THEORETICAL LIMITATIONS ON THE RATE OF TRANSMISSION OF INFORMATION

by

WILLIAM G. TULLER

S.B., S.M., Massachusetts Institute of Technology

(1942)

SUBMITTED IN PARTIAL FULFILLMENT OF THE

REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

DEPARTMENT OF ELECTRICAL ENGINEERING

(June 1948)

Signature of Author.....
Department of Electrical Engineering, June 1948

Signature of Thesis Supervisor

Signature of Chairman.....
Department Committee on Graduate Study

Acknowledgment

The philosophy of communication presented here is the outgrowth of discussions with many people over a period of time of some ten years. All cannot be identified here, but prominent among them are Dr. W. M. Hall, who first aroused my interest in the problem in classes at the Institute, Professor N. H. Frank, who encouraged a broad research program, Professors E. A. Guillemin and H. Wallman, who gave much helpful criticism, L. Katz, with whom were held many friendly discussions on the problem, G. E. Duvall, who helped teach me basic statistical theory, and others. Group credit is due the administrators of the Research Laboratory of Electronics, who provided a well-equipped laboratory in a research-stimulating atmosphere, and the Signal Corps, Air Forces, and Navy, who supported the work under Signal Corps Contract No. W-36-039 sc-32037. Credit is also due Mr. R. V. L. Hartley, whose work first posed the problem considered here, and Professor N. Wiener, whose application of statistics made possible an adequate treatment.

Abstract

Following a discussion of the previous history of the problem and a definition of the principal terms employed in the paper, an examination is made of the process of transmitting information by electrical means.

It is shown that previous theories on the limitations imposed on the rate of transmission of information by finite circuit bandwidth are incorrect. A practical system which violates these theories is described in detail and the practical limitation on its rate of transmission of information discussed. No theoretical limit is found in the noise-free-system considered.

The communication system with noise present is then analyzed. It is shown that noise limits the rate of transmission of information. Two types of system are analyzed and the expression for amount of information transmitted as a function of bandwidth, time, and carrier-to-noise ratio found in each case. It is shown that one class is superior to the other. The latter class comprises most of the commonly used systems. The superior system involves coding which is discussed in some detail. A discussion is given of the effect of internal correlation in the information and methods for removing this correlation are considered. A discussion of the application of the theory to practical communication systems is followed by consideration of systems other than those previously considered communications. Possible channels for future work are discussed.

Table of Contents

List of Illustrations

I. Introduction

There has been much written and said about various systems for the transmission of information. Such topics as frequency modulation, pulse modulation, etc., have all been discussed in considerable detail. The relative merits of the various systems as regards noise suppression have been analyzed with varying degrees of accuracy and the proponents of all the systems have written lengthy dissertations in favor of their particular favorite. However, there has actually been little published on the basic limitations underlying all these systems and the general problem of the transmission of information through noise that each modulation system attempts to solve. Further, at least two of the few papers written on this subject have contained conclusions believed erroneous, thus still further beclouding an already hazy scene. It is the purpose of this paper to consider this fundamental problem, to show that those analyses previously published that have neglected the effects of noise are to say the least incomplete, and to show how application of the complete theory, including noise, to various transmission problems might result in improved utilization of bandwidth, transmitter power, time, and receiver sensitivity.

The history of this investigation goes back to 1922 when J. R. 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 Küpfmüller[3],

[1]Numbered references refer to bibliography at end of paper.

working independently, showed that the number of telegraph signals that may be transmitted over a line is directly proportional to its bandwidth. In doing so however, they neglected to point out that knowledge of the transient response characteristics of the circuit may be used to overcome the limitation observed by them. 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 that 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. He further neglected noise.

In 1946, D. 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 and further fell into a pitfall that led him to set a quantitative limit to the amount of information that might be transmitted over a given band in a given period of time, in the absence of noise.

The workers whose papers have so far been discussed failed to 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 N. Wiener[6], who did a searching analysis of that problem which is a large part of the general one, the 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 deserves. Associates of Professor 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. Professor Wiener has himself done work parallel to that presented in this paper, but this work is as yet unpublished, and its existence was only learned of after the completion of substantially all the research reported on here. A group at the Bell Telephone Laboratories, including Dr. C. E. Shannon, has also done similar work, unpublished as yet, but the complete results of this work are not known to the writer. Both Wiener and the Bell Laboratories group have thought about the problem from the viewpoint of the statistician and are known to have included noise in their considerations, so their work should give comparable results, as is understood to be the case.

To come to the point of the present paper then, the problem to be considered is the transmission of information by electrical means, over a circuit of limited bandwidth, in a finite time, in the presence of noise. Before discussing the transmission of information, it is necessary to define one's terms, particularly information, with care. Accordingly considerable attention will be given to this matter.

Inasmuch as a great deal of the prior art, in fact all of the published portion of it, has omitted noise from the discussions of the transmission of information, it is believed worth devoting considerable attention to a demonstration of the fact that, in the absence of noise, transmission of information at an arbitrarily high rate is possible. Accordingly, a possible system is outlined for communication in a noise-free unquantized world at an arbitrarily high rate. The components of such a system are discussed in some detail and the theoretical justification for such a system shown. In connection with this problem, the matter of coding, or destroying the time character of a signal in such a manner as to use less of one transmission facility at the expense of another is brought up briefly to be taken up again in more detail in a later section of the paper.

Since communication in a noise-free, unquantized world is shown in this paper to be possible at an arbitrary rate, attention must be paid to a more complex case if one is to find any limit on the possible rate of transmission of information. A next higher approximation, the system with one source of noise, is therefore the next concern of this paper. This approximation is sufficiently close to the practical case to be immediately applicable to most communication systems. The result of the analysis of such a system shows that noise is the factor limiting rate of transmission of information, that it does so according to a perfectly calculable relation, and that any one facility, --time, bandwidth, or power-- may be decreased at the expense of an appropriate increase in either or both of the other two in the transmission of a given quantity of information.

So far nothing has been mentioned about systems of modulation, even though noise has been considered. It is shown in this paper that those systems of modulation now in use have essentially realized the maximum performance obtainable from them. The earlier wide band modulation systems such as frequency and pulse time modulation, trade bandwidth for power (or signal-to-noise ratio for a given power) according to a relation less efficient than the ideal one derived here, and applicable to the recently announced pulse code modulation. This is shown to be inherent in the general approach followed in realizing the two classes of modulation systems. The expression relating bandwidth, power, and time for the earlier systems is derived, and from it the theoretical disadvantages of these systems shown. Further, the possibility for modulation systems giving more efficient use of facilities is investigated and shown to be non-existent.

II. Definitions of Terms Frequently Used

Certain terms are used in the discussion to follow which are either so new to the art that accepted definitions for them have not yet been established, or have been coined for use in connection with the research here reported. The definitions used in this report for these 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", published by the Institute of Radio Engineers.

