Communication in the Presence of Noise

CLAUDE E. SHANNON, MEMBER, IRE

Classic Paper

A method is developed for representing any communication system geometrically. Messages and the corresponding signals are points in two "function spaces," and the modulation process is a mapping of one space into the other. Using this representation, a number of results in communication theory are deduced concerning expansion and compression of bandwidth and the threshold effect. Formulas are found for the maximum rate of transmission of binary digits over a system when the signal is perturbed by various types of noise. Some of the properties of "ideal" systems which transmit at this maximum rate are discussed. The equivalent number of binary digits per second for certain information sources is calculated.

I. INTRODUCTION

A general communications system is shown schematically in Fig. 1. It consists essentially of five elements.

1) An Information Source: The source selects one message from a set of possible messages to be transmitted to the receiving terminal. The message may be of various types; for example, a sequence of letters or numbers, as in telegraphy or teletype, or a continuous function of time $f(t)$, as in radio or telephony.

2) The Transmitter: This operates on the message in some way and produces a signal suitable for transmission to the receiving point over the channel. In telephony, this operation consists of merely changing sound pressure into a proportional electrical current. In telegraphy, we have a encoding operation which produces a sequence of dots, dashes, and spaces corresponding to the letters of the message. To take a more complex example, in the case of multiplex PCM telephony the different speech functions must be sampled, compressed, quantized and encoded, and finally interleaved properly to construct the signal.

3) The Channel: This is merely the medium used to transmit the signal from the transmitting to the receiving point. It may be a pair of wires, a coaxial cable, a band of radio frequencies, etc. During transmission, or at the receiving terminal, the signal may be perturbed by noise or distortion. Noise and distortion may be differentiated on the basis that distortion is a fixed operation applied to the signal, while noise involves statistical and unpredictable

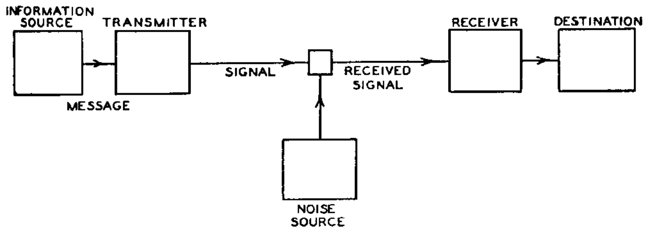

Fig. 1. General communications system.

perturbations. Distortion can, in principle, be corrected by applying the inverse operation, while a perturbation due to noise cannot always be removed, since the signal does not always undergo the same change during transmission.

4) The Receiver: This operates on the received signal and attempts to reproduce, from it, the original message. Ordinarily it will perform approximately the mathematical inverse of the operations of the transmitter, although they may differ somewhat with best design in order to combat noise.

5) The Destination: This is the person or thing for whom the message is intended.

Following Nyquist[1] and Hartley,[2] it is convenient to use a logarithmic measure of information. If a device has n possible positions it can, by definition, store $\log_b n$ units of information. The choice of the base b amounts to a choice of unit, since $\log_b n = \log_b c \log_c n$. We will use the base 2 and call the resulting units binary digits or bits. A group of m relays or flip-flop circuits has 2^m possible sets of positions, and can therefore store $\log_2 2^m = m$ bits.

If it is possible to distinguish reliably M different signal functions of duration T on a channel, we can say that the channel can transmit $\log_2 M$ bits in time T. The *rate* of transmission is then $\log_2 M/T$. More precisely, the *channel capacity* may be defined as

$$C = \lim_{T\to\infty} \frac{\log_2 M}{T}. \tag{1}$$

[1] H. Nyquist, "Certain factors affecting telegraph speed," *Bell Syst. Tech. J.*, vol. 3, p. 324, Apr. 1924.

[2] R. V. L. Hartley, "The transmission of information," *Bell Syst. Tech. J.*, vol. 3, p. 535–564, July 1928.

This paper is reprinted from the PROCEEDINGS OF THE IRE, vol. 37, no. 1, pp. 10–21, Jan. 1949.
Publisher Item Identifier S 0018-9219(98)01299-7.

0018–9219/98$10.00 © 1998 IEEE

DOCKET ALARM Find authenticated court documents without watermarks at docketalarm.com.

