Bounds on Communication

By DAVID SLEPIAN

(Manuscript received February 6, 1963)

Six parameters of importance in many communication systems are: (*a*) *the rate at which digital information is transmitted;* (*b*) *the bandwidth of the system;* (*c*) *the signal power of the transmitted signals;* (*d*) *the noise power of disturbances in transmission;* (*e*) *the error probability in digits recovered at the receiver output;* (*f*) *the length of time that the transmitter and receiver can store their inputs. These six parameters cannot assume arbitrary values: certain sets of values cannot be realized. In a series of curves, this paper describes the boundary between compatible and incompatible sets of parameter values. In the model studied, it is assumed that the disturbance is additive Gaussian noise with constant power density spectrum in the transmission band.*

I. INTRODUCTION

In comparing the performance of communication systems that transmit information by means of signals of limited bandwidth, six quantities descriptive of the system and its environment are of particular importance: (*i*) the rate at which the system transmits information; (*ii*) the bandwidth occupied by the transmission signals; (*iii*) a measure of the power of these signals; (*iv*) a measure of the ambient noise which perturbs the transmitted signals; (*v*) the delay time (caused by the transmitter and receiver) between the introduction of information at the input of the system and the emergence of useful information at the output of the system; (*vi*) a measure of the fidelity with which the information at the output of the system represents the information presented to the input of the system.

To compare the performance of two communication systems in a meaningful manner, it is usually necessary to consider the values of at least these six quantities for the two systems. In general, such a comparison will not yield a simple ordering of the two systems. Two systems may utilize the same bandwidth, introduce the same delay, and operate in the same noise environment. The first system may transmit information at a greater rate with somewhat better fidelity than the second, but may require much more signal power. Which system is to be judged better then depends on external considerations such as the economics of equipment and the purpose for which communication is being established. These external considerations allow the engineer to assign relative weights or costs to the six quantities in question.

Quite apart from these costs dictated by external considerations that may vary with every conceivable usage of a communication system, it is clearly desirable to know, in the first place, what mutual values of the six quantities can ever be obtained by any means. In order to provide such quantitative information it is necessary to particularize both the model of the communication system and the six descriptive parameters.

In all that follows we shall assume that a discrete message source presents independent equiprobable decimal digits for transmission at the uniform rate R decimal digits (or dits) per second. (The output of any other discrete source having entropy rate R can be encoded into this form.) A transmitter operates on these decimal digits to produce a continuous signal of average power S lying in the frequency band $(0,W)$ cycles/second. The signal produced by the transmitter is perturbed by the addition of independent Gaussian noise of total power N and constant power spectral density N/W in the band $(0,W)$ cycles/second. A receiver operates on the perturbed signal to produce decimal digits at an average rate R symbols/second. When the receiver output symbols and the transmitter input symbols are placed in proper correspondence, the average probability, P_e, that an output symbol be different from the corresponding input symbol will be taken as the measure of fidelity with which the system operates. To perform their coding functions, the transmitter and receiver may each require the internal storage of T seconds of their inputs. We use the dimensionless parameter

$$n = 2WT$$

(that is, T measured in Nyquist intervals) as a measure of the delay or complexity of encoding associated with transmitter and receiver.

Our concern henceforth is with the six quantities R, W, S, N, n, and P_e of this model and with the determination of the boundaries of the region of compatible values for these parameters. The famous capacity formula of Shannon[1] published in 1948, $C = W \log (1 + S/N)$, provides information about this boundary when $n \rightarrow \infty$, i.e., when arbitrarily complicated receiver and transmitter coding operations are allowed. The astonishing fact that P_e could be made arbitrarily small for certain finite nonzero values of R, W, and S/N by letting $n \rightarrow \infty$, promised the existence of most remarkable and previously unsuspected communication systems. This led Gilbert[2] and others to compute the values of R, W, S/N, and P_e obtainable with specific transmitters and receivers having fixed delay n and to compare these results with Shannon's formula. The results were disappointing. For all systems examined, even those permitting quite complex encodings ($n = 100$), it was found that to achieve practical values of P_e, S/N had to be at least 6 db more than that given by the capacity formula. The question arose: was this result due to the comparative poorness of the specific systems chosen, or is the approach to the ideal systems described by the capacity formula inherently very slow with increasing n? For a fixed finite value of n, what values of R, W, S/N and P_e are theoretically attainable?

Some information on this subject for large values of n was given by Rice[3] as early as 1950. The question was answered in considerable detail by Shannon in an important paper[4] in which he presented a number of inequalities that permit rather accurate determination of the region of attainable parameter values for all values of n. Shannon's primary interest here was again in the case of large delay, and he developed asymptotic forms for his inequalities in this case. For small delay, the inequalities involve quite complicated expressions and their numerical evaluation is not a simple matter.

The present paper describes in Appendix A a technique which, by means of an electronic computer, permits highly accurate evaluation of the quantities entering these inequalities. The technique has been used to map out bounds on the compatible region of the six quantities in question over a wide range of parameter values. The results of the computations are presented here in a number of curves which cross plot the quantities in various ways which we hope will be useful to the communication engineer.* In particular, the curves show quantitatively the improvement in communication systems that can be achieved with a given degree of coding (measured by delay). Considerable improvement can be obtained with a small amount of encoding, but to approach within a few db of the capacity formula in general requires extremely complicated systems. The curves also give numerical information concerning the trade-offs of the various parameters. They should provide useful references of comparison for existing communication systems.

* An application of these curves to the problem of determining the threshold in modulation systems that expand bandwidth is given in Ref. 5.

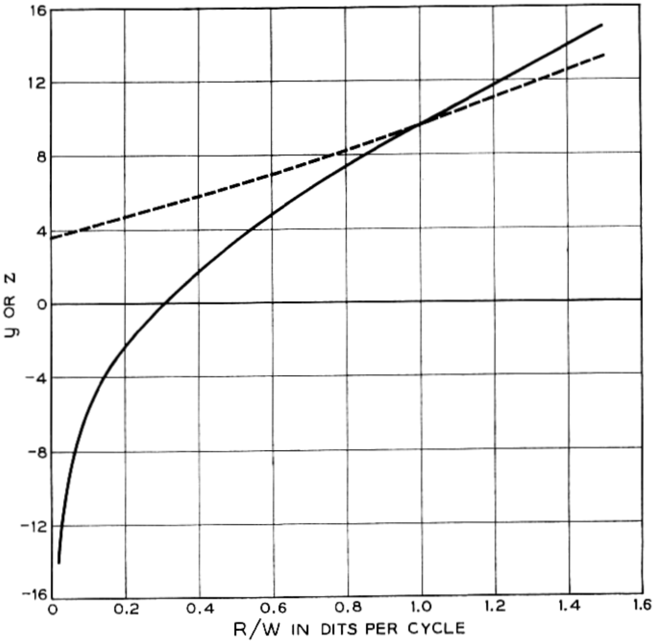

Fig. 1 — Relationship between signal parameters with arbitrarily complex encoding. Solid curve gives $y = 10 \log_{10} S/N$ vs R/W; dashed curve gives $z = 10 \log_{10} (SW/NR)$ vs R/W.

II. IDEAL SYSTEMS — UNRESTRICTED CODING

The solid curve on Fig. 1 shows a plot of the relation

$$R = W \log_{10} (1 + S/N) \tag{1}$$

in terms of the two dimensionless quantities

$$r = R/W, \qquad y = 10 \log_{10} (S/N).$$

This curve can be interpreted* as follows. For values of R, W, S and N corresponding to points above the curve, transmission with arbitrarily small positive values of P_e can be achieved by use of sufficiently complicated coding schemes (sufficiently large finite values of n). For values of R, W, S and N corresponding to points below the curve, P_e is bounded away from zero independently of n. For systems represented by these points, no amount of coding can make the error probability arbitrarily small.

