THE TRANSMISSION OF INFORMATION

ROBERT M. FANO

TECHNICAL REPORT NO. 65

MARCH 17, 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. 65 March 17, 1949

THE TRANSMISSION OF INFORMATION

Robert M. Fano

Abstract

This report presents a theoretical study of the transmission of information in the case of discrete messages and noiseless systems. The study begins with the definition of a unit of information (a selection between two choices equally likely to be selected), and this is then used to determine the amount of information conveyed by the selection of one of an arbitrary number of choices equally likely to be selected. Next, the average amount of information per selection is computed in the case of messages consisting of sequences of independent selections from an arbitrary number of choices with arbitrary probabilities of their being selected. A recoding procedure is also presented for improving the efficiency of transmission by reducing, on the average, the number of selections (digits or pulses) required to transmit a message of given length and given statistical character. The results obtained in the case of sequences of independent selections are extended later to the general case of non-independent selections. Finally, the optimum condition is determined for the transmission of information by means of quantized pulses when the average power is fixed.

THE TRANSMISSION OF INFORMATION

Introduction

It is the opinion of many workers in the field of electrical communications that the communication art is today at a major turning point of its development. The objective of almost all electrical communication systems has been, up to now, to eliminate distance in some form of human activity or relationships between men. Telegraph, telephone and television are typical examples of such communication systems. We may add to these teletype, telecontrol and telemetering. It is interesting to note that the names of all these communication systems involve the prefix tele, meaning "at a distance".

Although, for obvious reasons, forms of communication over distances much greater than the ranges of human senses and reach were first to receive attention, the magnitude of the distance involved is not of primary importance from a logical point of view in the concept of communication. Communication is basically any form of transmission of information, regardless of the distance between the transmitter and the receiver. In a broader sense, the field of communication includes any handling, combining, comparing or employing of information, since such processes involve and are intimately connected with the transmission of such information.

It is clear, then, that most human activities involve communication in a broad sense, and, in particular, those activities which are considered of higher intellectual type because they depend to a high degree on the process of "thinking". Thinking itself, in fact, involves a natural communication system of a complexity far beyond that conceivable for any man-made system.

The above considerations point clearly to a very wide field of useful applications of the communication art which has hardly been touched as yet. It is to be expected that each application should present problems of a higher order of complexity than those encountered in the past. Consequently, it is also to be expected that the solution of these problems should necessitate the use of more powerful analytical tools and, particularly, should require a more fundamental study of the process of transmission of information. As a matter of fact, the first and most significant step in the direction of such a study was made by Norbert Wiener (1) in connection with the development of predictors for antiaircraft fire control. The statistical nature of this problem led him to the realization that all communication problems are fundamentally of a statistical nature, and must be handled accordingly. He argued that the signal to be transmitted in a communication system can never be considered as a known function of time, because if it were *a priori* known it could not convey any new information and therefore would not need to be transmitted. On the other hand, what can be known a priori about a signal to be transmitted is its statistical character – that is, for instance, the probability distribution of its amplitude. In addition, it is equally clear, that noise, which plays such an important part in communication problems, can be described only in statistical terms. It follows that all communication problems are inherently statistical in nature, and that disregarding this fact may lead to unexplainable inconsistencies in addition to precluding a deeper understanding of such problems.

The statistical theory of optimum prediction and filtering developed by Wiener led further to the realization of the need for a basic and general criterion for judging the quality of communication systems. In fact, the mean-square error criterion used by Wiener in this part of his work is dictated by mathematical convenience rather than by physical considerations; consequently it may not be useful in certain practical problems. The search for a more appropriate criterion leads naturally to the question of what is the operation that a communication system must perform. If we take as an example a telegraph system, it might seem at first obvious that such a system must reproduce at the output each and every letter of the input message in the proper order. We may observe, however, that if one letter is received incorrectly, the word containing it is still perfectly understandable in most cases, and so, of course, is the whole message. Moreover, the message would still be comprehensible if, for instance, all the vowels were eliminated (which is what is done in written Hebrew). On the other hand, the incorrect transmission of a digit in a number would make the received message incorrect.

It appears therefore that the transmission of the information conveyed by a written message is what we wish to obtain and that this is not necessarily equivalent to the transmission of all the letters contained in the written message. More precisely, it appears that the different symbols, letters or figures contained in a written message do not contribute equally to the transmission of information – so much so, that some of them may be completely unnecessary. Similar conclusions are reached by considering other types of communication systems. In particular, the recent work on the Vocoder (2) and the clipping of speech waves (3) has provided considerable evidence in the same general direction.

The above considerations are relevant to another problem with which communication engineers are becoming more and more concerned, namely, that of bandwidth reduction. As a matter of fact, the Vocoder was developed primarily for the purpose of reducing the bandwidth required for speech transmission. It is clear that if different parts of a message are not equally important, some saving in bandwidth might be possible by providing transmission facilities which are proportional to the importance of these different parts. The bandwidth problem, in turn, is intimately connected with the noise-reduction problem. In fact, all the different types of modulation developed for the purpose of noise and interference reduction require a bandwidth wider than that required by amplitude modulation. This method of paying for an improved signal-to-noise ratio with an increased bandwidth appears to be the result of some fundamental limitation which, however, the conventional approach to communication problems has failed to clarify.

The above discussion of some of the problems confronting or likely to confront the communication engineer indicates clearly the necessity of providing a measure for the "thing" which is to be transmitted and which has been vaguely called "information". Such a measure will then permit a quantitative and more fundamental study of the process involved in the transmission of information which, in turn, will lead eventually to the design of better and more efficient communication devices. A considerable amount of work in this direction has already been done independently by Norbert Wiener (4) and Claude Shannon (5). The work of Wiener is particularly outstanding because of its philosophical profoundness and its importance in many branches of science other than communication engineering. Mention should be made also of the pioneering work of Hartley (6) and of the more recent work of Tuller (7).

This paper presents the work done by the author in the past year on the transmission of discrete signals through a noiseless channel. Although most of the results obtained have already been published by Wiener and Shannon, it is felt that the method of approach used here is sufficiently different to justify this redundant presentation.

I. Definition of the Unit of Information

In order to define, in an appropriate and useful manner, a unit of information, we must first consider in some detail the nature of those processes in our experience which are generally recognized as conveying information. A very simple example of such processes is a yes-or-no answer to some specific question. A slightly more involved process is the indication of one object in a group of N objects, and, in general, the selection of one choice from a group of N <u>specific</u> choices. The word "specific" is underlined because such a qualification appears to be essential to these information-conveying processes. It means that the receiver is conscious of all possible choices, as is, of course, the transmitter (that is, the individual or the machine which is supplying the information). For instance, saying "yes" or "no" to a person who has not asked a question obviously does not convey any information. Similarly, the reception of a code number which is supposed to represent a particular message does not convey any information unless there is available a code book containing all the messages with the corresponding code numbers.

Considering next more complex processes, such as writing or speaking, we observe that these processes consist of orderly sequences of selections from a number of specific choices, namely, the letters of the alphabet or the corresponding sounds. Furthermore, there are indications that the signals transmitted by the nervous system are of a discrete rather than of a continuous nature, and might also be considered as sequences of selections. If this were the case, all information received through the senses could be analyzed in terms of selections. The above discussion indicates that the operation of selection forms the basis of a number of processes recognized as conveying information, and that it is likely to be of fundamental importance in all such processes. We may expect, therefore, that a unit of information, defined in terms of a selection, will provide a useful basis for a quantitative study of communication systems.

Considering more closely this operation of selection, we observe that different informational value is naturally attached to the selection of the same choice, depending on how likely the receiver considered the selection of that particular choice to be. For example, we would say that little information is given by the selection of a choice which the receiver was almost sure would be selected. It seem appropriate, therefore, in order to avoid difficulty at this early stage, to use in our definition the particular case of equally likely choices — that is, the case in which the receiver has no reason to expect that one choice will be selected rather than any other. In addition, our natural concept of information indicates that the information conveyed by a selection increases with the number of choices from which the selection is made, although the exact functional relation between these two quantities is not immediately clear.

On the basis of the above considerations, it seems reasonable to define as the unit of information the simplest possible selection, namely, the selection between two equally likely choices, called, hereafter, the "elementary selection". For completeness, we must add to this definition the postulate, consistent with our intuition, that N independent selections of this type constitute N units of information. By independent selections we mean, of course, selections which do not affect one another. We shall adopt for this unit the convenient name of "bit" (from "binary digit"), suggested by Shannon. We shall also refer to a selection between two choices (not necessarily equally likely) as a "binary selection", and to a selection from N choices, as an N-order selection. When the choices are, *a priori*, equally likely, we shall refer to the selection as an "equally likely selection".

We can now proceed to develop ways of measuring the information content of discrete messages in terms of the unit just defined. Most of this paper will be devoted to the solution of this problem.

II. Selection from N Equally Likely Choices

Consider now the selection of one among a number, N, of equally likely choices. In order to determine the amount of information corresponding to such a selection, we must reduce this more complex operation to a series of independent elementary selections. The required number of these elementary selections will be, by definition, the measure in bits of the information given by such an N-order selection.