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.

Sampling -- Sampling, or Instantaneous Sampling as it is sometimes called, is the process of obtaining a sequence of instantaneous values of the information function. These values are called instantaneous samples or, more simply, samples.

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 equivalued increments.

Carrier-to-Noise Ratio -- The carrier-to-noise ratio is the ratio of the root mean square signal voltage to that of the noise after selection and before any non-linear process such as amplitude limiting and detection. Each voltage is measured in the absence of the other.

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). This time varying voltage will mathematically have associated with it a spectrum containing all frequencies from zero to infinity, but it is customary for the communications engineer to set an upper and lower limit to the range of frequencies transmitted by the communication system and to state that the portion of the spectrum of the information function lying between these limits is enough to convey adequately the information contained in the information function. This selection of the limits of the pass band of the system always represents a compromise between physical and psychological requirements, and is generally complicated by economic factors. This problem will not be considered here, but for the purposes of this paper it will be assumed that a width of pass band has been selected and the lower and upper limits of this band 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 need not be transmitted. These assumptions are, it is realized, somewhat arbitrary, since a system satisfying them would cause considerable transient distortion of certain types of information function.

Further, these assumptions are recognized as specifying a transmission characteristic that cannot be physically constructed and is even mathematically unrealizable. However, these assumptions will serve as a first approximation. The exact statement of the transmission characteristic assumed for the system is then:

> The phase shift of the system is assumed to be linear with respect to frequency for all frequencies from minus to plus infinity. The overall attenuation of the system is assumed to be zero decibels at all frequencies below a cut-off frequency f_c and is assumed to be so large for all frequencies above f_c that energy passing through the system at these frequencies is small in comparison with the unwanted disturbances or noise present in the output of the system. It is obvious that this characteristic can be made band pass or high pass by the well known transformations.

It is our task to transmit over such a system an information function whose power spectrum follows essentially the same specifications as the amplitude characteristics of the system, in such a manner that the function shall suffer as little distortion as possible and use as little facilities as is practical. Before considering the transmission problem however, we must consider the function itself, and our measure of information content.

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 that 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[9] 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 I, 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 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 the function

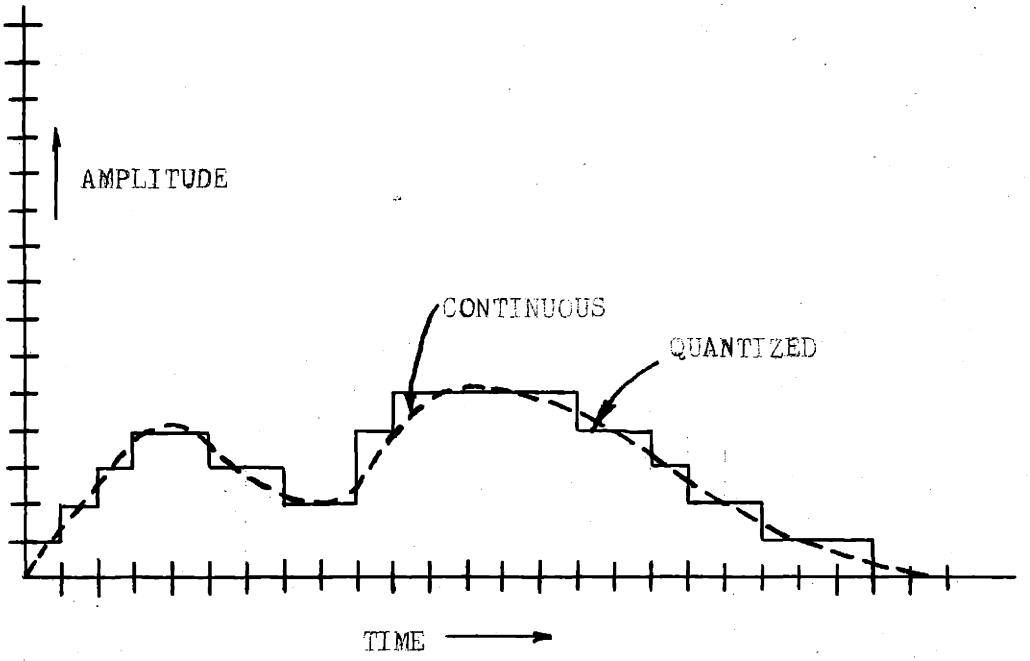


Figure I

The Information Function

of Figure I, quantized, and drawn in n, s space is given in Figure II.

The steps outlined in the preceding paragraph are useful in showing the physical significance of the parameters n and s to be used in the work below. The question now before us is "What is the information content of a function in the n, s plane such as that sketched in Figure II?" The answer of Hartley is the "quantity of information" given by

$$H = k\, n \log s \qquad (1)$$

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." 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 that 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.

Considering condition a) above, this means that

$$H \propto n \qquad (2)$$

Further, since H must be a function of n and s

$$H \propto nf(s) \qquad (3)$$

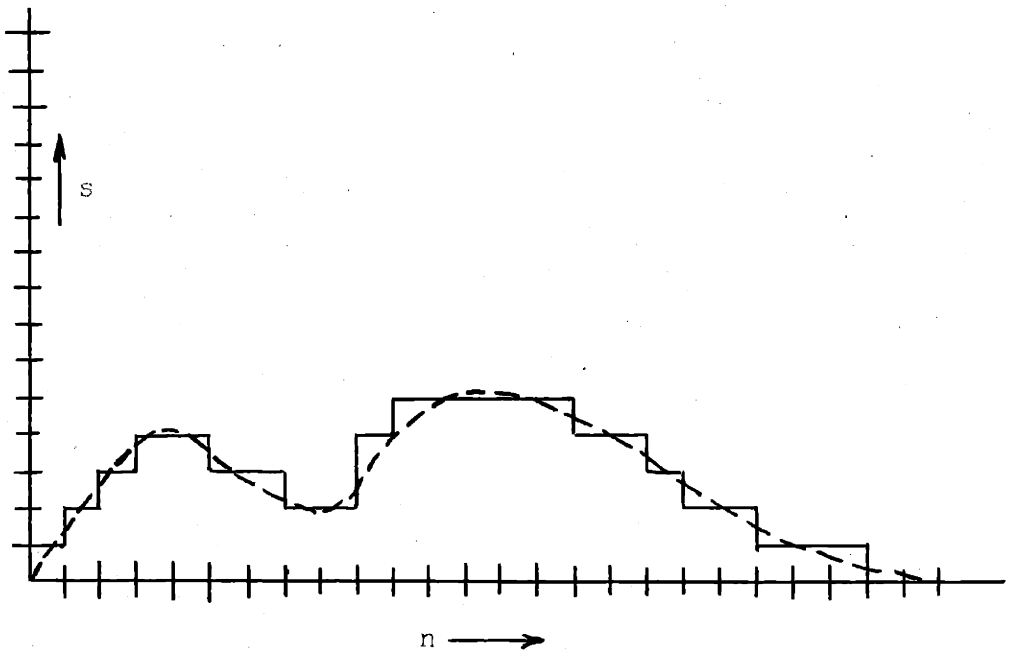


Figure II

The Information Function in n,s Space

where f(s) is some as yet undetermined function of s. Considering condition b), suppose we have two messages with quantity of information H and H_1, respectively. Let us assume

