#3

THEORETICAL LIMITATIONS ON THE RATE OF TRANSMISSION OF INFORMATION

WILLIAM G. TULLER

TECHNICAL REPORT NO. 114

APRIL 23, 1949

RESEARCH LABORATORY OF ELECTRONICS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

The research reported in this document was made possible through support extended the Massachusetts Institute of Technology, Research Laboratory of Electronics, jointly by the Army Signal Corps, the Navy Department (Office of Naval Research) and the Air Force (Air Materiel Command), under Signal Corps Contract No. W36-039-sc-32037, Project No. 102B; Department of the Army Project No. 3-99-10-022.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Research Laboratory of Electronics

Technical Report No. 114 April 23, 1949

THEORETICAL LIMITATIONS ON THE RATE OF TRANSMISSION OF INFORMATION*

William G. Tuller**

ABSTRACT

A review of prior work on the theory of the transmission of information is followed by a critical survey of this work and a refutation of the point that, in the absence of noise, there is a finite limit to the rate at which information may be transmitted over a finite frequency band. A simple theory is then developed which includes, in a first-order way, the effects of noise. This theory shows that information may be transmitted over a given circuit according to the relation $H < BT \log (1 + C/N)$, where H is the quantity of information, B the transmission link bandwidth, T the time of transmission, and C/N the carrier-to-noise ratio. Certain special cases are considered, and it is shown that there are two distinctly different types of modulation systems, one trading bandwidth linearly for signal-to-noise ratio, the other trading bandwidth logarithmically for signal-to-noise ratio.

The theory developed is applied to show some of the inefficiencies of present communication systems. The advantages to be gained by the removal of internal message correlations and analysis of the actual information content of a message are pointed out. The discussion is applied to such communication systems as radar relays, telemeters, voice communication systems, servomechanisms, and computers.

---

* This paper is based on a thesis submitted in partial fulfilment of the requirements of the degree of Doctor of Science at the Massachusetts Institute of Technology.

** Now at Melpar, Inc.

THEORETICAL LIMITATIONS ON THE RATE OF TRANSMISSION OF INFORMATION

I. Introduction

The history of this investigation goes back at least to 1922 when Carson [1], analyzing narrow-deviation frequency modulation as a bandwidth-reduction scheme, wrote "all such schemes are believed to involve a fundamental fallacy." In 1924, Nyquist [2] and Kupfmuller [3], working independently, showed that the number of telegraph signals which may be transmitted over a line is directly proportional to its bandwidth. Hartley [4], writing in 1928, generalized this theory to apply to speech and general information, concluding that "the total amount of information which may be transmitted . . . is proportional to the product of the frequency range which is transmitted and the time which is available for the transmission." It is Hartley's work which is the most direct ancestor of the present paper. In his paper he introduced the concept of the information function, the measure of quantity of information, and the general technique used in this paper. He neglected, however, the possibility of the use of the knowledge of the transient-response characteristics of the circuits involved. Further, he neglected noise.

In 1946, Gabor [5] presented an analysis which broke through some of the limitations of the Hartley theory and introduced quantitative analysis into Hartley's purely qualitative reasoning. However, Gabor also failed to include noise in his reasoning.

The workers whose papers have so far been discussed did not give much thought to the fact that the problem of transmitting information is in many ways identical to the problem of analysis of stationary time series. This point was made in a classical paper by Wiener [6], who made a searching analysis of that problem which is a large part of the general one – the

---

(1) J. R. Carson, "Notes on the Theory of Modulation," Proc. I.R.E., vol. 10, p. 57; February, 1922.

(2) H. Nyquist, "Certain Factors Affecting Telegraph Speed," Bell Sys. Tech. Jour., vol. 3, p. 324; April, 1924.

(3) K. Kupfmuller, "Transient Phenomena in Wave Filters," Elek. Nach. Tech., vol. 1, p. 141; 1924.

(4) R. V. L. Hartley, "Transmission of Information," Bell Sys. Tech. Jour., vol. 7, p. 535 – 564; July, 1928.

(5) D. Gabor, "Theory of Communication," Jour. I.E.E. (London), vol. 93, part III, p. 439; November, 1946.

(6) N. Wiener, "The Extrapolation, Interpolation and Smoothing of Stationary Time Series," National Defense Research Council, Section D2 Report, February, 1942.

problem of the irreducible noise present in a mixture of signal and noise. Unfortunately, this paper received only a limited circulation, and this, coupled with the fact that the mathematics employed were beyond the off-hand capabilities of the hard-pressed communication engineers engaged in high-speed wartime developments, has prevented as wide an application of the theory as its importance deserved. Associates of Wiener have written simplified versions of portions of his treatment [7],[8], but these also have as yet been little accepted into the working tools of the communication engineer. Wiener has himself done work parallel to that presented in this paper, but this work is as yet unpublished, and its existence was learned of only after the completion of substantially all the research reported on here. A group at the Bell Telephone Laboratories, including C. E. Shannon, has also done similar work [9],[10],[11].

II. Definitions of Terms

In the discussion to follow, certain terms are used which are either so new to the art that accepted definitions for them have not yet been established, or which have been coined for use in connection with the research here reported. The definitions used in this Report for those terms are given below for the convenience of the reader. No justification of the choice of terms or of the definitions will be given at this point, since it is hoped that this justification will be provided by the bulk of the paper. Terms used in the body of the paper which are not defined below and are peculiar to the jargon of radio engineers will be found in the various "Standards" reports published by The Institute of Radio Engineers [12].

---

(7) N. Levinson, "The Wiener (RMS) Error Criterion in Filter Design and Prediction," Jour. Math. Phys., vol. 25, no. 4, p. 261; 1947.

(8) H. M. James, "Ideal Frequency Response of Receiver for Square Pulses," Report No. 125 (v-12s), Radiation Laboratory, MIT, November 1, 1941.

(9) C. E. Shannon, "A Mathematical Theory of Communication," Bell Sys. Tech. Jour., vol. 27, pp. 379 – 424 and 623 – 657; July and October, 1948.

(10) C. E. Shannon, "Communication in the Presence of Noise," Proc. I.R.E., vol. 37, pp. 10 – 22; January, 1949.

(11) The existence of this work was learned by the author in the spring of 1946, when the basic work underlying this paper had just been completed. Details were not known by the author until the summer of 1948, at which time the work reported here had been complete for about eight months.

(12) In particular, "Standards on Antennas, Modulation Systems, and Transmitters: Definitions of Terms, 1948."