Let us assume for the moment that N is a power of two. In addition (just to make the operation of selection more physical), let us think of the N choices as N objects arranged in a row, as indicated in Figure 1.

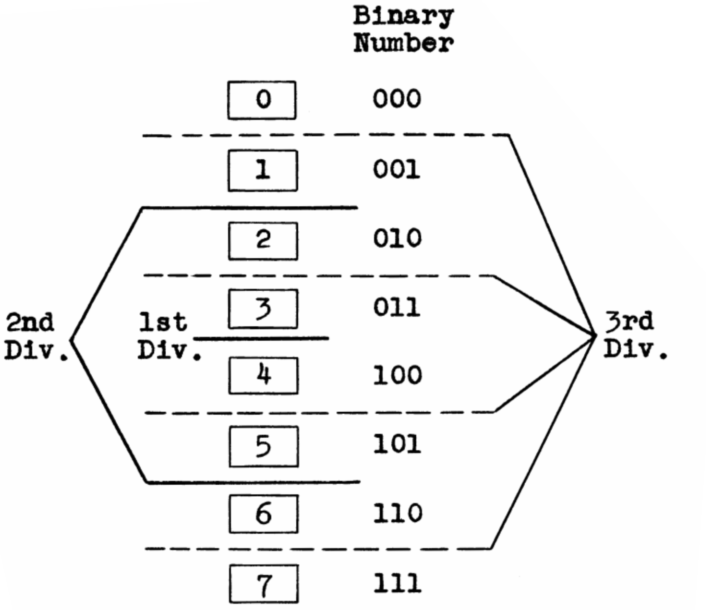

Fig. 1 Selection procedure for equally likely choices.

These N objects are first divided in two equal groups, so that the object to be selected is just as likely to be in one group as in the other. Then the indication of the group containing the desired object is equivalent to one elementary selection, and, therefore, to one bit. The next step consists of dividing each group into two equal subgroups, so that the object to be selected is again just as likely to be in either subgroup. Then one additional elementary selection, that is a total of two elementary selections, will suffice to indicate the desired subgroup (of the possible four subgroups). This process of successive subdivisions and corresponding elementary selections is carried out until the desired object is isolated from the others. Two subdivisions are required for $N = 4$, three for $N = 8$, and, in general, a number of subdivisions equal to $\log_2 N$, in the case of an N-order selection.

The same process can be carried out in a purely mathematical form by assigning order numbers from 0 to N-1 to the N choices. The numbers are then expressed in the binary system, as shown in Figure 1, the number of binary digits (0 or 1) required being equal to $\log_2 N$. These digits represent an equal number of elementary selections and, moreover, correspond in order to the successive divisions mentioned above. In conclusion, an N-order, equally likely selection conveys an amount of information

$$H_N = \log_2 N \quad . \qquad (1)$$

The above result is strictly correct only if N is a power of two, in which case H_N is an integer. If N is not a power of two, then the number of elementary selections required to specify the desired choice will be equal to the logarithm of either the next lower or the next higher power of two, depending on the particular choice selected. Consider, for instance, the case of $N = 3$. The three choices, expressed as binary numbers, are then

$$00 \ ; \ 01 \ ; \ 10 \quad .$$

If the binary digits are read in order from left to right, it is clear that the first two numbers require two binary selections — that is, two digits, while the third number requires only the first digit, 1, in order to be distinguished from the other two. In other words, the number of elementary selections required when N is not a power of two is equal to either one of the two integers closest to $\log_2 N$. It follows that the corresponding amount of information must lie between these two limits, although the significance of a non-integral value of H is not clear at this point. It will be shown in the next section that Eq.(1) is still correct when N is not a power of two, provided H_N is considered as an average value over a large number of selections.

III. Messages and Average Amount of Information

We have determined in the preceding section the amount of information conveyed by a single selection from N equally likely choices. In general, however, we have to deal with not one but long series of such selections, which we call messages. This is the case, for instance, in the transmission of written intelligence. Another example is provided by the communication system known as pulse-code modulation, in which audio waves are sampled at equal time intervals and then each sample is quantized, that is approximated by the closest of a number N of amplitude levels.

Let us consider, then, a message consisting of a sequence of n successive N-order selections. We shall assume, at first, that these selections are independent and equally likely. In this simpler case, all the different sequences which can be formed equal in number to

$$S = N^n , \tag{2}$$

are equally likely to occur. For instance, in the case of $N = 2$ (the two choices being represented by the numbers 0 and 1) and $n = 3$, the possible sequences would be 000, 001, 010, 100, 011, 101, 110, 111. The total number of these sequences is $S = 8$ and the probability of each sequence is 1/8. In general, therefore, the ensemble of the possible sequences may be considered as forming a set of S equally likely choices, with the result that the selection of any particular sequence yields an amount of information

$$H_S = \log_2 S = n \log_2 N. \tag{3}$$

In words, n independent equally likely selections give n times as much information as a single selection of the same type. This result is certainly not surprising, since it is just a generalization of the postulate, stated in Section II, which forms an integral part of the definition of information.

It is often more convenient, in dealing with long messages, to use a quantity representing the average amount of information per N-order selection, rather than the total information corresponding to the whole message. We define this quantity in the most general case as the total information conveyed by a very long message divided by the number of selections in the message, and we shall indicate it with the symbol H_N, where N is the order of each selection. It is clear that when all the selections in the message are equally likely and independent and, in addition, N is a power of two, the quantity H_N is just equal to the information actually given by each selection, that is

$$H_N = \frac{1}{n} \log_2 S = \log_2 N \quad . \tag{4}$$

We shall show now that this equation is correct also when N is not a power of two, in which case H_N has to be actually an average value taken over a sufficiently long sequence of selections.*

The number S of different and equally likely sequences which can be formed with n independent and equally likely selections is still given by Eq.(2), even when N is not a power of two. On the contrary, the number of elementary selections required to specify any one particular sequence must

* The author is indebted to Mr. T. P. Cheatham, Jr. (of this Laboratory) for the original idea on which is based both this proof and the corresponding recoding procedure (see Section IV).

be written now in the form

$$B_S = \log_2 S + d \quad , \tag{5}$$

where d is a number, smaller in magnitude than unity, which makes B_S an integer and which depends on the particular sequence selected. The average amount of information per N-order selection is then, by definition,

$$H_N = \lim_{n\to\infty} \frac{1}{n}(\log_2 S + d) \quad . \tag{6}$$

Since N is a constant and since the magnitude of d is smaller than unity while n approaches infinity, this equation together with Eq.(2) yields

$$H_N = \log_2 N \quad . \tag{7}$$

We shall consider now the more complex case in which the selections, although still independent, are not equally likely. In this case, too, we wish to compute the average amount of information per selection. For this purpose, we consider again the ensemble of all the messages consisting of n independent selections and we look for a way of indicating any one particular message by means of elementary selections. If we were to proceed as before, and divide the ensemble of messages in two equal groups, the selection of the group containing the desired message would no longer be a selection between equally likely choices, since the sequences themselves are not equally likely. The proper procedure is now, of course, to make equal for each group not the number of messages in it but the probability of its containing the desired message. Then the selection of the desired group will be a selection between equally likely choices. This procedure of division and selection is repeated over and over again until the desired message has been separated from the others. The successive selections of groups and subgroups will then form a sequence of independent elementary selections.

One may observe, however, that it will not generally be possible to form groups equally likely to contain the desired message, because shifting any one of the messages from one group to the other will change, by finite amounts, the probabilities corresponding to the two groups. On the other hand, if the length of the messages is increased indefinitely, the accuracy with which the probabilities of the two groups can be made equal becomes better and better since the probability of each individual message approaches zero. Even so, when the resulting subgroups include only a few messages after a large number of divisions, it may become impossible to keep the probabilities of such subgroups as closely equal as desired unless we proceed from the beginning in an appropriate manner as indicated below. The messages are first arranged in order of their probabilities, which can be easily computed if the probabilities of the choices are known. The divisions in groups and subgroups are then made successively without changing the order of the messages, as illustrated in Figure 2. In this manner, the smaller subgroups will contain messages with equal or almost equal probabilities, so that further subdivisions can be performed satisfactorily.

It is clear that when the above procedure is followed, the number of binary selections required to separate any message from the others varies

Probabilities of Groups Obtained by Successive Divisions									
I Div.	II Div.	III Div.	IV Div.	V Div.	VI Div.	Message	$P(1)$	Recoded Message	$P(1)B_g(1)$
0.49						00	0.49	0	0.49
0.51		0.14				01	0.14	100	0.42
	0.28	0.14				10	0.14	101	0.42
	0.23		0.07			02	0.07	1100	0.28
		0.14	0.07			20	0.07	1101	0.28
		0.09	0.04			11	0.04	1110	0.16
			0.05	0.02		12	0.02	11110	0.10
				0.03	0.02	21	0.02	111110	0.12
					0.01	22	0.01	111111	0.06
								$(B_g)_{av.}$ =	2.33

Fig. 2 Recoding of messages consisting of 2 third-order selections, for choice probabilities $p(0) = 0.7$, $p(1) = 0.2$, $p(2) = 0.1$, $H_3 = -[0.7 \log_2 0.7 + 0.2 \log_2 0.2 + 0.1 \log_2 0.1] = 1.157$.