$$H = H_1 \tag{4}$$

From condition b) and Equation (4),

$$s^n = s_1^{n_1} \tag{5}$$

Substituting in (3) gives

$$nf(s) = n_1 f(s_1) \tag{6}$$

Taking the logarithms of both sides of (5)

$$n \log s = n_1 \log s_1 \tag{7}$$

$$\frac{n}{n_1} = \frac{\log s_1}{\log s} \tag{8}$$

Substitute (8) in (6)

$$f(s) = \frac{f(s_1)}{\log s_1} \log s \tag{9}$$

or

$$f(s) = k \log s \tag{10}$$

$$H = nf(s) = k\, n \log s \tag{11}$$

This shows the validity of Hartley's expression for "quantity of information", on the basis of the assumptions made.

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, the previously published works on the transmission of information have neglected noise, and come out with 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 worthwhile 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. Inasmuch as information has been defined in the terms of the previous writers, the difficulty in previous theories is not one of definition but rather of overlooking the possibility of analytical determination of the output of a filter from the knowledge of the input driving function's wave form and the transient response of the filter. The basic fact that has been neglected in earlier analyses and which resulted in their errors is that the output wave form of a network is completely determined for all time by the input wave-form and the characteristics of the network. Knowing these data and measuring the amplitude of the output of the network at any two instants of time gives all the data needed to analytically determine the output of the network at all times past and future. 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. Other types of modulation could be used, but pulse modulation brings out most clearly the principles involved. In these pulse modulation systems the information is carried in a series of recurrent pulses. If these pulses are passed through a filter that one would suspect as being of too narrow band to faithfully reproduce the pulses, there will result intersymbol interference, as shown in Figure III. 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 of 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 be done since, as was mentioned above, the output amplitude is uniquely determined by the input amplitude. Knowing the wave form 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 wave form may be generated locally and subtracted electrically from the output of the filter. Alternatively, the output wave form may be

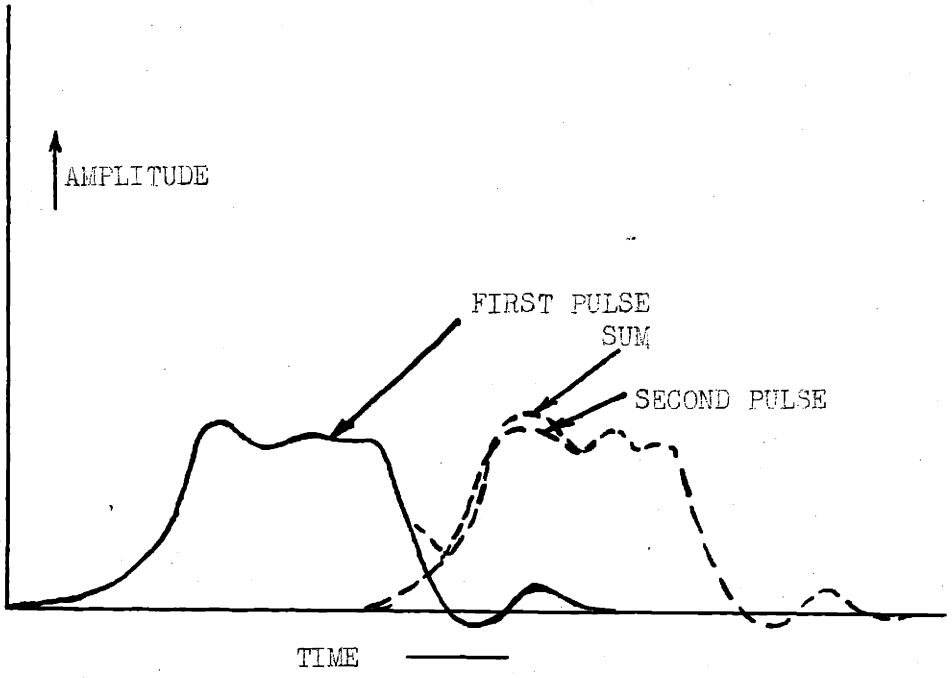


Figure III

Response of Communication System to Rectangular Pulses

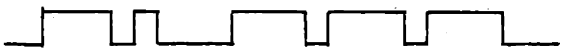
recorded graphically, the component due to the first pulse computed, and this component subtracted by graphical methods, to give the system output free of intersymbol interference caused by this pulse. Either method may be applied repeatedly to remove intersymbol interference from the following pulses. 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 pass band of filter, 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 IV, let us consider transmission of the message "NO". IVa shows the message in English, IVb in Continental Morse Code, IVc in binary digits, and IVd gives the number to the base ten that corresponds to our message. Figure IVe 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 IVe may be used then as our message.

There have been three types of pulses mentioned in the preceding two pages, perhaps causing a certain amount of confusion. To

a) Message N O

b) Coded Message
(Continental Morse Code)



c) Coded Message
(Binary Digits) 111010001110111011

d) Number (to base ten)
Corresponding to Message 477,147

e) Amplitude Modulated Pulse
Corresponding to Message

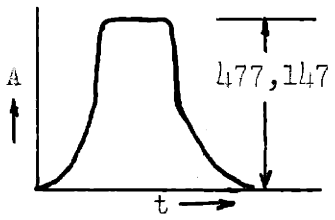


Figure IV

Representation of Message by an Integer

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 solely contained 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 that would have been present at this time ordinarily 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.

Considering a semi-practical arrangement for accomplishing this result, a possible arrangement of tubes and circuits is shown in Figure V, and Figure VI shows some of the waveforms involved in this transmission system, which we assume to be linear. It can be seen from Figure V that at the transmitter the incoming information function is first sampled and used to control the amplitude of a regularly recurring series of pulses. These pulses are of very short duration, for example, a fraction of a microsecond. The pulses are controlled in time by a synchronizing oscillator. The output of the generator of amplitude modulated pulses is to be applied to a low pass circuit whose pass band is much narrower than the reciprocal of the pulse duration and may in fact even be equal to or lower than the highest frequency contained in the spectrum of the information function. The output of this filter is then used to modulate a carrier wave in any of the usual ways, for example by amplitude modulation. The amplitude modulated carrier is then transmitted over a link, which may be wire or space radio, and received by a conventional

