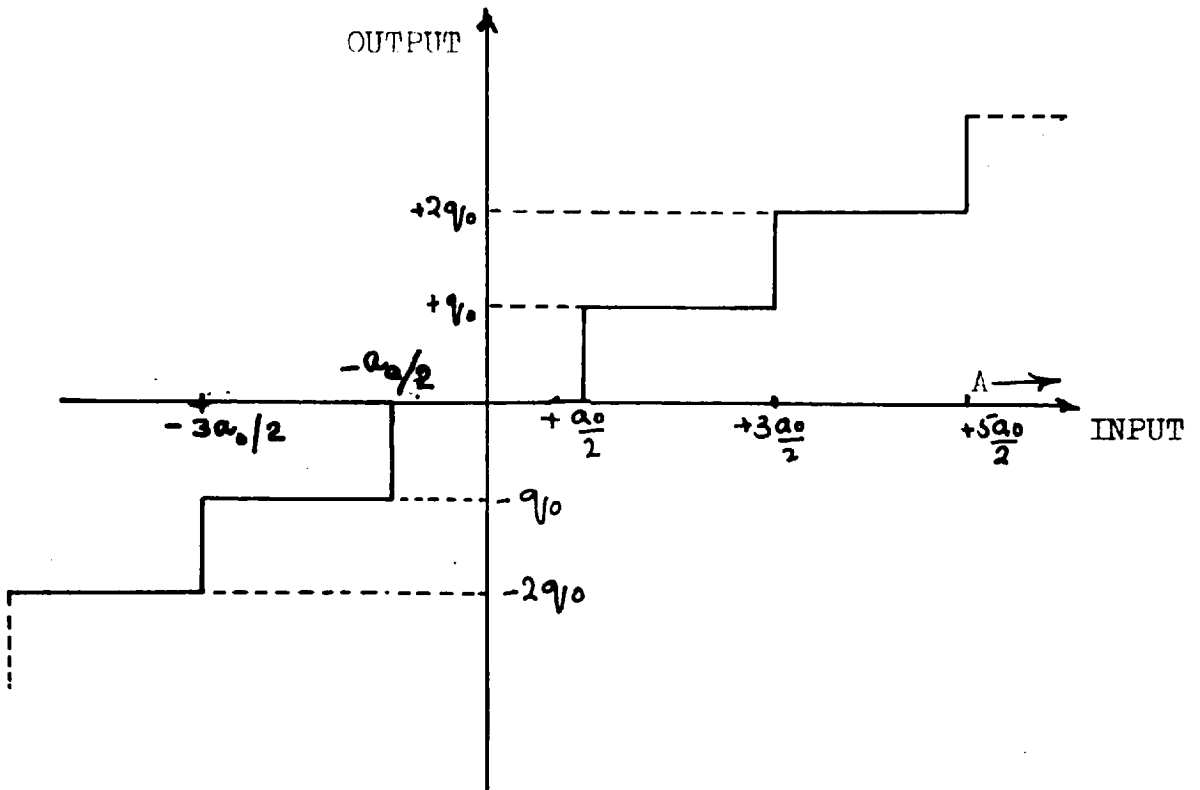


FIG. 7.15. STABILITY DIAGRAM FOR TERTIARY SYSTEM.



DESCRIBING FUNCTION:

$$N_l(A) = \frac{2q_0}{\pi A} \sum_{m=1}^I \left[4 - (2m-1)^2 \frac{a_0^2}{A^2} \right]^{1/2} \quad \text{for } 2A \geq a_0$$

where I is the largest integer contained in $\frac{1}{2} \left[\left| \frac{2A}{a_0} \right| + 1 \right]$,

$$N_l(A) = 0, \quad \text{for } 2A < a_0$$

FIG: 7.16: QUANTIZER NONLINEARITY : ℓ -LEVELS.

that is, if

$$a_0 \geq A_0, \quad \dots 7.104$$

then it is obvious that the loci of the two curves cannot intersect. Therefore, by proper choice of the dead-zone a_0 , for a given value of q_0 , the possibility of oscillation can be avoided in a ternary system.

The above result is especially important in considering the suitability of communication systems using 1-digit codes for telemetry or process-control applications.

7.7.3. Statistical Transfer Function of Quantization Nonlinearity.

Since, in a communication system, the input to the nonlinear element is a time series and not a sinusoidal function of given frequency and amplitude, the use of describing function techniques to study the overall performance of a system to a wide range of input signals, is of limited application. It has however been suggested by Booton (Ref.70) that the performance of a nonlinear element can be studied by assigning a 'linear gain function' to the element, the nature of the function being determined by the characteristics of the element as well as of the random input. If the exact functional dependence between the input $x(t)$ and the output $y(t)$ of the element is specified by the relation

$$y = f(x) \quad \dots 7.105$$

then it is assumed that the exact functional relationship can be replaced by a linear equation of the form

$$y(t) = k.x(t) + d(t) \quad \dots 7.106$$

if the constant of proportionality, k - called "the equivalent gain" of the element - is chosen such that the variance of the residual term $d(t)$ is a minimum for the type of input signal under consideration. In other words, if the probability density function of the input signal is $p(x)$,

then the constant k is chosen such that the integral

$$I = \int_{-\infty}^{+\infty} [f(x) - kx]^2 p(x) dx \quad \dots 7.107$$

$$= \int_{-\infty}^{+\infty} f^2(x) p(x) dx - 2k \int_{-\infty}^{+\infty} xf(x) p(x) dx + k^2 \int_{-\infty}^{+\infty} x^2 p(x) dx \quad \dots 7.108$$

is a minimum. The optimum value of k is obtained by solving

$$\frac{dI}{dk} = 0 \quad \dots 7.109$$

giving

$$k = \frac{\int_{-\infty}^{+\infty} x f(x) p(x) dx}{\int_{-\infty}^{+\infty} x^2 p(x) dx} = \frac{\int_{-\infty}^{+\infty} xf(x) p(x) dx}{\sigma_x^2} \quad \dots 7.110$$

where

σ_x^2 is the variance of the input signal $x(t)$.

If the assumption of linearity is again invoked, in a different context, to be able to assume that the probability distribution function of the output of the nonlinear element is substantially the same as that of the input except for a multiplication constant, then for a gaussian input to the system, the probability density function of the amplitude of the signal associated with any part of the feedback loop is also gaussian (Ref.70,71). Therefore, for a gaussian input

* If the number of levels of quantization is sufficiently large, with respect to the behaviour of the amplitude probability density function of the input (section 6), it is justifiable to assume that the shapes of the probability density functions of the input and the output are substantially the same.

to the communication system of fig. 7.1, the input to the quantizer is also gaussian. Under these conditions, the "equivalent gain" or the "statistical transfer function", k_3 , of a 3-level quantizer as specified by the relation

$$f(x) = \begin{cases} +q & x > +\frac{a_0}{2} \\ 0 & -\frac{a_0}{2} \leq x \leq +\frac{a_0}{2} \\ -q & x < -\frac{a_0}{2} \end{cases} \quad \dots 7.111$$

is given by

$$\begin{aligned} k_3 \sigma^2 &= \int_{+\frac{a_0}{2}}^{+\infty} q \cdot x \cdot \frac{1}{\sigma \sqrt{2\pi}} \cdot e^{-\frac{x^2}{2\sigma^2}} dx \\ &+ \int_{-\frac{a_0}{2}}^{+\frac{a_0}{2}} (0) \cdot x \cdot \frac{1}{\sigma \sqrt{2\pi}} \cdot e^{-\frac{x^2}{2\sigma^2}} dx \\ &+ \int_{-\infty}^{-\frac{a_0}{2}} \left(-\frac{a_0}{2}\right) (-q) \cdot x \cdot \frac{1}{\sigma \sqrt{2\pi}} \cdot e^{-\frac{x^2}{2\sigma^2}} dx \\ &= \int_{-\frac{a_0}{2}}^{+\infty} \frac{q}{\sigma \sqrt{2\pi}} \cdot x \cdot \frac{2\sigma^2}{2x} (-1) \cdot d \left\{ e^{-\frac{x^2}{2\sigma^2}} \right\} \\ &+ \int_{-\infty}^{-\frac{a_0}{2}} \frac{(-q)}{\sigma \sqrt{2\pi}} \cdot x \cdot \frac{2\sigma^2}{2x} (-1) \cdot d \left\{ e^{-\frac{x^2}{2\sigma^2}} \right\} \\ &= \sqrt{\frac{2}{\pi}} \cdot q \sigma e^{-\frac{(a_0/2)^2}{2\sigma^2}} \end{aligned}$$

Therefore,

$$k_3 = \sqrt{\frac{2}{\pi}} \cdot \frac{q}{\sigma} \cdot \exp \left[\frac{-a_0^2}{8\sigma^2} \right] \quad \dots 7.112.$$

It may be noted that for an ON-OFF nonlinearity (that is, if $a_0 = 0$) this expression becomes

$$k_2 = \sqrt{\frac{2}{\pi}} \cdot \frac{q}{\sigma} \quad \dots 7.113$$

which is the same as the expression given by Booton for the limiting case of a limiter whose range of operation in the linear mode tends to zero.

The statistical transfer function of an 1-level quantizer can be obtained in the same manner as for the 3-level quantizer, or by a superposition of the required number of 3-level nonlinearities of peak to peak amplitude (at output) of $2q$ and of dead-zones of magnitude $a_0, 3a_0, 5a_0, \dots, (l-2)a_0$:

$$k_l = \sqrt{\frac{2}{\pi}} \cdot \frac{q}{\sigma} \left[\sum_{m=0}^{\frac{l-3}{2}} e^{-\frac{[a_0(1+2m)]^2}{8\sigma^2}} \right] \dots 7.114$$

l is assumed to be odd.

By varying the magnitude of q_i as well as of a_i associated with any given 3-level nonlinearity $N_3(i)$ involved in the superposition operation, and using the corresponding statistical transfer function

$$k_3(i) = \sqrt{\frac{2}{\pi}} \cdot \frac{q_i}{\sigma} \cdot e^{-\frac{a_i^2}{8\sigma^2}} \dots 7.115$$

the above technique can be extended to cover the case of logarithmic quantization or of any other type of quantization (Ref.19) involving the use of nonlinear expansion - compression curves.

It may be mentioned here that the statistical transfer function of any nonlinear element possessing odd symmetry (including those not involving quantization) can be determined to any required degree of accuracy by adding the statistical transfer functions corresponding to the different 3-level nonlinear elements used to obtain, by superposition, an adequate approximation to the given nonlinearity. This is a much quicker method for obtaining the equivalent gain of the nonlinear element than the method suggested in Ref.72, which requires the evaluation of the very large number of integrals that are necessary to characterize the approximation obtained by splitting the given nonlinearity into an adequate number of **piecewise** linear segments.

Once the statistical transfer function of the quantizer is evaluated and the expected variance σ_{ϵ}^2 of the error signal at its input estimated (Ref.70), then the system as a whole can be considered as a linear network. Filtering and smoothing networks can now be introduced, into the system, if necessary (for instance, to discriminate against noise in the signal input to the system), making use of the standard techniques of linear network theory and random noise analysis (Ref.71, 72).

8. Tristable Circuits.

In a preliminary study of the problems involved in the instrumentation of a ternary system of pulse code modulation, one of the main problems that presented itself was that of the analogue-to-digital conversion of the magnitudes of the primary signal. It was proposed to make use of an encoder of the feedback type (see section 9) to perform this operation, a decoding network consisting of a suitable number of tristable circuit elements, each capable of being triggered to give any of three different levels of voltage across its output terminals, being used in the feedback loop of the encoder. Since the stability, accuracy and speed of operation of the encoder are directly dependent upon the corresponding characteristics of the decoder and therefore, of the tristable elements of which it is composed, it was found necessary to develop a reliable, fast tristable circuit conforming to the following specifications:

1. The circuit shall have 3 stable states.
2. The output voltage can assume any of one of three different, distinct levels, each voltage level corresponding to one stable state of the circuit.
3. By means of repeated triggering the circuit can be made to pass through the three stable states sequentially, so that a periodic staircase waveform is obtained at the output.
4. The circuit can be reset to any stable state by a single reset pulse, while the circuit is in either of the other two states.
5. The increments (or decrements) in voltage between successive output levels while the staircase is ascending (or descending), should be equal.
6. It should be possible to build up several circuits with almost identical characteristics, especially with respect to the magnitude of the output voltage levels, without

requiring the use of close matching of circuit components.

The various circuit configurations that were investigated may be classified under four main categories, namely

- A. Energy Storage Circuits (Ref.73)
- B. Combinations or matrices of bistable multivibrators.
- C. Tristable multivibrators based on the twin-triode Ecoles-Jordan trigger circuit. An additional nonlinear mode of operation, besides the two "conventional" modes of the bistable multivibrator, in which the loop gain of the circuit is less than unity, is obtained either through the introduction of extra nonlinearities, such as diodes, into the regenerative loop of the circuit or through the unconventional use of nonlinearities already inherent in the system, such as the grid-current nonlinearity associated with each multivibrating-valve. (Ref.74-77)
- D. Tristable circuits, using three valves, based upon the ring-counter principle.

Since the theory and development of these circuits have been discussed in detail by the present writer elsewhere (Ref:78), only a very brief outline of the theory of operation of the tristable circuit finally chosen for use in the encoder is given here.

The basic circuit is given in fig.8.1.

When the circuit is in state-2, valve V_1 is conducting and valves V_2 and V_3 are cut off. If a negative input pulse is now applied to the anodes P_1 , P_2 , P_3 of V_1 , V_2 and V_3 through the diodes D_1 , D_2 , D_3 , a negative pulse appears at P_2 and P_3 but not at P_1 since the anode of V_1 is at -27 volts which is considerably below the voltage level existing at the cathodes of D_1 , D_2 and D_3 . The negative pulse appearing at P_3 is capacitively coupled to the grid of V_1 which therefore begins to cut off. The positive voltage step appearing at the anode of V_1 appears at the grid of V_2 before it appears at the grid of V_3 , since there is capacitive coupling between P_1 and grid of V_2 , but there exists only resistive coupling between P_1 and grid

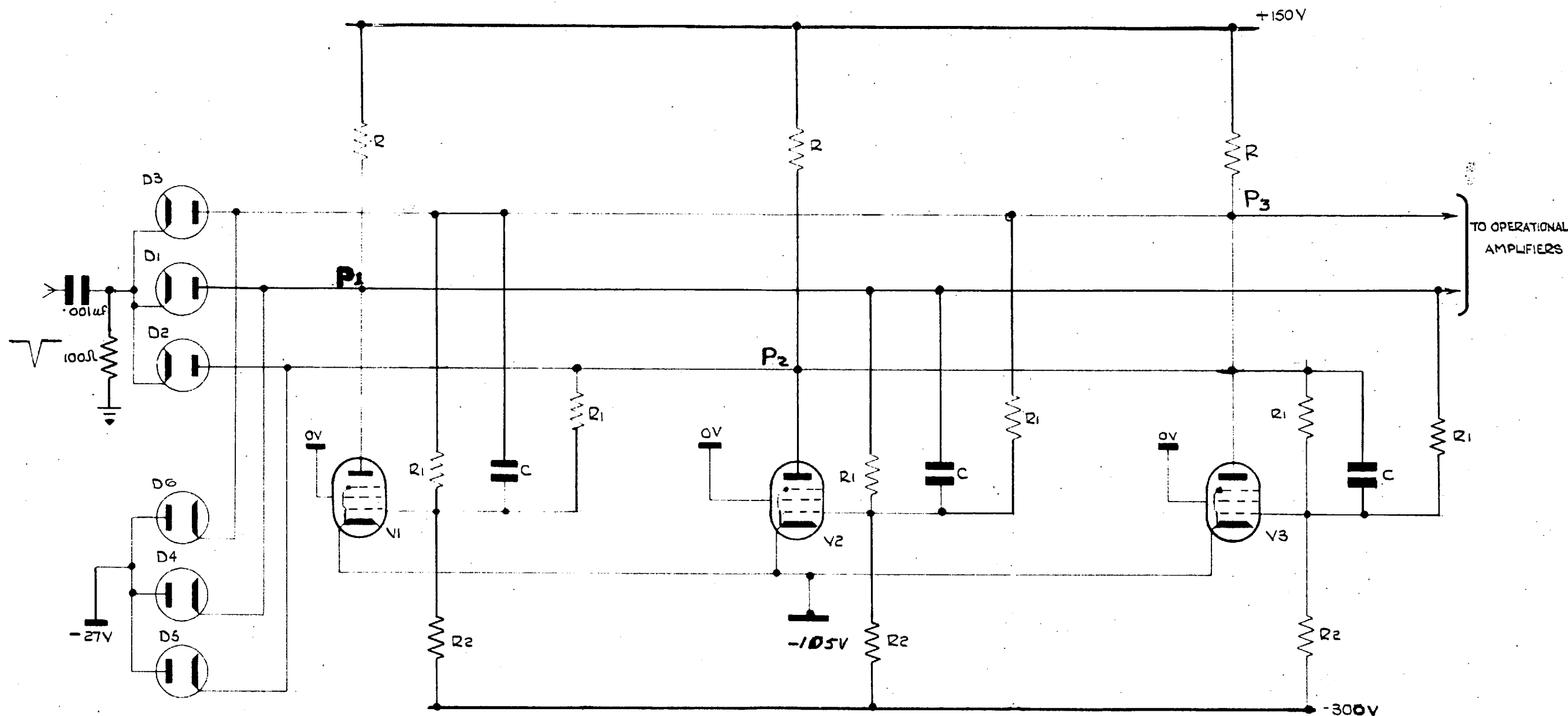
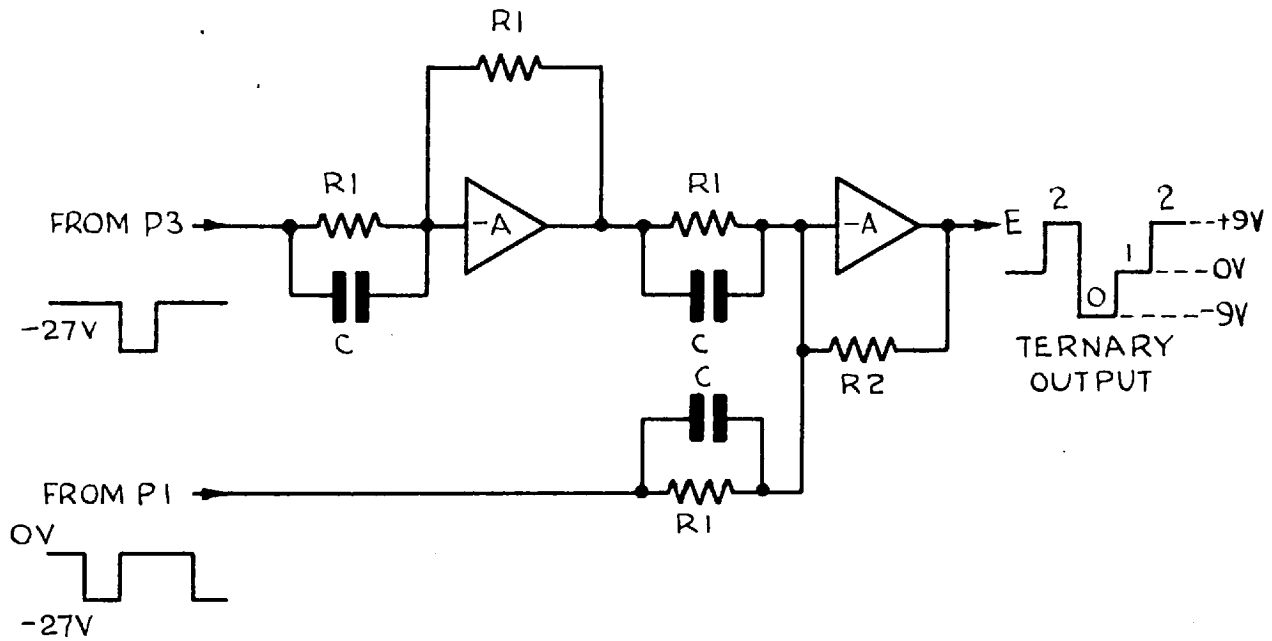