For original code $\eta = \frac{H_3}{\log_2 3} = 0.73$;

For new code $\eta = \frac{2H_3}{(B_g)_{av.}} = 0.993$.

from message to message. Messages with a high probability of being selected require less binary selections than those with lower probabilities. This fact is in agreement with the intuitive notion that the selection of a little-probable message conveys more information than the selection of a more-probable one. Certainly, the occurrence of an event which we know a priori to have a 99 per cent probability is hardly surprising or, in our terminology, yields very little information, while the occurrence of an event which has a probability of only 1 per cent yields considerably more information. More precisely, as shown below, if $P(i)$ is the probability of the i^{th} message, the number of binary selections required to indicate this message will be an integer $B_S(i)$ close to $-\log_2 P(i)$. In fact, $P(i)$ is just the probability of the last subgroup obtained by successively halving (approximately) the probability of the whole ensemble of messages (which is unity) a number of times equal to $B_S(i)$, so that $P(i) \simeq 2^{-B_S(i)}$. By making the messages sufficiently long — that is, the number n of N-order selections sufficiently large — the integer $B_S(i)$ can be made to differ in percentage from $-\log_2 P(i)$ by less than any desired amount. Hence, in this limiting case, we can write

$$B_S(i) = -\log_2 P(i) \quad . \tag{8}$$

Let us consider now a sequence of M selections of messages, each message consisting of n N-order selections (forming a sequence of nM selections). By making the number M sufficiently large, we can be practically sure that the i^{th} message will appear in the sequence with a frequency as close to $P(i)$ as desired. Therefore the number of binary selections required on the average to select one message, that is, "the mathematical expectation of B_S", will be

$$E(B_S) = \sum_{i=0}^{S-1} P(i)\, B_S(i) \quad . \tag{9}$$

The average amount of information per N-order selection is then, from Eqs. (8) and (9),

$$H_N = \lim_{n\to\infty} \frac{E(B_S)}{n} = \lim_{n\to\infty} - \left(\frac{1}{n}\right) \sum_{i=0}^{S-1} P(i) \log_2 P(i) \quad , \tag{10}$$

that is, the limit of the ratio of the number of binary selection required, on the average, to select one message to the number of N-order selections in the message.

Now let $p(k)$ be the probability of the k^{th} choice (of the N), and n_k

be the number of times the k^{th} choice is selected in the i^{th} message (sequence of n selections). The probability of the i^{th} message is

$$P(i) = \prod_{k=0}^{N-1} [p(k)]^{n_k(i)} \quad . \tag{11}$$

The number of binary selections required to indicate this message can be written as

$$B_s(i) = -\log_2 \left[\prod_{k=0}^{N-1} [p(k)]^{n_k(i)} \right] = - \sum_{k=0}^{N-1} n_k(i) \log_2 p(k) \tag{12}$$

with any degree of accuracy desired. In the limit when n approaches infinity these binary selections become elementary selections, that is, binary selections between equally likely choices. We must now compute $E(B_s)$ according to Eq.(9). The number of sequences of selections, that is, messages, to which correspond the same values of $P(i)$ and $B_s(i)$, is equal to the number of different permutations of the choices selected in the i^{th} sequence; that is, to

$$\frac{n!}{\prod_{k=0}^{N-1} n_k(i)!}$$

It follows that the average value of $B_s(i)$ is given by

$$E(B_s) = -\sum \left\{ \left[\frac{n!}{\prod_{k=0}^{N-1} n_k!} \right] \left[\prod_{k=0}^{N-1} [p(k)]^{n_k} \right] \times \left[\sum_{k=0}^{N-1} n_k \log_2 p(k) \right] \right\} \quad , \tag{13}$$

where the n_k and $p(k)$ are always positive and subject to the conditions

$$\sum_{k=0}^{N-1} n_k = n \quad , \tag{14}$$

$$\sum_{k=0}^{N-1} p(k) = 1 \quad . \tag{15}$$

The overall summation in Eq.(13) is made over all possible combinations of integral positive values of the n_k which satisfy Eq.(14).

In order to compute the values of $E(B_s)$ we begin by expressing the factorials in Eq.(13) by means of Stirling's formula (8)(9).

$$n! = \sqrt{2\pi n}\; n^n\, e^{-n} \quad , \tag{16}$$

valid for large values of n. We obtain then

$$\frac{n!}{\prod\limits_{k=0}^{N-1} n_k!} \prod_{k=0}^{N-1} [p(k)]^{n_k}$$

$$= \frac{\sqrt{2\pi n}\; n^n\, e^{-n}}{\prod\limits_{k=0}^{N-1} \sqrt{2\pi n_k}} \prod_{k=0}^{N-1} \left\{ \left[\frac{p(k)}{n_k}\right]^{n_k} e^{n_k} \right\} = n^{-(N-1)}\, f(x) \quad , \tag{17}$$

where

$$f(x) = \left(\frac{n}{2\pi}\right)^{(N-1)/2} \left\{ \prod_{k=0}^{N-1} \left[\frac{p(k)}{x_k}\right]^{x_k} \right\}^n \left\{ \prod_{k=0}^{N-1} (x_k)^{-1/2} \right\} \quad . \tag{18}$$

The variables $x_k = n_k/n$ are always positive, smaller than unity and subject to the constraint

$$\sum_{k=0}^{N-1} x_k = 1 \quad . \tag{19}$$

It is convenient, at this point, to consider the function $f(x)$ as a continuous, rather than a discontinuous, function of the x_k and to transform the summation of Eq.(13) into an integral. We observe, in this regard, that when n_k varies from zero to n,x_k varies from zero to one. It follows that to a unit increment of n_k (n_k takes only integral values) corresponds an increment of x_k equal to $1/n$. Therefore, when n approaches infinity, to the unit increments of the n_k correspond the differentials $dx_k = 1/n$. In conclusion, the summation of Eq.(13) can be transformed (10) into an integral and Eq.(10) then becomes

$$H_N = -\lim_{n\to\infty} \int dx_1 \int dx_2 \ldots \int dx_{N-1} \left[f(x) \sum_{k=0}^{N-1} x_k \log_2 p(k) \right] \tag{20}$$

The integration is extended over the region of the hyperplane defined by

Eq.(19), in which all the x_k are positive and smaller than one. It will be noted that in Eq.(20) x_o is considered as a function of all the other x_k.

$$x_o = 1 - \sum_{k=1}^{N-1} x_k \quad , \qquad (21)$$

so as to limit the integration to the above-mentioned hyperplane.

To compute the integral appearing in Eq.(20), we observe first that the integral of $f(x)$ alone over the same region represents the summation of the probabilities of all possible messages consisting of n selections, provided, of course, that n is sufficiently large. Therefore, the integral of $f(x)$ must be equal to unity for all large values of n. On the other hand, as shown in Appendix I, $f(x)$ has a peak at a point which approaches $x_k = p(k)$ when n approaches infinity. The height of this peak is proportional to $(N-1)/n^2$. It follows that when n approaches infinity, $f(x)$ becomes a delta-function, or unit impulse, located at $x_k = p(k)$. The integral of Eq.(20) is, therefore, equal to the value for $x_k = p(k)$ of the rest of the integrand, that is, of the summation. Eq.(20) yields finally

$$H_N = - \sum_{k=0}^{N-1} p(k) \log_2 p(k) \quad , \qquad (22)$$

which is then the average amount of information per N-order selection.

The conclusions which can be reached from the evaluation of the integral in Eq.(20) extend far beyond Eq.(22). It is easy to see that if the function

$$\sum_{k=0}^{N-1} x_k \log_2 p(k)$$

were any other finite function of the x_k, the limiting value of the integral would still be equal to the value of the function for $x_k = p(k)$. In other words, the expectation (or average value) of any function of the x_k is equal to the value of the function itself for $x_k = p(k)$. From a physical point of view, we can say that the ensemble of possible sequences of selections can be divided in two groups. The first group consists of sequences for which the frequencies x_k of occurrence of the different choices differ from the probabilities $p(k)$ of the choices by less than amounts which approach zero as $1/\sqrt{n}$ when n approaches infinity. The total probability of the sequences in this group approaches unity when n increases indefinitely, and therefore the number of sequences in this group approaches

$$M = \prod_{k=0}^{N-1} [p(k)]^{-np(k)} = 2^{nH_N} \quad . \qquad (23)$$

The second group consists of all other sequences, and its total probability approaches zero when n approaches infinity.

The sequences of the first group are all equally probable and, therefore, the selection of one of them out of the group requires a number of binary selections equal to

$$\log_2 M = nH_N \quad . \qquad (24)$$

In other words, the sequences of the first group can be represented by means of sequences of n H_N binary digits, that is H_N digits per N-order selection. All the other sequences together, regardless of the way in which they are represented, cannot increase by any finite amount, beyond H_N, the number of binary digits required on the average per N-order selection.