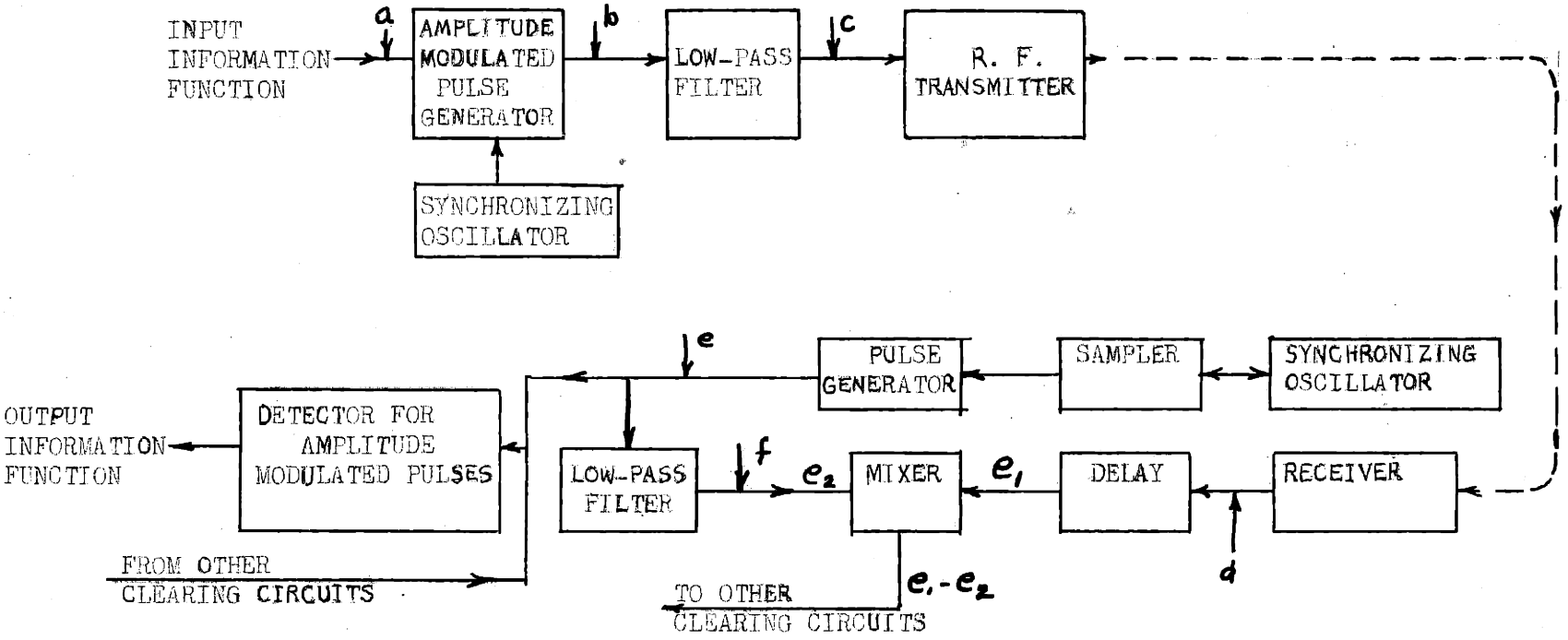


Figure V

Block Diagram of Narrow Band Communication System

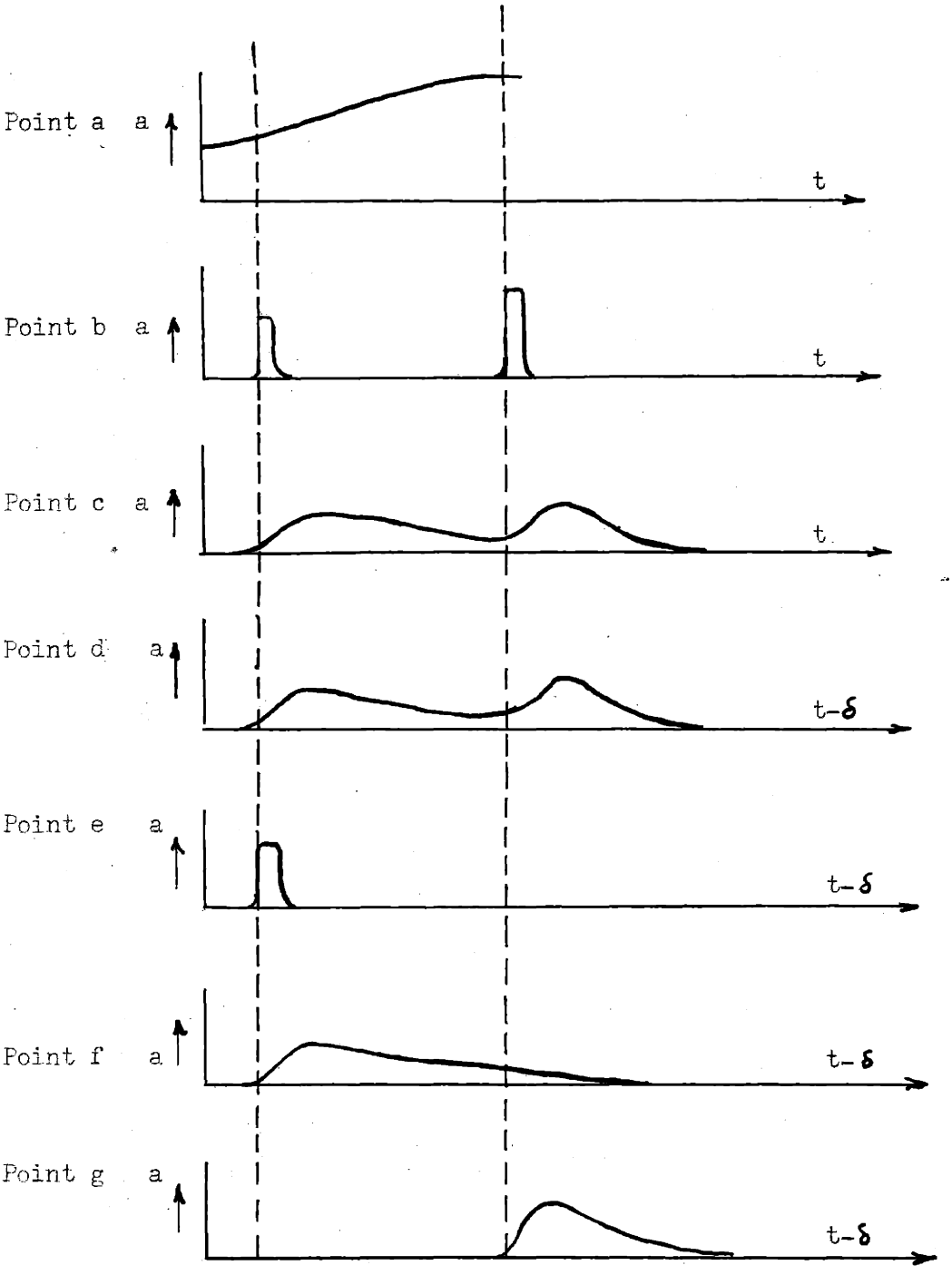


Figure VI

Waveforms in Narrow Band Communication System

receiver. It is the output of this receiver on which we must operate to recover the information which was contained in the original information function. This function was subsequently distorted by having been sampled and transmitted through a narrow band filter. The devices used to perform this operation are shown in Figure V as a sampler, pulse generator, low-pass filter, and mixer. The function of the last named device is to mix the locally generated out of phase intersymbol interference with the incoming signal, after the latter has been suitably delayed.