* There are many subtle and thorny points in the argument that permits one to apply the capacity formula to communication systems transmitting continuous signals. Some of these points are discussed in Appendix B.

In many communication situations, the quantity

$$Z = \frac{S/N}{R/W} = \frac{S/R}{N/W}$$

is a useful system parameter. This quantity is the signal energy per dit divided by the noise power per unit bandwidth. From (1),

$$Z = (10^r - 1)/r. \tag{2}$$

The dashed curve of Fig. 1 shows a plot of

$$z = 10 \log_{10} Z$$

vs r as determined by (2). For a given value of r, values of z above the curve are attainable with arbitrarily small positive P_e and finite delay; arbitrarily small positive values of P_e cannot be obtained for z values below the curve with finite delay.

The curves on Fig. 1 describe the relations between R, W, S and N along the intersection of the planes $P_e = 0$, $n = \infty$ with the boundary of the region of mutual compatibility of the six parameters. The intersection of any two other planes, say $P_e = c_1$ and $n = c_2$, with this boundary also determines a curve in the y-r or z-r plane. Unfortunately, the exact form of these curves is not known at present.

III. FINITE n AND NONZERO P_e

To understand fully the assumptions implicit in the remaining curves to be presented here, it is necessary to recall the approach taken by Shannon in Refs. 4 and 6.

Since the signal produced by the transmitter is limited in frequency to the band $(0,W)$ cycles/second, it can (according to the sampling theorem) be thought of as generated by the application of a train of impulses as input to an ideal low-pass filter with cutoff frequency W. The impulses are spaced $1/(2W)$ seconds apart and are of varying amplitude. During a fixed time T, $n = 2WT$ such impulses are applied to the filter. During this same time T, the information source can produce one of $M = 10^{RT}$ different messages. One method, then, of determining from the output of the information source the train of impulses to be applied to the filter is to provide a dictionary that lists for each of the possible M messages a corresponding sequence of n impulses. The transmitter examines the source output for T seconds and determines which of the M messages was produced. The dictionary is then consulted to obtain the corresponding sequence of n impulses. These impulses are applied at a uniform rate to the filter during the next T seconds. At the end of this time, the source has produced another message from the list of M messages and the process is repeated. This method of encoding the source is known as block coding of length n.

In a block coding scheme of length n, the average power of the signal produced at the output of the filter depends on the amplitudes of the impulses listed in the encoding dictionary. It is easy to show that each word of the dictionary, i.e., each sequence of n impulses, contributes an energy $d^2/2W$ to the transmitted signal. Here d^2 is the sum of the squares of the amplitudes of the n impulses in question. Since one word is transmitted every T seconds, one method of achieving average power S for the transmitted signal is to require that $d^2 = nS$ for each word of the dictionary. We shall refer to dictionaries of this sort as equal energy block codes.

In Ref. 4, Shannon presents explicit formulae for functions $Q_n(r,Y)$ and $\bar{Q}_n(r,Y)$ which have the following significance. For the communication model under discussion, there exist transmitters and receivers using equal energy block codes of length n such that

$$P_e \leqq \bar{Q}_n(R/W,S/N).$$

For every equal energy block code of length n, the system parameters satisfy the inequality

$$P_e \geqq Q_n(R/W,S/N).$$

Here P_e is the probability that a transmitted word of the dictionary be decoded incorrectly. The functions Q_n and $\bar{Q}_n$ and their numerical evaluation are discussed further in Appendix A.

Consider now a relationship such as

$$Q_{101}(R/W,S/N) = 10^{-4} \tag{3}$$

which serves to determine S/N as a function of R/W. This relation could be plotted on Fig. 1 with S/N measured in db to yield a curve lying above the solid-line capacity curve shown there. For our purposes, the vertical difference between these two curves is of primary interest. This difference is shown by the bottom solid curve of Fig. 2. Explicitly, the bottom curve of Fig. 2 is a plot of

$$y = 10 \log_{10} (S/N) - 10 \log_{10} (10^{R/W} - 1)$$

vs R/W, where S/N is given in terms of R/W by (3). The bottom dashed curve of Fig. 2 is an analogous display of the relation defined by

$$\bar{Q}_{101}(R/W,S/N) = 10^{-4}.$$

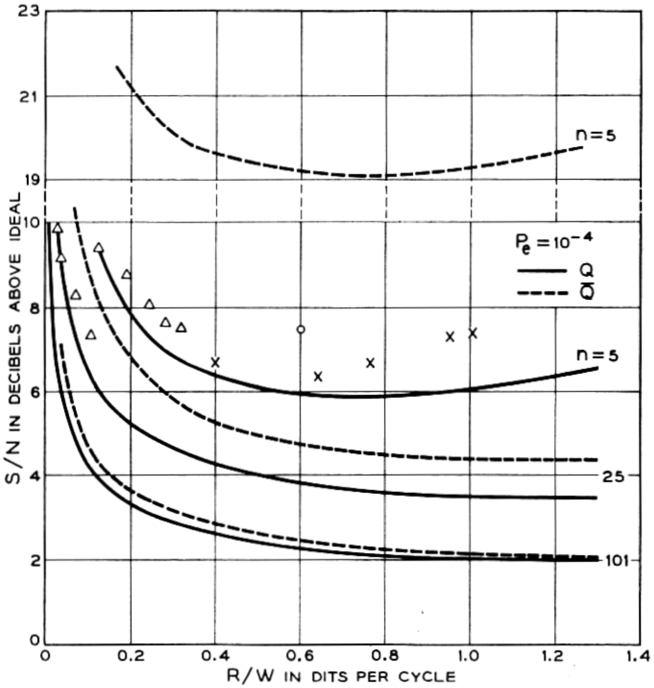

Fig. 2 — Upper and lower bounds (dashed and solid curves, respectively) on S/N needed to achieve word-error probability of 10^{-4} for various values of $n = 2WT$. Circle, triangles, and crosses give performance of some known codes.

The two bottom curves on Fig. 2 have the following significance. For a given value of R/W, there exist equal energy block codes of length 101 that will achieve an error probability of $P_e = 10^{-4}$ with as small a value of S/N as that given by the ordinate of the dashed curve. On the other hand, every equal energy block code of length 101 that achieves an error probability of 10^{-4} must operate with a value of S/N at least as large as the ordinate of the solid curve. The curves thus serve to bound the minimal signal-to-noise ratio with which an error probability of 10^{-4} can be achieved when equal energy block codes of length 101 are employed. The bounds are plotted in db above the signal-to-noise ratio given by the capacity formula, and thus measure the penalty in signal-to-noise ratio that must be paid for restricting the coding ($n = 101$).

The remaining curves on Fig. 2 give analogous results for $n = 5$ and $n = 25$. It is to be noted that the solid and dashed curves are much closer together for large n, than for small n. This effect is shown more clearly on Fig. 3, which was obtained from a cross plot of many curves

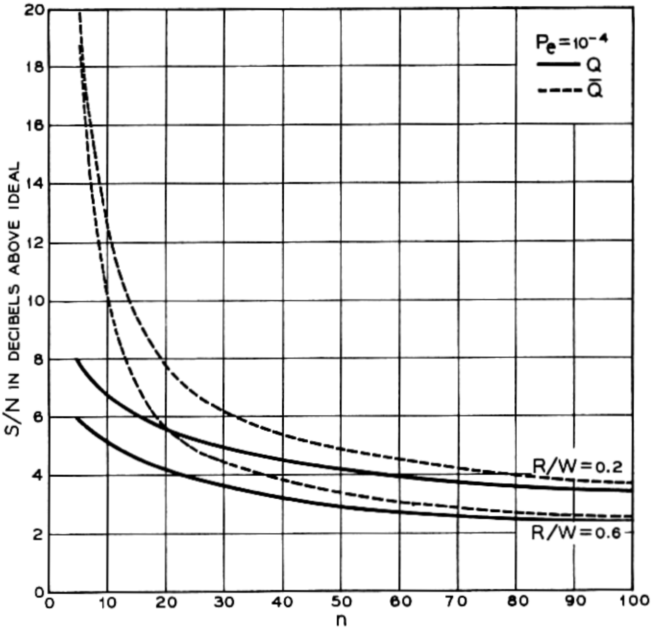

Fig. 3 — Cross-sections of Figure 2 taken for $R/W = 0.2$ and 0.6.

of the sort shown on Fig. 2. For $n = 101$, there is little practical difference between the two bounds. For small values of n, however, the disparity is great, and the question naturally arises: does the solid curve, or the dashed curve, more nearly represent the minimal signal-to-noise ratio needed to obtain $P_e = 10^{-4}$ with an equal energy block code of fixed length n?