$$\begin{aligned} R &= 18 \text{ K} \\ R_1 &= 100 \text{ K} \\ R_2 &= 100 \text{ K} \\ C &= 15 \text{ pF} \end{aligned}$$

FIG: 8.1

D1, D2, D3, D4, D5, D6: 1/2 GALS



 = OPERATIONAL AMPLIFIER
 OF GAIN (-A)

R_1 = 100 K

R_2 = 33.3 K

C = 0.5 pF

FIG: 8.2

of V_3 . V_2 therefore begins to conduct and the negative step appearing at P_2 is immediately applied to the grid of V_3 through the coupling RC network, thus ensuring that valve V_3 remains cut-off. After the transients have decayed, the circuit settles down into the new stable state (State 1) in which valve V_2 is conducting and valves V_1 and V_3 are cut-off.

Applying the same reasoning as given above, it is easily shown that the application of another negative pulse would cause the circuit to switch to State-0, in which valve V_3 alone conducts. The circuit can now be switched back to State-2 by applying another negative pulse, the circuit thus passing through one complete cycle of ternary states.

The positive and negative plate voltage excursions at P_1 , P_2 and P_3 are limited to +0 volts and -27 volts respectively, using the clamping diodes D_1 , D_2 , D_3 , D_4 , D_5 , D_6 .

The circuit can be reset to any required state by applying a positive pulse to the grid of the appropriate valve.

A ternary output, defined by three distinct voltage levels, can be obtained by feeding the outputs of valves V_1 and V_3 into the system of operational amplifiers shown in fig.8.2.

The necessity for operational amplifiers is avoided in the tristable circuit given in fig.8.3. The output of this circuit when in State-2, is the same as the supply voltage, +150V, since the conditions at the output are unaffected by the conduction in valve V_1 . In State-1, however, valve V_2 conducts causing current to flow through the parallel resistive network $(r) (r + r)$ in series with the plate resistor R of V_2 . Since only half as much current passes through $(r + r)$ as through r , the voltage at the output is exactly half that at the junction of r and the plate resistor R of V_2 . In State-0, valve V_3 conducts, Now, however, the full voltage drop across the parallel resistive network $(r) (r + r)$ is available as the output, thus ensuring that the voltage difference between the outputs for State 0 and State 2 is twice that for State 0 and State 1.

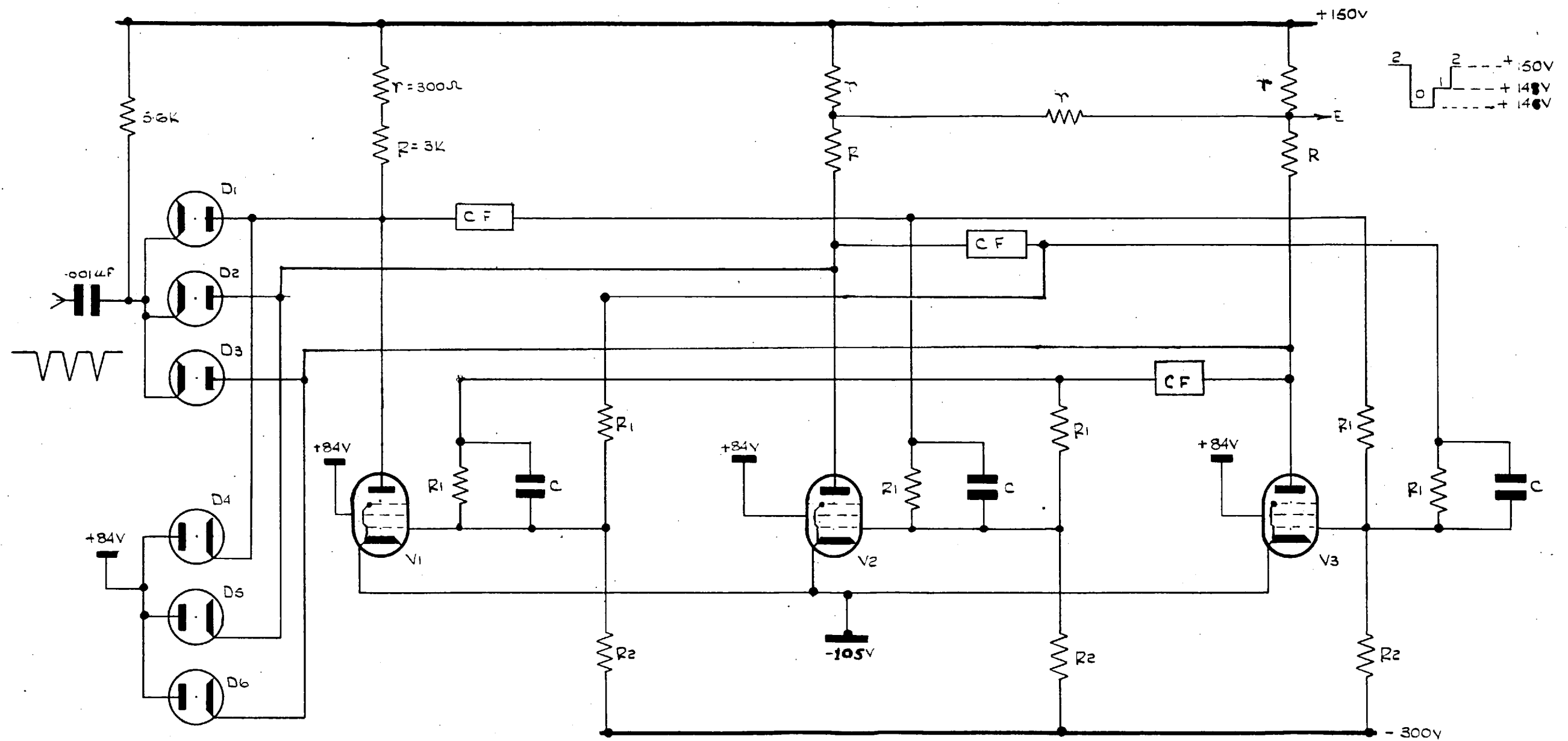
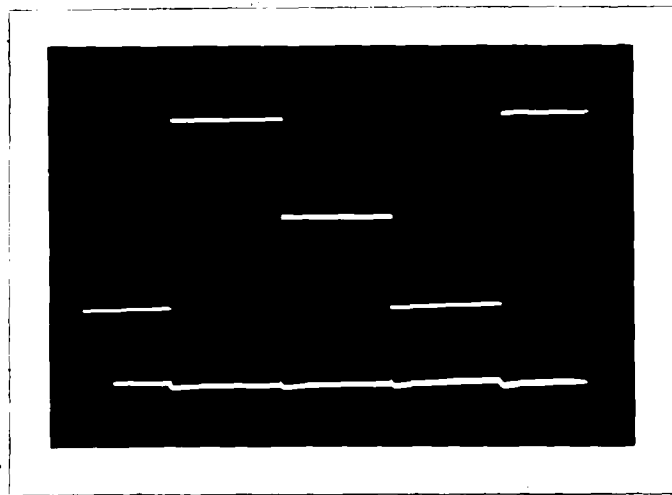


FIG: 8.3

V1, V2, V3 : EF91
 D1, D2, D3, 1/2 6AL5
 D4, D5, D6
 CF: CATHODE FOLLOWER



OUTPUT
WAVEFORM OF
FLIP-FLAP-FLOP.

TRIGGER
PULSES.

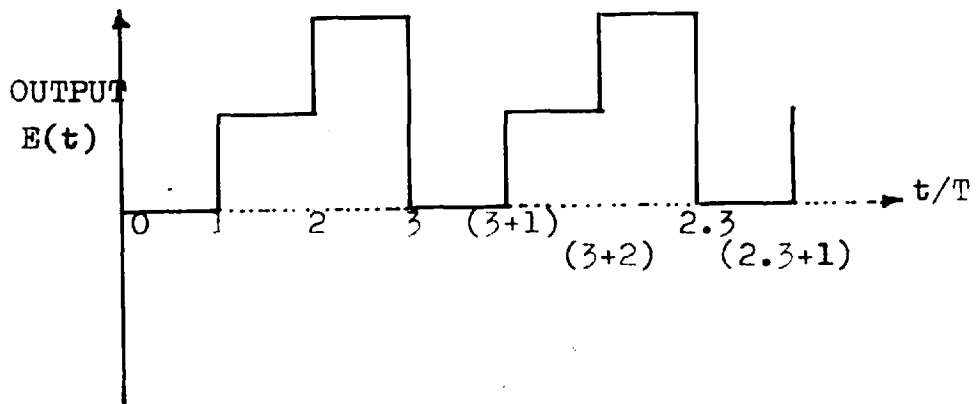
TIME SCALE : 5 MICROSECONDS PER CM.

Fig: 8.4

TRISTABLE CIRCUIT WAVEFORMS

The coupling between the plate resistors and the resistance capacitance voltage divider networks is made through cathode followers. This not only ensures faster switching but also defines the output voltage for State-2 precisely, since no ~~add~~ current is now supplied to the R-C voltage divider networks. The maximum negative excursion at the anodes of V_1, V_2, V_3 is limited to $+84^V$ by means of the clamping diodes D_4, D_5, D_6 . This ensures that the current through the plate resistor R of the conducting valve is the same for all three states.

The ternary output waveform of the circuits given in figures 8.1 and 8.2, is shown in fig.8.4.



$E(t)$ = Remainder of the division of the variable integer $[t]$ by the fixed integer 3, where $[t]$ means the greatest integer contained in t .

$$\begin{aligned}
 &= [t] - 3 \cdot [t/3] \\
 &= \sum_{k=1}^{k=2} -\frac{1}{2} \frac{\sin \{ [t] + (1/2) \} \cdot (2\pi k/3)}{\sin (\pi k/3)}
 \end{aligned}$$

FIG: 8.5: OUTPUT OF TRISTABLE CIRCUIT.

9. Ternary Encoder.

Three methods are available for encoding an analogue signal into digital form:

- A. Counting. The magnitude of the sample is determined by counting the number of units contained in it one by one, until the difference between the magnitude of the sample and that corresponding to the total count is less than half a unit. (Ref.7,10).
- B. Feedback Subtraction. The magnitude of the sample is compared with that corresponding to one code digit after another, proceeding from the most significant to the least, and subtracting an amount equal to the magnitude of the given digit from that of the sample each time the digit value is found to be smaller than the magnitude of the sample or of its residue from the previous subtraction. The procedure continues until the residue is less than half a unit. (Ref. 8,81).
- C. Coding Tube Method. Magnitude of the sample is measured in a single step by comparison with a set of standard values (Ref. 11, 82).

Of the three methods mentioned, the last is by far the fastest. It is also possible, using this technique, to present the output with the coded digits arranged in ascending order of significance with the least significant digit occurring first and the most significant last. This makes it possible to introduce considerable simplification in the decoding process (Ref: 79, 80). However, this method is rather more costly to implement than the other two. Therefore, it was decided to base the development of the ternary encoder on the feedback principle, since this method is much faster and more flexible than the counting method (Ref:81).

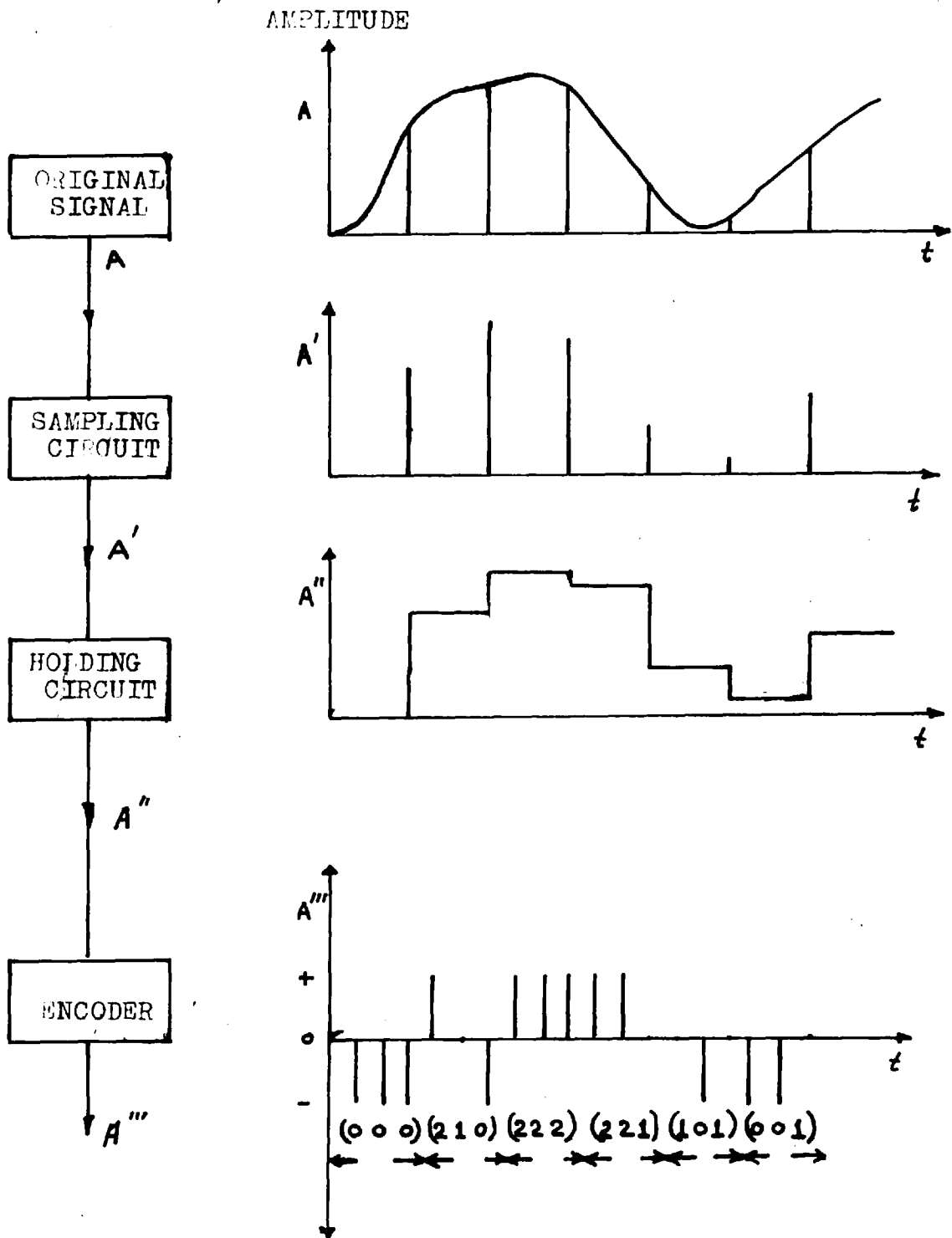


FIG.9-1: TERNARY ENCODING.

Two different ternary encoders, Encoder I and Encoder II, each capable of encoding the amplitude of a rectangular voltage pulse into a sequence of ternary digits, are described. Of these, only a modified version (using 2 coded digits per sample) of Encoder II was actually constructed and tested.