A precise meaning will be given later to the requirement of reliable resolution of the M signals.

II. The Sampling Theorem

Let us suppose that the channel has a certain bandwidth W in cps starting at zero frequency, and that we are allowed to use this channel for a certain period of time T. Without any further restrictions this would mean that we can use as signal functions any functions of time whose spectra lie entirely within the band W, and whose time functions lie within the interval T. Although it is not possible to fulfill both of these conditions exactly, it is possible to keep the spectrum within the band W, and to have the time function very small outside the interval T. Can we describe in a more useful way the functions which satisfy these conditions? One answer is the following.

Theorem 1: If a function $f(t)$ contains no frequencies higher than W cps, it is completely determined by giving its ordinates at a series of points spaced 1/2 W seconds apart.

This is a fact which is common knowledge in the communication art. The intuitive justification is that, if $f(t)$ contains no frequencies higher than W, it cannot change to a substantially new value in a time less than one-half cycle of the highest frequency, that is, 1/2W. A mathematical proof showing that this is not only approximately, but exactly, true can be given as follows. Let $F(\omega)$ be the spectrum of $f(t)$. Then

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{i\omega t}\,d\omega \tag{2}$$

$$= \frac{1}{2\pi}\int_{-2\pi W}^{+2\pi W} F(\omega)e^{i\omega t}\,d\omega \tag{3}$$

since $F(\omega)$ is assumed zero outside the band W. If we let

$$t = \frac{n}{2W} \tag{4}$$

where n is any positive or negative integer, we obtain

$$f\left(\frac{n}{2W}\right) = \frac{1}{2\pi}\int_{-2\pi W}^{+2\pi W} F(\omega)e^{i\omega\frac{n}{2W}}\,d\omega. \tag{5}$$

On the left are the values of $f(t)$ at the sampling points. The integral on the right will be recognized as essentially the nth coefficient in a Fourier-series expansion of the function $F(\omega)$, taking the interval $-W$ to $+W$ as a fundamental period. This means that the values of the samples $f(n/2W)$ determine the Fourier coefficients in the series expansion of $F(\omega)$. Thus they determine $F(\omega)$, since $F(\omega)$ is zero for frequencies greater than W, and for lower frequencies $F(\omega)$ is determined if its Fourier coefficients are determined. But $F(\omega)$ determines the original function $f(t)$ completely, since a function is determined if its spectrum is known. Therefore the original samples determine the function $f(t)$ completely. There is one and only one function whose spectrum is limited to a band W, and which passes through given values at sampling points separated 1.2 W seconds apart. The function can be simply reconstructed from the samples by using a pulse of the type

$$\frac{\sin 2\pi Wt}{2\pi Wt}. \tag{6}$$

This function is unity at $t = 0$ and zero at $t = n/2W$, i.e., at all other sample points. Furthermore, its spectrum is constant in the band W and zero outside. At each sample point a pulse of this type is placed whose amplitude is adjusted to equal that of the sample. The sum of these pulses is the required function, since it satisfies the conditions on the spectrum and passes through the sampled values.

Mathematically, this process can be described as follows. Let x_n be the nth sample. Then the function $f(t)$ is represented by

$$f(t) = \sum_{n=-\infty}^{\infty} x_n \frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)}. \tag{7}$$

A similar result is true if the band W does not start at zero frequency but at some higher value, and can be proved by a linear translation (corresponding physically to single-sideband modulation) of the zero-frequency case. In this case the elementary pulse is obtained from $\sin x/x$ by single-side-band modulation.

If the function is limited to the time interval T and the samples are spaced 1/2W seconds apart, there will be a total of $2TW$ samples in the interval. All samples outside will be substantially zero. To be more precise, we can define a function to be limited to the time interval T if, and only if, all the samples outside this interval are exactly zero. Then we can say that any function limited to the bandwidth W and the time interval T can be specified by giving $2TW$ numbers.

Theorem 1 has been given previously in other forms by mathematicians[3] but in spite of its evident importance seems not to have appeared explicitly in the literature of communication theory. Nyquist,[4,5] however, and more recently Gabor,[6] have pointed out that approximately $2TW$ numbers are sufficient, basing their arguments on a Fourier series expansion of the function over the time interval T. This given TW and $(TW + 1)$ cosine terms up to frequency W. The slight discrepancy is due to the fact that the functions obtained in this way will not be strictly limited to the band W but, because of the sudden starting and stopping of the sine and cosine components, contain some frequency content outside the band. Nyquist pointed out the fundamental importance of the time interval $1/2W$ seconds in connection with telegraphy, and we will call this the Nyquist interval corresponding to the band W.