We believe that the bound obtained from Q is quite close to the minimal attainable S/N even for small n. Indeed, for $n = 5$, we have been able to construct explicit equal energy block codes with a variety of rates whose parameters plot close to the top-most solid line of Fig. 2 when S/N was adjusted to guarantee an error probability not greater than 10^{-4}. The five right-most triangles in the figure locate the performance of certain block codes known as simplex codes [the $(D,D + 1)$ codes of Ref. 2]. The crosses locate the performance of certain new codes to be described in a later paper. The circle gives the performance of 5-bit PCM. The four left-most triangles locate the performance of some simplex codes of block length 25. Apart from these explicit examples that plot near the bounds obtained from Q, there are theoretical considerations which show that $\bar{Q}$ is a very weak bound for small values of n. Henceforth, in this paper we shall deal only with bounds obtained from Q and shall treat the relationship

$$Q_n(R/W,S/N) = P_e \tag{4}$$

as the defining equation of the boundary of the region of compatible values of R, W, S, N, P_e and n for equal energy block codes.

IV. DISCUSSION OF RESULTS

Figs. 4, 5 and 6 give plots of S/N vs R/W as determined from (4) for various values of P_e and n. The ordinates here, as in Fig. 2, are given in db above capacity, i.e., in db above the solid curve of Fig. 1. One advantage of this representation is that the ordinates of Figs. 4, 5 and 6 may also be interpreted as values of Z, the latter now being measured in db above the capacity value given by the dashed line of Fig. 1.

From Figs. 4, 5 and 6, it is apparent that for a fixed rate and fixed error probability modest amounts of coding (small values of n) can produce a significant reduction in signal power, but that the return for increased encoding diminishes rapidly. This is seen more clearly from the cross plot given on Fig. 7.

The improvement in performance that can be obtained by encoding can also be expressed in terms of decreased error probability for a fixed rate and signal-to-noise ratio as is shown in Fig. 8.

An interesting feature of Figs. 4, 5 and 6 is the minimum value clearly

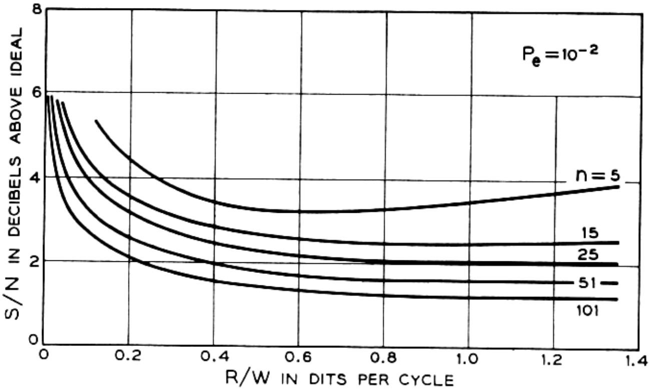

Fig. 4 — Minimum possible S/N to attain word-error probability of 10^{-2} for various values of R/W and n.

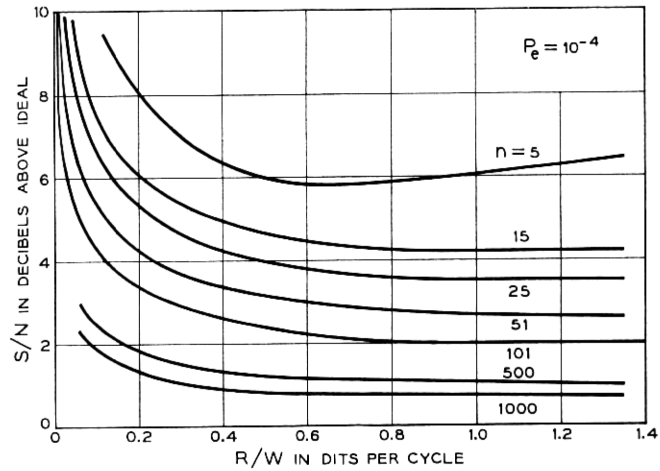

Fig. 5 — Minimum possible S/N to attain word-error probability of 10^{-4} for various values of R/W and n.

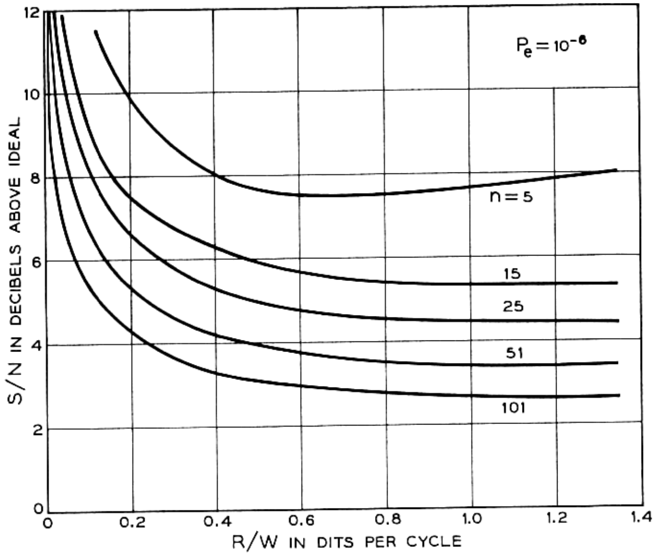

Fig. 6 — Minimum possible S/N to attain word-error probability of 10^{-6} for various values of R/W and n.

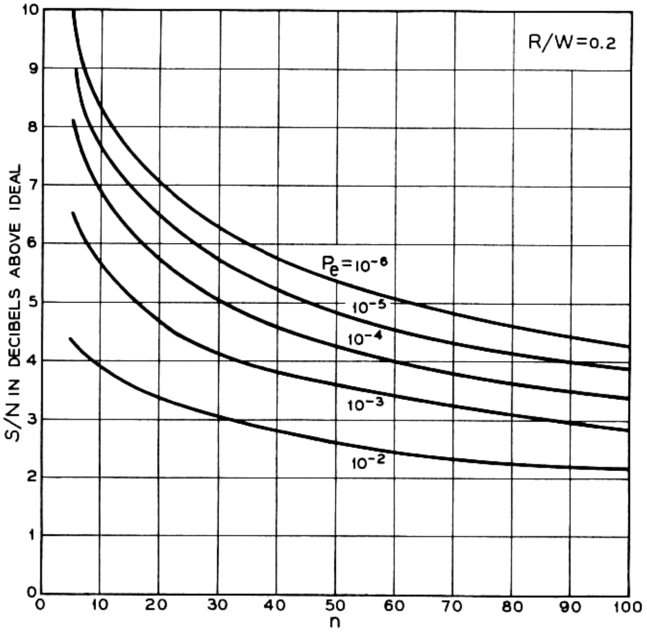

FIG. 7 — Cross-plot of Figs. 4, 5 and 6 showing (for $R/W = 0.2$) decrease in S/N needed to achieve a given word-error probability as n is increased.

evident on the $n = 5$ curves. It is not hard to show (see Appendix C) that for all values of n, the curves obtained from (4) as plotted on these figures must rise indefinitely with increasing R/W. For equal energy block codes, there is, for any fixed P_e and n, a best value of R/W in the sense of minimizing the additional signal-to-noise ratio needed above that given by the channel capacity formula. When the curves of Figs. 4, 5 and 6 are plotted on a graph such as Fig. 1 with absolute S/N as ordinate, the curves are monotone increasing but eventually for large R/W depart further and further above the capacity formula curve. This phenomenon is due to the restriction imposed here that all code words of the dictionary have the same energy, a restriction likely to be realized in practice. This point is discussed further in Appendix D.