9.1. Encoder I.

The operation of Encoder I is described below:

Samples from the holding circuit arrives at the input A of the difference amplifier DA, at the rate of 10 Kc/s. Outputs of the tristable circuits TSC I, TSC II and TSC III are added together in the ratio 9:3:1 respectively, and this combined output is applied to the input B of the difference amplifier. The voltage developed at B is proportional to the magnitude of the ternary number represented by the states of the three tristable circuits I, II and III. For instance, if the tristable circuits I, II and III are in states 2, 0 and 1 respectively, the output voltage at B would be equal to

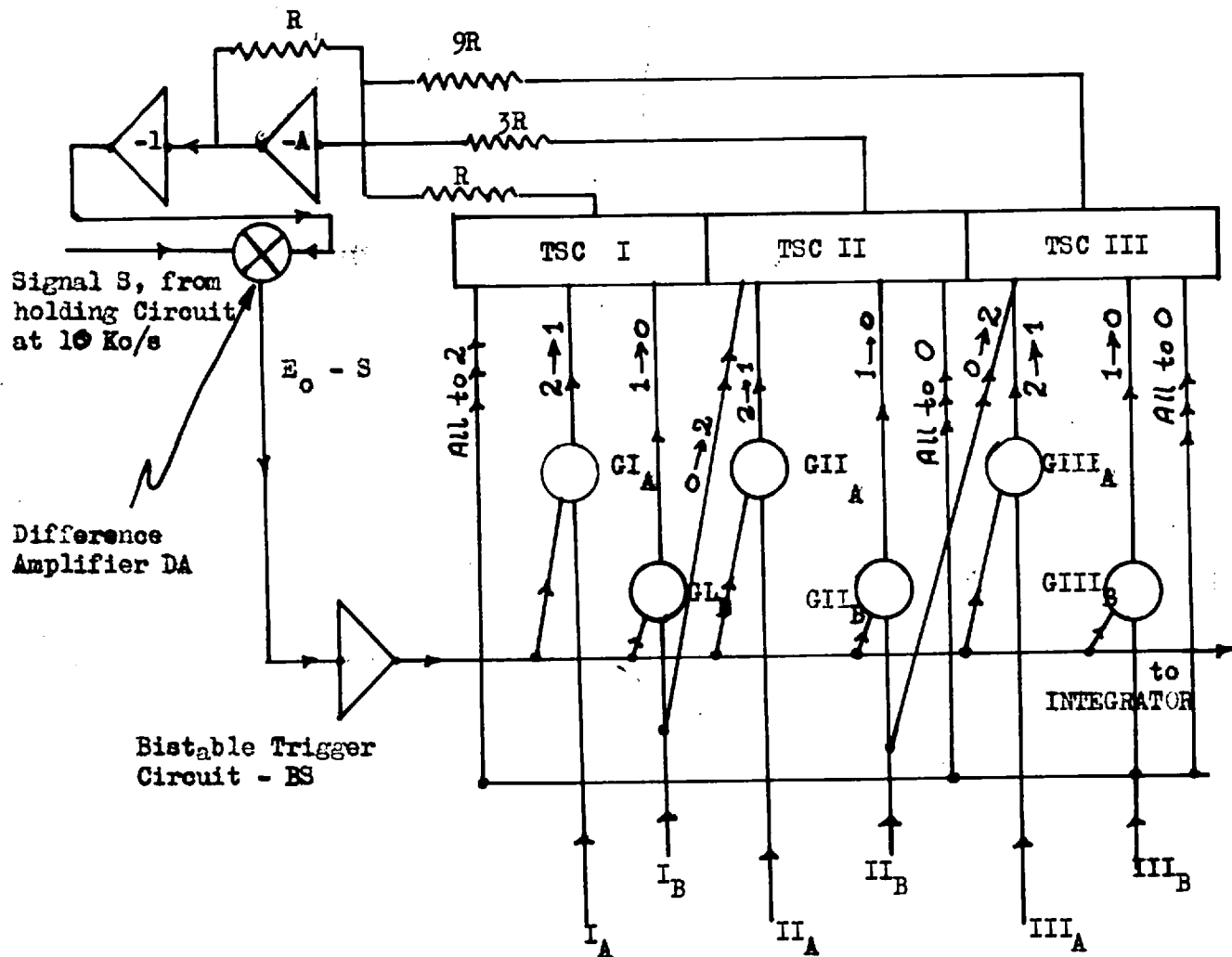
$$= k(2 \times 3^2 + 0 \times 3^1 + 1 \times 3^0)$$

$$= k(18 + 0 + 1)$$

$$= 19k,$$

where k is a constant of proportionality.

Encoding is performed by triggering the tristable circuits in such a manner as to reduce to a minimum the difference between the signal voltage at A and the output voltage at B of the three tristable circuits. This is accomplished by making use of the output voltage of the difference amplifier DA to ^{control the gating of} the trigger pulses I_A , I_B , II_A , II_B , III_A and III_B which occur at a fixed repetition frequency of 10 Kc/s, with an interval of $1/(60 \times 10^3) = 16.6 \mu \text{ sec.}$ between successive pulses, as described below.



$G_{IA}, G_{IIA}, \dots, G_{IIIA}, G_{IIIB}$ = Gate
 TSCI, TSCII, TSCIII = Tristabl
 Circuits

— = Gating pulses

$\Rightarrow \Rightarrow \Rightarrow$ = Reset pulses

Fig. 9.2: Encoder I

Trigger pulses at 10 ko/s.

Interval between pulses = $1/60 \times 10^3$ sec

= 16.6 μ Sec.

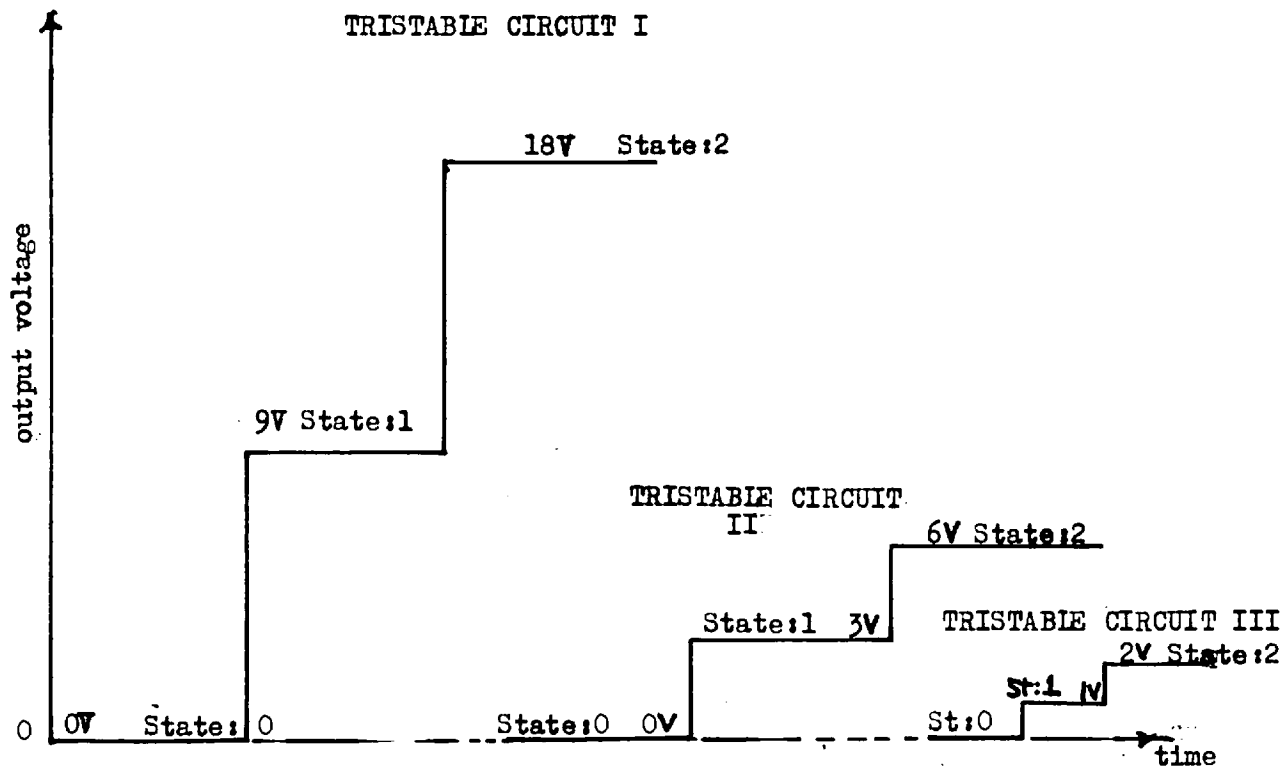
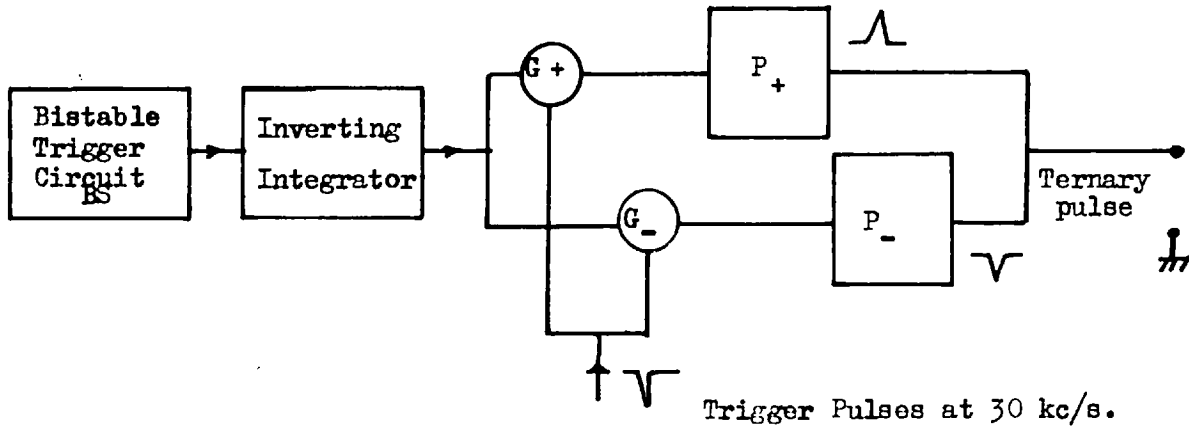


Fig. 9.3.

Output voltages developed by TSCI, TSCII, and TSCIII at input B of Difference Amplifier DA.

State 0 potentials are assumed to be at earth potential, so that the ternary number 000 corresponds to zero, 111 corresponds to $9+3+1=13$; 222 corresponds to $18+6+2=26$



BS = Bistable circuit of Fig. 9.3.

G_+ = Gate, open for +ve gating voltage

G_- = Gate, open for -ve gating voltage

P_+ = Positive polarity pulse generator

P_- = Negative polarity pulse generator

Fig. 9.4 : Ternary Pulse Generator [ENCODER I]

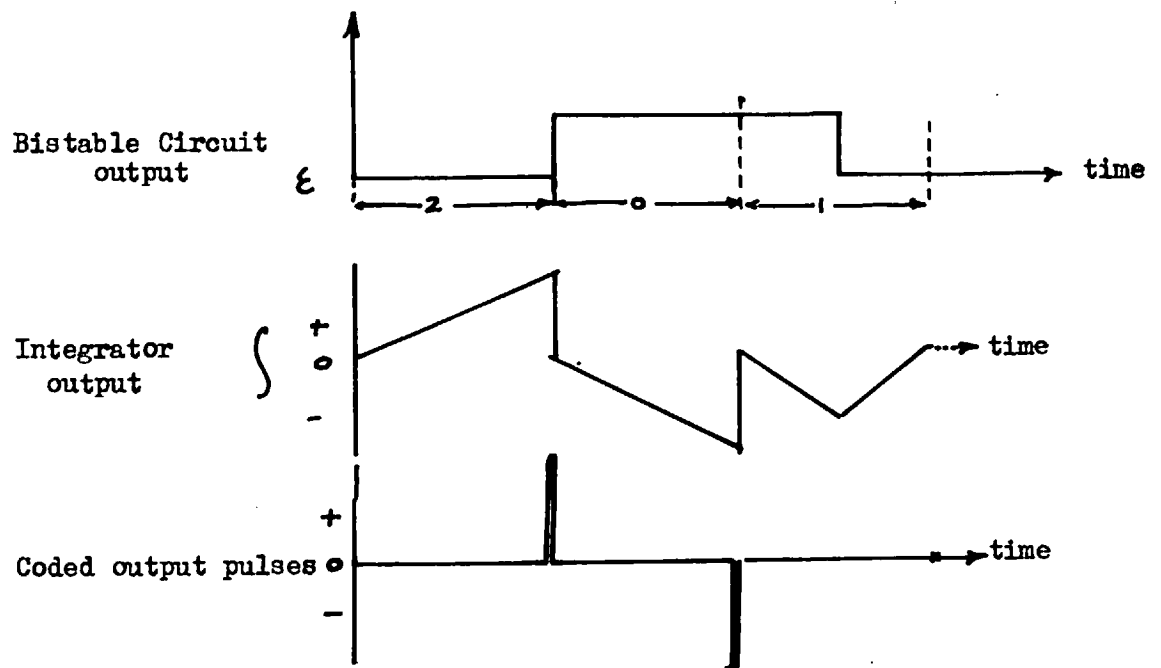


Fig. 9.4A : Ternary Coding of Sample of Amplitude 19^V

Initially, TSC I is set to state 2 and both TSC II and TSC III to state 0. The input at B would therefore be proportional to the magnitude of the ternary number 200. If the voltage at B is greater than the magnitude of signal at A, a positive difference signal is developed by DA. This triggers the bistable circuit BS, producing a positive step at the output of BS which is coupled to the input of the gates $G I_A$, $G I_B$, ..., $G III_A$ and $G III_B$, enabling the gates to pass any trigger pulse that may be applied to them now. Therefore, when the trigger pulse I_A occurs, it passes through the gate $G I_A$ and sets back TSC I from state 2 to state 1. If the signal voltage lies between 000 and 022, the voltage at B would be still greater than the signal voltage at A, so that the error voltage would continue to be positive. Consequently, when pulse I_B occurs, it will pass through gate $G I_B$ which remains open, and set back TSC I to state 0 from state 1.

On the other hand, if the signal voltage at A is greater than the magnitude of the ternary number 200 as represented by the voltage at B, the error signal developed by DA would be negative. Consequently, the bistable circuit BS would not trigger with the effect the pulse I_A does not pass through the gate I_A which remains closed. Therefore, TSC I remains in state 2. As the error signal continues to be negative, the gate $G I_B$ would not be open when the pulse I_B arrives. Therefore, TSC I remains in state 2, unaffected, signifying that the signal voltage is greater than 200.

As a third alternative, suppose that the signal voltage lies between 100 and 200. In this case, at the beginning of operations, the bistable circuit BS would trigger applying a positive voltage to gate $G I_A$. Therefore, the pulse I_A would trigger TSC I from state 2 to state 1. But thereafter the voltage at B would be less than the signal voltage. Consequently, gate $G I_B$ would not be open to let pulse I_B through. Therefore, TSC I remains in state 1.

From the above discussion it can be seen that at the output of the bistable circuit BS there would be

1. No change, if the signal voltage is greater than 200.
2. A positive rectangular pulse followed by a region of no change of equal duration, if the signal voltage lies between 100 and 200 and
3. Two positive rectangular pulses merging into each other if the signal voltage is less than 100.

The output of the bistable circuit is now passed through a phase - inverting, integrator circuit. If the reference datum for the output of the integrator is taken as that corresponding to case 2, the integrator output, at the beginning of pulse I_B , would be

positive for case 1,
 zero for case 2 and
 negative for case 3.

The output of the integrator is used to gate trigger pulses to two pulse generators, a positive pulse being obtained in case 1 and a negative pulse in case 3.

The pulse I_B , besides triggering TSC I, if necessary, from state 1 to state 0, also resets TSC II to state 2 from state 0, thus enabling the encoder to commence the coding operations for obtaining the second ternary digit, which are the same as for the first digit.

Similarly, at the end of the encoding for the second digit, the pulse II_B also resets TSC III to state 2 from state 0, so that the encoding for the third digit may follow.

The circuit shown encodes the signal only correct to three ternary digits. The signal has been completely encoded when pulse III_B is over. This pulse is therefore also used to reset the various tristable circuits to their initial states, namely, to reset TSC I to state 2, and TSC II and TSC III to state 0. At the same instant, a new sample

arrives at the input A of the difference amplifier, from the holding circuit, and the encoding processes are repeated for the new sample.

From the above discussion it can be seen that two separate operations are necessary to encode each ternary digit. In other words, to encode a sample into a sequence of (n) ternary digits, (2n) trigger pulses, equally spaced in time, would be necessary. This is because the present system makes use of two binary decisions to determine the magnitude of each ternary digit, each of the binary decisions being made individually and sequentially, instead of simultaneously.