Information Function: The information function is the function (generally instantaneous amplitude of a current or voltage as a function of time) to be transmitted by electrical means over the communication systems to be analyzed.

Intersymbol Interference: Intersymbol interference is the disturbance present in a signal caused by the energy remaining in the transient following the preceding signal.

Coding: Coding is the representation of the information function by a symbol or group of symbols bearing a definite mathematical relation to the original function, and containing all the information contained in the original function.

Binary Coding: Binary coding is coding in which the instantaneous amplitude of the information function is represented by a sequence of pulses. The presence or absence of these pulses at certain specified instants of time represents a digit (either one or zero) in the binary system of numbers.

Clearing Circuit: The clearing circuit is a circuit which will clear intersymbol interference from the output of a filter.

Quantized: A variable is said to be quantized when it varies only in discrete increments.

III. Communication-System Transmission Characteristics

In general, the information which one wishes to transmit over a communication system is supplied to the system (neglecting any transducers which may be present to transfer the energy from its source to the electrical system) in the form of a time-varying voltage (or current), lasting for a period T.

For the purposes of the paper, it will be assumed that a width of pass band f_c has been selected, and the lower and upper limits of this band have been fixed. It will be assumed that all frequency components of the information function lying within this pass band are to be transmitted without distortion of any type, and that all frequencies outside the limits of the pass band are completely unimportant and may be suppressed an arbitrary large amount, enough to make them negligible. These assumptions are, it is realized, somewhat arbitrary, since a system satisfying them would cause considerable transient distortion of certain types of information function. However, these assumptions will serve as a first approximation.

IV. Definition of Quantity of Information

It must, of course, be recognized that any physical transmission system will have an upper bound to the amplitude of information function it will

transmit. The accuracy of specification of the information function at any given time may be specified in terms of this maximum value. Thus, within the range of possible values of the information function, there will be a number s of these values which are significant. Similarly, if the information is examined over a period of time of length T, there will be a number n of times at which samples of the function may be taken and yet the information will be unchanged, since the function may be recreated from a knowledge of its values at these intervals. It is known from the work of Bennett [13] and others, that n must be greater than $2f_cT$, using the nomenclature of the previous sections, in order to recreate exactly any arbitrary function. Considering the continuous information function shown in Figure 1, the second statement given above permits us to consider its values only at specified, and in this particular case equispaced, intervals of time, as is shown in the solid stairstep curve. The statement of the finite number of

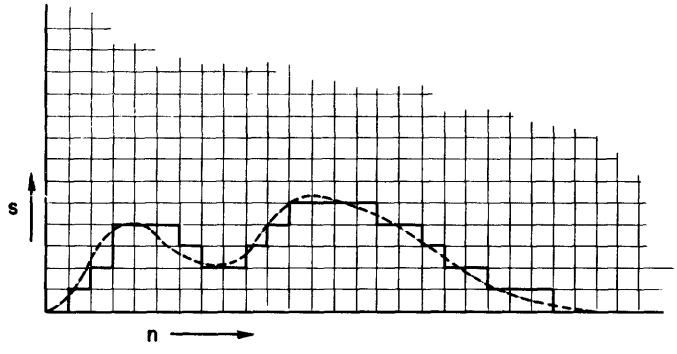

Fig. 1 The information function in n,s space.

significant values of the function allows us to consider only certain discrete amplitudes, separated from each other by twice the error of specification. The information function may thus be redrawn so as to follow only certain lines in a rectangular coordinate system. Such a function is called quantized, since it takes on values chosen from a discrete set. A plot of such a function quantized, and drawn in n,s space, is also given in Figure 1.

The question now before us is, "What is the information content of a function in the n,s plane?" The answer of Hartley is the "quantity of information" given by

$$H = kn \log s \quad , \tag{1}$$

where k is a proportionality constant. The reasons for Hartley's choice may be expressed in a straightforward manner on the basis of two fundamental requirements of a definition of "quantity of information."

(13) W. R. Bennett, "Time-Division Multiplex Systems," Bell Sys. Tech. Jour., vol. 20, pp. 199 – 222; April, 1941.

These are:

(a) Information must increase linearly with time. In other words, a two-minute message will, in general, contain twice as much information as a one-minute message.

(b) Information is independent of s and n if s^n is held constant. This states that the information contained in a message in a given n,s plane is independent of the course of the information function in that plane, allowing only single-valued functions. With this restriction, the number of different messages which may occupy a given n,s plane is s^n. Transmission of one of these messages corresponds to making one of s^n choices. Stating that information is independent of s and n if s^n is constant means that we gauge quantity of information by the number of possible alternatives to a given message, not by its length or number of possible values at any given instant of time.

On the basis of these two requirements, it can be shown that (1) is the only possible definition of quantity of information.

V. Transmission of Information In A Noise-Free Universe

The preceding discussion has been set up on the basis of a system with noise, and, indeed, this is the most practical arrangement. However, most of the previously published works on the transmission of information have neglected noise, and thus have reached the conclusion that even in the absence of noise there is a limit, in any system containing elements capable of storing energy, to the rate at which information may be transmitted. This theory has been widely — in fact, almost universally — accepted by communication engineers. It is, therefore, believed worth while to show by example that this theory is incorrect, even though the correct theory to be derived later in this paper shows implicitly the error in the previous theories. The basic fact which has been neglected in earlier analyses, and which resulted in their errors, is that the output waveform of a network is completely determined for all time by the input waveform and the characteristics of the network. A method of utilizing this effect in practice is outlined in the following paragraphs.

Suppose that we choose to transmit the information by a series of modulated pulses according to any of the well-known methods of pulse modulation. If these pulses are passed through a filter which one would suspect as having too narrow a pass band to reproduce faithfully the pulses, there will result intersymbol interference, as shown in Figure 2; that is, energy stored in the filter from the first pulse will appear at the output of the filter during the time at which the second, third, and all succeeding pulse outputs are present. However, if we know the shape

of the pulses and the transient response of the filter, we may transmit our intelligence in the following manner, theoretically. Let us first transmit the pulse to be used as a standard of comparison. The output from the filter resulting from this pulse will be measured for a period of time sufficient to determine exactly the amplitude of the initial pulse. This may