Another way of presenting (4) that shows the departure from the ideal system of the capacity formula that results with equal energy block codes of restricted length is shown in Fig. 9. Fix P_e and n. Then from (4), a given value of $r = R/W$ determines a corresponding signal-to-noise ratio, S/N. From the capacity formula, using this value of S/N it is possible to achieve any desired P_e with a rate per bandwidth $\bar{r} =$

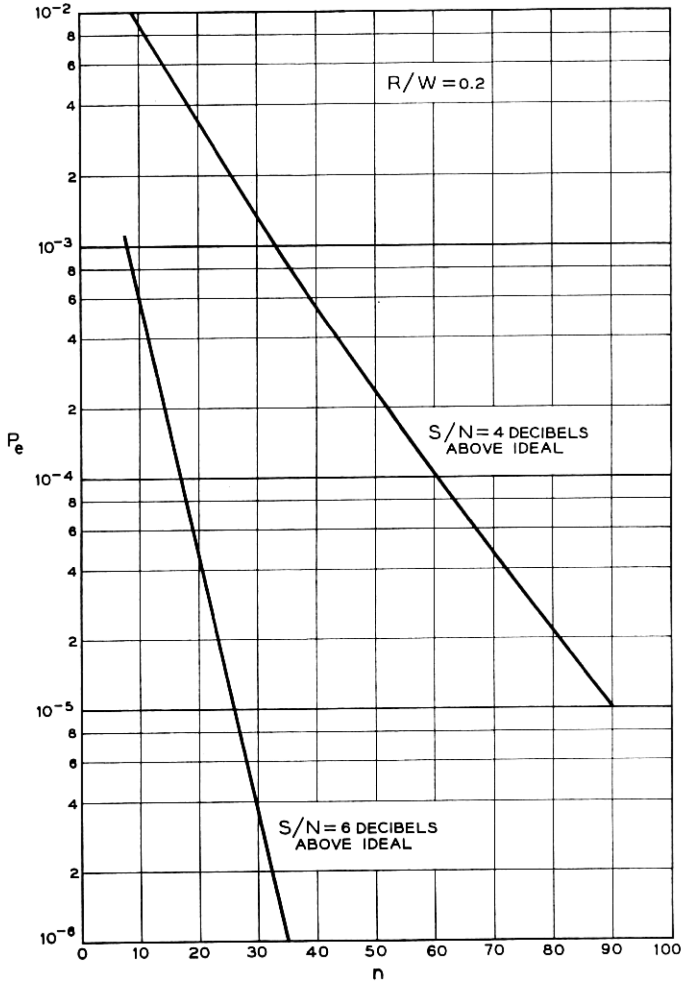

Fig. 8 — Word-error probability vs n for given R/W and S/N above ideal for best possible equal-energy codes.

$\log_{10}(1 + S/N)$ by sufficiently complex encoding. The ratio $r/\bar{r}$ then measures the price paid in lost rate due to restricting the amount of encoding. The solid curves on Fig. 9 were obtained from Q and give upper bounds on $r/\bar{r}$ for equal energy block codes; the dashed curves derived from $\bar{Q}$ give lower bounds for this ratio. It can be shown (see Appendix C) that the solid curves approach $(n - 1)/n$ asymptotically with increasing R/W.

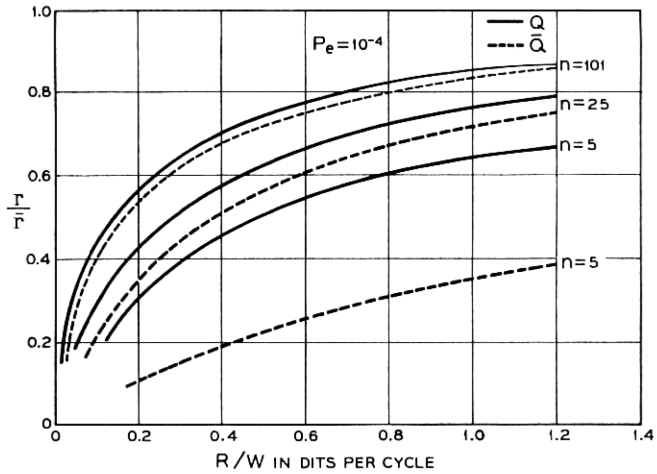

Fig. 9 — Upper and lower bounds (solid and dashed curves, respectively) on fractional loss in rate, $r/\bar{r}$, due to finite encoding. Loss plotted vs R/W for fixed n and P_e.

Yet another way of viewing the bounds is given on Fig. 10. Here, for a fixed signal-to-noise ratio and a fixed error probability, the improvement in signaling rate that can be obtained by increasing the length of equal energy codes is shown. It is seen, for example, that even with signal to noise ratios as high as 20 db, one cannot achieve 75 per cent of the ideal rate with equal energy codes of length less than 15 when the prescribed error probability is 10^{-6}. The $S/N = \infty$ curve is given by $r/\bar{r} = (n - 1)/n$. That this limiting curve is different from unity is again due to the fact that the bounds used here are those for equal energy codes. If restricted energy codes were used, (see Section V) the limiting curve corresponding to $S/N = \infty$ would be $r/\bar{r} = 1$.

V. CONCLUDING REMARKS

The exact computation of Q_n that was carried out here allows one to test the range of validity of Shannon's asymptotic expressions for this quantity. On plots such as Figs. 4, 5 and 6, his formula* (4) of Ref. 4 gives curves in very close agreement with those shown for $n = 101$. At $n = 25$ the error is about 0.1 db at large rates and 0.3 db at small rates. This formula was used to compute the curves for $n = 500$ and 1000 shown on Fig. 5. Although it involves only elementary functions,

* This formula contains a misprint. The printed version must be multiplied by $-G$ to be corrected.

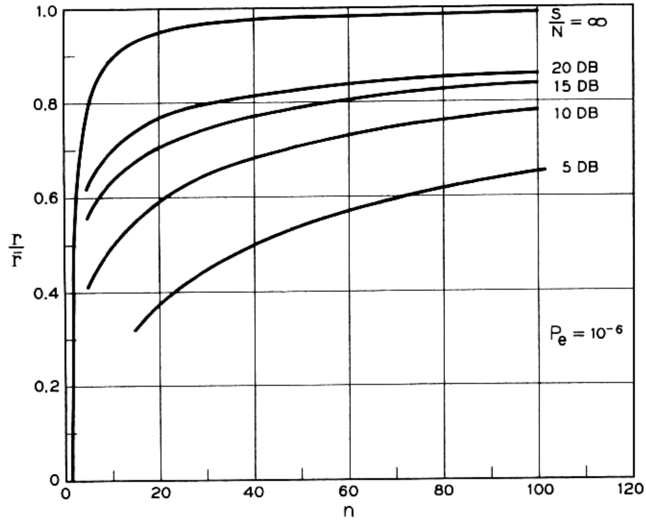

Fig. 10 — Rate loss of Fig. 9 plotted vs n for fixed S/N and P_e .

the formula is quite complicated, and for extensive computations machine methods are desirable. For moderate or small values of n, exact values of Q can be obtained by the method of Appendix A with comparable ease. Shannon's elementary asymptotic formula (73) of Ref. 4 has also been evaluated. For $n = 500$ and 1000, it gives values that agree with the curves of Fig. 5 to about 0.1 db for $R/W > 0.5$. For small rates it gives values 0.5 db too large. The accuracy of the formula diminishes rapidly as n is decreased below 100.