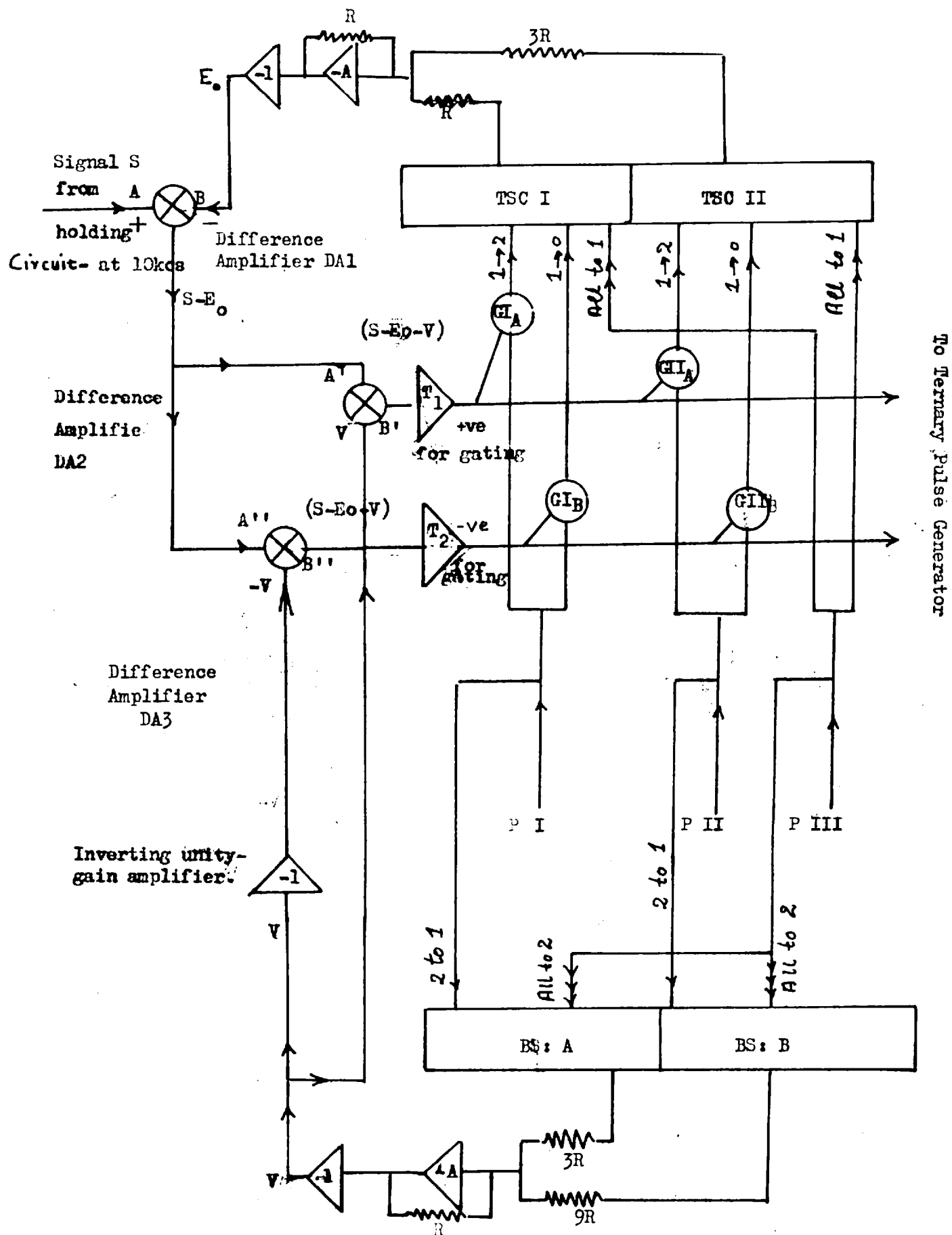
A more efficient system is described in the next section.

9.2. Encoder II.

In this encoder, the position of zero potential of the signal is taken to be in the centre of the dynamic range of the signal. In other words, if the ternary number 222 corresponds to the extreme positive excursion of the signal, and 000 corresponds to the extreme negative excursion, then the number 111 corresponds to no-signal conditions. This arrangement makes it possible to determine the magnitude of a ternary digit - namely whether it is greater, equal to or less than 1- in a single step as described below.

If the first digit of a ternary number corresponding to a given sample is 1, then the maximum and minimum values the number can have are 122 and 100 respectively; if the maximum value of the number is greater than 122, the first digit is 2, and if the minimum value is less than 100, the first digit is zero. In other words, if the magnitude of the first digit is 1, then the maximum excursion of the signal from the zero or earth potential (corresponding to no-signal conditions) cannot be greater than $\pm (122 - 11)$ or $\pm (122)^*$

* Note: According to the present choice of scale, $(122 - 111) \neq 011$, as given in the discussion for Encoder I, but is equal to 122.



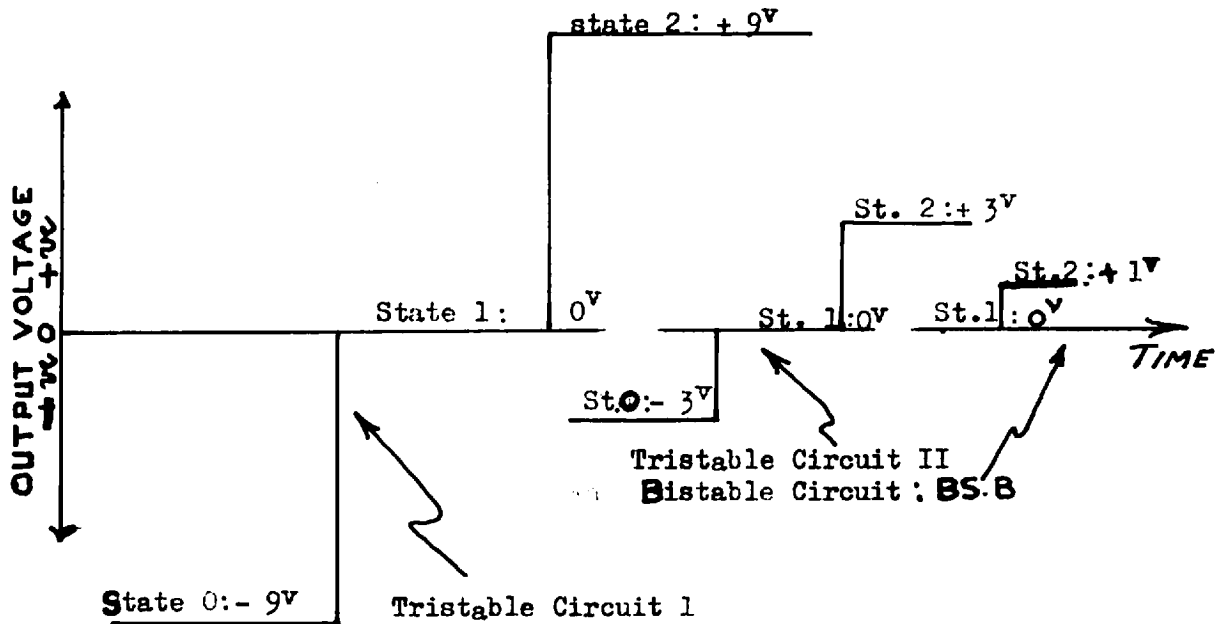
PI, PII, PIII = Trigger pulses at 10 kc/s.

Interval between pulses = 33.3 μ Sec

T₁, T₂ = Trigger Circuits

DA1, DA2, DA3 = Unity Gain Difference Amplifiers

9.5.00 ENCODER II



Output voltages of TSCI & TSC II:

Measured at point B of fig. 9.5

All voltages are measured assuming state - 1 potentials of the Tristable circuits to be at earth potential, so that the ternary number 111 corresponds to zero

$$222 \quad " \quad " \quad 9 + 3 + 1 = 13v$$

$$000 \quad " \quad " \quad -9 - 3 - 1 = -13v$$

FIG. 9. 6.

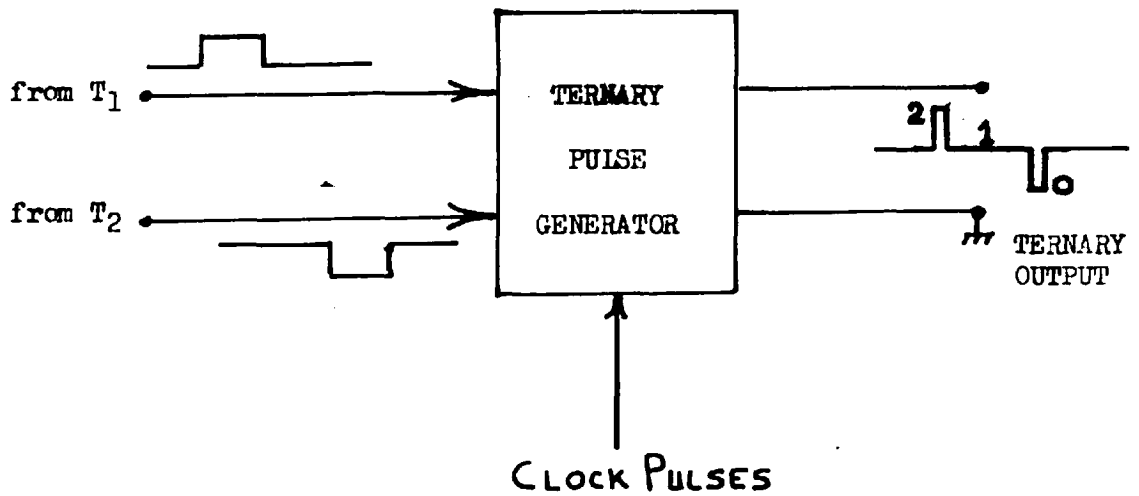


FIG. 9.7. TERNARY PULSE GENERATOR [ENCODER II]

Accordingly, in the present Encoder, the incoming signal is compared with a voltage of magnitude 122. If the signal is greater than 122, obviously, the first digit of the number is 2, not 1. Simultaneously, the signal is compared with a voltage of magnitude $-(122) = 100$, which is obtained by inverting the comparison voltage (+122). If the signal is less than 100, then the first digit of the number is zero and not 1. On the other hand, if the signal voltage lies between + 122 and - 122, that is between 122 and 100, then the first digit of the ternary number corresponding to the signal, is 1.

Once the first digit has been evaluated, the second digit can be determined as before after a voltage corresponding to the magnitude of the first digit has been subtracted from the original signal voltage. For instance, if the first digit is 2, $3^2 = 9$ volts is subtracted from the signal; if the first digit is 0, $(-3^2) = (-9)$ volts is subtracted, if, however, the digit is 1, no voltage need be subtracted as 1 corresponds to zero voltage. This subtraction process is necessary to ensure that the residual voltage would not be greater than the magnitude corresponding to a first digit, so that comparison of voltages may again be performed at zero or state -1 level. If the second digit is 1, this residual voltage cannot be greater than $(112 - 111) = 112$ or less than $(110 - 111) = -112$. Therefore, this residual voltage is simultaneously compared against voltages of magnitude 112 and - 112 to determine the second digit.

When the second digit is evaluated, the process is repeated by subtracting from the residual signal a voltage proportional to the value of the second digit and comparing this against voltages of $(111 + \delta/2)$ and $(111 - \delta/2)$ where $\delta = (112 - 111) = (111 - 110)$. $\int_{\tau_{ct}}$ In a practical encoder, such as given in fig.9.5, it is necessary to use an additional tristable circuit to provide the values of $+\delta/2$ and $-\delta/2$; this is done by properly setting the triggering levels of the trigger circuits T_1 and T_2 —7

A circuit which performs the above operations is shown in fig.9.5. The output (S) from the holding circuit is applied to the input A of the difference amplifier DA 1. The outputs (E_0) of the tristable circuits TSC I and TSC II are added together in the ratio 3 : 1 respectively, and applied to input B of DA 1. Initially, both TSC I and TSC II are in state 1. Therefore the output of DA 1 is equal to $(S - E_0) = S$, namely, the signal only. This is presented to the inputs A' and A'' of the differential amplifiers DA 2 and DA 3. The outputs (V) of the bistable circuits BS A and BS B are added together in the ratio 9 : 3 respectively, and applied to the input B' of DA 2 directly. The output V of BS:A and BS:B is also applied after inversion through a unity gain amplifier, to the input B'' of DA 3. Initially, both BS:A and BS:B are in state 2* and their combined output V corresponds to the ternary number 122. The outputs (S-V) and (S+V) of the differential amplifiers will therefore be positive and negative respectively according as the signal is larger than 122 or smaller than 100. The output of DA 2 is applied to the trigger circuit T_1 which triggers if $(S-V) > (\delta/2)$. The output of DA 3 is applied to the trigger circuit T_2 which triggers if $(S+V) < (\delta/2)$. The output of T_1 is a positive step while that of T_2 is a negative step.

The outputs of T_1 and T_2 are coupled to the gates ϕI_A and ϕI_B respectively, so that pulse P I is allowed to pass through ϕI_A if T_1 has triggered, or through ϕI_B if T_2 has triggered. If the signal is greater than 122, T_1 triggers, as explained earlier, and PI, when it arrives, triggers TSC I from state 1 to state 2, thereby subtracting a voltage of

* The two possible states of BS:A and BS:B are denoted by State 2 and State 1, the former indicating the positive output condition, while the latter denotes the absence of any voltage at the output terminals.

magnitude (9^V) corresponding to the value of the first digit (2) from the signal. On the other hand, if signal is smaller than (-122), T_2 triggers and P I sets TSC I to state 0, thereby subtracting a voltage of magnitude (-9^V) corresponding to the value to the first digit 0, from the signal. The same pulse PI resets BS:A to state 1, so that the residual signal is compared now with voltages equal to 112 and -112, instead of 122 and -122, as before.

The magnitude of the first ternary digit is found out by sensing the outputs of the trigger circuits T_1 and T_2 . A positive step at T_1 indicates that the first digit is 2, while a negative step at T_2 indicates that the digit is 0. The outputs of T_1 and T_2 are fed into a ternary pulse generator which gives a positive output pulse if the output of T_1 is positive, a negative output pulse if the output of T_2 is negative and no output at all if the outputs of T_1 and T_2 are negative and positive respectively.

The second and third digits are determined in the same manner as the first digit. At the end of the coding of the second digit, the pulse P II resets BS:B to state 1, while at the end of the coding of the third digit, the pulse P III resets both TSC I and TSC II to state 1 and BS:A and BS:B both to state 2.

The actual waveforms at various points of the encoder are given in figures 9.8, 9.9 and 9.10.

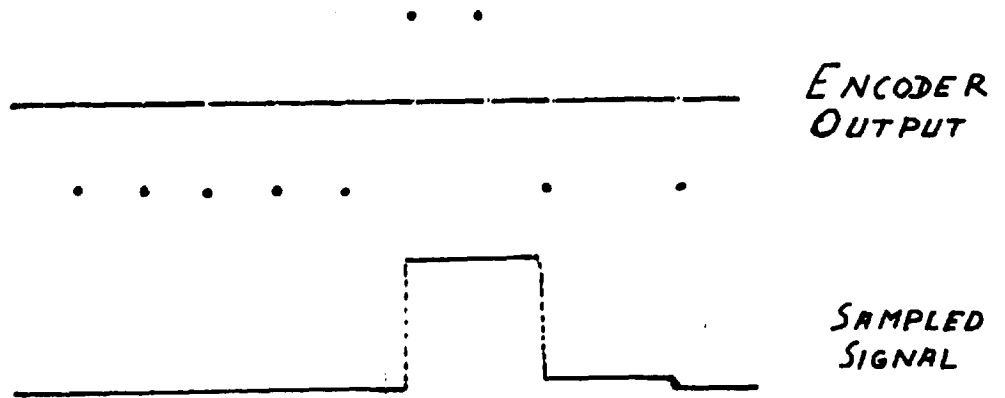
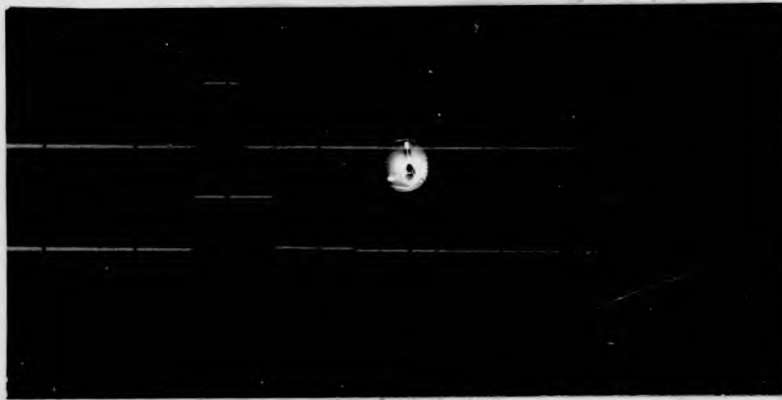


Fig: 9.8.