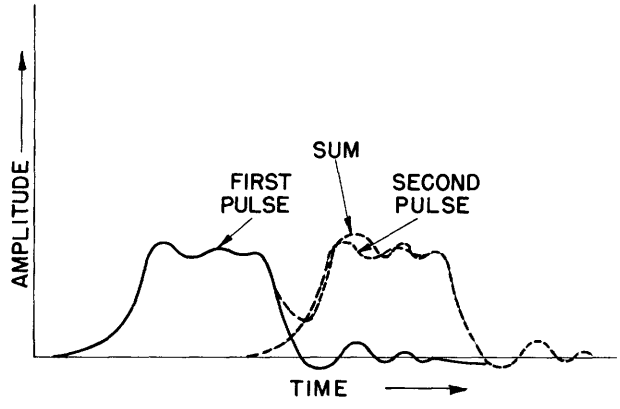

Fig. 2 Response of communication system to rectangular pulse.

be done since, as was mentioned above, the output amplitude is uniquely determined by the input amplitude. Knowing the waveform of the output from our knowledge of system characteristics and the amplitude of the output from measurements, we may compute the exact output voltage to be obtained from the system at any time in the future. This voltage waveform may be generated locally and subtracted electrically from the output of the filter. Alternatively, the output waveform may be recorded graphically, the component due to the first pulse computed, and this component subtracted by graphical methods, to give the system output free of intersymbol interference caused by this pulse. Either method may be applied repeatedly to remove intersymbol interference from the following pulses. Other methods may, no doubt, be used, but these two serve to indicate the problem involved.

To formalize the argument, suppose we are given an arbitrary signal $f(t)$ and an arbitrary filter pass band f_c. We wish to transmit an arbitrarily long message at the rate of one per second. Let us assign to the m^{th} message we wish to transmit an integer M_m characterizing the message in a one-to-one manner. This number M_m may, for example, be derived from the number in the binary system of units corresponding to the message as transmitted in ordinary continental Morse Code, using a one to correspond to a signal of unit length in this code and a zero to correspond to a space of unit length. In this fashion, a one-to-one correspondence may be realized between some number (in the binary system of units in this case) and our message. Referring to Figure 3, let us consider transmission of the message "NO". Figure 3(a) shows the message in English, 3(b) in Continental

Morse Code, 3(c) in binary digits, and 3(d) gives the number to the base ten which corresponds to our message. Figure 3(e) shows that all the information contained in the message may be contained in one pulse, of arbitrary duration and amplitude 477,147 units. The pulse of Figure 3(e) may be used then as our message.

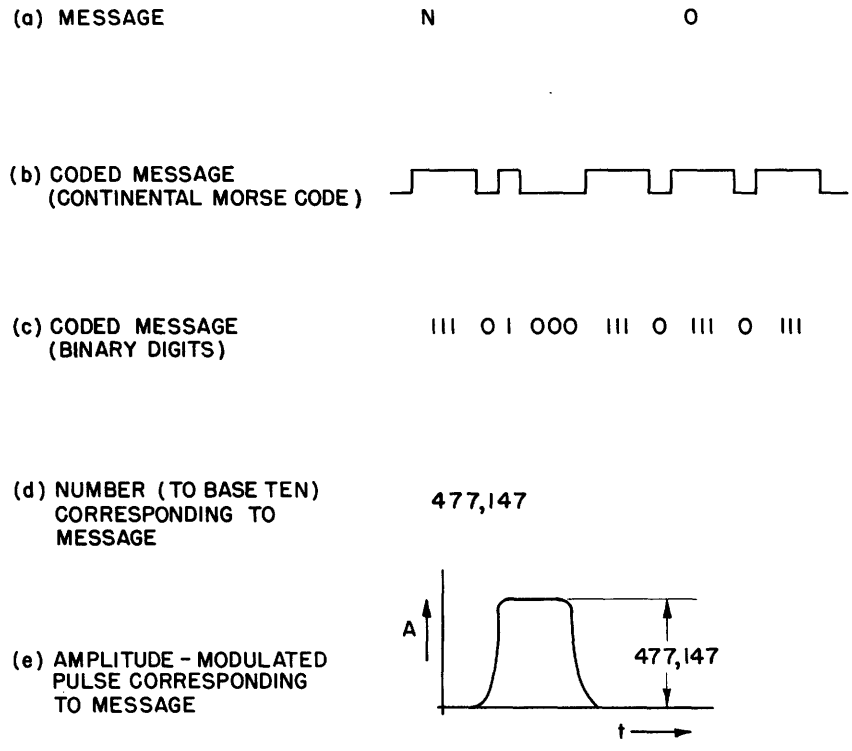

Fig. 3 Representation of message by an integer.

Three types of pulses have been mentioned in the preceding paragraphs, perhaps causing a certain amount of confusion. To distinguish among them, let us reconsider for a moment. The first pulse mentioned was a typical pulse in a pulse-modulation system, used as an example to show how intersymbol interference might be eliminated. The second pulses mentioned were those forming our message in Morse Code, used in a typical process for obtaining a one-to-one correspondence between a message and an integer M_m. The third pulse mentioned was one of arbitrary duration and M_m units in amplitude. It may, therefore, be used as our message. It may, further, be used as a channel pulse in a pulse-amplitude-modulation communication system, since the information it carries is contained solely in its amplitude.

To return to our original argument, we are given a waveform f(t). Let F(t) be the response of our channel, of bandwidth f_c, to f(t). At time t = 0 we transmit $M_o f(t)$ where M_o is a known calibrating amplitude. To calibrate our system, we measure the voltage in the receiver channel at

some time t_o. We may call this voltage $n_o F(t_o)$. Since we know F(t) for all values of t, we know it for t_o. From this and our measurement, we obtain n_o, the demodulated voltage at the output of the system corresponding to a modulating voltage M_o.

We now introduce the voltage $-n_o F(t)$, for t greater than t_o, into the output of our system. This clears the channel completely. We may then take a voltmeter reading at $t_o + 1$, from which we get M_1. This second message may then be cleared from our system by the same procedure as used previously and the same process repeated. This may now go on until $t = t_m$. At this time we measure the voltage $n_m F(t_m)$. All energy which would ordinarily have been present at this time, because of intersymbol interference, has been eliminated by the clearing process. We know F(t) for all values of t, so we know it for t_m. From this and the previously measured relation between n_o and M_o, we may obtain successively n_m and M_m. Thus our message M_m is obtained regardless of the intersymbol interference present and, therefore, regardless of f_c, providing only that we may measure to negligible error and that our system characteristics are accurately known.

A semipractical arrangement for accomplishing this is shown in block form in Figure 4. Figure 5 shows some of the waveforms involved.