The expression for H_N obtained above indicates that H_N can be considered as the expectation of $\log_2 [1/p(k)]$. In other words, we may say that the selection of a particular choice k conveys an amount of information equal to the logarithm-base-two of the reciprocal of its probability. This interpretation is fundamental. It will be shown later to apply also to the general case of non-independent selections, in which case p(k) will be substituted by the conditional probability that the k^{th} choice will be selected, based on the knowledge of all preceding selections.

It is easy to see from Eq.(22) that H_N vanishes only when all but one of the p(k) are equal to zero, in which case the one different from zero must be equal to unity. In other words, H_N vanishes only when the choice which will be selected is known a priori with unity probability. In this instance, it is intuitively clear that no information is being transmitted. On the other hand, H_N is a maximum (as shown in Appendix I), when all the p(k) are equal, that is, when there is no a priori knowledge at all about the selections. Under these circumstances, Eq.(22) reduces to Eq.(7), since $p(k) = 1/n$. The manner in which H_N varies with the probabilities of the choices is illustrated in Figure 3, for the particular case of N = 2.

The amount of information conveyed by a message of given length was defined above as the number of independent elementary (binary, equally likely) selections required, on the average, to specify such a message. The notion of a minimum number of binary selections required did not enter the definition. It should be intuitively clear, however, that the minimum number of binary selections required, on the average, to specify a message is equal to the average information conveyed, or, in other words, the number of

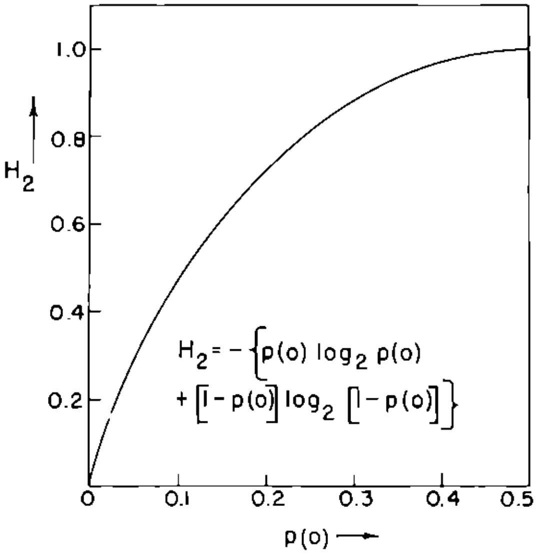

Fig. 3 The amount of information per binary selection as a function of the probability of either choice.

binary selections becomes a minimum when the selections are equally likely and independent. To prove this identity, we observe that the amount of information conveyed by a sequence of independent binary selections is a maximum when the selections are equally likely. Conversely, therefore, it is always possible to represent any sequence of m binary, not equally likely selections with a number of elementary selections smaller, on the average, than m. It follows that no binary representation of a message can be obtained with a number of selections smaller than the amount of information conveyed. It is clear, of course, that all message representations, which employ independent equally likely selections,require,on the average,the same number cf selections. It will be shown later that a larger number of selections is required whenever non-independent selections are used.

It is appropriate to point out here that the mathematical form of Eq.(22) suggests a very interesting analogy between information and entropy, as expressed in statistical mechanics. In fact, H_N appears formally as the entropy of a system whose possible states have probabilities p(k). For a physical interpretation of this analogy, the reader is referred to the work of Norbert Wiener (Ref. 1).

IV. Codes and Code Efficiency

The preceding sections have been devoted to the definition of the unit of information and to the computation of the average amount of information per selection in the case of messages consisting of sequences of independent N-order selections. It was pointed out in Section III that H_N represents the minimum number of binary selections required, on the average, to perform an N-order selection with given choice probabilities. Therefore, if we take the number of binary selections employed as a basis for comparing different methods of conveying the same information, H_N represents a theoretical limit corresponding to maximum efficiency.

The knowledge of such a theoretical limit is extremely important, but perhaps even more important is the ability to approach this limit in practice. In our case, fortunately, the procedure followed in computing H_N (that is, the theoretical limit) indicates a convenient method for approaching this limit in practice. Let us consider again all the sequences of n N-order selections (in which, however, n may be a small integer), and arrange them in order of increasing probability. If we wish to separate any one particular sequence from the others by means of successive division in almost equally probable groups, as discussed in the preceding section, the number of divisions required, on the average, that is, $E(B_S)$, will be larger than nH_N. However, if we increase n, that is, the length of the sequences, we find that $E(B_S)/n$ keeps decreasing and approaches H_N when n approaches infinity. It must be kept in mind, in this regard, that $E(B_S)/n$ does not decrease necessarily in a monotonic manner, but may have an oscillatory behavior as a function of n.* It follows that an increase of n may actually produce an increase of $E(B_S)/n$. For instance (as shown in Figure 4), in the case of $N = 2$, $p(0) = 0.7$, $p(1) = 0.3$, the value of $E(B_S)/n$ is 0.905 for $n = 2$, 0.909 for $n = 3$, and 0.895 for $n = 4$, the limiting value being $H_2 = 0.882$.

The above discussion indicates that, in transmitting a message consisting of a large number of selections, we should transmit the selections not individually, but in sequences of n as units, the number n being as large as permitted by practical considerations. The transmission of each of these units is then performed by means of sequences of binary selections corresponding in order to the successive divisions of the ensemble of all possible sequences of n N-order selections, as indicated in Figures 2, 4, and 5. It will be noted that, although the sequences of binary selections are not equal in length, it is always possible to identify the end of any of them in a long message. In fact, the first m selections of any sequence of length larger than m are always different from any of the sequences consisting of exactly m selections.

If it is desired to perform the transmission by means of N'-order selections (N' being any integer), we can proceed in the same manner as in the case of binary selections, the only difference being that we must divide successively the ensemble of all possible sequences in N' groups instead of just two. After each division, the groups containing the desired sequence

* This fact was first pointed out to me by L. G. Kraft of this Laboratory.

Original Message	P(i)	Recoded Message	$P(i)B_s(i)$
00	0.49	0	0.49
01	0.21	10	0.42
10	0.21	110	0.63
11	0.09	111	0.27

$E(B_s) = 1.81$
$E(B_s)/2 = 0.905$
$\eta = 0.975$

Original Message	P(i)	Recoded Message	$P(i)B_s(i)$
000	0.343	00	0.686
001	0.147	01	0.294
010	0.147	100	0.441
100	0.147	101	0.441
011	0.063	1100	0.252
101	0.063	1101	0.252
110	0.063	1110	0.252
111	0.027	1111	0.108

$E(B_s) = 2.726$
$E(B_s)/3 = 0.909$
$\eta = 0.972$

Original Message	P(i)	Recoded Message	$P(i)B_s(i)$
0000	0.2400	00	0.480
0001	0.1030	010	0.309
0010	0.1030	011	0.309
0100	0.1030	100	0.309
1000	0.1030	1010	0.412
0011	0.0441	1011	0.1764
0110	0.0441	11000	0.2205
1100	0.0441	11001	0.2205
0101	0.0441	11010	0.2205
1001	0.0441	11011	0.2205
1010	0.0441	11100	0.2205
0111	0.0189	11101	0.0945
1011	0.0189	111100	0.1134
1101	0.0189	111101	0.1134
1110	0.0189	111110	0.1134
1111	0.0081	111111	0.0486

$E(B_s) = 3.5812$
$E(B_s)/4 = 0.895$
$\eta = 0.985$

Fig. 4 Recoding of binary messages for $n = 2, 3, 4$, $p(0) = 0.7$, $p(1) = 0.3$, $H_2 = 0.882$.

Original Message	P(i)	Recoded Message	$P(i)B_s(i)$
00	0.81	0	0.81
01	0.09	10	0.18
10	0.09	110	0.27
11	0.01	111	0.03

$E(B_s) = 1.29$
$\eta = 0.725$

Original Message	P(i)	Recoded Message	$P(i)B_s(i)$
000	0.729	0	0.729
001	0.081	100	0.243
010	0.081	101	0.243
100	0.081	111	0.243
011	0.009	11100	0.045
101	0.009	11101	0.045
110	0.009	11110	0.045
111	0.001	11111	0.001

$E(B_s) = 1.594$
$\eta = 0.882$

Original Message	P(i)	Recoded Message	$P(i)B_s(i)$
0000	0.0550	0	0.6550
0001	0.0729	100	0.2187
0010	0.0729	101	0.2187
0100	0.0729	110	0.2187
1000	0.0729	1110	0.2916
0011	0.0081	111100	0.0486
0110	0.0081	1111010	0.0567
1100	0.0081	1111011	0.0567
0101	0.0081	1111100	0.0567
1010	0.0081	1111101	0.0567
1001	0.0081	1111110	0.0567
0111	0.0009	111111100	0.0081
1011	0.0009	111111101	0.0081
1101	0.0009	111111110	0.0081
1110	0.0009	1111111110	0.0090
1111	0.0001	1111111111	0.0010

$E(B_s) = 1.9691$
$\eta = 0.95$

Fig. 5 Recoding of binary messages for $n = 2, 3, 4$; $p(0) = 0.9$, $p(1) = 0.1$, $H_2 = 0.468$.

will then be indicated by means of an N'-order selection.