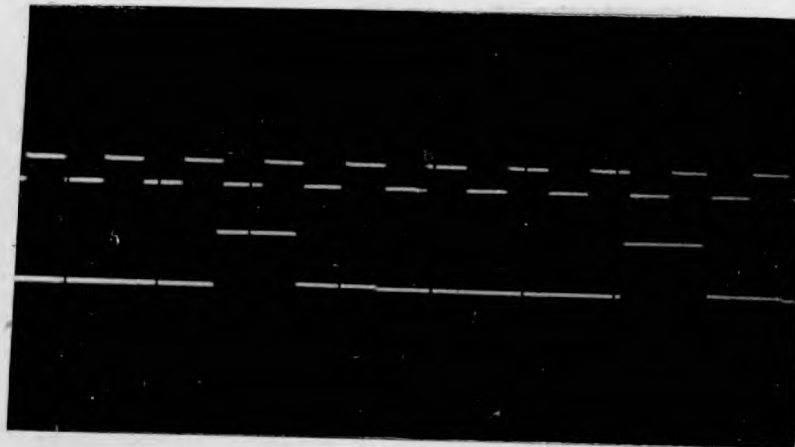
OUTPUT OF ENCODER II

(TRACING FROM PHOTOGRAPH)



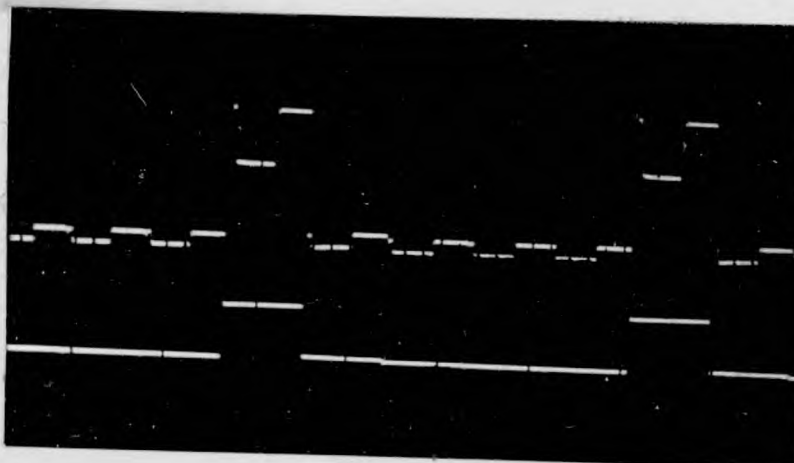
E

SAMPLED SIGNAL: S

FIG. 9.9A

V

SAMPLED SIGNAL: S

FIG. 9.9B $(S - E + V)$

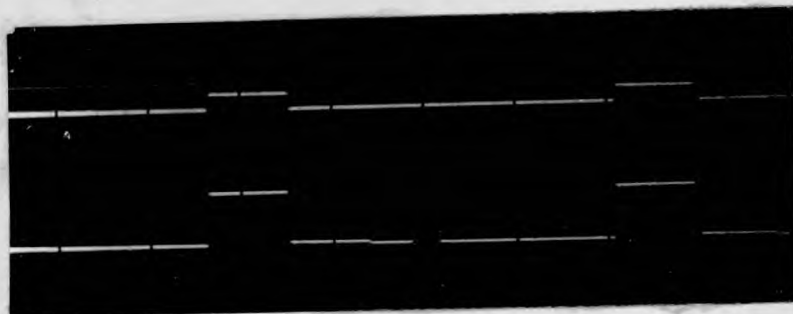
SAMPLED SIGNAL: S

FIG. 9.9C



(S-E-V)

SAMPLED SIGNAL: S

FIG. 9.9D(S-E+V):
AFTER TRIGGERING

SAMPLED SIGNAL: S

FIG. 9.9E(S-E-V):
AFTER TRIGGERING

SAMPLED SIGNAL: S

FIG. 9.9F

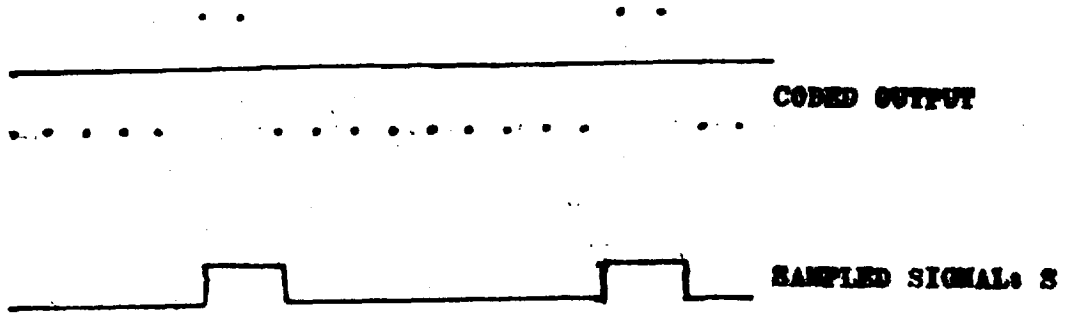
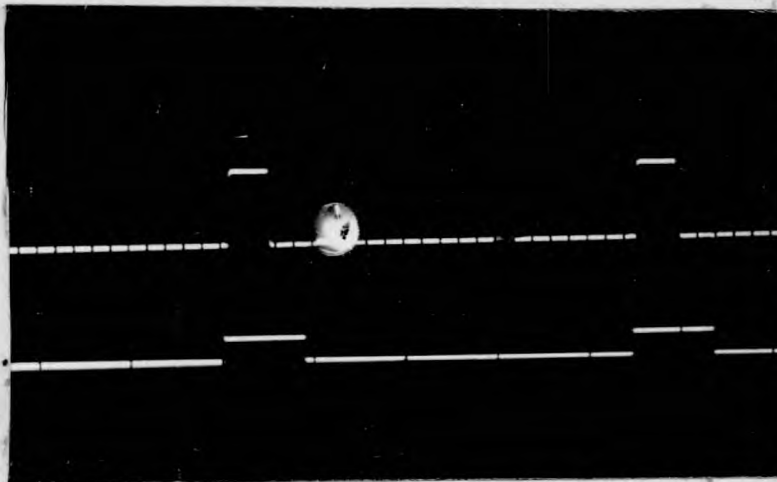


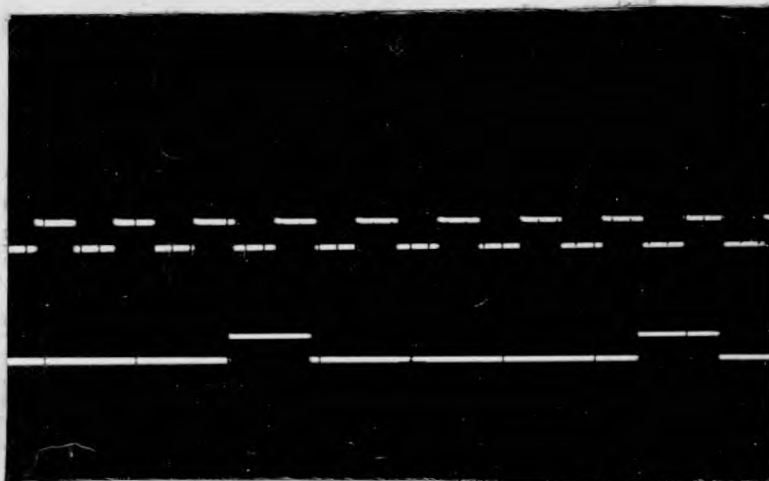
FIG: 9.98. CODING OF TERNARY CODE GROUP (2 2)

[TRACING FROM PHOTOGRAPH]



E

SAMPLED SIGNAL: S

FIG. 9.10 A.

V

SAMPLED SIGNAL: S

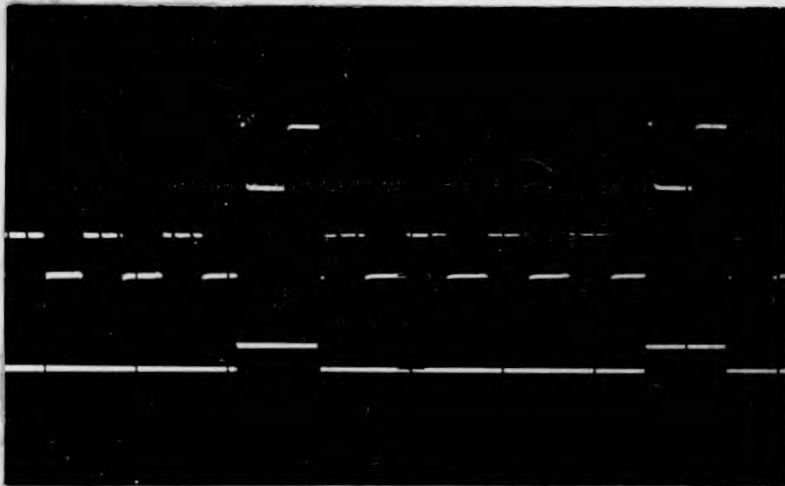
FIG. 9.10 B.

(S-E+V)

SAMPLED SIGNAL: S

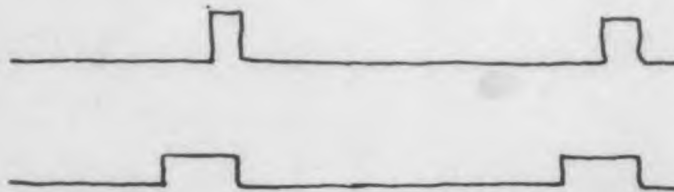
FIG. 9.10 C.

(S-E-V)



SAMPLED SIGNAL: S

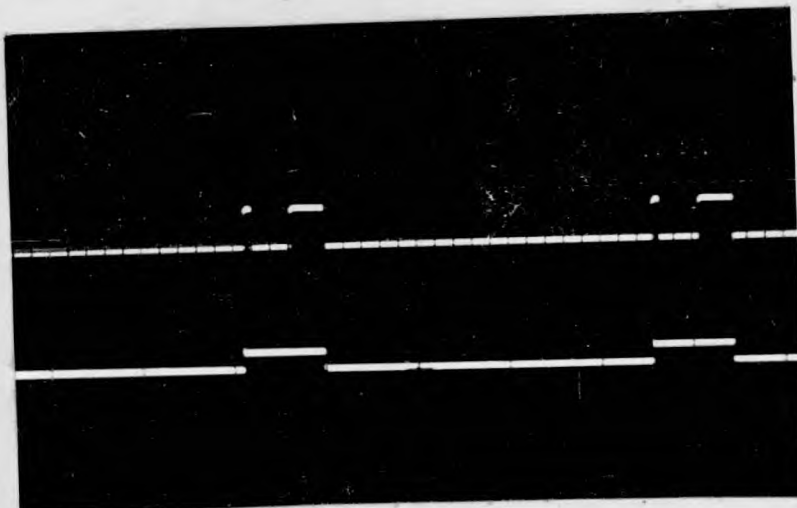
FIG. 9.10 D.

(S-E+V)
AFTER TRIGGERING

SAMPLED SIGNAL: S

[TRACING FROM PHOTOGRAPH]

FIG. 9.10 E.

(S-E-V)
AFTER TRIGGERING

SAMPLED SIGNAL: S

FIG. 9.10 F.

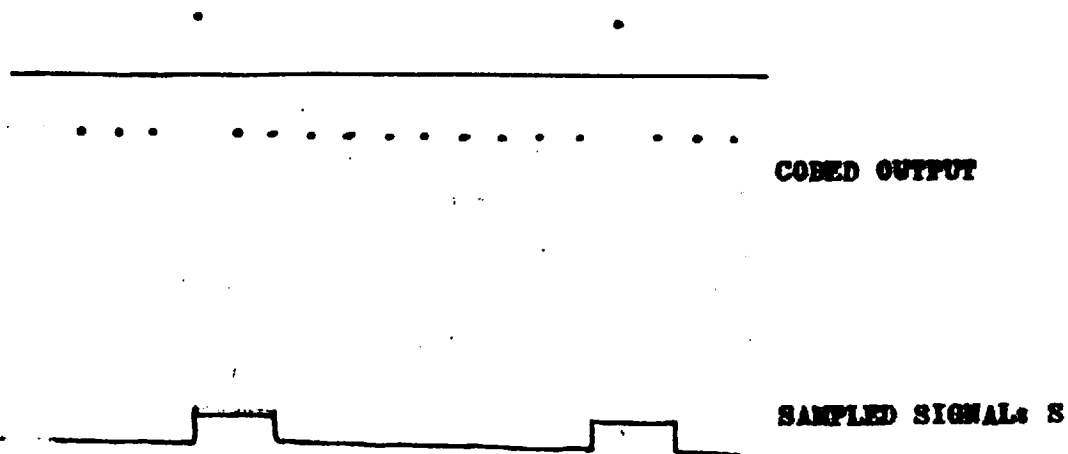


FIG. 9-10 G. CODING OF TERNARY CODE GROUP (2 0)

[TRACING FROM PHOTOGRAPH]

10. Ternary Decoder.

If the output of the encoder had been such that the least significant digit of the code group occurs first and the most significant digit last, then it is possible to use a modified form of the Shannon Decoder (Ref.79,80) to decode the received coded signal, as shown in fig. 10.1. The received noisy signal is regenerated and applied to a ternary current generator circuit which supplies an instantaneous pulse of either (+I), 0 or (-I) units of current to an RC network, depending upon whether the received code digit is in state 2, 1 or 0 respectively. The time constant RC is such that any charge on the capacitor C decays to exactly one third its initial value during the interval of time T allocated to one code digit. Thus, one coding interval T after the arrival of the complete code group, the total charge remaining on the capacitor is equal to the sum of the contributions due to the various digits weighted according to a ternary scale, the weight of code digits 1, 2, ... n being 3^0 , 3^1 , ... 3^{n-1} , respectively. The desired signal output is obtained by sampling the output of the RC circuit at this instant, and applying it to a holding capacitor and a correction network to compensate for the aperture effect introduced by the holding circuit.

By alternate use of two decoders, one for odd - numbered code groups and the other for the even - numbered groups, an idle interval equal to the time taken for the transmission of a code group, is made available between successive code groups, which can be made use of to allow the result of one decoding operation to decay to insignificant proportions before another code group is decoded.

However, since the output of the encoder actually used in the experiment provides the most significant digit first and the least significant digit last, decoding cannot be effected in the present case using the ternary form of

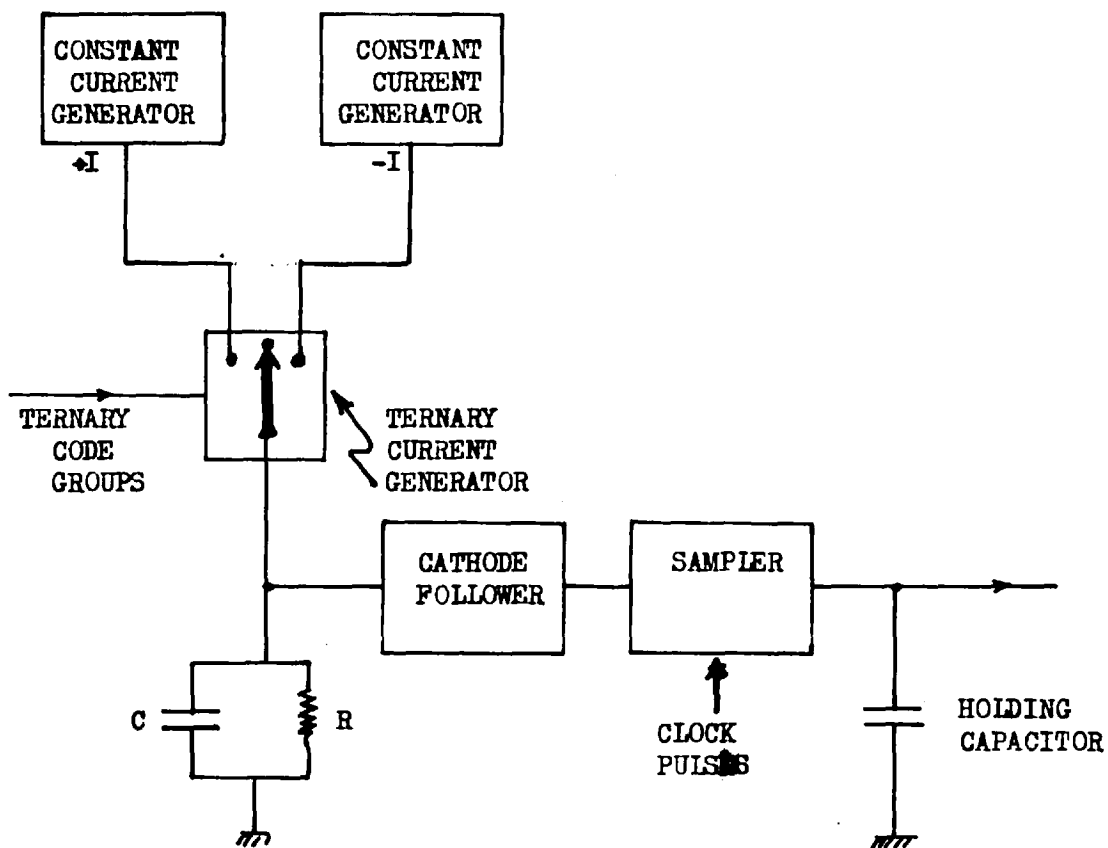


FIG. 10.1. TERNARY DECODER: EXTENSION OF THE SHANNON DECODER

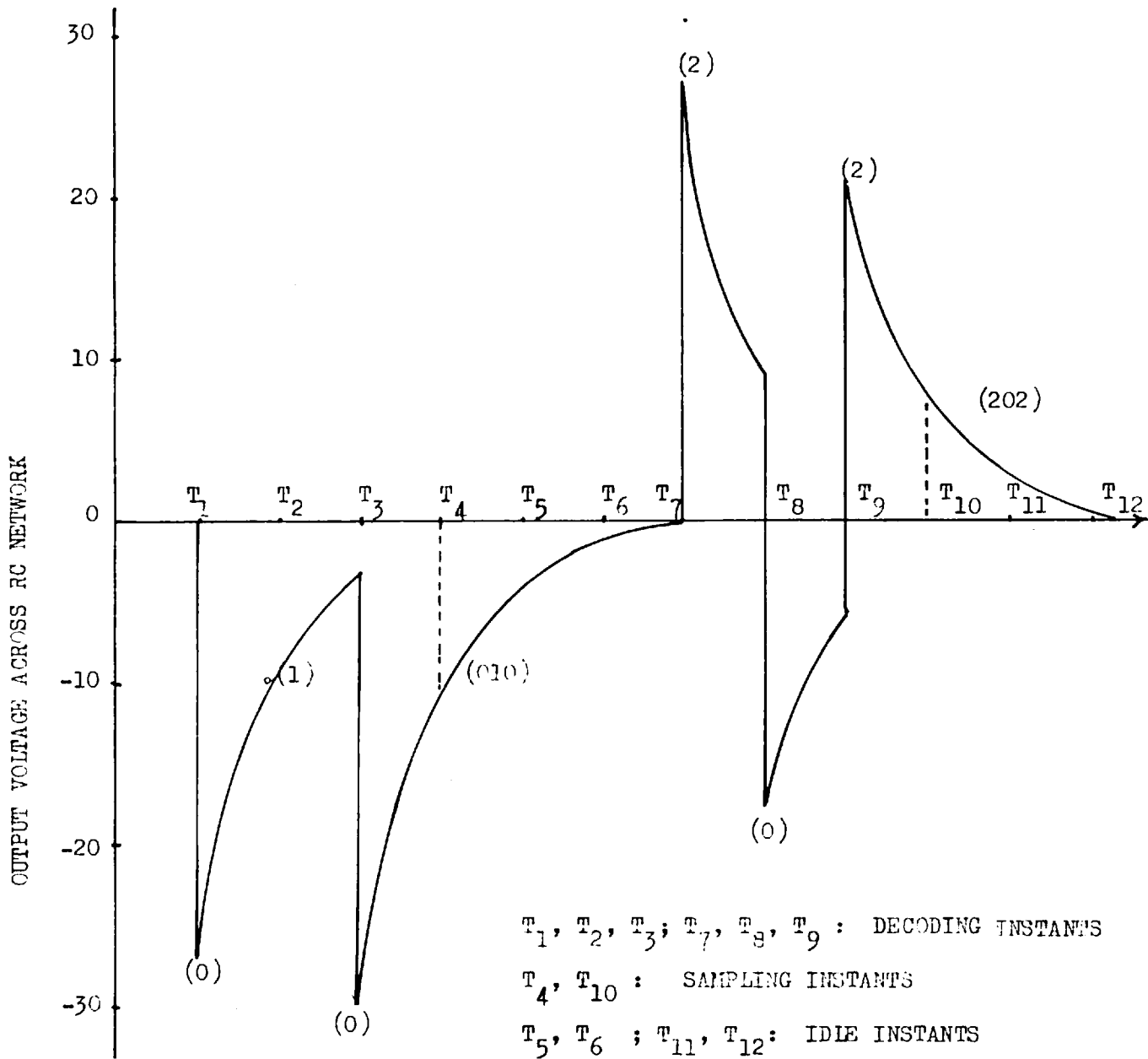


FIG. 10.2. TERNARY DECODER WAVEFORMS

Shannon decoder mentioned above. Since only two digits are used per code group in the experimental encoder, it was decided to use the simple switching scheme outlined below, for decoding.