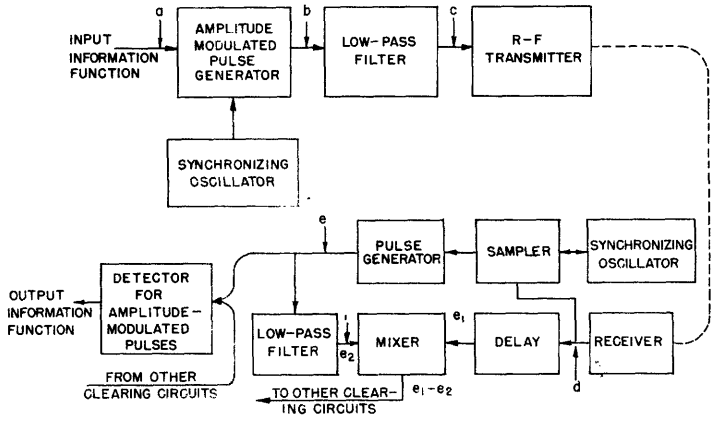

Fig. 4 Block diagram of narrow-band communication system.

It has been observed by critics of the system proposed above that this scheme is theoretically unworkable, since in the limit each pulse will have an infinite precursor which will not be eliminated by the clearing circuits. However, in this limiting case, the system has an infinite delay also. Therefore, the precursor begins at the time of transmission of our first pulse, and, though small, may be measured at this time. The components due to other pulses cannot exist before these pulses are impressed on the system, and hence cannot affect the measurement of the precursor to the first pulse. Once the precursor to the first pulse of the system has been measured, the amplitude of this first pulse is known, and its effects may be cleared from the system. The system proposed here

does not, it should be emphasized, depend on the early results of Nyquist, who pointed out that systems which are self-clearing at certain times can be constructed, and hence can transmit a very large amount of information at these times. The system proposed here can transmit information at a rate limited only by the delay one wishes to tolerate. As has been pointed

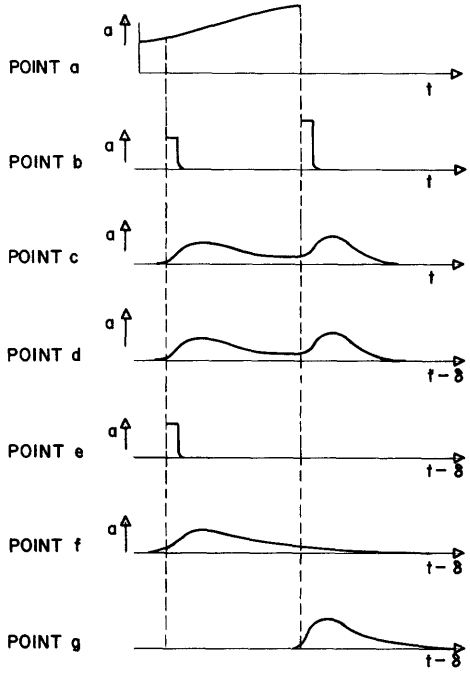

Fig. 5 Waveforms in narrow-band communication system.

out by several workers, long delay is an unavoidable concomitant of any system conserving bandwidth to the utmost. This does not, however, affect the rate of transmission of information, the quantity under consideration here.

As a result of the considerations given above, we are led to the conclusion that the only limits to the rate of transmission of information in a noise-free circuit are economic and practical, not theoretical.

VI. Transmission of Information in the Presence of Noise

In some ways, the discussion of the section immediately preceding this represents a digression in the main argument. It may be well, therefore, to review the main argument at this point, and to indicate the direction it is to take. So far, Hartley's definition of information has been investigated and shown adequate for this analysis. The previous theories of transmission of information have been refuted. In the portion of the work which follows, a modified version of the Hartley law, applicable to a system in which noise is present, is derived. This is done for the general case and for two special types of wide-band modulation systems, uncoded and coded systems. As a result of these analyses, the fundamental

relation between rate of transmission of information and transmission facilities is derived.

Since we have shown that intersymbol interference is unimportant in limiting the rate of transmission of information, let us assume it is absent. Let S be the rms amplitude of the maximum signal which may be delivered by the communication system. Let us assume, a fact very close to the truth, that a signal amplitude change less than noise amplitude cannot be recognized, but a signal amplitude change equal to noise is instantly recognizable [14]. Then, if N is the rms amplitude of the noise mixed with the signal, there are 1 + S/N significant values of signal which may be determined. This sets s in the derivation of Eq. (1). Since it is known (13) that the specification of an arbitrary wave of duration T and maximum component f_c requires $2f_cT$ measurements, we have from (1) the quantity of information available at the output of the system:

$$H = kn \log s = k2f_cT \log (1 + S/N). \qquad (2)$$

This is an important expression, to be sure, but gives us no information, in itself, as to the limits which may be placed on H. In particular, f_c is the bandwidth of the overall communication system, not the bandwidth of the transmission link connecting transmitter and receiver. Also S/N may not at this stage of the analysis have any relation to C/N, the ratio of the maximum signal amplitude to the noise amplitude as measured before such nonlinear processes as demodulation which may occur in the receiver. It is C/N which is determined by power, attenuation, and noise limitations, not S/N. Similarly, it is bandwidth in the transmission link which is scarce and expensive. It is, therefore, necessary to bring both these quantities into the analysis and go beyond Eq. (2).

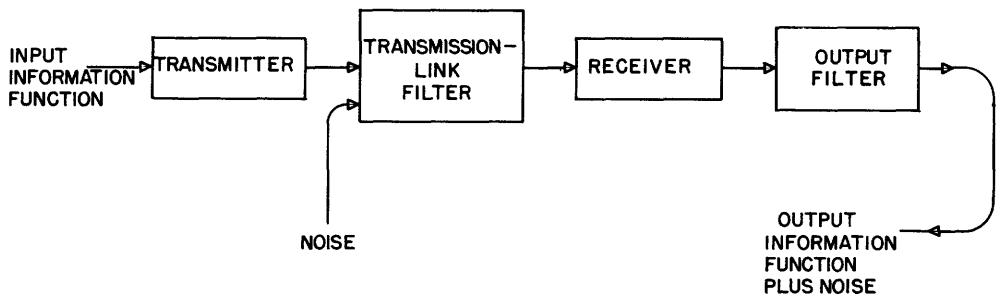

Fig. 6 Block diagram of simplified communication system used in analysis.

The transmission system assumed for the remainder of this analysis is shown in the block diagram of Figure 6. The elements of this system

(14) This assumption ignores the random nature of noise to a certain extent, resulting in a theoretical limit about 3 to 8 db above that actually obtainable. The assumption is believed worth while in view of the enormous simplification of theory obtained. For a more precise formulation of the theory, see footnote references 9 and 10.

may be considered separately. The transmitter, for example, is simply a device which operates on the information function in a one-to-one and reversible manner. The information contained in the information function is preserved in this transformation.