The operation described above is, effectively, a change of code, that is, we may say, of the conventional language in which the message is written. Therefore this operation will be referred to as "message recoding". The advantage resulting from this recoding is conveniently expressed in terms of the code efficiency

$$\eta = \frac{H_N}{\log_2 N} \quad , \qquad (25)$$

that is, the ratio of the information transmitted on the average per selection, to the information which could be transmitted with an equally likely selection of the same order. The efficiency of a binary code resulting from the recoding of sequences of N-order selections can be computed most conveniently in the form

$$\eta = \frac{nH_N}{E(B_S)} \quad , \qquad (26)$$

where n is the number of N-order selections used in the recoding operation. Note that nH_N is the average amount of information per sequence of n N-order selections and $E(B_S)$ represents the amount of information which could be transmitted, on the average, by one of the sequences of binary selections in which the original sequences are recoded, if these binary selections were equally likely. If the new code is of N' order, we must substitute for $E(B_S)$ the product of $\log_2 N'$ by the number of N'-order selections required, on the average, to specify a sequence of n N-order selections.

A final remark must be made regarding the recoding operation. Since the process of successive divisions of an ensemble of sequences into equally probable groups cannot be carried out exactly, it is not clear at times whether one sequence should be included in one group or in another. Of course, we wish to perform all divisions in such a way as to obtain at the end the most efficient code. Unfortunately, no general rule could be found for determining at once how the divisions should be made in doubtful cases in order to obtain maximum code efficiency. However, so long as the divisions are made in a reasonable manner the resulting code efficiency will not differ appreciably from its maximum value.

We have implicitly assumed in the foregoing discussion that we know a priori the probabilities p(k) of the choices for a message still to be transmitted. It seems appropriate at this point to discuss in some detail this assumption, since the practical value of the results obtained above depends entirely on its validity. When we state that the probability of a particular choice has a value p(k) we mean that the frequency of occurrence of that choice in a message originating from a given source is expected to be close to p(k). The longer is the message, the closer we expect the frequency to approach p(k). It must be clear, however, that we have no assurance that the frequency of occurrence will not differ considerably from the probability even in the case of a very long message, although such a situation is very unlikely to arise.

In practice, p(k) must be estimated experimentally following the reverse process, that is, by inference from the measurement of the frequency in a number of sample messages. If the frequencies in the sample messages are reasonably alike, or, more precisely, if their values are scattered in the manner which might be expected on the basis of the length of the messages used, we may feel relatively safe in taking their average value as a good estimate of the probability. In other words, we may expect that the frequency in any other message originating from the same source will be reasonably close to the average value obtained. If this is the case, the source of such messages is said to have a stationary statistical character. We can conceive the case, however, in which the frequencies in the sample messages available are so widely scattered that hardly any significance can be attributed to their average value. Such a result may mean that the source has not a stationary statistical character, at least for practical purposes, in which case the concept of probability loses any physical significance. Fortunately, however, the sources of interest appear to have a stationary character for any practical purpose. In addition, the estimates of the probabilities of the choices do not need to be too close. It should be clear, in this respect, that the fact that a code has been designed for a particular set of choice probabilities does not mean that only messages with the same statistical character can be transmitted. It means only that such a code will transmit most efficiently, that is, with the smallest number of selections — messages with the choice frequencies equal to the assumed probabilities. Moreover, we can expect that the efficiency of transmission will not depend in a critical manner on the actual frequencies of the messages to be transmitted. A proof that this is actually the case is given below.

Suppose that a code which is optimum for a set of choice probabilities p'(k) is used to transmit messages with choice probabilities p(k). If we consider again all possible sequences of n selections, the expression for the number of binary selections required, on the average, to indicate one particular sequence, $E(B_S')$, is still given by Eq.(13), where, however, the p(k) which appear in the form $\log_2 p(k)$ should be changed into p'(k). It follows that, in the limit when n approaches infinity, the number of binary selections per N-order selection will approach, according to Eq.(22), the value

$$H_N' = -\sum_{k=0}^{N-1} p(k)\ \log_2 p'(k) \quad . \qquad (27)$$

It is clear from this equation that H'_N varies rather slowly with any one of the p'(k), unless the corresponding p(k) is close to zero or unity. H'_N is, of course, a minimum when p'(k) = p(k). The case of N = 2 is illustrated in Figure 6 for p(0) = 0.5 and p(0) = 0.7. We may conclude, therefore, that the statistical characteristics assumed a priori can be rather different from those of the messages actually transmitted, without the efficiency being lowered too much.

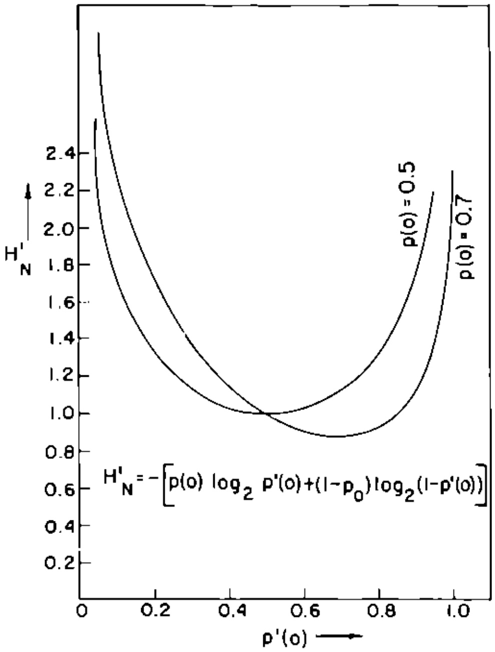

Fig. 6 Behavior of H'_N as a function of p'(0) for binary messages.

V. The Case of Non-Independent Selections

Thus far we have been considering only messages of a particularly simple type, namely, messages consisting of sequences of independent selections. Obviously, the statistical character of most practical messages is much more complex. Any particular selection depends generally on a number of preceding selections. For instance, in a written message the probability that a certain letter will be an "h" is highest when the preceding letter is a "t". In a television signal the light intensity of a certain element of a scanning line depends very strongly on the light intensities of the corresponding elements in the preceding lines and in the preceding frames. In fact, the light intensity is very likely to be almost uniform over wide regions of the picture and to remain unchanged for several successive frames.

The simplifying assumption that any one selection is independent of the preceding selections, although quite unrealistic, does not invalidate completely the results obtained in the preceding sections, but merely reduces their significance to that of first approximations. Intuitively, the average amount of information conveyed by a sequence of given length is decreased by the a priori knowledge of any correlation existing between successive selections. Therefore, the value given by Eq.(22) will always be larger than the correct value for the average amount of information per selection, and the same is true of the code efficiency given by Eq.(25). Similarly, any recoding operation performed in the manner discussed in Section IV will result in a higher efficiency of transmission, but not so high as could be obtained by taking into account the correlation between successive selections.

The procedure for computing the average amount of information per selection and for recoding messages is still essentially the same as that used in Sections III and IV, even when the dependence of any selection on the preceding selections is taken into account. The only difference is that the probability of a particular sequence will not be equal simply to the product of the probabilities of the choices in it, since these are no longer independent. We must still arrange all the possible sequences of given length n in order of probability, and separate the desired sequence by successive divisions of the ensemble of sequences in groups as equally probable as possible. The number of divisions required, on the average, divided by the number n of selections will approach H_N when n approaches infinity.

Let $P_n(i)$ be the probability of the i^th sequence of n selections, and $H_S(n)$ the average amount of information per sequence of n selections when successive sequences are assumed to be independent. We have then

$$H_S(n) = -\sum_{i=0}^{N^n-1} P_n(i) \log_2 P_n(i) \quad . \qquad (28)$$

Let us consider next a sequence of n+1 selections and let $P_{n+1}(i;k)$ be the conditional probability that the i^th sequence (of the $S = N^n$ sequences of n selections) is followed by the k^th choice (of the N). We have then

$$H_S(n+1) = -\sum_{k=0}^{N-1} \sum_{i=0}^{N^n-1} P_n(i)\, P_{n+1}(i;k) \log_2 P_n(i)\, P_{n+1}(i;k) \quad , \qquad (29)$$

which, since

$$\sum_{k=0}^{N-1} P_{n+1}(i;k) = 1 \quad , \qquad (30)$$

becomes

$$H_S(n+1) = H_S(n) - \sum_{k=0}^{N-1} \sum_{i=0}^{N^n-1} P_n(i)\, P_{n+1}(i;k) \log_2 P_{n+1}(i;k) \quad . \qquad (31)$$

The increment of information resulting from the $(n+1)^{th}$ selection is then, on the average,

$$H_N(n+1) = -\sum_{k=0}^{N-1} \sum_{i=0}^{N^n-1} P_n(i)\, P_{n+1}(i;k)\, \log_2 P_{n+1}(i;k) \quad . \qquad (32)$$

Expressing now $H_S(n)$ in terms of the successive increments, we obtain

$$H_S(n) = \sum_{m=1}^{n} H_N(m) \quad . \qquad (33)$$

The final correct value of the average amount of information per selection can then be written in the form

$$H_N = \lim_{n\to\infty} (1/n) \sum_{m=1}^{n} H_N(m) \quad . \qquad (34)$$

To proceed further in our analysis, we must distinguish between two types of statistical character of practical importance. We shall say that the output of a certain source is statistically uniform if each and any selection depends in the same manner on the m^{th} preceding selection, as seems to be the case in a written message. We shall say that the output is periodically discontinuous if it is possible to divide any output sequence in sub-sequences of fixed and equal length, so that each and any selection depends in the same manner on the m^{th} preceding selection of the same sub-sequence but is independent of all selections of the preceding sub-sequences. This is the case when messages transmitted in succession are similar in character and equal in length but entirely unrelated to one another, as, for example, in facsimile transmission. The above differentiation of statistical character is not an exhaustive classification but only a characterization of two special cases of practical interest in which different results are obtained.