The bounds presented here were obtained for communication systems using equal energy block codes of fixed length. It is, of course, possible to signal using block codes that have words of differing energy. One code of this sort of particular interest that is treated by Shannon in Ref. 4, Section XIII is the *restricted energy block code*. In these codes, each word of the dictionary contributes energy ST or less to the transmitted signal, i.e., for each code word $d^2 \leqq nS$. Note that for these codes S is no longer the average signal power, but rather the maximum contribution to the signal power by any code word.

For any communication system with parameters R, W, S, N using a restricted energy block code of length n, Shannon showed that the average error probability, P_e', for a decoded word is bounded below by

$$P_e' \geqq Q_{n+1}\left(\frac{n}{n+1}\frac{R}{W}, \frac{S}{N}\right). \qquad (5)$$

For any fixed value of R/W, as n becomes large this lower bound approaches the one already given for equal energy block codes, and so asymptotically (in n) one can do no better with restricted energy codes than with equal energy codes. However, for any fixed value of n, as R/W becomes large the lower bounds for the two classes of codes behave very differently, and indeed it is easy to argue that in this limit restricted energy codes are superior to equal energy codes. This point is discussed further in Appendix D.

The solid curves of Fig. 11 are those already shown in Fig. 6. The dashed curves were obtained from the lower bound (5) for restricted energy block codes. These dashed curves approach the horizontal asymptotes indicated at the right. From the figure it is seen that for $R/W < 0.6$ and $n \geqq 25$ the bounds for restricted energy codes differ from those for equal energy codes by less than 0.2 db. For small values of n, the dashed curves lie below the solid ones even for small rates.

It should be pointed out in closing that the error probability P_e used throughout these calculations is the probability that a word of the block code be improperly identified when a maximum likelihood receiver is used. This is not in general the probability that an individual decoded decimal digit be in error but rather an upper bound to this quantity. For large n, a single code word is decoded into many decimal digits

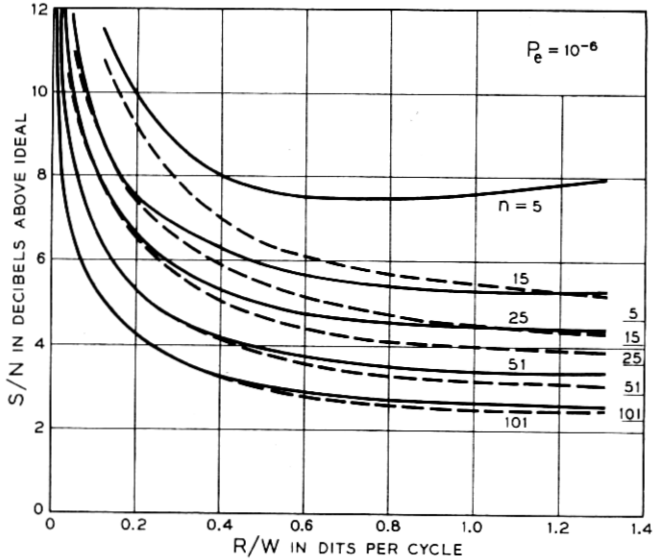

Fig. 11 — Comparison of bounds for equal-energy codes and restricted-energy codes.

The received code word may be incorrectly identified and yet decoded into a block of decimal digits many of which are correct. When large block codes are used and P_e is small, errors in the decoded stream of decimal digits are not distributed uniformly. Many successive groups of decimal digits, each containing RT digits, will be error free. Then a single block of RT digits will be produced that contains from one to RT erroneous digits. This bunching of errors may, in certain applications, be a serious drawback to the use of block coding.

VI. ACKNOWLEDGMENT

All of the key ideas contained in this paper are due to C. E. Shannon. The author's contribution has been one of exposition and the reduction to numerical form of certain inequalities contained in Ref. 4. In this latter task he was greatly aided by Miss Helen Arnold, who programmed the algorithm of Appendix A for the IBM 7090 and oversaw the computations. The author is indebted to E. Wolman for many helpful discussions of this work.

APPENDIX A

Computation of Q and $\bar{Q}$

Our notation is similar to Shannon's[4] and we here adopt his geometrical point of view:

S = signal power (each signal vector is of length $\sqrt{nS}$);

N = noise power (variance N in each dimension);

A = $\sqrt{S/N}$ = signal-to-noise "amplitude" ratio;

n = number of dimensions;

M = number of signal vectors;

$\Omega(\theta)$ = solid angle in n-space of a cone of half-angle θ, or area of unit n-sphere cut out by the cone;

$Q(\theta)$ = probability of a point X in n-space, at distance $A\sqrt{n}$ from the origin, being moved outside a circular cone of half-angle θ with vertex at the origin O and axis OX (the perturbation is assumed spherical Gaussian with unit variance in all dimensions);

θ_1 = angle such that $M\Omega(\theta_1) = \Omega(\pi)$.

Shannon shows [his equation (20)] that

$$Q(\theta_1) \leqq P_e \leqq Q(\theta_1) - \frac{1}{\Omega(\theta_1)} \int_0^{\theta_1} \Omega(\theta)\, dQ(\theta),$$

where P_e is the error probability of the best equal energy M-vector code in n-space used with signal-to-noise ratio A. We proceed to discuss the evaluation of these quantities.

As shown by Shannon [his equation (21)]

$$\Omega(\theta) = \frac{(n-1)\pi^{(n-1)/2}}{\Gamma\left(\frac{n+1}{2}\right)} \int_0^\theta \sin^{n-2}\xi \, d\xi. \tag{6}$$

The surface $\Omega(\pi)$ of the unit n-sphere has area

$$\Omega(\pi) = \frac{n\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)}.$$

A change of variable $\sin^2 \xi = t$ shows that

$$\begin{aligned} \frac{\Omega(\theta)}{\Omega(\pi)} &= \frac{1}{2} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n-1}{2}\right)} \int_0^{\sin^2\theta} t^{1/2(n-1)-1}(1-t)^{\frac{1}{2}-1}\, dt \\ &= \frac{1}{2}\, I_{\sin^2\theta}\left(\frac{n-1}{2}, \frac{1}{2}\right), \end{aligned}$$

where $I_\alpha(p,q)$ is Pearson's incomplete beta function.[7] Thus θ_1 is given by

$$\frac{2}{M} = I_{\sin^2\theta_1}\left(\frac{n-1}{2}, \frac{1}{2}\right). \tag{7}$$

The rate is related to θ_1 by

$$\frac{R}{W} = \frac{2}{n} \log_{10} M. \tag{8}$$

To evaluate $Q(\theta)$, it is convenient to use n-dimensional cylindrical coordinates with origin located on the axis of the cone at a distance

$$l \equiv \sqrt{n}A$$

from the vertex and within the cone. The z- or rotational axis of the coordinate system coincides with the axis of the cone and is oriented so that the vertex of the cone has z-coordinate $-\,l$. Denote distance from the z-axis by r. Then an element of "area" distant r from the axis and having radial dimension dr and axial dimension dz sweeps out volume

$$\frac{(n-1)\pi^{(n-1)/2} r^{n-2}\, dr\, dz}{\Gamma\left(\frac{n+1}{2}\right)},$$

when rotated about the z-axis. One has therefore

$$Q_n = Q(\theta) = \int_0^\infty \int_{-\infty}^{r\alpha - l} dz \, \frac{\exp\left[-\frac{1}{2}(r^2 + z^2)\right]}{(2\pi)^{n/2}} \, \frac{(n-1)\pi^{(n-1)/2} r^{n-2}}{\Gamma\left(\dfrac{n+1}{2}\right)}, \quad (9)$$

where we have set

$$\alpha = \cot \theta.$$

Now set

$$c_n = \sqrt{\frac{2}{\pi}} \, \frac{1}{2^{(n-1)/2}\, \Gamma\left(\dfrac{n-1}{2}\right)}.$$