The receiver is the mathematical inverse of the transmitter; that is, in the absence of noise or other disturbance, the receiver will operate on the output of the transmitter to produce a signal identical with the original information function. The receiver, like the transmitter, need not be linear.

It is assumed throughout the remainder of this analysis, however, that the difference between two carriers of barely discernible amplitude difference is N, regardless of carrier amplitude. This corresponds to an assumption of overall receiver linearity, but does not rule out the presence of nonlinear elements within the receiver. This assumption is convenient but not essential. If it does not hold, the usual method of assuming linearity over a small range of operation and cascading these small ranges to form the whole range may be used in an entirely analogous analysis with essentially no change in method and only a slight change in definition of C/N and S/N, here assumed to be amplitude-insensitive.

The filter at the output of the receiver is assumed to set the response characteristic of the transmission system. (It should be noted that when reference is made to "transmission system", all the elements shown in Figure 6 are included. "Transmission link" refers only to those elements between the output of the transmitter and the input to the receiver.) The transmission characteristics of this filter are, therefore, those previously given for the overall transmission system. Coming now to the elements of the transmission link, consider first the filter which sets the link's transmission characteristics. The phase shift of this filter is assumed to be linear with respect to frequency for all frequencies from minus to plus infinity. The overall attenuation is assumed to be zero decibels at all frequencies less than B, and is assumed to be so large for all frequencies above B that energy passing through the system at these frequencies is small in comparison with the unwanted disturbance (or noise) present in the output of the system. It should be obvious that this characteristic may be made band-pass or high-pass by the well-known transformations.

The noise is assumed to have a power spectrum whose amplitude is uniform over the range of frequencies passed by the filter in the transmission link. The noise spectrum is, therefore, identical with that of the pass band of the receiver.

From the above discussion, it is apparent that the transmission system sketched in Figure 6 is a close approximation to most communications systems

in which only one source of noise is important. The transmitter can be anything from a pair of wires connecting input and output up to and beyond a pulse-code-modulation generator modulating a high-frequency carrier. The lumping of noise into one generator, lumping the transmission characteristics of the link into one filter, and lumping the transmission characteristics of the overall system into one other filter, as well as the assumption of linearity, are admitted to be unreal assumptions which, however, come reasonably close to the true facts — close enough for engineering purposes in many instances. The special shapes of the transmission characteristics assumed are, as has been mentioned, chosen for convenience and not necessity.

In addition to the general case, it is interesting to consider two special cases:

1. Uncoded: To every specified and unique point in the information function, there corresponds one specified and unique point in time of the information function as transformed by the transmitter. There are the same number of points in the transformed as in the original function. Information is conserved. The overall time taken for transmission may, but need not, be equal at the input and output of the transmitter.

2. Coded: While information is conserved in this case also, one point in the transformed information function may, in this case, be specified so accurately as to contain all the information contained in a whole series of rather inaccurately specified points in the original information function, or vice versa. The transformation again must be reversible.

The uncoded transmission corresponds to direct transmission of the information function, transmission of an amplitude-modulated carrier, transmission of a frequency-modulated carrier, and the like. Coded transmission corresponds to pulse-code modulation or similar systems.

The points to be shown about the three types of transmission are as follows:

1. In general, for large signal-to-noise ratios, the signal-to-noise ratio may be equal to or less than the carrier-to-noise ratio raised to the power B/f_c. (See Eq. (7).)

2. Coded transmission is capable of realizing the fullest capabilities of the general system; i.e., signal-to-noise ratio may equal carrier-to-noise ratio raised to the power B/f_c, for large signal-to-noise ratios. (See Eq. (17).)

3. In uncoded transmission, the signal-to-noise ratio may be equal to or less than the carrier-to-noise ratio multiplied by the factor B/f_c, for large signal-to-noise ratios. (See Eq. (19).)

These points will be discussed separately in the order given.

Let us first consider the general system. In this case, making the assumption that a change in carrier voltage equal to rms noise is just detectable, and applying the reasoning that led to Eq. (2) to the receiver input, we have, for the "quantity of information" at the receiver input,

$$H_{in} = k \cdot 2BT \log (1 + C/N) \quad . \tag{3}$$

We know from (2) that the "quantity of information" at the output of the receiver is

$$H_{out} = k \cdot 2f_c T \log (1 + S/N) \quad . \tag{4}$$

The receiver cannot be a source of information. By this we imply that to every value of C/N there corresponds one and only one value of S/N. This must be true if the system is to operate in the absence of noise, since otherwise there might correspond more than one value of S for a given C, an unworkable situation. We may, however, lose information in the receiver; i.e., it may not be perfect. Allowing for this,

$$H_{out} \leqslant H_{in} . \tag{5}$$

Substituting (3) and (4) in (5) and clearing like quantities and logarithms from both sides of the inequality gives

$$(1 + S/N) \leqslant (1 + C/N)^{B/f_c} \quad . \tag{6}$$

Or, if $C/N \gg 1$ and $S/N \gg 1$,

$$S/N \leqslant (C/N)^{B/f_c} \quad . \tag{7}$$

Let us now consider coded transmission. In this case, as will be shown by an example, the equals sign of Eq. (6) may be achieved. Suppose, for example, we wish to transmit the message of Figure 3(e). Clearly, this requires a carrier-to-noise ratio of at least 477,146 if it is to be transmitted as an amplitude-modulated pulse. Suppose, however, a carrier-to-noise ratio of but unity is available, so that the best we can do is distinguish between carrier off and carrier on. In this case, we may still transmit the message in the form of Figure 3(b), essentially coding it in binary digits. Here, if we wish the message to be transmitted in the same time, we must transmit it in nineteen times as much bandwidth, since we must transmit nineteen time units during the duration of our message, instead of the original single pulse, or time unit. At the receiver, the pulses of Figure 3(b) may be deciphered to form the single pulse of Figure 3(e). One must be careful how he uses this data to avoid error. The various quantities of (6) might erroneously be considered to be, for