Considering now in more detail the increments of information $H_N(n+1)$, our intuition indicates that the average amount of information conveyed by any additional selection can be, at most, equal to the value obtained when the selection is independent of all preceding selections. Mathematically, it must be

$$H_N(n+1) \leqslant H_N(1) = -\sum_{k=0}^{N-1} p(k)\, \log_2 p(k) \quad . \qquad (35)$$

A proof of this inequality is given in Appendix II. In addition, it is intuitively clear also that, in the case of uniform statistical character, the average amount of information conveyed by the $(n+1)^{th}$ selection of a sequence can be, at most, equal to the amount of information conveyed by the n^{th} selection, since the latter has less preceding selections on which to depend. Mathematically, we expect that, for statistically uniform sequences,

$$H_N(n+1) \leqslant H_N(n) \quad . \tag{36}$$

A proof of this inequality is given also in Appendix II. Eq.(36) is satisfied with the equal sign when the $(n+1)^{th}$ selection, and therefore any following selection, depends only on the n–1 preceding selections.

Eq. (36) shows that the limit in Eq.(34) is approached in a monotonic manner. In addition, we expect $H_N(m)$ to approach monotonically a limit with increasing m, since the dependence of any selection on the preceding selections cannot extend, in practice, over an indefinitely large number of selections. Suppose, for instance, that this dependence extends only over the n_o–1 preceding selection. Then $H_N(m)$ becomes constant and equal to $H_N(n_o)$ when m is larger than n_o, and Eq.(34) yields

$$H_N = H_N(n_o) \quad . \tag{37}$$

This result is correct, of course, only in the case of statistically uniform sequences.

In the case of a periodically discontinuous statistical character, Eq.(36) is valid only when the n^{th} and the $(n+1)^{th}$ selections belong to the same sub-sequence. If this is not the case, the $(n+1)^{th}$ selection must be the first selection of a sub-sequence, and therefore is independent of all preceding selections. It follows that $H_N(m)$ is a periodic function of m with period equal to the length n_o' of the sub-sequences, and that the limit of Eq.(34) is approached in an oscillatory manner. If we compute this limit by increasing n in steps equal to n_o', it is easily seen that Eq.(34) yields

$$H_N = (1/n_o') \sum_{m=1}^{n_o'} H_N(m) \quad , \tag{38}$$

a value larger than that given by Eq.(37), as was expected.

The recoding procedure in the case of messages consisting of non-independent selections is still the same as in the case of independent selections. The efficiency of transmission, still given by Eq.(25), increases (although not necessarily monotonically), with the number of selections used as units in the recoding process, and approaches unity when the number increases indefinitely. It is worth emphasizing that in the recoding process any sequence, even if statistically uniform, is considered as periodically discontinuous. In fact, the groups of selections recoded as units are effectively sub-sequences which are treated as though they were totally unrelated. It follows that, if the recoding operation of a statistically uniform sequence is performed on groups of n_o selections, the efficiency of transmission after recoding can be at most equal to

$$\eta(n_o) = \frac{n_o H_N(n_o)}{\sum_{m=1}^{n_o} H_N(m)} \quad . \tag{39}$$

In the case of statistically discontinuous sequences, it would seem reasonable to make the number of selections in the recoding groups an integral fraction or multiple of the length of the sub-sequences.

A final remark is in order regarding the fitting of the recoding procedure to the statistical character of the messages to be transmitted. It may happen, as it does in the case of television signals, that the dependence of any one selection on the m^{th} preceding selection does not decrease monotonically when m increases, but behaves in an oscillatory manner. In this case, one should first reorder the selections before recoding, in such a manner that selections which are closely related take positions close to one another in the sequence. This idea of reordering the selections in the sequence can be generalized as follows. Any type of transmission of information can be considered as the transmission, in succession, of patterns in a two-dimensional or multi-dimensional space, time being one of the dimensions. Then the problem of ordering selections in an appropriate manner can be generalized to the problem of how best to scan these patterns. It is clear, on the other hand, that such a scanning problem is also at the root of the problem of reducing the bandwidth required by television signals. The generalized scanning problem seems to be, therefore, of fundamental practical, as well as theoretical, importance. However, no work can yet be reported on this subject.

VI. Practical Considerations

The main purpose of this paper was to provide a logical basis for the measurement of the rate of transmission of information. It has been shown that an appropriate measure for the rate of transmission in the case of sequences of selections can be provided by the minimum number of binary selections required, on the average, to indicate one of the original selections. We were then led naturally to consider the problem of actually performing the transmission of the original sequences by means of as few binary or higher-order selections as possible. We did not consider, however, the physical process corresponding to such selections — that is, their transmission by electrical means.

A convenient way of transmitting binary selections in a practical communication system is by means of pulses with two possible levels, one and zero. This is just the technique employed in pulse-code modulation. The maximum rate at which information can be transmitted in this case is simply equal to the number of pulses per second which can be handled by the electrical system — which we know to be proportional to the frequency band available. However, as soon as we start dealing with electrical pulses rather than logical operations like selections, an additional item must be considered in the problem, namely, the power required for the transmission. In the case of two-level pulses, the average power corresponding to the maximum rate of transmission of information is equal to one-half the pulse power, since the zero and one levels are equally probable.

If pulses with N rather than two levels equally spaced in voltage are used, the maximum rate of transmission is equal to $\log_2 N$ times the number of pulses per second which can be handled by the system. The average power required becomes, in this case,

$$W = \left(\frac{W_o}{N}\right)\sum_{k=0}^{N-1} k^2 \quad , \tag{40}$$

where W_o is the power corresponding to the lowest (non-zero) voltage level.

The theoretical limit stated above for the rate of transmission of information certainly has practical significance when the limiting factors in the physical problem are the frequency band available and the number of pulse levels permitted by technical and economical considerations. It is to be noted, in this regard, that the effect of noise is here taken into account, to a first approximation, by setting a lower limit to the voltage difference between pulse levels, and therefore to W_o. For a detailed discussion of the effect of noise, the reader is referred to the work of Shannon (5).

Eq.(40) shows, on the other hand, that the average power increases approximately as N^2, while the rate of transmission is proportional only to $\log_2 N$. It follows that, if no limitation is placed on the frequency band employed, the smallest value of N should be used — that is, two. This value has, in addition, the very important practical advantage that the receiver is not required to measure a pulse, but only to detect the existence or the lack of a pulse. It might happen, on the other hand, that the frequency band and the average power are the limiting factors, while any reasonable number of pulse levels can be allowed. This case represents quite a different problem from those considered above, and the maximum rate of transmission of information is no longer obtained by making the pulse levels (that is, the choices) equally probable, as one might think at first. For example, more than one unit of information per pulse can be transmitted with an average power $W = W_o/2$, by using pulses with three levels not equally probable. It seems worth while, therefore, to determine the maximum amount of information which can be transmitted per pulse, for a given average power W, a minimum level power W_o, and an unlimited number of pulse levels equally spaced in voltage.

Let, therefore, p(0) be the probability of occurrence of the zero level (no pulse), and p(k) the probability of the k^{th} level. The amount of information per pulse is given by

$$H = -\sum_{k=0}^{\infty} p(k) \log_2 p(k) \quad , \tag{41}$$

and the average power by

$$W = W_o \sum_{k=0}^{\infty} p(k)\, k^2 \quad . \tag{42}$$

We wish to maximize H with respect to the p(k), subject to the condition expressed by Eq.(42) and, of course, the usual condition

$$\sum_{k=0}^{\infty} p(k) = 1 \quad . \tag{43}$$

The maximization procedure is carried out in Appendix III, and yields

$$H_{max.} = -\frac{W}{W_o}\left[\log_2\left(\frac{p(1)}{p(0)}\right) + \log_2 p(0)\right] \quad ; \tag{44}$$

$$\frac{p(k)}{p(0)} = \left(\frac{p(1)}{p(0)}\right)^{k^2} \quad . \tag{45}$$

The values of p(1)/p(0) and p(0) are plotted in Figure 7 as functions of W/W_o. The value of $H_{max.}$ is plotted as a function of the same variable in Figure 8. The latter curve shows, for instance, that the maximum amount of information per pulse for $W=W_o/2$ is 1.14, that is, 14 per cent higher than the value obtained by using two equally probable levels.