One then has

$$\begin{aligned}
\frac{Q_n}{c_n} &= \int_0^\infty dr \, r \exp\left(-\tfrac{1}{2}r^2\right) \left\{ r^{n-3} \int_{-\infty}^{r\alpha - l} dz \exp\left(-\tfrac{1}{2}z^2\right) \right. \\
&= -\exp\left(-\tfrac{1}{2}r^2\right) \left\{ r^{n-3} \int_{-\infty}^{r\alpha - l} dz \exp\left(-\tfrac{1}{2}z^2\right) \right\} \Bigg|_0^\infty \\
&\qquad + (n-3) \int_0^\infty dr \exp\left(-\tfrac{1}{2}r^2\right) r^{n-4} \int_{-\infty}^{r\alpha - l} dz \exp\left(-\tfrac{1}{2}z^2\right) \\
&\qquad + \alpha \int_0^\infty dr \exp\left(-\tfrac{1}{2}r^2/2\right) r^{n-3} \exp\left[-\frac{(\alpha r - l)^2}{2}\right] \\
&= (n-3)\frac{Q_{n-2}}{c_{n-2}} + \alpha J_{n-2}, \qquad n > 3, \qquad (10)
\end{aligned}$$

on integrating by parts. Here

$$\begin{aligned}
J_n &= \int_0^\infty dr \, r^{n-1} \exp\left[-\frac{(1+\alpha^2)r^2 - 2\alpha l r + l^2}{2}\right] \\
&= \frac{1}{1+\alpha^2} \int_0^\infty dr \, r^{n-2}\left[(1+\alpha^2)r - \alpha l\right] \exp\left[\frac{(1+\alpha^2)r^2 - 2\alpha l r + l^2}{2}\right] \\
&\qquad + \frac{\alpha l}{1+\alpha^2} \int_0^\infty dr \, r^{n-2} \exp - \left[\frac{(1+\alpha^2)r^2 - 2\alpha l r + l^2}{2}\right] \\
&= \frac{\alpha l}{1+\alpha^2} J_{n-1} - \frac{1}{1+\alpha^2} r^{n-2} \exp\left[-\frac{(1+\alpha^2)r^2 - 2\alpha l r + l^2}{2}\right] \Bigg|_0^\infty \\
&\qquad + \frac{n-2}{1+\alpha^2} \int_0^\infty dr \, r^{n-3} \exp\left[-\frac{(1+\alpha^2)r^2 - 2\alpha l r + l^2}{2}\right]
\end{aligned}$$

$$= \frac{\alpha l}{1+\alpha^2} J_{n-1} + \frac{n-2}{1+\alpha^2} J_{n-2}, \qquad n > 2. \tag{11}$$

Now set

$$G_n = c_{n+2} J_n csc\, \theta, \qquad b_n = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}, \qquad \xi = \frac{l}{\sqrt{2}}.$$

One has from (10) and (11)

$$\left.\begin{aligned} Q_n &= Q_{n-2} + \cos\theta\, G_{n-2}, && n > 3 \\ G_n &= \xi \cos\theta \sin\theta\, b_n G_{n-1} + \frac{n-2}{n-1} \sin^2\theta\, G_{n-2}, && n > 2 \\ b_n &= \frac{n-2}{n-1} b_{n-2}, && n > 2. \end{aligned}\right\} \tag{12}$$

The initial values

$$b_1 = \sqrt{\pi}, \qquad b_2 = \frac{2}{\sqrt{\pi}},$$

$$G_1 = \tfrac{1}{2} \exp(-\xi^2 \sin^2\theta) \operatorname{erfc}(-\xi \cos\theta)$$

$$G_2 = \frac{1}{\pi} \sin\theta\, e^{-\xi^2} + \frac{2\xi}{\sqrt{\pi}} \sin\theta \cos\theta\, G_1,$$

$$Q_3 = \tfrac{1}{2} \operatorname{erfc}(\xi) + \cos\theta\, G_1,$$

with

$$\operatorname{erfc}(x) \equiv \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2}\, dt$$

permit one to compute $Q_n(\theta)$ for odd n from the recurrence (12). Since $0 \leqq \theta \leqq \pi/2$, all quantities involved are positive.

The curves of Figs. 4, 5, and 6 were obtained as follows. With θ_1 fixed in value Q_5, Q_{15}, Q_{25}, Q_{51} and Q_{101} were determined as functions of ξ by repeated application of the recurrence. A given $Q_n(\theta_1)$ was then expressed as a function of the signal-to-noise ratio, A^2, by the relation $\xi = A\sqrt{n/2}$. Values of A^2 for which $Q_n(\theta_1)$ took the values 10^{-2}, 10^{-4}, 10^{-6} were determined graphically. The corresponding rate was found from (7) and (8). Repetition of the process for different values of θ_1 permits plotting the curves.

An integration by parts and (6) allow Shannon's upper bound to be written in the form

$$\bar{Q} = \frac{M}{\sqrt{\pi} b_{n-1}} \int_0^{\theta_1} Q_n(\theta) \sin^{n-2} \theta \, d\theta. \tag{13}$$

Curves based on $\bar{Q}$, such as shown on Fig. 2, were obtained by using the recurrence (12) to obtain values of $Q_n(\theta)$ for a fixed ξ. The integral in (13) was evaluated numerically using a trapezoidal formula with 150 points of evaluation for the integrand. Values of ξ and θ_1 were expressed in terms of R/W and S/N as already explained.

APPENDIX B

The theorems and formulae of Shannon's Information Theory are statements about certain *mathematical* constructs. In order to make useful inferences from these formulae about *physical* communication systems, it is necessary to examine the sense in which the mathematical model approximates the behavior of the key elements of the physical system. At best, the correspondence between mathematical and physical entities is only a close approximation: the "true" theorems of the mathematical model, when stated in physical terms, are only "partial truths."

The formula

$$C = (\alpha/2) \log_{10} (1 + S/N) \text{ dits/second} \tag{14}$$

gives the capacity of the following *mathematical* channel. Real numbers are chosen at a transmitting point at the rate α numbers per second. Each number chosen is transmitted to the receiving point, but is perturbed by an additive Gaussian variate, so that the ith transmitted real number, s_i, is received as $s_i + x_i$. The x_i are assumed independent Gaussian random variables with the same variance N. The transmitted sequence satisfies the constraint

$$\lim_{K\to\infty} \frac{1}{2K} \sum_{-K}^{K} s_i^2 = S.$$

(The reader should consult Ref. 8, Chapter 9, for a more careful, rigorous definition of this channel and a precise mathematical interpretation of the capacity formula.)

The foregoing description of the channel is essentially that given by Shannon in Ref. 4. The channel is discrete in time; there is no mention of bandlimited continuous functions of a time variable defined on the real line. Within the mathematical theory, there is no question of the validity of (14) for the capacity of the discrete time channel described nor of the validity of Shannon's bounds for the error probability attainable with block codes of finite length. The problem is to justify the application of these formulae derived for a discrete time mathematical channel to physical communication systems employing "continuous" signals of "bandwidth" W.