The received noise - free ternary signal is applied simultaneously to two resistor networks, one of gain unity and the other of gain equal to $1/3$. The outputs of these networks are then selectively switched using a diode switching circuit (Fig. 10.4), to give a common output, the switching being such that the output of the unity gain network is connected to the output circuit during odd multiples $T, 3T, \dots (2n+1)T$ of the digit interval T , while the $1/3$ gain network is connected to the output during even multiples $2T, 4T, \dots (2n)T$ of T . This ensures that the two digits of the code group are weighted in the required proportions. The output of the switching circuit is then applied to an integrator (Fig. 10.5) which sums the contributions from both digits of the group. Immediately after the contribution due to the second digit has been operated upon by the integrator the integrator output is sampled (Fig. 10.6) and applied to a holding capacitor. The output of this holding circuit is then amplified and applied to a low pass filter to provide the required smoothing for the output signal.

A short duration after the output of the integrator is sampled, but before the next ternary pulse arrives at the input of the integrator, the integrator capacitor is completely discharged using a diode resetting circuit (Fig. 10.5) to prepare it to operate on the next code group.

The circuit is designed to operate over a wide band of sampling frequencies from approximately 4 Kc/s to above 40 Kc/s.

The main flow diagram of the receiver as well as the detailed circuit diagrams are given in figures 10.3 to 10.6.

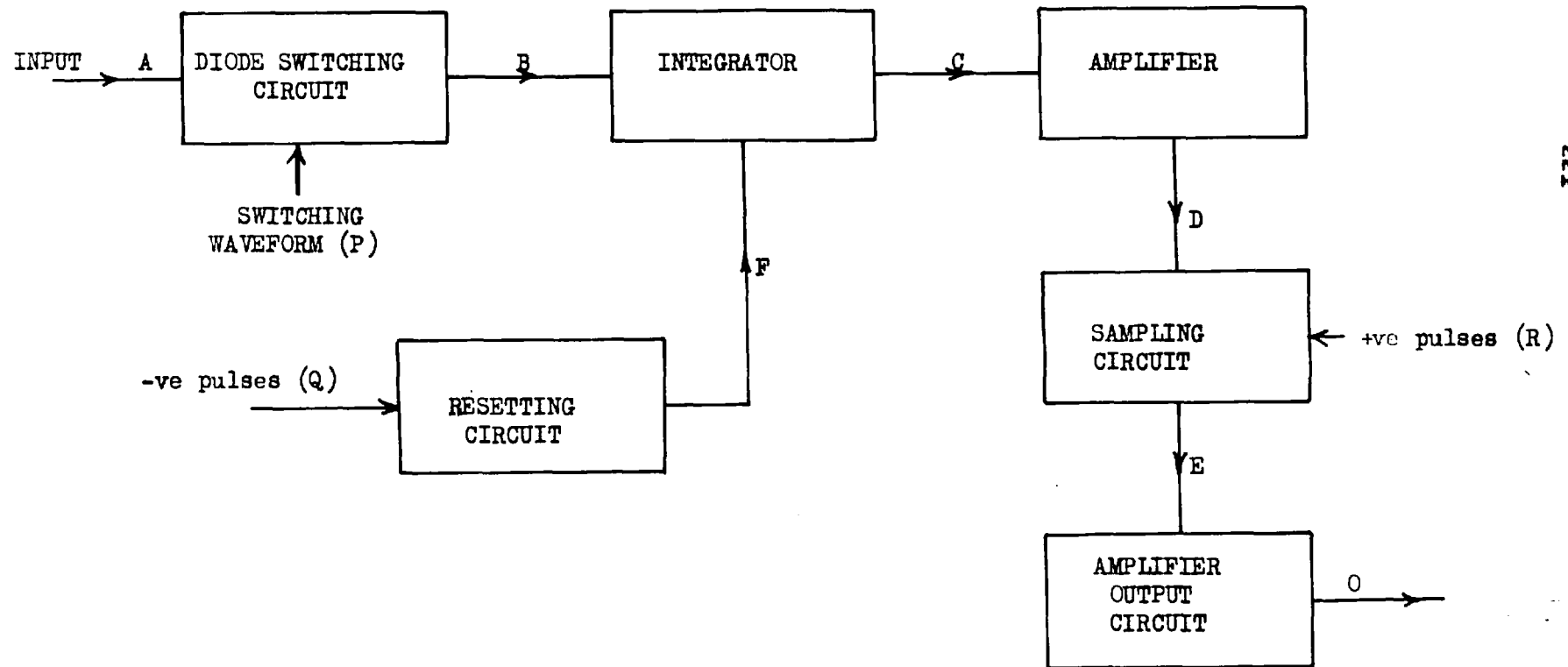


FIG. 10.3 RECEIVER UNIT

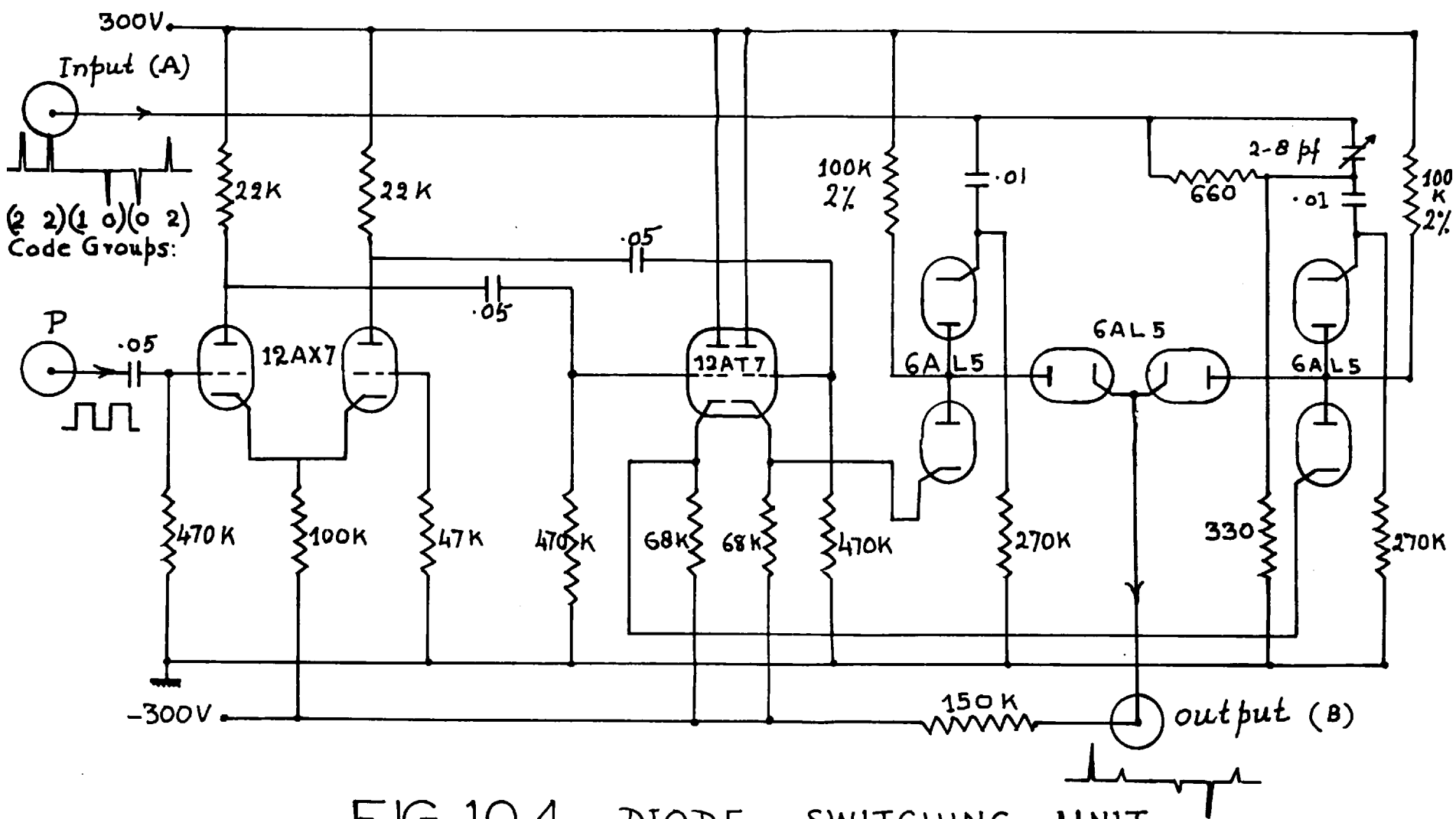


FIG. 10.4 DIODE SWITCHING UNIT

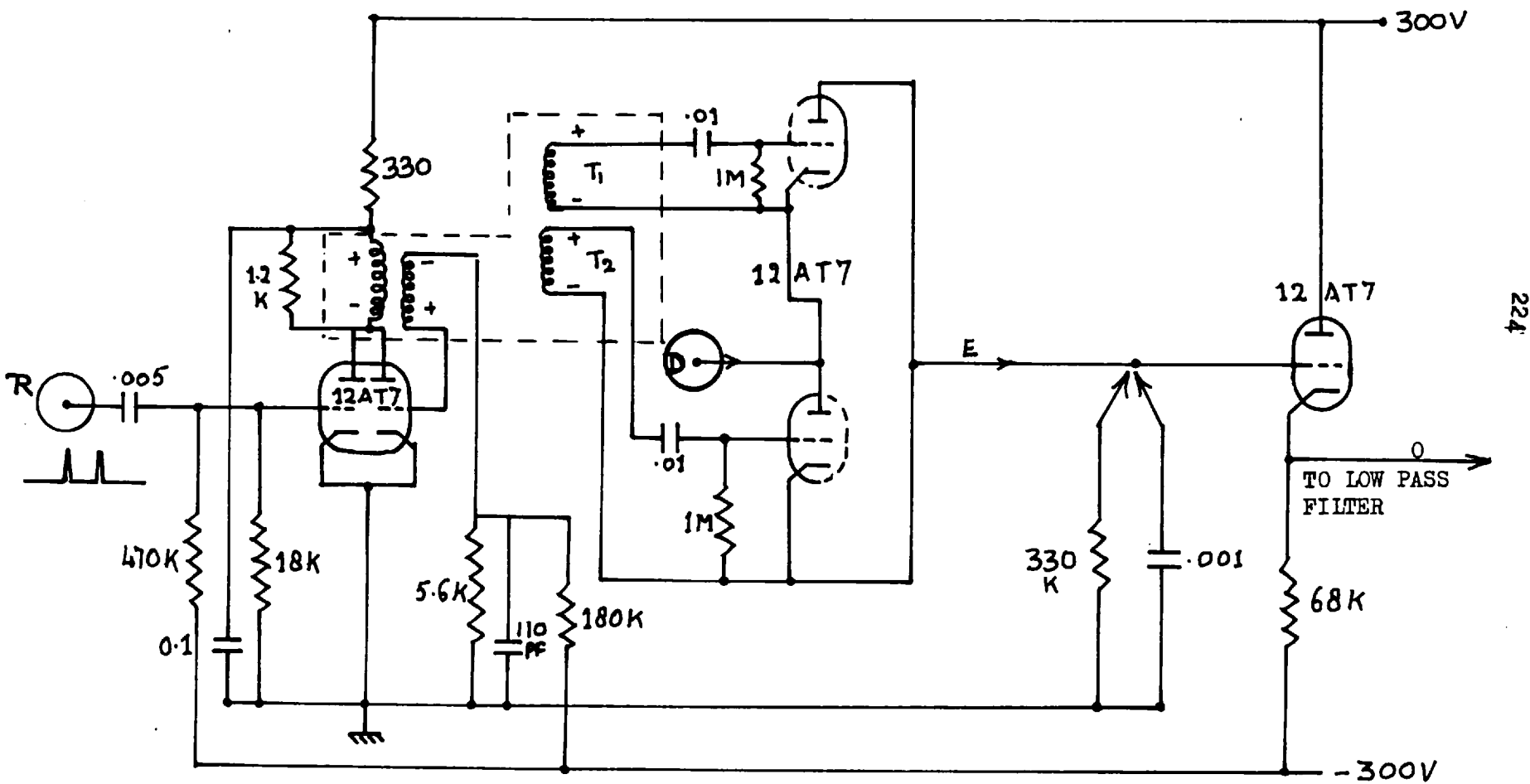


FIG. 10.6. SAMPLING CIRCUIT AND AMPLIFIER.

11. Summary of Conclusions

The thesis, especially the part dealing with theoretical investigations, is mainly concerned with the development of analytical techniques for evaluating and optimising the performance of discrete data transmission systems in general rather than with the detailed study of any particular communication system. However, on account of the importance of ternary systems, special emphasis has been placed on the study of ternary PCM systems.

Introducing the concept of information capacity per state of a symbol, it is shown in section 2 that ternary systems have the highest coding efficiency of all physically realizable systems. Methods are then suggested by which the efficiency of certain well known statistical coding techniques, such as the Shannon-Fano code, can be further improved.

In section 3, the transmission efficiency of communication systems, in which the transmission channel is perturbed by noise, is defined. The transmission efficiencies of binary and ternary PCM systems are evaluated assuming an equiprobable distribution for the various states of the transmitted symbol. It is seen that except for very low signal to noise ratios of the order of (-30) db or less, the efficiency of a ternary system is always considerably higher, the difference being especially noticeable when the signal to noise ratio is large. It is shown that the concept of channel capacity as defined by Shannon is inadequate to evaluate the performance of actual systems and that a more general formulation of the concept is necessary, in which the channel capacity is defined in terms of the parameters of the actual system including the receiver, and of the distribution of the perturbing noise.

The variation in the form of the functional dependence between transmission efficiency and the channel signal to noise ratio, as the characteristics of channel noise change, is next studied.

In section 4, equivocation in binary and ternary systems is computed and shown to be less in a binary system.

The signal to false-pulse noise ratio in a ternary system is evaluated in section 5 making use of the techniques of symbolic calculus, and shown to be less than in a binary system. The dependence of the signal to false-pulse noise ratio on the probability of occurrence of the various states of the transmitted ternary symbol, is studied and methods are then suggested for improving the signal to false-pulse noise ratio. The contributions to the total false-pulse noise error power from the various digits of a code group are estimated and analytical techniques are presented for minimizing the total false-pulse noise error power.

The problem of optimum detection of ternary signals in the presence of noise is formulated and solved. Means for graphical representation of the general ternary detection problem are also investigated.

The magnitude of quantization noise is computed in section 6 making use of certain well known techniques of Applied Statistics. A correlation technique, capable of general application, is also presented for evaluation of quantization noise.

An analytical expression is derived for minimizing quantization distortion by non-uniform spacing of quantization levels.

The effect of quantization on entropy of the signal is studied and a method is suggested for minimizing the reduction in the entropy of the signal, caused by quantization, in which linear and nonlinear shaping networks are used to operate upon the power spectrum as well as the probability distribution of amplitude of the signal. An important result in information theory on which the above optimizing technique is based, is also derived.