The transmitter of Figure V is a quite conventional pulse amplitude modulated transmitter except for the low-pass filter. It is assumed for the purposes of this discussion that within the pass band of this filter no other component in the transmitter or receiver has appreciable amplitude or phase distortion. The filter, in other words, controls the frequency response of the transmission system. Wave forms resulting from the generation of a single pulse of the amplitude modulated train of pulses are shown in Figure VI. Figure VIa shows the information function at times close to the sampling time. Figure VIb shows the amplitude modulated pulse generated from the information function, and Figure VIc shows the result of passing this pulse through the low-pass filter. Under the assumptions made about the system, Figure VIc is also the output of the receiver as a function of time. However, this output occurs at a somewhat later time than that at which the pulse was originally generated, thanks to time delays in the transmitter, receiver, and medium of transmission. This fact is shown by Figure VId, the output of the receiver plotted

on the same scale as that used in earlier portions of Figure VI.

The receiver output is now sampled at a time when, according to the receiver's synchronizing oscillator, the output resulting from the single transmitted pulse is a maximum. Since the shape of the pulse and the transient response of the filter are known, the output of the receiver resulting from the first pulse to enter it uniquely defines the amplitude of this pulse. The sampler output may, therefore, be used to control the output of a pulse generator so as to make the pulse generator produce a pulse that is an excellent replica of that generated initially at the transmitter. This pulse may be fed through a low-pass filter having identical characteristics to the one in the transmitter, and the output of this filter will, of course, be an exact replica, delayed in time, of the wave used to modulate the transmitter. If this locally generated wave is combined with the received signal delayed by a proper interval of time, and the difference between the two taken, this difference will be identically zero. We can therefore conclude that the effects of the first pulse to be transmitted over the system have been eliminated and that no intersymbol interference from this pulse remains. That is, the receiver has been cleared of the effects of the first pulse. The output of the mixer, however, will contain components of power from all succeeding pulses. The earliest of these, the second pulse ever transmitted over this system, can be removed from this jumble by another identical process of sampling, pulse generation, filtration and mixing with the delayed output of the receiver. The third pulse can be handled in a similar manner and so on for as long as one has

patience to construct clearing circuits. Obviously there will eventually come a time when the output of the receiver due to the first pulse would have been negligible even without the intersymbol interference eliminator. At this time the first clearing circuit can again be pressed into service so that an infinite chain of clearing circuits is not necessary. Only enough need be provided to cancel out components of intersymbol interference during all the time that the interference produced by the first pulse is important. After this component becomes negligible, the clearing circuits can be reused in order, since the interference from succeeding equispaced pulses will become negligible at just the time when the clearing circuit cancelling out this interference is needed for cancellation of interference produced by a new incoming pulse.

Additional channels can be superimposed upon this system, without requiring the use of additional bandwidths and to a number limited by the resolution of the system (set by the length of pulses generated by the pulse generators) and the number of clearing circuits one is willing to employ, since each additional channel requires the interleaving of as many additional clearing circuits as were used for the first channel.

Suppose just for the sake of discussion that the pulses formed by the pulse generator in the system just described are each a microsecond long. Suppose further that they are repeated at a rate (sampling frequency) of 10 kilocycles per second. The low-pass filter might have a cut-off frequency of 10,000 cycles per second or higher. The output of the mixer in the first clearing circuit would then be zero

from the beginning of the first pulse to come out of the receiver to the beginning of the second, 100 microseconds later in the single channel case. If therefore a second channel were introduced into this 100 microsecond interval by time division multiplex, the output of the first clearing loop would contain, at an appropriate time, a signal coming only from this second channel. This channel could be wiped out by another clearing circuit, inserted in the system between what were initially the first and second. The second pulse transmitted from the second channel would have to be removed by a clearing circuit inserted between what were the second and third loops of the single channel system, etc. This same general scheme could be applied to an m channel system, but if each channel of the system required n clearing circuits, then the m channel system would require a total of mn circuits, usually a rather large number. If one had no objections to providing this number of circuits, he would still reach a limit to the possible number of channels in the time division multiplexing system. For the case described and making reasonable assumptions based on the present state of the art, about 40 or 50 channels could be handled without undue complication. One of these channels would have to be used for synchronization, but the rest could carry information with frequency components up to four or five thousand cycles, resulting in a possible information bandwidth of about 200,000 cycles transmitted in a 10,000 cycle band. Further compression could be realized by the use of shorter pulses, the number of channels being almost inversely proportionated to the pulse duration. If one-twentieth microsecond pulses were used, a four megacycle

channel could be transmitted over a 10,000 cycle band. This would be achieved by two multiplexing processes, assuming the information were contained throughout the whole four megacycle channel. The first of these would be frequency division multiplex, in which each four or five thousand cycle portion of the spectrum of the information function would be filtered out and transferred so that one edge fell at zero frequency. These separate channels would then be superposed by time division multiplex and used to modulate the transmitter. If one were able to assume that the intersymbol interference from any one pulse were negligible ten repetition periods after the pulse (if the power can be assumed to be reduced to about 1% of its original value at this time) then ten clearing circuits would be required per channel, or a total of about 10,000 clearing circuits. This degree of compression is probably more than it would be economic to try in practice, but will serve to illustrate the numbers which one might encounter in such systems.

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 on a noise-free circuit are economic and practical, not theoretical. In other words, one can use the concept of sampling the information function n times per second, having an arbitrary large number of possible values available at each sampling time and therefore being able to transmit an arbitrarily large amount of information per second, limited only by the complications and apparatus one wishes to consider. A counter-example has thus been given to Hartley's law that only a limited amount of information can be transmitted over a given band in a given period of time. This proves that

Hartley's law is false. Other methods of compressing information can, and no doubt will, be devised, some of them simpler than that here discussed.

The next question that may well be asked is if bandwidth does not limit the speed of transmission of information, what does? Considering the various phenomena that limit physical processes one might expect random processes, that is, fluctuations of the microscopic components of our macroscopic world, to be very logical candidates for the role of limiting factors. Detailed analysis will show that this is indeed so.

VI. Transmission of Information in the Presence of Noise

In some ways the discussion of the section immediately preceding this one represents a digression in the main argument to be continued below. 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 that follows, a modified version of the Hartley law applicable to a system in which noise is present is derived. This is done for the two general types of wide band modulation systems, uncoded and coded systems. Since this brings up the question of coding, this process, and its advantages and disadvantages is analyzed in some detail. As a result of these analyses the fundamental relation between rate of transmission of information and transmission facilities is derived. The results are then applied to some known transmission systems and shown to yield results consistent with practice.