this example,

$$1 + S/N = 477,147 \quad , \tag{8}$$

$$1 + C/N = 2 \quad , \tag{9}$$

$$B/f_c = 19 \quad . \tag{10}$$

We find, however, that

$$2^{19} = 522,288 \quad , \tag{11}$$

and, therefore, in this case

$$(1 + S/N) < (1 + C/N)^{B/f_c} \quad . \tag{12}$$

This does not correspond to the earlier statement that the equals sign of Eq. (6) can be realized. However, one message which could be sent over our system would be a long dash of 19 time-units duration. The number corresponding to this dash would be 522,287. This fact gives a clue to the application of (6). In using this formula, $(1 + S/N)$ must be the number of possible allowed states of the receiver output at any one time, and $(1 + C/N)$ the number of possible allowed states of the receiver input for any one instant of time. In other words, $(1 + S/N)$ is actually s, measured at the output of the receiver, and $(1 + C/N)$ is s, measured at the input to the receiver. Considering things in this correct manner, we have

$$(1 + S/N) = 522,288 \quad , \tag{13}$$

$$(1 + C/N) = 2 \quad , \tag{14}$$

$$B/f_c = 19 \quad , \tag{15}$$

$$2^{19} = 522,288 \quad , \tag{16}$$

$$(1 + C/N)^{B/f_c} = (1 + S/N). \tag{17}$$

If $C/N \gg 1$ and $S/N \gg 1$,

$$(C/N) = (S/N)^{B/f_c} \quad . \tag{18}$$

Thus, we see that in this manner, at least, the equals sign of Eq. (16) may be realized.

Now let us consider uncoded transmission. In this case, one and only one functional value is specified for each original time value; that is to say, the total number of samples taken of the wave is maintained constant. However, to each unit, or quantum, of time in the original information function there correspond B/f_c resolvable units, or quanta, of time on the transmitted information function. This was, of course, also true in the case of the coded transmission. Now, however, we specify that during all but one of these B/f_c units the function be zero. During each of these

periods, we may have the carrier-to-noise ratio C/N and hence a possible range of values (1 + C/N).

Corresponding to our original possibility of one point on the original information function with any of (1 + C/N) possible significant amplitudes, we now have a possibility of one point [15] with any one of B/f_c times (1 + C/N) possible significant values. This comes about because the point may have (1 + C/N) possible significant amplitudes and (in consequence of the improvement in system resolution capability by the factor B/f_c) B/f_c possible independent time values. We have, therefore, increased the number of degrees of freedom of each point by the factor B/f_c. This argument is illustrated in Figure 7, showing the original and transformed signals in

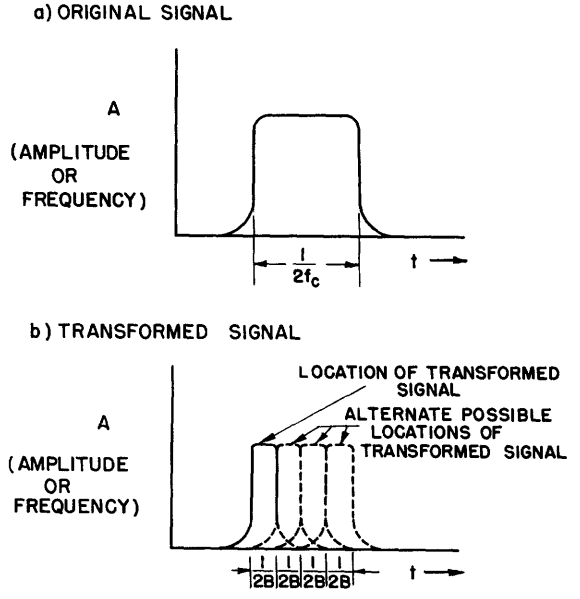

Fig. 7 Uncoded transformations.

amplitude-time and frequency-time space. We can, therefore, make one of B/f_c times C/N possible choices for this point. We have, therefore, as the number of possible independent states for the receiver over any period of $1/2f_c$ seconds,

$$(1 + S/N) \leqslant B/f_c \cdot (1 + C/N). \tag{19}$$

Or, again, if $S/N \gg 1$ and $C/N/N \gg 1$,

$$S/N \leqslant B/f_c \cdot (C/N). \tag{20}$$

Again the equality can be realized, as in the well-known frequency-modulation system [16].

(15) In contrast to the coded case, we are here forbidden to employ more than one point by the very definition of uncoded transmission.

(16) The absence of the $\sqrt{3}$ from Eq. (20) in the frequency-modulation case can be shown to be due to differences in definition of bandwidth between that assumed here (for nonrectangular spectra) and that generally used in the conventional FM analyses.

It is interesting to apply the results of Eqs. (18) and (19) to the noise-free system. In this case, C/N is infinite. Therefore, if H is a very large but finite quantity, either T or B may equal zero. Thus we are led directly to the refutation of the early theories of transmission of information.

VII. The Use of Coded Information Functions

Binary coding is the only type of coding which has received attention in published papers. This type of coding represents the maximum sacrifice in bandwidth requirements and maximum decrease in required carrier-to-noise ratio, since the latter is reduced to unity. Time of transmission is unchanged in this coding system, as is true in most systems designed for two-way transmission.

The publicity given to binary coding, and its value in systems employing regenerative repeaters, have overshadowed the potentialities of other coding schemes. For example, consider the national radio-broadcasting system. The standard system employing amplitude modulation has been recently supplemented by a frequency-modulation service offering increased freedom from noise at a cost (not considering the additional transmitted audio-frequency range provided) of five times the spectrum width per station. Since frequency modulation is an uncoded transmission system, one expects that the signal-to-noise ratio would be improved about five times by this band widening. Suppose we wish to obtain this improvement by coding. We might choose a double-band code system, in which each point on the original information function was replaced by two points on the transformed information function. Taking a signal-to-noise ratio of 1000 (60 db) as a reasonable figure, this could be accomplished with a carrier-to-noise ratio of but 32 (30 db). A frequency-modulation system similar to the present standard would require a carrier-to-noise ratio of about 120 (42 db)to accomplish this, or about four times the transmitter power, and, moreover, would use 2.5 times the bandwidth [17]. If higher signal-to-noise ratios were required, the difference between the two systems would be even more spectacular.