Section 7 is concerned with the development of more sophisticated forms of pulse code modulation in which any available a-priori knowledge of the statistical structure of the signal is made use of to improve the rate of transmission of information of the system. A system, suitable for this purpose, that incorporates an extrapolator is described. Problems involved in the synthesis of the extrapolator as well as the relevant physical and statistical considerations that determine the choice of the parameters of the extrapolator, are discussed. The expected improvement in the performance of the suggested system over conventional PCM systems, is evaluated for different types of statistical inputs.

The deterioration in the accuracy of extrapolation and therefore, in the performance of the system, caused by quantization is next studied. The effect on the stability of the system due to the presence of the quantization nonlinearity within the feedback loop, is also investigated. It is demonstrated that whereas the use of a symmetrical binary quantizer with 2^n levels causes unavoidable system instability, a stable mode of operation can be obtained in a properly designed system that employs a symmetrical ternary quantizer with 3^n levels.

The statistical transfer function of a symmetrical 1-level quantizer is derived making use of a simple technique that easily lends itself to the derivation of the statistical transfer function of any nonlinearity possessing odd symmetry. The importance of the use of statistical transfer functions of nonlinearities in the analysis of nonlinear communication or control systems is stressed.

A brief outline of the experimental work that was carried out, is presented in the next three sections. It may be mentioned in this connection that the experimental studies outlined in this thesis were undertaken with the idea of investigating the physical realizability of ternary systems in general rather than to obtain experimental verification of the theoretical results presented earlier.

In section 8, a short outline of the theory of operation of a certain type of tristable circuits is given. No original ideas are presented in this section; the emphasis is on the development of reliable and fast practical tristable circuits for use in the instrumentation of the feedback encoder mentioned in section 9.2. Section 9 describes the operation of two different types of ternary feedback encoders, of which only the one described in section 9.2. was actually built and tested. A ternary decoder based upon Shannon's binary PCM decoder, is described in section 10. The design of a simple decoder (different from that mentioned above) that was built to decode code groups consisting of not more than two ternary digits, is also given.

Except for the development work briefly set out in section 8, the research work described in sections 2 to 10 of this thesis, as reviewed above in the present section, represents original contributions by the writer to communication theory.

APPENDIX 2.1

Entropy per Symbol:

$$H = - \sum_{i=1}^b p(i) \log_e p(i) \text{ units} \quad \dots A_2.1.1$$

Average entropy per state of a symbol

$$= \frac{1}{b} \left[- \sum_{i=1}^b p(i) \log_e p(i) \right] \text{ units} \quad \dots A_2.1.2$$

Information capacity per state of a symbol, for a given value of b :

$$\begin{aligned} h_b &= \frac{1}{b} \left[\text{max.} - \sum_{i=1}^b p(i) \log_e p(i) \right] \text{ units} \\ &= - \left\{ \frac{1}{b} \log_e \left(\frac{1}{b} \right) \right\} \text{ units} \quad \dots A_2.1.3 \end{aligned}$$

To obtain information capacity per state of a symbol for optimum value of b ($= b_{\text{opt}}$), differentiate eqn. $A_2.1.3$ with respect to b , and equate to zero:

$$\frac{d}{db} \left[\frac{1}{b} \log_e \frac{1}{b} \right] = 0$$

i.e.

$$-\frac{1}{b^2} - \frac{1}{b^2} \log_e \frac{1}{b} = 0$$

i.e.

$$\log_e \left[\frac{1}{b} \right] = -1$$

$$\therefore b_{\text{opt}} = e$$

$\dots A_2.1.4$

Since the number of possible states (b) per digit has to be finite and an integral multiple of unity, the value of e is inadmissible for b_{opt} . The function

$$h_b = - \left[\frac{1}{b} \log \frac{1}{b} \right]$$

is continuous and has but one stationary point. Therefore, the optimum value of b has to be either or both of the two integers, a and b , specified by the inequality

$$a < e < b \quad \dots A_2.1.5$$

namely, the integers 2 and 3 .

Substituting the values 2 and 3 in the expression for h_b , we find

$$h_2 = - \frac{1}{2} \log \frac{1}{2}$$

$$h_3 = - \frac{1}{3} \log \frac{1}{3} \quad \dots A_2.1.6$$

Therefore, the optimum number of states b per digit for which the average information capacity per state of a symbol is maximum is 3 .

APPENDIX 2.2

The magnitude of l_0 is calculated as follows:

Let the total number (λ) of simple elements in the binary and ternary systems be λ_b and λ_t respectively. Also, let the corresponding number of messages that can be specified in the two systems be l_b and l_t respectively. Therefore,

$$l_b = {}_2 \frac{\lambda_b}{2} \quad \dots A_2.2.1$$

$$l_t = {}_3 \frac{\lambda_t}{3} \quad \dots A_2.2.2$$

A physically realizable ternary system would require less number of simple elements than the corresponding binary system, under all conditions, if the total number of messages is so large that

$$l_t \geq 2 l_b \quad \dots A_2.2.3$$

for the same number of simple elements in both systems,

$$\lambda_0 = \lambda_b = \lambda_t \quad \dots A_2.2.4$$

The minimum value of l_b for which equations $A_2.2.3$ and $A_2.2.4$ are simultaneously satisfied gives l_0 .

Combining equations $A_2.2.3$ and $A_2.2.4$,

$${}_3 \frac{\lambda_0}{3} \geq 2 \left[{}_2 \frac{\lambda_0}{2} \right] \geq \frac{\lambda_0 + 2}{2} \quad \dots A_2.2.5$$

Taking logarithms,

$$\frac{\lambda_0}{3} \log 3 \geq \frac{\lambda_0 + 2}{2} \log 2$$

$$\text{i.e. } \frac{\lambda_0}{\lambda_0 + 2} \geq \frac{3}{2} \frac{\log 2}{\log 3}$$

$$\therefore \lambda_0 \geq 35.3$$

= 36, approximating to the next
largest integer.

$\therefore l_0$ corresponds to a binary sequence of $(36/2)$
or 18 binary digits

$$\therefore l_0 = \underline{2^{18} = 262,144}$$

APPENDIX 2.3

The relationships obtained in Section 2.3.1 concerning the relative coding efficiencies η_B and η_T of binary and ternary systems, can be derived by solving the two simultaneous equations

$$3^n \equiv r \pmod{2^m} \quad \dots A_{2.3.1}^*$$

and

$$3n = 2m \quad \dots A_{2.3.2}$$

where m , n and r are integers.

Solutions to these equations have, however, to be obtained by inspection since there is no general method for solving them. The symbolism is, however, useful in understanding the relationships derived in Section 2.3.1.

For instance, it is seen that η_B is never equal to η_T for a given number of simple elements λ , since there is no solution to the equation

$$3^n \equiv 0 \pmod{2^m}$$

for any integral, non-zero values of m and n .

* The relationship "a congruent b, modulo p",

$$a \equiv b \pmod{p}$$

signifies the integral relationship that

$$\frac{a - b}{p} \text{ is an integer,}$$

given that a , b and p are integers.

APPENDIX 3.1

Capacity of a Binary Channel

Let the probability of transmitting the first and second symbols be

$$\left. \begin{aligned} p(1) &= P \\ p(2) &= Q \end{aligned} \right\} \dots A_3.1.1$$

respectively, and of receiving them

$$\left. \begin{aligned} p'(1) &= P' \\ p'(2) &= Q' \end{aligned} \right\} \dots A_3.1.2)$$

respectively. Assuming that the noise has a symmetrical amplitude distribution and that the detection level at the receiver is zero,

$$\left. \begin{aligned} p_1(1) &= p_2(2) = r \\ p_1(2) &= p_2(1) = s \end{aligned} \right\} \dots A_3.1.3$$

The rate of transmission R

$$\begin{aligned} R &= H(Y) - H_X(Y) \\ &= - \left[P' \ln P' + Q' \ln Q' \right] + \\ &\quad \left[P + Q \right] \left[r \ln r + s \ln s \right] \\ &= - \left[P' \ln P' + Q' \ln Q' \right] + \text{constant} \end{aligned} \dots A_3.1.4$$

since $P + Q = 1$

is a maximum when $H(Y)$

$$H(Y) = - \left[P' \ln P' + Q' \ln Q' \right]$$

is maximum, because of equation A₃.1.4 . It is easily shown that this occurs when

$$P' = Q' = \frac{1}{2}$$

Therefore,

$$\begin{aligned} \begin{Bmatrix} P \\ Q \end{Bmatrix} &= \left\| \tilde{p}_x(y) \right\|^{-1} \begin{Bmatrix} P' \\ Q' \end{Bmatrix} \\ &= \left\| \begin{array}{cc} r & s \\ s & r \end{array} \right\|^{-1} \begin{Bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{Bmatrix} \\ &= \frac{1}{\left| \begin{array}{cc} r & s \\ s & r \end{array} \right|} \left\| \begin{array}{cc} r & -s \\ -s & r \end{array} \right\| \begin{Bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{Bmatrix} \\ &= \frac{1}{(r-s)} \left\| \begin{array}{cc} r & -s \\ -s & r \end{array} \right\| \begin{Bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{Bmatrix} \\ &= \begin{Bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{Bmatrix} \end{aligned}$$

APPENDIX 3.2Determination of the Optimum Detection Levels, $\pm Z$

The optimum detection levels are determined by the relation

$$\frac{dR}{dZ} = 0 \quad \dots 3.28A$$

where

R is the rate of transmission

$$\therefore - \sum_y [-1 + \ln p(y)] \frac{d}{dZ} (p(y)) + \frac{1}{3} \sum_x \sum_y [-1 + \ln p_x(y)] \frac{d}{dZ} [-p_x(y)] = 0 \quad \dots 3.28B$$

$$\begin{aligned} \text{i.e.} \quad & - 2 \left[1 + \ln \left\{ \frac{1}{2} \frac{\psi(2\alpha-\beta) - \psi(\beta) - \psi(2\alpha+\beta)}{6} \right\} \right] \\ & \times \left[\frac{1}{3} \left\{ \psi(2\alpha-\beta) + \psi(\beta) + \psi(2\alpha+\beta) \right\} \right] \\ & - \left[1 + \ln \left\{ \frac{\psi(2\alpha+\beta) + \psi(\beta) - \psi(2\alpha-\beta)}{3} \right\} \right] \\ & \times \left[\frac{2}{3} \left\{ \psi(2\alpha+\beta) + \psi(\beta) + \psi(2\alpha-\beta) \right\} \right]^{(*)} \\ & + \frac{2}{3} \left[1 + \ln \frac{1 + \psi(2\alpha-\beta)}{2} \right] \left[-\psi(2\alpha-\beta) \right] \\ & + \frac{2}{3} \left[1 + \ln \frac{\psi(2\alpha+\beta) - \psi(2\alpha-\beta)}{2} \right] \left[\psi(2\alpha+\beta) + \psi(2\alpha-\beta) \right] \\ & + \frac{2}{3} \left[1 + \ln \frac{1 - \psi(2\alpha+\beta)}{2} \right] \left[-\psi(2\alpha+\beta) \right] \\ & + \frac{2}{3} \left[1 + \ln \frac{1 - \psi(\beta)}{2} \right] \left[-\psi(\beta) \right] \\ & + \frac{1}{3} \left[1 + \ln \psi(\beta) \right] \left[2\psi(\beta) \right] = 0 \quad \dots A_3.2.1 \end{aligned}$$

(*) $\times \left[\frac{2}{3} \left\{ \psi(2\alpha+\beta) + \psi(\beta) + \psi(2\alpha-\beta) \right\} \right]$

where

$$\Psi(\alpha) = \int_{-\alpha}^{+\alpha} \psi(x) dx$$

$$\psi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Using the approximations

$$\Psi(V+\delta V) = 2\delta V\psi(V) + \Psi(V) \quad \dots A_3.2.2A$$

and

$$\psi(V+\delta V) = \psi(V) - \delta V \cdot V \cdot \psi(V) \quad \dots A_3.2.2B$$

we get

$$\begin{aligned} & \frac{2}{3} \left[1 + \ln \left\{ \frac{1}{2} - \frac{\psi(3\alpha)}{6} - \frac{1}{3} \delta\alpha (2\psi(\alpha) + \psi(3\alpha)) \right\} \right] \\ & \quad \times \left[2\psi(\alpha) + \psi(3\alpha) - \delta\alpha \{ 3\alpha \cdot \psi(3\alpha) \} \right] \\ & - \frac{2}{3} \left[1 + \ln \left\{ \psi(3\alpha) + \frac{2}{3} \delta\alpha \cdot \overline{2\psi(\alpha) + \psi(3\alpha)} \right\} \right] \\ & \quad \times \left[2\psi(\alpha) + \psi(3\alpha) - \delta\alpha \{ 3\alpha \cdot \psi(3\alpha) \} \right] \\ & - \frac{2}{3} \left[1 + \ln \left\{ \frac{1 + \psi(\alpha)}{2} - \delta\alpha \cdot \psi(\alpha) \right\} \right] \left[\psi(\alpha) + \delta\alpha \{ \alpha \cdot \psi(\alpha) \} \right] \\ & + \frac{2}{3} \left[1 + \ln \left\{ \frac{\psi(3\alpha) - \psi(\alpha)}{2} + \delta\alpha \cdot \overline{\psi(\alpha) + \psi(3\alpha)} \right\} \right] \\ & \quad \times \left[\psi(\alpha) + \psi(3\alpha) + \delta\alpha \{ \alpha \cdot \psi(\alpha) - 3\alpha \cdot \psi(3\alpha) \} \right] \\ & - \frac{2}{3} \left[1 + \ln \left\{ \frac{1 - \psi(3\alpha)}{2} - \delta\alpha \cdot \psi(3\alpha) \right\} \right] \left[\psi(3\alpha) - \delta\alpha \{ 3\alpha \cdot \psi(3\alpha) \} \right] \\ & = \frac{2}{3} \left[1 + \ln \left\{ \frac{1 - \psi(\alpha)}{2} - \delta\alpha \cdot \psi(\alpha) \right\} \right] \left[\psi(\alpha) - \delta\alpha \{ \alpha \cdot \psi(\alpha) \} \right] \\ & - \frac{2}{3} \left[1 + \ln \left[\psi(\alpha) + 2\delta\alpha \cdot \psi(\alpha) \right] \right] \left[\psi(\alpha) - \delta\alpha \{ \alpha \cdot \psi(\alpha) \} \right] \\ & = 0, \end{aligned}$$

where

$$\delta\alpha = \beta - \alpha$$

$\dots A_3.2.3$

Equation A₃.2.3 can be expressed in the form

$$f + g \delta\alpha + h (\delta\alpha)^2 = 0$$

...A₃.2.4

where

$$f = 2[\gamma(\alpha) + \gamma(3\alpha)] \left[\ln \frac{3-2\Psi(3\alpha)}{2} - \ln \Psi(3\alpha) \right]$$

$$+ [\gamma(\alpha)] \left[\ln \frac{1+\Psi(\alpha)}{2} + \ln \Psi(\alpha) \right]$$

$$+ [\gamma(\alpha) + \gamma(3\alpha)] \left[1 + \ln \frac{\Psi(3\alpha) - \Psi(\alpha)}{2} \right]$$

$$- [\gamma(3\alpha)] \left[1 + \ln \frac{1-\Psi(3\alpha)}{2} \right]$$

$$- [\gamma(\alpha)] \left[1 + \ln \frac{1-\Psi(\alpha)}{2} \right]$$

$$g = [2\gamma(\alpha) + \gamma(3\alpha)]^2 \left[\frac{1}{3-2\Psi(3\alpha)} - \frac{1}{\Psi(3\alpha)} \right] [2]$$

$$+ [3\alpha \gamma(3\alpha)] \left[2 + \ln \frac{3-2\Psi(3\alpha)}{6} + \ln \frac{\Psi(3\alpha)}{3} \right]$$

$$+ [\alpha \gamma(\alpha)] \left[-2 - \ln \frac{1+\Psi(\alpha)}{2} - \ln \Psi(\alpha) \right]$$

$$+ [\gamma(\alpha)]^2 \left[\frac{1}{1+\Psi(\alpha)} + \frac{1}{\Psi(\alpha)} \right] [2]$$

$$+ [\alpha \gamma(\alpha) - 3\alpha \gamma(3\alpha)] \left[1 + \ln \frac{\Psi(3\alpha) - \Psi(\alpha)}{2} \right]$$

$$+ [\gamma(\alpha) + \gamma(3\alpha)] \left[\frac{\gamma(\alpha) + \gamma(3\alpha)}{\Psi(3\alpha) - \Psi(\alpha)} \right] [2]$$

$$+ [3\alpha \gamma(3\alpha)] \left[1 + \ln \frac{1-\Psi(3\alpha)}{2} \right]$$

+

[continued overleaf]

$$\begin{aligned}
& + \left[\gamma(3\alpha) \right] \left[\frac{\gamma(3\alpha)}{1-\gamma(3\alpha)} \right] \left[2 \right] \\
& + \left[\alpha \gamma(\alpha) \right] \left[1 + \frac{1-\gamma(\alpha)}{2} \right] \\
& + \left[\gamma(\alpha) \right] \left[\frac{\gamma(\alpha)}{1-\gamma(\alpha)} \right] \left[2 \right]
\end{aligned}$$

and

$$\begin{aligned}
h = & \left[3\alpha \gamma(3\alpha) \right] \left[2\gamma(\alpha) + \gamma(3\alpha) \right] \left[\frac{1}{3-2\gamma(3\alpha)} - \frac{1}{\gamma(3\alpha)} \right] \left[2 \right] \\
& + \left[\gamma(\alpha) \right]^2 \left[\frac{1}{1+\gamma(\alpha)} - \frac{1}{\gamma(\alpha)} \right] \left[2\alpha \right] \\
& + \left[\alpha \gamma(\alpha) - 3\alpha \gamma(3\alpha) \right] \left[\frac{\gamma(\alpha) + \gamma(3\alpha)}{\gamma(3\alpha) - \gamma(\alpha)} \right] \left[2 \right] \\
& - \left[3\alpha \gamma(3\alpha) \right] \left[\frac{\gamma(3\alpha)}{1-\gamma(3\alpha)} \right] \left[2 \right] \\
& - \left[\alpha \gamma(\alpha) \right] \left[\frac{\gamma(\alpha)}{1-\gamma(\alpha)} \right] \left[2 \right]
\end{aligned}$$

Since f , g and h are functions of α only, $(\delta\alpha)$ can be determined by solving the quadratic equation $A_3.2.4$

f , g and h are plotted as functions of α in Fig. 3.5.