To consider the first point, we have shown that intersymbol interference is unimportant in limiting the rate of transmission of information. Let us therefore assume it absent, since the purpose of this portion of the paper is to show the underlying factor limiting rate of transmission of information. Let S be the r.m.s. amplitude of the maximum signal that 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. Then if N is the r.m.s. amplitude of the noise mixed with the signal, there are $1 \neq S/N$ significant values of signal that may be determined. This sets s in the derivation of Section IV. It is known[9] that the specification of an arbitrary wave of duration T and maximum frequency component f_c requires $2f_cT$ measurements. This sets n in the derivation of Section IV. We have from equation (11) the quantity of information available at the output of the system

$$H = kn \log s = k2f_c\ T \log (1 \neq S/N). \qquad (12)$$

This is an important expression, to be sure, but gives us no information in itself as to the limits that may be placed on H. In particular, f_c is the bandwidth of the overall communications 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 non-linear processes as demodulation that may occur in the receiver. It is C/N that is determined by power, attenuation, and noise limitations, not S/N. Similarly, it is bandwidth in the transmission link that is scarce and expensive. It is therefore necessary to bring both these quantities into the analysis and go beyond equation (12).

The transmission system assumed for the remainder of this analysis is shown in block diagram in Figure VII. As is shown in this figure, the transmission system considered contains an input for the information wave, a transmitter, an attenuating transmission

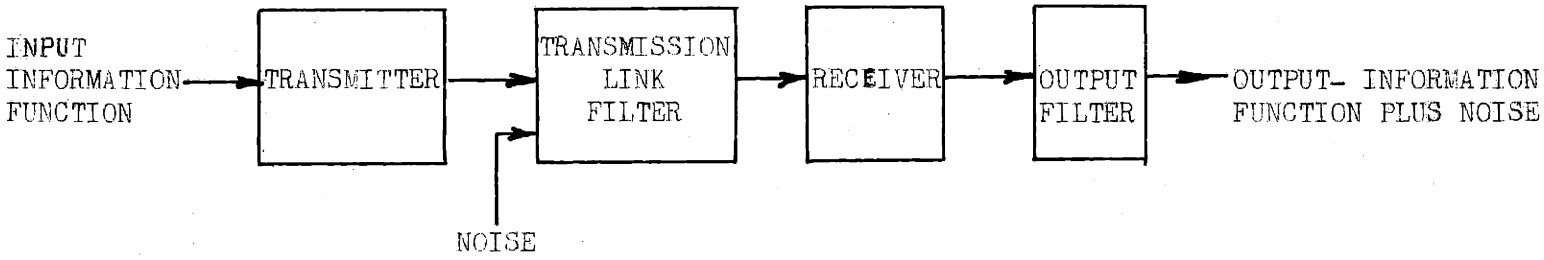


Figure VII

Block Diagram of Simplified Communication System Used in Analysis

medium, a source of noise, a filter in which are assumed lumped all the frequency selective properties of the transmission link, a receiver, and a second filter assumed to have the characteristics described as the overall frequency selective characteristics of the system. The elements of this system may be considered separately.

The transmitter, for example, is simply a device that 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, however, throughout the remainder of this analysis 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 non-linear 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 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 "transmission system" is referred to, all the elements shown in Figure VII 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 given for the overall transmission system previously. 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 bandpass or highpass 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 passband of the receiver.

From the above discussion, it is apparent that the transmission systems sketched in Figure VII 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. Both are equally well delineated by the above statements. 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 must again 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 so on. Coded

transmission corresponds to pulse code modulation or other similar systems, and is a very effective mode of transmission, as can be seen from the following discussion.

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 Equation '17').
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 Equation '27').
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 Equation '29').

These points will be shown 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 r.m.s. noise is just detectable, and applying the reasoning that led to Equation (12) to the receiver input, we have for the "quantity of information" at the receiver input

$$H_{in} = k \cdot 2BT \log (1 + C/N) \qquad (13)$$

We know from (12) that the "quantity of information" at the output of the receiver is

$$H_{out} = k \cdot 2f_c \; T \log (1 \nearrow S/N) \qquad (14)$$

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

$$H_{out} \leqq H_{in} \qquad (15)$$

Substituting (13) and (14) in (15) and clearing like quantities and logarithms from both sides of the inequality gives

$$(1 \nearrow S/N) \leqq (1 \nearrow C/N)^{B/f_c} \qquad (16)$$

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

$$S/N \leqq (C/N)^{B/f_c} \qquad (17)$$

Let us now consider coded transmission. In this case, as will be shown by an example, the equals sign of (16) may be achieved. Suppose for example, we wish to transmit the message of Figure IVe. 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 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 IVb, essentially coding it in binary digits. In this case, if we wish the message to be transmitted in the same time, we must transmit it in 19 times as much bandwidth, since we must transmit 19 time units during the duration of

our message, instead of the original single pulse, or time unit. At the receiver the pulses of Figure IVb may be deciphered to form the single pulse of Figure IVe. One must be careful how he uses this data to avoid error. The various quantities of (16) might erroneously be considered to be, for this example,

$$1 \nearrow S/N = 477,147 \tag{18}$$

$$1 \nearrow C/N = 2 \tag{19}$$

$$B/f_c = 19 \tag{20}$$

We find, however, that

$$2^{19} = 522,288 \tag{21}$$

and therefore in this case

$$(1 \nearrow S/N) < (1 \nearrow C/N)^{B/f_c} \tag{22}$$

This does not correspond to the earlier statement that the equals sign of (16) can be realized. However, one message that 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 (16). In using this formula, $(1 \nearrow S/N)$ must be the number of possible allowed states of the receiver output at any one time, and $(1 \nearrow C/N)$ the number of possible allowed states of the receiver input for any one instant of time. In other words $(1 \nearrow S/N)$ is actually s, measured at the output of the receiver, and $(1 \nearrow C/N)$ is s, measured at the input to the receiver. Considering things in this correct manner we have

$$(1 \nearrow S/N) = 522,288 \tag{23}$$

$$(1 \nearrow C/N) = 2 \tag{24}$$

$$B/f_c = 19 \tag{25}$$

$$2^{19} = 522,288 \tag{26}$$

$$(1 \not\perp C/N)^{B/f_c} = (1 \not\perp S/N) \tag{27}$$

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

$$(C/N) = (S/N)^{B/f_c} \tag{28}$$

Thus we see that in this manner, at least, the equals sign of (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 $(1 \not\perp C/N)$.