I have placed quotation marks around the words continuous and bandwidth to call attention to the fact that these two concepts have no well-accepted operational definitions in terms of experiments in the real world. They are again part of another strictly *mathematical* model that is used to describe signals of the physical world. The elements of this mathematical model are the real number continuum, functions and Fourier analysis. The correspondence between these elements and observables of the laboratory (meter readings, etc.) is again an approximation — a very good one in many circumstances, but a poor one in many others. It is meaningless to ask if the reading of a meter in the laboratory is a rational number or an irrational one, or if the trace seen on an oscilloscope is a continuous function in the sense used in the mathematical model. Within the mathematical model, there are many notions introduced for which one cannot easily find meaningful counterparts in the real world of the laboratory. The asymptotic behavior of spectra at infinity is such an example. One must be very suspicious of the utility of applying in the real world formulae derived from the mathematical models which are sensitive to assumptions about those concepts of the model that have no operationally defined counterparts in the laboratory.

It is evident that a good case for applying (14) to real communication systems can be made if one can justify the statement

"In the laboratory, using signals of duration T and bandwidth W, we can communicate about $2WT$ numbers and no more." (15)

Perhaps it would be simplest to take this statement as a basic axiom for practical communication engineering and justify it by experiment (with "bandwidth," "number," etc. suitably defined in operational terms). It is intellectually more satisfying, however, to be able to derive it from the mathematical models that have served so well to describe signals in other circumstances.

The approach taken by Shannon in Ref. 6 and paraphrased here at the beginning of Section III is one method of deriving statements in the spirit of (15) from the usual mathematical model of signals and spectra. This approach is reasonably satisfactory in justifying the fact that for very large T one can transmit $2WT$ numbers using signals of (mathematical) bandwidth W and nominal duration T. From it one can argue rather convincingly that rates arbitrarily close to those given by the capacity formula can be achieved with arbitrarily small error probability using (mathematically) bandlimited functions for signaling. Using this approach, however, it is difficult to make a convincing argument that one cannot exceed capacity or that Shannon's bounds Q_n and $\bar{Q}_n$ have any significance for channels employing (mathematical) bandlimited functions.

The difficulty here lies in the fact that mathematical bandlimited functions are entire functions and hence perfectly predictable for all time from knowledge over any finite interval. If one allows all the usual mathematical operations, the receiver, on the basis of observing the bandlimited signal plus noise in an arbitrarily short time interval, could extrapolate this function for all time and obtain sample values at an arbitrarily great rate.

The heart of the dilemma presented here lies in the fact that the mathematical specification that a signal be bandlimited is a statement about concepts of the model that have no well defined physical counterpart — namely, the behavior of spectra at infinity. The sampling theorem, unfortunately, requires an assumption about this nonphysically interpretable part of the mathematical model.

Yet, one feels that in the real world something like (15) holds with laboratory meanings for bandwidth. If so, this should be derivable from the mathematical model of functions and Fourier analysis without making assumptions in the model about such nonphysical entities as the behavior of spectra at infinity. A result of this sort is indeed the content of an important theorem recently published by Pollak and Landau.[9] Their results are too complex to discuss in detail here. The main point is that within the classical model of functions and Fourier analysis they define a suitable class of functions that are "limited" in both time and frequency. The definition of this class does not entail specification of spectral behavior at infinity. The specification, when translated to physical terms, involves only an assumption about one's ability to measure energy, and the correspondence between their class and laboratory bandlimited signals defined in an operational way is easy to make. They prove that in an appropriate sense this class of functions is $2WT$-dimensional. From this, a form of statement (15) results which is, I believe, the best justification on theoretical grounds to date of this important postulate.

Quite apart from this difficulty of justifying (15), there are, of course, many other ways in which the mathematical model only approximates the behavior of equipment in the laboratory: measurement errors prevent one from specifying real numbers meaningfully by more than a finite number of significant figures; disturbances are not truly Gaussian; etc., etc. The attainment of arbitrarily small error by sufficient encoding in the mathematical theory entails a delicate balance between many quantities which only approximate their physical counterparts. One should not believe that real communication systems can be built which will signal at fixed rates with arbitrarily small error. Somewhere, for large enough n, the mathematical model fails to describe adequately the physical realities. How large is this n? This is a very difficult question. My engineering judgment is that the results given on the curves of this paper for n up to 100 might conceivably be achieved with real communication systems. Until we have learned to describe and instrument optimal codes of this size, I am safe from experimental contradiction. Today, this time seems remote.

APPENDIX C

We show here that if

$$Q_n(R/W,S/N) = P_e \tag{16}$$

and

$$\bar{r} = \log (1 + S/N) \tag{17}$$

then, with n and P_e fixed $(0 < P_e < 1)$,

$$\lim_{R/W\to\infty} \frac{R/W}{\bar{r}} = \frac{n-1}{n}.$$

Referring to (7) and (8) we see that if $R/W \to \infty$, then $\theta_1 \to 0$. Indeed, for small values of θ_1, one can easily develop the incomplete beta function to obtain

$$\frac{R}{W} = \frac{2}{n}\left[\ln (n-1)\beta\left(\frac{n-1}{2},\frac{1}{2}\right) \right.$$
$$\left. - (n-1)\ln \sin \theta_1 + 0\,(\theta_1^2)\right]\log_{10} e. \tag{18}$$

Here $\beta(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$ as usual.

It is now convenient to write equation (9) as

$$\frac{Q_n(\theta_1)}{d} = \int_0^\infty dr \int_{-\infty}^{r \cot \theta_1 - \sqrt{n}A} dz\, r^{n-2} \exp\left[-(r^2 + z^2)/2\right]$$

where

$$d = \frac{(n-1)\,\pi^{(n-1)/2}}{(2\pi)^{n/2}\,\Gamma\left(\frac{n+1}{2}\right)}$$

and as before we adopt the abbreviation $A = \sqrt{S/N}$. The region of integration is shaded in Fig. 12.

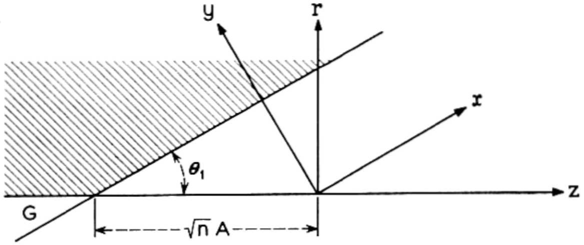

Fig. 12 — Integration region and coordinate transformation.

To investigate the behavior of Q_n as $\theta_1 \to 0$, it is convenient to transform the integral by the rotation

$$z = x \cos \theta_1 - y \sin \theta_1$$

$$r = x \sin \theta_1 + y \cos \theta_1$$

and to write the result as the integral over the region $y \geqq \sqrt{n}A \sin \theta_1$ minus the integral over the region G indicated in the figure.

$$\frac{Q_n}{d} = \int_{\sqrt{n}A \sin \theta_1}^{\infty} dy \int_{-\infty}^{\infty} dx (x \sin \theta_1 + y \cos \theta_1)^{n-2} \exp\left[-\frac{x^2+y^2}{2}\right] - \int_G \int dy\, dx (x \sin \theta_1 + y \cos \theta_1)^{n-2} \exp\left[-\frac{x^2+y^2}{2}\right].$$

With $A \geqq 0$, the integral over G vanishes as $\theta_1 \to 0$, so

$$\frac{Q_n}{d} \to (\cos \theta_1)^{n-2} \int_{\sqrt{n}A \sin \theta_1}^{\infty} dy \int_{-\infty}^{\infty} dx (y + x \tan \theta_1)^{n-2} \exp\left[-\frac{x^2+y^2}{2}\right]$$

$$\to \int_{\sqrt{n}A \sin \theta_1}^{\infty} dy\, y^{n-2} \exp\,(-y^2/2) \int_{-\infty}^{\infty} dx \exp\,(-x^2/2)$$

$$= \sqrt{2\pi} \int_{\sqrt{n}A \sin \theta_1}^{\infty} dy\, y^{n-2} \exp\,[-y^2/2] + 0(\theta_1).$$

One thus finds that if $A\theta_1 \to \infty$, $Q_n/d \to 0$ whereas if $A\theta_1 \to 0$, $Q_n \to 1$.