Coding may also be used to reduce bandwidth or time of transmission at the expense of carrier-to-noise ratio. The bandwidth required for transmission, for example, may be halved by combining the information contained in two points in the information function into one point on the transformed

---

(17) If additional advantage were taken of single-sideband operation, the coded signal could be accommodated in a standard broadcast channel and give the same signal-to-noise ratio as frequency modulation with one-fourth the power and one-fifth the required spectrum.

information function. This requires that carrier-to-noise ratio be the square of that required without coding, but does accomplish the required results. For example, if each point were expressed originally to an accuracy of 1 part in 10, the first might be nine units in amplitude and the second three. A system capable of transmission accuracy of 1 part in 100 could transmit 1.93 units high, which is easily recognized as containing all the information present in the previous two. The transmission of this one point might use the time previously required to transmit the previous two, and thus would require but half the bandwidth. There is no theoretical advantage to be gained in using bandwidth for transmission in excess of that required for the transmission of the coded information function; i.e., wide-band modulation systems using uncoded transformations are inherently inefficient in their utilization of spectrum. If a carrier is unnecessary and the signal-to-noise ratio attained by simply transmitting the information function directly is just adequate, then this is the most efficient utilization of spectrum. If a carrier must be used, single-sideband amplitude modulation, narrow-deviation frequency modulation, or some other "narrow-band" modulation system should be used. Coding may be used as desired to gain in one parameter at a sacrifice in some other or others without any loss in efficiency.

It should be pointed out that economic factors not considered in any of the analysis above modify these theoretical considerations. In particular, in the present state of the art, coded transmission is more suitable for point-to-point communication in which the number of transmitters and receivers are equal, rather than for broadcasting, where one transmitter services many receivers. In the latter case, an uncoded "brute force" scheme may be desirable, putting the burden on a big transmitter in order to permit the simplest possible receivers. The existence of such a situation should, however, only point the way to needed improvements in the art.

VIII. The Function With Maximized Information

Until now we have considered a rather general type of information function, limited only by a finite width of spectrum. It is of some importance to consider the amount of actual irreducible information contained in such a function. The transmission of information involves the transmission of one of a set of possible alternative choices. If certain analytic properties of the information function make the selection of a particular choice mandatory at some time, no actual choice is made, since no alternatives can exist. Continuation of this line of reasoning, as is shown below, leads to the possibility of reducing the bandwidth required to

transmit many types of information function. Further, this analysis shows that one particular type of information function, here called the function with maximized information, conveys the maximum intelligence from one point to another for a given set of transmission facilities. This function has the general characteristics of filtered random noise, except for its distribution function.

To derive the characteristics of this function, let us first consider the definition of "quantity of information." This definition was arrived at by considering a series of n selections, each made from a set of s possible choices. It should be obvious that if, in some selections, we choose from only $s - j$ possible choices, we transmit less information than if s choices were available each time a selection were made. Nothing has been said previously about this point, but it should be recognized that s need not be a constant during a message, but may vary with time.

If s is not a constant, we must provide system facilities adequate to transmit the maximum value of s ever realized in the message. Considering the transmission link of Figure 1, we have, therefore,

$$C/N \geqslant s_{max} - 1 \quad . \tag{21}$$

The actual quantity of information contained in a system with variable s is

$$H = kn \log s_{av} \quad , \tag{22}$$

where s_{av} is the average value of s, obtained in the conventional manner. Thus, in this case, since

$$s_{max} \geqslant s_{av} \quad , \tag{23}$$

the formula

$$H = k2BT \log (1 + C/N) \tag{24}$$

no longer holds, and, in fact, we have

$$H \leqslant k2BT \log (1 + C/N) \quad . \tag{25}$$

To realize the equals sign in Eq. (25), we must achieve the equals sign in (23). This can be done only when s is a constant, since only in this case does $s_{max} = s_{av}$.

We have, therefore, the fact that if s is not constant during a message, the transmission of that message will require more time, bandwidth, or power than would be necessary to transmit the same quantity of information in a form in which s were constant. We now ask the implications of this statement. These are that unless, at any instant of sampling, the

sample is equally likely to take on any of its allowed significant values, we are wasting time, bandwidth, or power; and further, that every message should be examined in detail for possible long-time-interval coherences before transmission. This implies delay and storage in the transmission of a message, since we can make sure that s is constant only by examining every portion of the complete message.

To carry the argument a step further, suppose the future amplitude of the function may not be exactly determined in the future from a knowledge of the function in the past, but that it may be determined to a certain probability. Then only the range of values having high probability need be transmitted, since, by omitting that range having low probability, power is made available to transmit the high-probability range with greater accuracy. To give a numerical example, suppose that from some knowledge of the information function it is determined that the amplitude of the function will be within 10 per cent of the possible amplitude range at a given instant of sampling, to a probability of 0.9. Suppose the system has an s of 100. In this case, a range of ten possible significant amplitudes will have a probability of occurrence of 0.9 and the remaining range of ninety has a probability of 0.1. We may then let the more probable range occupy 50 per cent of the scale, by a prearranged scheme, and express this range in fifty significant steps, rather than in just ten. The accuracy of reproduction, so far as this most-probable region is concerned, is thereby increased by a factor of five, at the expense of a similar reduction of accuracy of the remainder of the scale. Therefore, 90 per cent of the time the effect is to transmit with s five times as great as was formerly the case; 10 per cent of the time the effect is to reduce s by a factor of five. The average effect is roughly to increase s by 4.5, and the quantity of information which may be transmitted over the system by the logarithm of this quantity.

Another way of stating this requirement of maximized information is to state that there must be no possibility of analytic continuation of the information function to an accuracy of better than one part in s for the duration of the interval between samples. This may readily be arrived at by consideration of the arguments above.

To summarize, any information function which is not one of maximized information will require more time, bandwidth, or power to transmit a given quantity of information than will the maximized information function. It is, therefore, extremely important that, in any transmission requiring maximum efficiency in utilization of these three parameters, one make sure the input wave is one of maximized information: the spectrum of such a wave, if the interval between samples is constant, is that of white noise

passed through an ideal low-pass filter. This is a convenient, although not sufficient, method of assuring that such a function has been obtained.

IX. Application To Other Fields

The point of view developed in the work described above has already been very useful in the analysis of systems not generally considered as belonging to the communications family, but which, as several people have recently come to believe, should be so classified. Typical general fields in which information-transmission problems occur and, in fact, may completely govern system design are radar, radar relay, telemetering, servomechanisms, and computing mechanisms of the digital type. Application of the viewpoint here developed can show possible simplification in system design, unrealized information-handling capability, or the use of a system inefficient in that it supplies more information than is required.