Corresponding to our original possibility of one point on the original information function with any of $(1 \not\perp C/N)$ possible significant amplitudes, we now have a possibility of one point* with any one of B/f_c times $(1 \not\perp C/N)$ possible significant values. This comes about because the point may have $(1 \not\perp C/N)$ possible significant amplitudes

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

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 VIII showing the original and transformed signals in amplitude-time and frequency-time space. We can therefore make one of B/f_c times C/N possible choices for this point. We therefore have as the number of possible independent states for the receiver over any period of $\frac{1}{2f_c}$ seconds,

$$(1 \nearrow S/N) \leq B/f_c \cdot (1 \nearrow C/N) \tag{29}$$

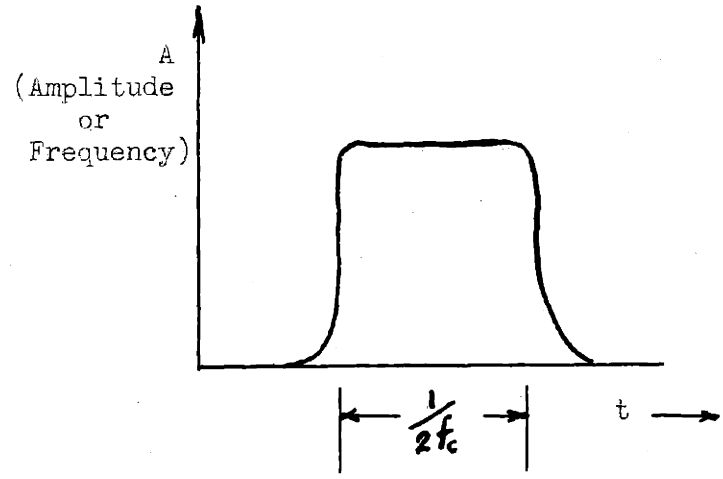
Or again if $S/N \gg 1$ and $C/N \gg 1$

$$S/N \leq B/f_c \cdot C/N \tag{30}$$

Again the equality can be realized, as in the well known frequency modulation system, to be discussed later.

It is interesting to apply the results of (28) and (29) 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. The fact implied by this approach to the transmission of a large amount of information in zero time and finite bandwidth is that a signal whose amplitude is known to infinite precision contains an infinite amount of information, and that in the absence of noise we may measure an amplitude to infinite precision in zero time. This last point has been brought out in the discussion of the earlier parts of this paper. The system described in the first part of the paper should be thought of merely, therefore, as one way of realizing this theoretical possibility. Others are of course possible.

a) ORIGINAL SIGNAL



b) TRANSFORMED SIGNAL

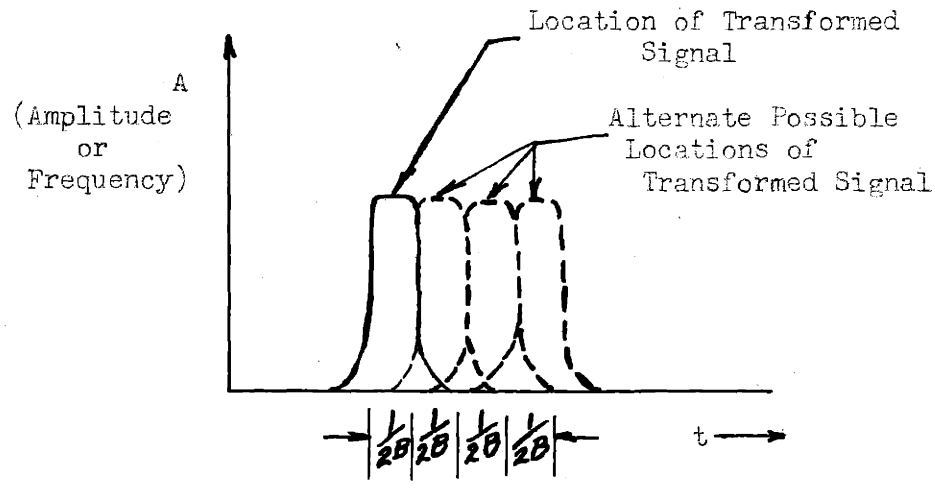


Figure VIII

Uncoded Transformation

VII. Coding

The question of coding information has been mentioned several times in the discussions above. With the general theorems out of the way, one may discuss this interesting new development in the art in a manner that shows its place in the overall picture. For the purposes of this paper, we may make the following general definition:

> Coding is altering the information function in such a manner as to alter the bandwidth, time, and accuracy of definition required for transmission without changing the quantity of information contained in the function.

Binary coding is the only type of coding that 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 five times by this band widening. (Actually signal-to-

noise ratio is improved $5\sqrt{3}$ times (about 18 db) not counting the gain from the use of pre-emphasis and de-emphasis, which tends to make voice and music waveforms have more nearly uniform spectra. The factor of $\sqrt{3}$ comes about because of the approximations made in defining B in the analysis given above. In particular a "rectangular" spectrum was used in the analysis. This is not present in the FM case). 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 one-thousand (60 db) as a reasonable figure, this could be accomplished with a carrier-to-noise ratio of but thirty-two (30 db). The standard frequency modulation system would require a carrier-to-noise ratio of about one hundred and twenty (42 db) to accomplish this, and moreover would use 2.5 times the bandwidth.* If higher signal-to-noise ratios were required, the difference between the two systems is 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 information function. This requires

*If additional advantage were taken of single side band operation, the resultant signal could be accommodated in a standard broadcast channel and give the same signal-to-noise ratio as frequency modulation with less power and one-fifth the required spectrum.

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 one part in ten, the first might be nine units in amplitude and the second three. A system capable of transmission accuracy of one part in one hundred could transmit one point ninety-three 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. We may conclude from the above discussion that coded systems are capable of realizing the maximum possible rate of transmission of information for the available facilities -- bandwidth, time, power, noise, and circuit attenuation. There is no 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 adequate, then this is the most efficient utilization of spectrum. If a carrier must be used, single side band 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 may modify these theoretical considerations. In particular, in the present state of the art coding requires a complex

receiver. This makes coded transmission 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 makes 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 VII therefore we have

$$C/N \geq s_{max} - 1 \tag{31}$$

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

$$H = k\ n \log s_{ave} \tag{32}$$

where s_{ave} is the average value of s, obtained in the conventional manner.

Thus in this case, since

$$s_{max} \geq s_{ave} \tag{33}$$

the formula of

$$H = k2BT \log (1 + C/N) \tag{34}$$

no longer holds, and in fact we have

$$H \leq k2BT \log (1 + C/N) \tag{35}$$

To realize the equals sign in (35) we must achieve the equals sign in (33). This can only be done when s is a constant, since only in this case does $s_{max} = s_{ave}$.

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 only make sure that s is constant by examining every portion of the complete message. Such a delay does not necessarily involve a decrease in the rate of transmission of information, but only a dead period before transmission begins.