The procedure for approaching in practice the theoretical limit obtained above by appropriate recoding of the messages is very similar to that discussed in Section IV. It differs only in that the ensemble of all sequences of given length must now be divided in groups with probabilities p(0), p(1)... p(k)..., instead of in equally probable groups. The number of pulse levels to be used in practice (it should be infinite in theory) must be selected

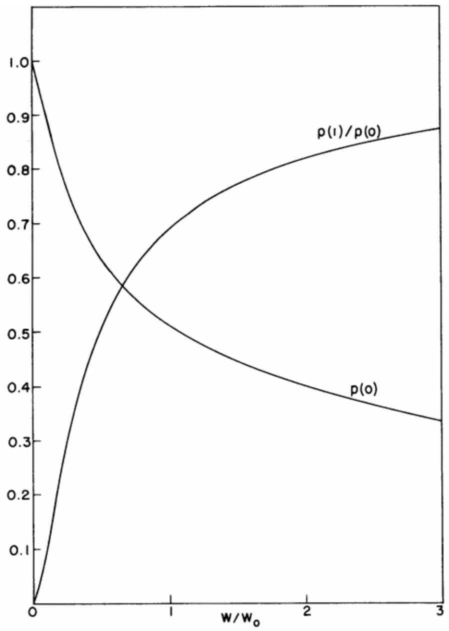

Fig. 7 Behavior of $p(1)/p(0)$ and $p(0)$ as functions of W/W_o.

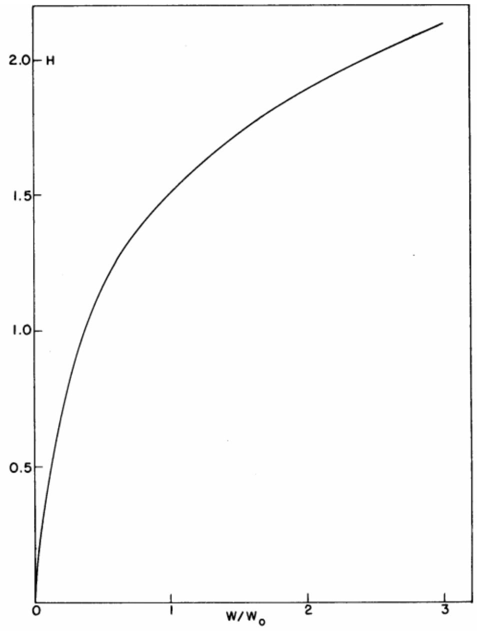

Fig. 8 Maximum information per pulse, $H_{max.}$, as a function of W/W_o.

on a compromise basis, and the values of the p(k) must then be readjusted, accordingly to make

$$\sum_{k=0}^{N-1} p(k) = 1 \quad .$$

In addition to the effect of limitations on the average power, another important practical consideration has been neglected in the preceding sections. All the types of recoding procedures suggested, for approaching in practice the theoretical limits derived above, require the use of devices capable of storing the information for a certain length of time in both the transmitter and the receiver. Such storage devices are needed to stretch or compress the time scale according to the probability of the group of original selections being recoded for transmission.

Satisfactory storage units are not yet available. In addition, even were they available, their use would undoubtedly add considerably to the complexity of communication systems. On the other hand, any substantial increase of transmission efficiency is fundamentally based on time stretching. In fact, since the logarithm of the probability of the choice or sequence of choices selected is a measure of the information conveyed by the selection (see p. 14), the time rate at which information is conveyed in actual signals may vary considerably with time. Even so, a communication system must be able to handle at any time the peak rate which may be present in the signal. It follows that any system not employing storage devices to stretch or compress the time scale is bound to have an efficiency lower than the ratio of the average rate to the peak rate at which information is fed to it. It is worth mentioning in this connection that in certain types of communications, such as telegraph and television, the input and output signals do not have inherently fixed time scales. This is the same as saying that such forms of communication inherently incorporate storage devices. In the case of the telegraph, the written messages at the input and at the output are effectively storage devices. In the case of television, the image to be televised and the cathode-ray tube perform the same function.

Although no reduction of frequency band for a given noise level can be obtained without storage devices, appropriate coding may lead to some reduction of average power. This reduction can be obtained by assigning sequences of pulses requiring the smallest energy to the most probable messages, and vice versa. In the particular case of pulse-code modulation, for instance, this can be done as follows. We arrange all digit combinations in order of increasing amount of energy required and the sampling levels in order of decreasing probability. We assign then the digit combinations to the sampling levels in the resulting order. Such a coding method requires, however, more flexible coding and decoding units than those used in present-day systems.

Before concluding this section, it should be made clear that the improvement of transmission efficiency discussed above and the resulting possible reduction of bandwidth requirements for a given signal power have little to do with the bandwidth reduction obtained by means of the Vocoder or other similar schemes. The Vocoder (2), for instance, does not improve the efficiency of transmission, but achieves a reduction in bandwidth by eliminating that part of the speech signal which is not strictly necessary for the mere understanding of the words spoken. Obviously, the recoding of messages according to their statistical character and the elimination of unnecessary information represent fundamentally different but equally important contributions to the solution of the bandwidth-reduction problem.

Appendix I

Maximization of f(x)

In determining the values of the x_k for which $f(x)$, as given in Eq.(18), is a maximum, it is more convenient to operate on the function

$$\varphi(x) = \ln f(x) \qquad \text{(I-1)}$$

whose maxima and minima at non-singular points coincide with those of $f(x)$. The x_k are the variables in the maximization process, but are subject to the constraint

$$\sum_{k=0}^{N-1} x_k = 1 \quad . \tag{I-2}$$

Using Lagrange's method, we equate to zero the partial derivatives, with respect to the x_k of the function

$$\varphi(x) + \lambda \sum_{k=0}^{N-1} x_k \quad , \tag{I-3}$$

where λ is a constant to be determined later. We obtain then N equations of the form

$$n[\ln p_k - (1 + \ln x_k)] - \frac{1}{2x_k} + \lambda = 0 \quad . \tag{I-4}$$

It is clear that when n approaches infinity these equations can be satisfied simultaneously only when $x_k = p_k$, in which case Eq.(I-2) is also satisfied. In addition, the function $f(x)$ is neither discontinuous nor a minimum at the point $x_k = p_k$, so that the existence of a maximum at this point does not require any further mathematical proof.

Maximization of H_N

The function H_N given by Eq.(22) must be maximized with respect to the p(k) which are, of course, subject to the constraint

$$\sum_{k=0}^{N-1} p(k) = 1 \quad . \tag{I-5}$$

Following the same method as above, we obtain N equations of the form

$$\frac{\partial}{\partial p(k)} \left[H_N + \lambda \sum_{k=0}^{N-1} p(k) \right] = - \frac{1}{\ln 2} [1 + \ln p(k)] + \lambda = 0. \tag{I-6}$$

This set of equations can be satisfied only if all the p(k) are equal. Again it is clear that H_N is neither discontinuous nor a minimum when all the p(k) are equal, and therefore it must be a maximum.

Appendix II

Proof That $H_N(n+1) \leqslant H_N(1)$

We wish to show, first, that the increment of the amount of information

$$H_N(n+1) = -\sum_{k=0}^{N-1} \sum_{i=0}^{N^n-1} P_n(i)\ P_{n+1}(i;k)\ \log_2 P_{n+1}(i;k) \qquad \text{(II-1)}$$

is a maximum when $P_{n+1}(i;k) = p(k)$, the probability of the k^{th} choice, that is, when the additional selection is independent of all preceding selections. Mathematically, we must maximize the function $H_N(n+1)$ with respect to the N^{n+1} variables $P_{n+1}(i;k)$, subject to the conditions

$$\sum_{i=0}^{N^n-1} P_n(i)\ P_{n+1}(i;k) = p(k), \qquad \text{(II-2)}$$

and

$$\sum_{k=0}^{N-1} P_{n+1}(i;k) = 1 \quad . \qquad \text{(II-3)}$$

Following Lagrange's method, we equate to zero the derivates with respect to the $P_{n+1}(i;k)$ of the function

$$H_N(n+1) + \sum_{i=0}^{N^n-1} \sum_{k=0}^{N-1} \lambda_i\ P_{n+1}(i;k) + \mu_k\ P_n(i)\ P_{n+1}(i;k) \qquad \text{(II-4)}$$

where the λ_i and μ_k are constants to be determined later. We obtain then, for each pair of values of i and k, an equation of the form

$$P_n(i)\ [1 + \ln P_{n+1}(i;k)] - \lambda_i - \mu_k P_n(i) = 0 \quad . \qquad \text{(II-5)}$$

The solution of the N^{n+1} equations of this type, together with Eqs.(II-2) and (II-3), is clearly

$$\lambda_i = P_n(i) \quad , \qquad \text{(II-6)}$$

$$\mu_k = \ln P_{n+1}(i;k) = \ln p(k) \quad . \qquad \text{(II-7)}$$

Therefore, the increment of information H_N is a maximum for $P_{n+1}(i;k) = p(k)$, since this is the only point at which a maximum can exist and a maximum must exist at some point. This result can also be stated in the form

$$H_N(n+1) \leqslant H_N(1) \quad , \tag{II-8}$$

where

$$H_N(1) = -\sum_{k=0}^{N-1} p(k) \log_2 p(k) \tag{II-9}$$

is the average amount of information per selection, that is, the average increment of information, when each selection is independent of all preceding selections.