Let us consider the radar problem. At the moment, we shall be concerned only with radar search in two dimensions, azimuth and range, although expansion of the theory to three-dimensional search systems offers only slight additional complications. The problem to be solved by the radar is the determination of the existence or nonexistence of a reflecting body at any point within the range of the equipment. A refinement which might be useful, although not always essential, consists in knowing the "electrical size" of the target, i.e., the strength of the reflected signal. Service codes for reporting echo signal strength recognize only five possible strengths, realizing the difficulty of estimating by eye the amplitude of a constantly fluctuating signal on the face of a cathode-ray tube. Therefore five digits and a zero are enough to tell all the significant facts about signal strength. If the maximum range of the equipment is R, and the desired range accuracy $\pm$ r, then there will be a total number R/2r ranges at which a target may be said to be located. Similarly, if search is to be carried out over 360°, to an azimuth accuracy of $\pm$ B degrees, there are 360/2B possible azimuth positions in which a target may be located. The total number of integers which must be transmitted to the operator for each complete scan is, therefore, (R/2 + 360/2B). Each integer has six possible values, ranging from zero to five. A five-to-one carrier-to-noise ratio is, therefore, all which is required. This assumes separate search in range and azimuth. An alternative way of searching is to examine separately each elemental area bounded by the concentric range-accuracy circles and the radial angular-accuracy lines. This involves the transmission of (R/2r $\times$ 360/2B) integers, each having six possible values. Since the quantities involved are such that the second system of scanning always results in the transmission of more information than the first, we may say

at once that the first scheme of scanning looks more efficient than the second, and ask why. The answer is not long in coming. The first system of scanning is adequate so long as there are never two targets in the same range-accuracy strip, or the same azimuth-accuracy wedge. If at any time there are two targets in such an area, the system will become confused and will report but one. Taking typical numbers to see the cost of this degree of data separation, R might be 100 miles, r, 1/4 mile, and B, 1 degree. Then the first system of scanning calls for the transmission of $200 + 180 = 380$ integers, while the second calls for the transmission of $200 \times 180 = 36{,}000$ integers. Therefore the second system of scanning should require almost one hundred times the bandwidth or transmission time of the first, holding the signal-to-noise ratio constant at five. This is the cost of the freedom from confusion. It would seem that the first type of scanning could be used advantageously for early-warning systems, so sited that target confusion is unlikely. The resultant decrease in information-handling capacity required of the system could be utilized in system design to make possible the use of lower power or narrow bands or decreased search time. On the other hand, these parameters could be held constant and the effective range of the system increased until its information-handling capacity was fully utilized.

If we assume that target confusion is highly probable, then it may be worth while to examine the second system of scanning in some detail, to see if modern radar systems are as efficient as they might be. As we have observed, some 36,000 integers must be transmitted during each complete scan. Each of these integers must be transmitted during each complete scan. Each of these integers has one of six possible values. Suppose, as is reasonable, that we wish to scan the complete area under surveillance once per minute. The information must then be transmitted at a rate of 600 integers per second. A bandwidth of 300 cps is all which is required to transmit this information at a five-to-one signal-to-noise ratio. Other signal-to-noise ratios may be accommodated, or taken advantage of, by coding, with a resultant change in required bandwidth. The high speed of scan in range, forced by the velocity of propagation of radio waves, is immutable in radar systems of the pulsed type and, therefore, one must accept a wide band of frequencies. It should not be necessary, however, to force the indicator circuitry to respond at this speed. If a delay and storage circuit is provided, it could accept signal information in short bursts and disgorge this information at the uniform rate of 600 integers per second, for use by the indicator and operator.

Considering now a remote indicator or relay for the radar system outlined above, one could be provided with a bandwidth of but 300 cps and a

minimum signal-to-noise ratio acceptable for good results of but five, using techniques already at hand. The storage could, perhaps, be provided by the conventional long-persistence-screen cathode-ray tube used as a radar-system indicator, with pick-off of data provided by television or facsimile methods at very low scan rates. The cases considered here do not, it should be mentioned, contemplate relaying more information than that present on the local indicator, in contrast to some other proposed systems.

The telemetering problem is similar in many respects to the radar-relay problem but simpler in that only one-dimensional information need be transmitted, usually a meter reading. If 1 per cent accuracy is required, the problem is that of transmitting one of fifty integers, at whatever rate is desired. The bandwidth required is, therefore, half the desired rate, and the signal-to-noise ratio a minimum of fifty. If this signal-to-noise ratio is not attainable under the most unfavorable conditions, best spectrum utilization is obtained, not by resorting to uncoded modulation schemes, which have been shown to be inefficient, but by coding the information. This can be carried out to the point where a signal-to-noise ratio of unity is adequate for the accuracy of data transmission required; but, in these circumstances, the bandwidth or time of transmission required must be increased by a factor of almost six, for most cases. To illustrate the gain in efficiency of this method over the various conventional but uncoded wide-band schemes, it need only be noted that any of them would require a bandwidth increase of fifty to accomplish the same results. Narrowest band of transmission and least time of transmission are, of course, always obtained by operating the system at the lowest usable signal-to-noise ratio.

A servomechanism may be regarded as a communication system. Its function is to communicate the position of some object, such as a rotatable shaft, to a distant point, and there to cause another object to move in accordance with the motions of the first. The motion of the first object may be, but seldom is, known with absolute accuracy; the motion of the second must always be specified to within certain definite limits. Uncertainties arise in the link between transmitter and receiver; these may be due to backlash, electrical or mechanical noise, instrument imperfections, and the like. The sum of these uncertainties in the transmission link corresponds to the noise discussed in the theory outlined above. The position of the second object, the output member, corresponds to the information required to be transmitted over the system. This information may be considered as a group of integers, each corresponding to a possible output-member position. If the static position of the output member is to

be specified to 1 per cent, then one of fifty possible integers must be transmitted every time the input member moves through one one-fiftieth of its possible range. It may be that certain elements in the transmission link limit the accuracy of the system (its effective noise-to-signal ratio) to 5 per cent. The conventional thing to do in this case is to transmit the data at a higher rate with multiple-speed data-transmission systems. This is actually a coding scheme since, effectively, the single integer required to specify the position of the output member is transmitted as a two-digit integer, where the first digit transmits the rough data and the second the more precise. In this case, therefore, accepted practice coincides with the most efficient.

A feature of the scheme of analysis presented in this Report is the breakdown of a continuous smooth curve of data into a series of equispaced points, the value of each point being restricted to one of a number of integral values. Since this technique of analysis is exactly that used in the solution of problems by digital computing machines, one might expect to find correlation between the problems found in this field and those discussed above. To some degree, this is true at first glance, and a more thorough study would perhaps prove fruitful. For example, most new computing machinery uses coding to express numbers in the binary system of units; we have seen here that coding in the binary system of units obtains the maximum signal-to-noise ratio for a given carrier-to-noise ratio, consistent with the amount of frequency spectrum utilized. We may say, therefore, that the accuracy of the machine is least affected by noise and perturbations introduced in any of its various transmission links if the data is transmitted by the binary system. Therefore this is logically sound.

* * *