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 from some knowledge of the information function it is determined that the amplitude of the function will be within ten 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 then may let the more probable range occupy fifty per cent of the scale, by a pre-arranged scheme, and express this range in fifty significant steps, rather than 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, ninety

per cent of the time the effect is to transmit with s five times as great as was formerly the case; ten 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 that 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 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 in any transmission requiring maximum efficiency in utilization of these three parameters, that one make sure his 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. The only requirement that must be met by these systems is that they contain an information transmission problem. 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 mechanism 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 only be concerned 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 non-existence of a reflecting body at any point within the range of the equipment. A refinement that might be useful, although not always essential, consists in knowing the "electrical size" of the target, i.e., the strength of the reflected signal. As stated in this manner a triple infinity of information is required, one in azimuth angle to a certain tolerance, target range to another, and target

magnitude to a third. Taking these requirements in inverse order, service codes for reporting echo signal strength only recognize 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 that must be transmitted to the operator each complete scan are, therefore, (R/2r $\times$ 360/2B). Each integer has six possible values, ranging from zero to five. A five to one carrier-to-noise ratio is therefore all that is required. This assumes separate search in range and azimuth. An alternative way of searching is to examine each elemental area bounded by the concentric range accuracy circles and the radial angular accuracy lines separately. This involves the transmission of (R/2r x 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 the report but one. Taking typical numbers to see the cost of this degree of data separation, R might be 100 miles, r $\frac{1}{4}$ mile, B one degree. Then the first system of scanning calls for the transmission of 200 / 180, or 380 integers, while the second calls for the transmission of 200 x 180 x 36,000 integers. Therefore, the second system of scanning should require almost one hundred times the bandwidth or transmission time of the first holding 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 advantageously be used for early warning systems, sited so that target confusion is unlikely. The resultant decrease in information handling capacity required of the system could be taken advantage of 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 worthwhile 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, 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 cycles is all that 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. A comparison of the video bandwidths of modern radar equipment reveals nothing remotely resembling this bandwidth, values of one to four megacycles being more common, although for the radar system described above, 100 kc is adequate. The average radar indicator system therefore is extremely inefficient in its use of spectrum. 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 however, be necessary to force the indicator circuitry to respond at this speed. If a delay and storage circuit is provided, it could accept signal information at 200,000 integers per second in short bursts and disgorge this information at the uniform rate of 600 integers per second, for use by the indicator and operator. Although such a delay, storage, and integrator device may seem complex in comparison with the problem of constructing a 100 kilocycle wide video amplifier it must be remembered that a 100 kilocycle bandwidth video amplifier is simple only because it has been made many times in the past and simple, acceptable solutions are known. It is reasonable to expect that the delay and storage circuit can in the future become almost as simple as the present video amplifier.

Considering now a remote indicator or relay for the radar

system outlined above, one could be provided with a bandwidth of but 300 cycles 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. This represents a considerable simplification over the wide band relay systems now in use or projected, systems whose bandwidth is the same or even greater than that of the radar whose information they relay. 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 one percent 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 in coding the information. This can be carried out down to the point where a signal-to-noise ratio of unity is adequate for the accuracy of data transmission required, but in this case 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. Application of the principles outlined in this paper results in optimum utilization of available power, receiver sensitivity, bandwidth, and transmission time.

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, etc. 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% then one of 100 possible integers must be transmitted every time the input member moves through one one-hundredth 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%. 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 most efficient.

The servo-mechanism field, incidentally, furnishes one of the best examples of the comparison of practice with the theory developed here outside of the various pulse communication schemes, since in one type of relay servomechanism the data relating input to output is sampled at equispaced discrete intervals, and the correction needed by the system supplied according to the resultant data.

A feature of the scheme of analysis presented in this paper 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. Therefore, we may say 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. What is perhaps not logically sound is the transmission of the data through the computing machine in the form of a group of pulses separated by intervals during which no pulses exist. The faithful reproduction of such a pulse series requires considerably more bandwidth than is required to transmit the series of values of the function actually represented by the pulses. For example, the transmission of a series of equispaced pulses with 50% duty cycle at a rate of p pulses per second requires a bandwidth of about 4p cycles. Transmission of the required information could actually be accomplished with a bandwidth of p/2 with the proper corrective network, making possible a saving in bandwidth, or, what is usually of more importance to the designers of high speed computers, transmission and computation time, of a factor of eight. A gain of this magnitude, obtainable without any sacrifice in accuracy, should be worth serious consideration.

X. Bibliography

1. Carson, J. R. -- "Notes on the Theory of Modulation", Proc. I.R.E., vol. 10, no. 1, p. 57 (1922).

2. Nyquist, H. -- "Certain Factors Affecting Telegraph Speed", B.S.T.J., vol. 3, p. 324 (1924).

3. Küpfmüller, K. -- "Transient Phenomena in Wave Filters", E.N.T., vol. 1, p. 141 (1924).

4. Hartley, R. V. L. -- "Transmission of Information", B.S.T.J., vol. 7, p. 535 (1928).

5. Gabor, D. -- "Theory of Communication", I.E.E. (London) Journal, vol. 85, Part III, p. 429 (1946).

6. Wiener, N. -- "The Extrapolation, Interpolation and Smoothing of Stationary Time Series", N.D.R.C., Section D_2 Report, Feb., 1942.

7. Levinson, N. -- "The Wiener (RMS) Error Criterion in Filter Design and Prediction", Journal of Mathematics and Physics, vol. XXV, No. 4, p. 261 (1947).

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

9. Bennett, W. R. -- "Time-Division Multiplex Systems", B.S.T.J., vol. 20, p. 199 (1941).

10. Kretzmer, E. R. -- "Analysis of Step Approximation to a Continuous Function", Technical Report No. 12, Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Autobiography of Author

William Gordon Tuller was born in Rutherford, New Jersey, on September 8, 1918. He attended the Rutherford public schools and entered the Massachusetts Institute of Technology in 1935. He was accepted in the co-operative course in Electrical Engineering in 1937, and had two works assignments -- at the Western Electric Company's Kearny works, Engineer of Manufacture Division, mica condenser development section, and at the Bell Telephone Laboratories, open wire carrier telephone development group. He was a member of the Honors Group during his Junior and Senior years at Technology.

In 1939 Mr. Tuller became a Research Assistant in the Ultra-high Frequency Laboratory of the Institute, subsequently advancing to the grades of Research Associate and Staff Member, Radiation Laboratory. He left on leave of absence to the Raytheon Mfg. Co. in 1941, where he did radar receiver design and headed Raytheon's microwave components development laboratory.

He rejoined Technology in 1945, in the Research Laboratory of Electronics, to work on research problems associated with communication systems, continuing contact with Raytheon as a consultant. He is at present employed by Melpar, Inc. as a project engineer.

Mr. Tuller is a member of Sigma Xi and the Institute of Radio Engineers, and is a licensed professional engineer in the Commonwealth of Massachusetts. He has served on the Committee on Modulation Systems of the I.R.E. since 1946.