Proof That $H_N(n+1) \leqslant H_N(n)$

Let us consider a sequence of n selections as consisting of a first selection followed by a sequence of n-1 selections. Let $P_n(h;j)$ be the conditional probability that the selection of the h^{th} choice is followed by the selection of the j^{th} sequence from the N^{n-1} possible sequences of n-1 selections. Let also $P_{n+1}(h,j;k)$ be the conditional probability that the k^{th} choice is selected after the h^{th} choice and the j^{th} sequence. We shall still indicate with p(k) the probability of the k^{th} choice and, similarly, with p(h) the probability of the h^{th} choice. Using these new symbols, Eq.(II-1) becomes

$$H_N(n+1) =$$

$$-\sum_{h=0}^{N-1} \sum_{j=0}^{N^{n-1}-1} \sum_{k=0}^{N-1} p(h)\, P_n(h;j)\, P_{n+1}(h,j;k) \log_2 P_{n+1}(h,j;k) \quad . \tag{II-10}$$

We wish to show that, for a statistically uniform sequence, $H_N(n+1)$ is a maximum when $P_{n+1}(h,j;k)$ is independent of h. Mathematically, we must again maximize the function $H_N(n+1)$ with respect to the N^{n+1} variables $P_{n+1}(h,j;k)$, subject to the conditions

$$\sum_{k=0}^{N-1} P_{n+1}(h,j;k) = 1 \quad , \tag{II-11}$$

and

$$\sum_{h=0}^{N-1} p(h)\, P_n(h;j)\, P_{n+1}(h,j;k) = P_{n-1}(j)\, P_n(j;k) \quad , \tag{II-12}$$

where $P_{n-1}(j)$ is the probability of the j^{th} sequence of n-1 selections, and $P_n(j;k)$ is the conditional probability that the k^{th} choice will be selected after the j^{th} sequence. These two probabilities must, in turn, satisfy the condition

$$\sum_{j=0}^{N^{n-1}-1} P_{n-1}(j)\ P_n(j;k) = p(k) \quad , \qquad \text{(II-13)}$$

which, however, does not concern us, since it does not involve directly the $P_{n+1}(h,j;k)$. It must be clear, on the other hand, that the $P_n(j;k)$ are kept constant in the maximization process. In other words, the dependence of the $(n+1)^{th}$ selection on the n-1 preceding selection is fixed in this case, while in the case discussed previously it was allowed to vary. In addition, since we are dealing with a statistically uniform sequence, the $(n+1)^{th}$ selection depends on the n-1 preceding selections in the same manner as the n^{th} selection depends on its n-1 preceding, that is, on all the preceding selections.

Proceeding in the same manner as in the proof that $H_N(n+1) \leqslant H_N(1)$, we find that, for given $P_n(j;k)$, the $P_{n+1}(h,j;k)$ make $H_N(n+1)$ a maximum when they are independent of h, that is, of the first selection of the sequence. Mathematically speaking, the maximum occurs when $P_{n+1}(h,j;k) = P_n(j;k)$. It follows that Eq.(II-10) yields, with the help of Eq.(II-11),

$$H_N(n+1)_{max.} = - \sum_{j=0}^{N^{n-1}-1} \sum_{k=0}^{N-1} P_{n-1}(j)\ P_n(j;k)\ \log_2 P_n(j;k) = H_N(n) \qquad \text{(II-14)}$$

This result can also be stated in the form

$$H_N(n+1) \leqslant H_N(n) \quad . \qquad \text{(II-15)}$$

It must be clear that, in the case of non-statistically uniform sequences, $P_n(j;k)$ may be an entirely different function than that representing the dependence of the n^{th} selection on the first n-1 selections of the sequence, since, for instance, the $(n+1)^{th}$ selection can be entirely independent of the preceding selections while the n^{th} selection is not. It follows, in this latter case, that Eq.(II-14) is not valid, and $H_N(n+1)$ can be as large as $H_N(1)$.

Appendix III

We wish to maximize the average amount of information per pulse, H, for a given average power and an unlimited number of pulse levels equally spaced in voltage. Mathematically, this amounts to maximizing the function given by Eq.(41), subject to the conditions imposed by Eqs.(42) and (43). Following Lagrange's method, as in Appendices I and II, we obtain an infinite set of equations of the form

$$1 + \ln p(k) = \lambda + k^2\mu \quad , \tag{III-1}$$

where λ and μ are indeterminate constants. The first of these constants, λ, can be eliminated by subtracting the equation with k=0 from all the other equations of the set, which take then the form

$$\ln p(k) - \ln p(0) = k^2\mu \quad . \tag{III-2}$$

The remaining constant, μ , is then eliminated by subtracting k^2 times Eq.(III-2) – with k=0 – from the other equations of the same set. We obtain in this manner a set of equations of the form

$$[\ln p(k) - \ln p(0)] - k^2 \, [\ln p(1) - \ln p(0)] = 0 \tag{III-3}$$

It follows that

$$\frac{p(k)}{p(0)} = \left[\frac{p(1)}{p(0)}\right]^{k^2} \quad . \tag{III-4}$$

Eqs.(42) and (43) can now be written in the forms

$$p(0) \sum_{k=1}^{\infty} k^2 \left[\frac{p(1)}{p(0)}\right]^{k^2} = \frac{W}{W_o} \quad , \tag{III-5}$$

and

$$p(0) \sum_{k=1}^{\infty} \left[\frac{p(1)}{p(0)}\right]^{k^2} = 1 \quad . \tag{III-6}$$

The values of p(1)/p(0) are plotted in Figure 7 as functions of W/W_o. From these values, the p(k) are immediately obtained by means of Eq.(III-4).

The maximum value of the average amount of information H can now be obtained without difficulty by substituting for the p(k) in Eq.(41) the values determined above. We have then, after appropriate manipulation of the equation,

$$\begin{aligned}
H_{max.} &= - p(0) \left\{ \sum_{k=0}^{\infty} \left[\frac{p(k)}{p(0)}\right] \left[\log_2 \frac{p(k)}{p(0)} + \log_2 p(0)\right] \right\} \\
&= - \log_2 p(0) - p(0) \sum_{k=1}^{\infty} \left[\frac{p(1)}{p(0)}\right]^{k^2} \log_2 \left[\frac{p(1)}{p(0)}\right]^{k^2} \qquad \text{(III-7)} \\
&= - \log_2 p(0) - p(0) \left\{ \sum_{k=1}^{\infty} k^2 \left[\frac{p(1)}{p(0)}\right]^{k^2} \right\} \log_2 \frac{p(1)}{p(0)} \quad .
\end{aligned}$$

Using now Eq.(III-5), we obtain finally

$$H_{\text{max.}} = -\left[\frac{W}{W_o} \log_2 \frac{p(1)}{p(0)} + \log_2 p(0)\right] \quad . \qquad \text{(III-8)}$$

The value of $H_{\text{max.}}$ is plotted in Figure 8 as a function of W/W_o, using the values of $p(1)/p(0)$ and $p(0)$ given in Figure 7.

Acknowledgment

The author wishes to express his gratitude to Dr. Wiener and Dr. Shannon for discussing with him their work and allowing him to read their manuscripts before publication. In addition, he wishes to thank his colleagues in the Communications Group of the Research Laboratory of Electronics for their constructive criticisms as well as for their very helpful suggestions.

References

(1) N. Wiener, "The Extrapolation, Interpolation, and Smoothing of Stationary Time Series", N.D.R.C. Report, M.I.T. (Feb. 1, 1942) (being reprinted by the Technology Press).

(2) H. Dudley, J. Acous. Soc. Am., 11, 169(1939).

(3) J. C. R. Licklider and I. Pollack, J. Acous. Soc. Am., 20, 42 (1948).

(4) N. Wiener, "Cybernetics", New York, The Technology Press, John Wiley and Sons (1948).

(5) C. E. Shannon, "A Mathematical Theory of Communication", B.S.T.J. 27 (July and Oct. 1948).

(6) R. V. L. Hartley, "Transmission of Information", B.S.T.J. (July 1928).

(7) W. Tuller, "Theoretical Limitations on the Rate of Transmission of Information", Sc. D. Thesis in E.E. Dept., M.I.T. (June 1948).

(8) J. V. Uspensky, "Introduction to Mathematical Probability", McGraw-Hill, New York (1937) App. I, Sec. 2.

(9) T. C. Fry, "Probability and Its Engineering Uses" (Van Nostrand, New York, 1928), p. 103.

(10) Uspensky, op. cit., App. I, Sec. 1.

***