The above procedure gives only a very rough approximation of the true value of $\delta\alpha$. A better approximation can be obtained by substituting the approximate value for α in equation $A_3.2.3$ and repeating the procedure given above.

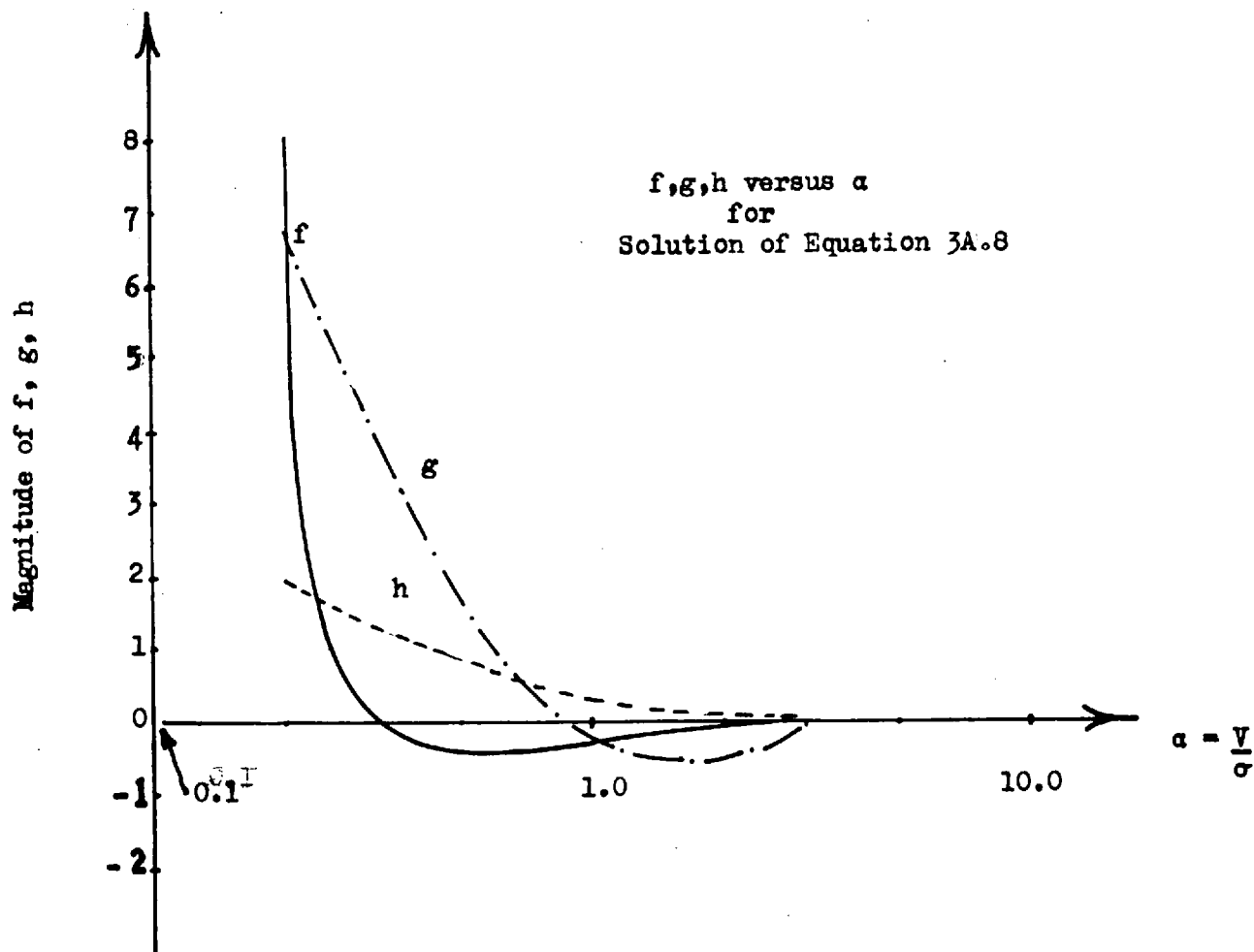


FIG. 3.5.

REFERENCES.

1. Armstrong, E.H., "A Method of Reducing Disturbances in Radio Signalling by a System of Frequency Modulation", Proceedings of the Institute of Radio Engineers, New York, Vol. 24, May 1936, pp.689 - 740.
2. Nyquist, H., "Certain Topics in Telegraph Transmission Theory". Transactions of the American Institute of Electrical Engineers. New York, Vol.47, April 1928, pp.617 - 644.
3. Hartley, R.V.L. "Transmission of Information", The Bell System Technical Journal, New York, Vol.7, No.3, July 1928, pp.535 - 563.
4. Shannon, C.E. "A Mathematical Theory of Communication," The Bell System Technical Journal, New York, Vol.27, No.3, July 1948, pp.379 - 423; and Vol.27, No.4, October 1948, pp.623 - 656.
5. Greig, D.D., and Gallay, H., "Pulse-Time-Modulated Multiplex Radio-Relay System-Radio. Frequency Equipment," Electrical Communication, New York, Vol.24, 1947, pp.141 - 158.
6. "Multi-Channel Pulse Modulation", Details of Army Wireless Station No.10, Wireless World, London, Vol.52, 1946. pp.187 - 192.
7. Reeves, A.H., French Patent 825, 185, October 23, 1939.
8. Goodall, W.M. "Telephony by Pulse Code Modulation," The Bell System Technical Journal, New York, Vol.26, No.3, July 1947, pp.395 - 409.
9. Greig, D.D. "Pulse Count Modulation" Electrical Communication, Vol,24, 1947, pp.287 - 296.
10. Black, H.S., and Edson, J.O., "P.C.M. Equipment", Electrical Engineering, Vol.66, 1947, pp.1123 - 1125.
11. Meacham, L.A. and Peterson, E. "An Experimental Multichannel Pulse-Code Modulation System of Toll Quality", The Bell System Technical Journal, New York, Vol.27, No.1, January 1948, pp.1 - 43.
12. Black, H.S.; "Modulation Theory". D. Van Nostrand Co.Inc.. New York, 1953.

13. Shannon, C.E., "Communication in the Presence of Noise", Proceedings of the Institute of Radio Engineers, New York, Vol.37, No.1, January 1949, pp.10 - 21.
14. Oliver, B.M., Pierce, J.R., and Shannon, C.E., "The Philosophy of PCM," Proceedings of the Institute of Radio Engineers, Vol.36, November 1948, pp.1324-1331.
15. Feldman, C.B. and Bennett, W.R., "Bandwidth and Transmission Performance", The Bell System Technical Journal, New York, Vol.28, No.3, July 1949, pp.490 - 595.
16. Bennett, W.R., "Spectra of Quantized Signals", The Bell System Technical Journal, New York, Vol.27, No.3, July 1948, pp.446 - 472.
17. Mallinckrodt, C.O., "Instantaneous Compandors", The Bell System Technical Journal, New York, Vol.30, No.3, July 1951, pp.706 - 720.
18. Lozier, J.C. "Instantaneous Compandors on Narrow-Band Speech Channels", The Bell System Technical Journal, New York, Vol.30, No.4, Part II, October 1951, pp.1214 - 1220.
19. Smith, B. "Instantaneous Companding of Quantized Signals", The Bell System Technical Journal, New York, Vol.36, May, 1957, pp.653 - 709.
20. Rice, S.O. "Mathematical Analysis of Random Noise", The Bell System Technical Journal, New York, Vol.23, No.3, July 1944, pp.282 - 332; and Vol.24, No.1, January 1945, pp.46 - 156.
21. Clavier, A.G., Panter, P.F., and Dite, W., "Signal to Noise Ratio Improvement in a PCM System," Proceedings of the Institute of Radio Engineers, New York, Vol.37, April 1949, pp.355 - 359.
22. Bidrosian, B. "Weighted PCM", Transactions of the Institute of Radio Engineers on Information Theory, Vol.IT - 4, No.1, March 1958, pp.45 - 49.
23. Flood, J.E., "Noise Reducing Codes for Pulse-Code Modulation", Proceedings of the Institution of Electrical Engineers, London, Vol.105, Part C, September 1958, pp.391 - 397.
24. Goldman, S., "Information Theory", Constable and Co.Ltd., London, 1953.

25. Richards, R.K., "Arithmetic Operations in Digital Computers", D. Van Nostrand, New York, 1955.
26. Fano, R.M., Technical Report, No.65, Research Laboratory of Electronics, M.I.T. March 1949.
27. Woodward, P.M. "Probability and Information Theory with Applications to Radar," Pergamon Press Ltd., London, 1953.
28. Leifer, M., and Schreiber, W.F., "Communication Theory" in "Advances in Electronics", Vol.III, Edited by Marton, Academic Press Inc., New York, 1951, pp.305 - 343.
29. Billings, A.R., "The Rate of Transmission of Information in Pulse-Code-Modulation Systems", Proceedings of the Institution of Electrical Engineers, London, Vol. 105, Part C, September 1958, pp.444 - 447.
30. Jelonck, Z. "A Comparison of Transmission Systems", Symposium on Information Theory, Edited by Willis Jackson, Butterworths Scientific Publications, 1953, pp.44 - 81.
31. Mayer, H.F., "Principles of Pulse Code Modulation" in "Advances in Electronics", Vol.III, Edited by Marton, Academic Press Inc., New York, 1951, pp.221 - 260.
32. Riordan, J. "Combinatorial Analysis", John Wiley and Sons, Inc., New York, 1958.
33. MacMahon, P.A., "Combinatory Analysis", Vol.I, 1915, Vol II, 1916, London.
34. Bell, E.T., "Postulational Bases for the Umbral Calculus", American Journal of Mathematics, Vol.62, 1940, pp. 717 - 724.
35. Bell, E.T., "Exponential Polynominals", Annals of Mathematics, Vol.35, 1934, pp.258 - 277.
36. Yates - Fish, N.L., and Fitch, E., "Signal/Noise Ratio in Pulse Code Modulation," Proceedings of the Institution of Electrical Engineers, London, Vol.102, Part B, No.2, March 1955, pp.204 - 210.
37. Wald, A., "Statistical Decision Functions", John Wiley & Cons, Inc., New York, 1950.

38. Neyman, J., and Pearson, E.S., "On the Problems of the Most Efficient Tests of Statistical Hypotheses", Philosophical Transactions of the Royal Society, Series A, Vol. 231, 1931, page 289.
39. Wald, A. and Wolfowitz, S., "Characterization of the Minimal Complete Class of Decision Functions when the number of Distributions and Decisions is Finite", Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley and Los Angeles, 1951, pp. 149 - 157.
40. Milne Thomson, L.M., "The Calculus of Finite Differences" Macmillan and Co. Ltd.,
41. Jordan, C., "Calculus of Finite Differences", Chelsea Publishing Co.,
42. Fisher, R.A., "On the Mathematical Foundations of Theoretical Statistics", Philosophical Transactions of the Royal Society, Series A, Vol. 222, 1922, page 309.
43. Aitken, A.C., "Statistical Mathematics," Interscience Publishers Inc., New York, 1947.
44. Panter, P.F., and Dite, W., "Quantization Distortion in Pulse-Count Modulation with Nonuniform Spacing of Levels," Proceedings of the Institute of Radio Engineers, New York, Vol. 39, January 1951, pp. 44 - 48.
45. Beranek, L.L., "Acoustic Measurements", John Wiley and Sons, Inc., New York, 1949.
46. Dunn, H.K., and White, S.D., "Statistical Measurements on Conversational Speech", Journal of the Acoustical Society of America, Vol. II. 1948, page 278.
47. Holbrook, B.D., and Dixon J.T., "Load Rating Theory for Multichannel Amplifiers", The Bell System Technical Journal, New York, Vol. 18, No. 4, October 1939, pp. 624 - 644.
48. Reference 12, page 324.
49. Reference 12, Chapter 4.
50. Hamming, R.W., "Error Detecting and Error Correcting Codes", The Bell System Technical Journal, New York, Vol. 29, No. 2, April 1950, pp. 147 - 160.
51. Gilbert, E.K., "A Comparison of Signalling Alphabets", The Bell System Technical Journal, New York, Vol. 31, No. 3, May 1952, pp. 504 - 522.

52. Scarborough, J.B., "Numerical Mathematical Analysis", The Johns Hopkins Press, Baltimore, 1955.
53. Porter, A, and Stoneman, F.W., "A New Approach to the Design of Pulse-Monitored Servo Systems", Proceedings of The Institution of Electrical Engineers, London, Vol.97, Part II, 1950, pp.597 - 608.
54. Cramer H, "Mathematical Methods of Statistics". Princeton University Press, Princeton, 1945.
55. Jager, F. de., "Deltamodulation - A Method of PCM Transmission Using the 1-Unit Code", Symposium on Applications of Communication Theory, edited by Willis Jackson, Butterworths Scientific Publications, London, 1953, pp.119 - 137.
56. Reference 54. Chapter 21.
57. Feller, W. "An Introduction to Probability Theory and Its Applications", Volume I, John Wiley and Sons Inc., New York, 1952, Chapter 15.
58. Reference 40. Chapter 13.
59. Reference 54, Chapter 23.
60. Weg, H. van de, "Quantizing Noise in a Single Integration Delta Modulation System with an N-Digit Code", Phillips Research Reports, Vol.8, No.5, October 1953, pp.367 - 385.
61. Zetterberg, L-H., "A Comparison between Delta and Pulse Code Modulation", Ericsson Technics, Vol.11, No.1, 1955, pp.95 - 153.
62. Widrow, B., "A Study of Rough Amplitude Quantization by Means of Nyquist Sampling Theory", Transactions of The Institute of Radio Engineers on Circuit Theory, New York, Vol. CT-3, No.4, December, 1956, pp.266-276.
63. Freeman, J.J., "Principles of Noise", John Wiley and Sons Inc., New York, 1958.
64. Kochenburger, R.J. "Frequency Response Method for Analyzing and Synthesizing Contactor Servomechanisms", Transactions of the American Institute of Electrical Engineers, Vol. 69, Part I, 1950, pp.270-284.
65. Truxal, J.G. "Automatic Feedback Control System Synthesis" McGraw Hill, New York, 1955.

66. Ragazzini, J.R., and Franklin, G.F., "Sampled Data Systems", McGraw Hill, Book Co.Inc., New York, 1958, Page 125.
67. Grensted, P.E.W. "Frequency Response Analysis of Non-Linear Systems", Proceedings of the Institution of Electrical Engineers, London, Vol.102, Part C, 1955, page 244.
68. Andronov, A.A., and Chaikin, C.E., "Theory of Oscillations" Princeton University Press, 1949.
69. Minorsky, N. "Introduction to Nonlinear Mechanics", J.W. Edwards, Ann Arbor, Michigan, 1947.
70. Booton, R.C.Jr., "The Analysis of Nonlinear Control Systems with Random Inputs", "Proceedings of the Symposium on Nonlinear Circuit Analysis", Polytechnic Institute of Brooklyn, New York, 1953. pp369-391.
71. Wiener, N, "The Interpolation, Extrapolation and Smoothing of Stationary Time Series", John Wiley and Sons, Inc., New York, 1949.
72. Barrett, J.F., and Coales, J.F., "An Introduction to the Analysis of Non-Linear Control Systems with Random Inputs", Proceedings of the Institution of Electrical Engineers, London, Vol.103, Part C, March 1956, pp.190-199.
73. Chance, B., et al, "Waveforms" Massachusetts Institute of Technology Radiation Laboratory Series, Vol.19, McGraw Hill Book Co.Inc., New York, 1949.
74. Booth, A.D., and Ringrose J., "A Three State Flip-Flop", Electronic Engineering, Vol, 23, April 1951.
75. Johnson, K.C., "A Three State Flip-Flop", (Letters to the Editor), Electronic Engineering, Vol.23, June, 1951.
76. Mackay, R.S., and MacIntyre, R., "Ternary Counters", Institute of Radio Engineers Transactions on Electronic Computers, New York, Vol. EC4, December 1955
77. Weissman, R. "High Speed Counter Uses Ternary Notation", Electronics, Vol.25, October, 1952.
78. Sidhan, T.C., "Analysis and Development of Tristable Circuits", Thesis submitted for Diploma of Imperial College, 1957.

79. Carbrey, R.L., "Decoding in PCM", Bell Laboratories Record, New York, Vol.26, No.11, November 1948, pp.451-455.
80. Johnson, A.E., "Timing Control for PCM", Bell Laboratories Record, New York, Vol.27, No.1, January 1949, pp.10-15.
81. Smith, B.D. "Coding by Feedback Methods", Proceedings of the Institute of Radio Engineers, New York, Vol.41, August 1953.
82. Soars, R.W., "Beam Deflection Tube for Coding in PCM", Bell Laboratories Record, New York, Vol.26, No.12, December 1948, pp.487 - 490.