To maintain (16), therefore, we must have $A\theta_1 = \alpha + 0(\theta_1)$ where $0 < \alpha < \infty$, or

$$A \sim (\alpha/\theta_1). \tag{19}$$

Equations (17) and (18) now give

$$\lim_{R/W\to\infty} \frac{R/W}{\bar{r}} = \lim_{\theta_1\to 0} \frac{\frac{2}{n}\left[\ln\,(n-1)\beta\left(\frac{n-1}{2},\frac{1}{2}\right) - (n-1)\ln\sin\theta_1\right]}{\ln\left(1+\frac{\alpha^2}{\theta_1^2}\right)}$$

$$= \frac{n-1}{n}$$

as was to be shown.

The preceding considerations also allow one to show directly that the curves of Figs. 4, 5 and 6 rise indefinitely as $R/W \to \infty$. For a given R/W, denote by A_i^2 the corresponding signal-to-noise ratio obtained from the capacity formula, so that $R/W = \log\,(1 + A_i^2)$. Then $A_i^2 \sim 10^{R/W}$ as $\theta_1 \to 0$. From (18) one finds,

$$A_i^2 \sim \left[\frac{(n-1)\beta}{\sin^{n-1}\theta_1}\right]^{2/n}.$$

Using (19), one then has

$$A^2/A_i^2 \sim c/\theta_1^{2/n}$$

with c a positive constant. As $R/W \to \infty$, $\theta_1 \to 0$ and $A^2/A_i^2 \to \infty$. The logarithm of this latter ratio is plotted on Figs. 4, 5, and 6.

APPENDIX D

Each word of a block code dictionary is a sequence of n real numbers which may be regarded as a point in an n-dimensional Euclidean space. The points of an equal energy block code all lie on the surface of a hypersphere of radius $\sqrt{nS}$ with center at the origin. The words of a restricted energy block code all lie on the surface or within such a sphere. In this geometric picture, the effect of the noise in the channel can be visualized by surrounding each word of the code by a sphere of radius $\sqrt{nN}$ centered at the word. Due to the noise on the channel, a received word lies on the average at a distance $\sqrt{nN}$ from the corresponding transmitted word. If the code is to have a small average error probability, the noise spheres surrounding the words of the code must not overlap too much. On the other hand, to achieve a large rate, it is necessary to have many words in the dictionary.

The volume of a sphere of radius r in n-space is proportional to r^n. The fraction of the volume of such a sphere that lies external to the concentric sphere of radius αr, $0 < \alpha < 1$ is therefore

$$\frac{r^n - (\alpha r)^n}{r^n} = 1 - \alpha^n.$$

For large enough n, then, almost all the volume of the sphere lies near its surface. For example, if $n \geqq 460$, then at least 99 per cent of the volume of the sphere lies within a thin skin of the surface whose thickness is 1 per cent of the radius of the sphere.

Suppose now that N and S are fixed, and consider the problem of placing code words on or within the sphere of radius $\sqrt{nS}$ so that the spheres of radius $\sqrt{nN}$ surrounding each code word do not overlap appreciably. The radius of these noise spheres is a fixed fraction, $\sqrt{N/S}$, of the radius of the large sphere of radius $\sqrt{nS}$. As n becomes large, almost all of the volume of the large sphere lies within a skin of the surface of fractional thickness much less than $\sqrt{N/S}$. It is not surprising, then, that little is to be gained by placing code words interior to the large sphere. Indeed, Shannon's bounds prove that in the limit $n \to \infty$ restricted energy block codes give no better performance than equal energy block codes.

In contrast now consider the situation when n and S are fixed and R/W becomes large. As we seek to place more and more code words on or within the sphere of radius $\sqrt{nS}$, the noise power N must be continuously decreased to prevent the noise spheres surrounding the code words from overlapping. Ultimately, for large enough rates, N must be made so small that the radii of these noise spheres is very small compared to the thickness of the skin of the sphere of radius $\sqrt{nS}$ containing most of its volume. It then becomes possible to pack appreciable numbers of code words interior to this sphere and restricted energy codes then give better performance than equal energy codes.

The asymptotic behavior of the dashed curves of Fig. 11 can readily be deduced from the bound (5) and the material of Appendix C. The curves are given by

$$P_e = Q_{n+1}\left(\frac{n}{n+1}\frac{R}{W}, \frac{S}{N}\right).$$

To maintain $0 < P_e < 1$, we find as in the derivation of (19) that

$$\mathrm{A} \sim (\alpha/\theta_1)$$

where α is given by

$$P_e = \frac{n\pi^{n/2}}{(2\pi)^{n/2}\Gamma\left(\frac{n}{2}+1\right)} \sqrt{2\pi} \int_{\alpha\sqrt{n+1}}^{\infty} dy\, y^{n-1} \exp\left(-y^2/2\right)$$

$$= \frac{1}{\Gamma\left(\frac{n}{2}\right)} \int_{\frac{\alpha^2(n+1)}{2}}^{\infty} dt\, t^{(n/2)-1}\, e^{-t}\,.$$

In the right member of (18), replace n by $n + 1$; in the left member, replace R/W by $[n/(n + 1)](R/W)$. There results

$$\frac{R}{W} \sim \log \frac{\left[n\beta\left(\frac{n}{2}, \frac{1}{2}\right)\right]^{2/n}}{\sin^2 \theta_1}\,.$$

It follows then that

$$A^2/A_i^2 \sim \frac{\alpha^2}{\left[n\beta\left(\frac{n}{2}, \frac{1}{2}\right)\right]^{2/n}}$$

so that

$$10 \log \frac{A^2}{A_i^2} \sim 20 \left\{\log \alpha - \frac{1}{n} \log \left[n\beta\left(\frac{n}{2}, \frac{1}{2}\right)\right]\right\}.$$

This latter value is the horizontal asymptote for the dashed curves of Fig. 11.

REFERENCES

1. Shannon, C. E., A Mathematical Theory of Communication, B.S.T.J., **27,** July and October, 1948, pp. 379–424 and 623–657.
2. Gilbert, E. N., A Comparison of Signalling Alphabets, B.S.T.J., **31,** May, 1952, pp. 504–522.
3. Rice, S. O., Communication in the Presence of Noise — Probability of Error for Two Encoding Schemes, B.S.T.J., **29,** January, 1950, pp. 60–93.
4. Shannon, C. E., Probability of Error for Optimal Codes in a Gaussian Channel, B.S.T.J., **38,** May, 1959, pp. 611–656.
5. Slepian, D., The Threshold Effect in Modulation Systems that Expand Bandwidth, I.R.E. Trans. Inf. Theory, Vol. **IT-8,** No. 5, September, 1962, pp. 122–127.
6. Shannon, C. E., Communication in the Presence of Noise, Proc. I.R.E, **37,** January, 1949, pp. 10–21.
7. Pearson, K., *Tables of the Incomplete Beta-Function*, Cambridge University Press, 1934.
8. Wolfowitz, J., *Coding Theorems of Information Theory*, Springer-Verlag, Berlin, 1961.
9. Landau, H. J., and Pollak, H. O., Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty—III: The Dimension of the Space of Essentially Time- and Band-Limited Signals, B.S.T.J., **41,** July, 1962, pp. 1295–1336.