[3] J. M. Whittaker, *Interpolatory Function Theory*, Cambridge Tracts in Mathematics and Mathematical Physics, no. 33. Cambridge, U.K.: Cambridge Univ. Press, ch. IV, 1935.

[4] H. Nyquist, "Certain topics in telegraph transmission theory," *AIEE Trans.*, p. 617, Apr. 1928.

[5] W. R. Bennett, "Time division multiplex systems," *Bell Syst. Tech. J.*, vol. 20, p. 199, Apr. 1941, where a result similar to Theorem 1 is established, but on a steady-state basis.

[6] D. Gabor, "Theory of communication," *J. Inst. Elect. Eng.* (London), vol. 93, pt. 3, no. 26, p. 429, 1946.

DOCKET ALARM
Find authenticated court documents without watermarks at docketalarm.com.

The $2TW$ numbers used to specify the function need not be the equally spaced samples used above. For example, the samples can be unevenly spaced, although, if there is considerable bunching, the samples must be known very accurately to give a good reconstruction of the function. The reconstruction process is also more involved with unequal spacing. One can further show that the value of the function and its derivative at every other sample point are sufficient. The value and first and second derivatives at every third sample point give a still different set of parameters which uniquely determine the function. Generally speaking, any set of $2TW$ independent numbers associated with the function can be used to describe it.

III. Geometrical Representation of the Signals

A set of three numbers x_1, x_2, x_3, regardless of their source, can always be thought of as coordinates of a point in three-dimensional space. Similarly, the $2TW$ evenly spaced samples of a signal can be thought of as coordinates of a point in a space of $2TW$ dimensions. Each particular selection of these numbers corresponds to a particular point in this space. Thus there is exactly one point corresponding to each signal in the band W and with duration T.

The number of dimensions $2TW$ will be, in general, very high. A 5-Mc television signal lasting for an hour would be represented by a point in a space with $2 \times 5 \times 10^6 \times 60^2 = 3.6 \times 10^{10}$ dimensions. Needless to say, such a space cannot be visualized. It is possible, however, to study analytically the properties of n-dimensional space. To a considerable extent, these properties are a simple generalization of the properties of two- and three-dimensional space, and can often be arrived at by inductive reasoning from these cases. The advantage of this geometrical representation of the signals is that we can use the vocabulary and the results of geometry in the communication problem. Essentially, we have replaced a complex entity (say, a television signal) in a simple environment [the signal requires only a plane for its representation as $f(t)$] by a simple entity (a point) in a complex environment ($2TW$ dimensional space).

If we imagine the $2TW$ coordinate axes to be at right angles to each other, then distances in the space have a simple interpretation. The distance from the origin to a point is analogous to the two- and three-dimensional cases

$$d = \sqrt{\sum_{n=1}^{2TW} x_n^2} \tag{8}$$

where x_n is the nth sample. Now, since

$$f(t) = \sum_{n=1}^{2TW} x_n \frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)} \tag{9}$$

we have

$$\int_{-\infty}^{\infty} f(t)^2 dt = \frac{1}{2W} \sum x_n^2 \tag{10}$$

using the fact that

$$\int_{-\infty}^{\infty} \frac{\sin \pi(2Wt - m)}{\pi(2Wt - m)} \frac{\sin \pi(2Wt - n)}{\pi(Wt - n)} dt = \begin{cases} 0 & m \neq n \\ \frac{1}{2W} & m = n. \end{cases} \tag{11}$$

Hence, the square of the distance to a point is $2W$ times the energy (more precisely, the energy into a unit resistance) of the corresponding signal

$$\begin{aligned} d^2 &= 2WE \\ &= 2WTP \end{aligned} \tag{12}$$

where P is the average power over the time T. Similarly, the distance between two points is $\sqrt{2WT}$ times the rms discrepancy between the two corresponding signals.

If we consider only signals whose average power is less than P, these will correspond to points within a sphere of radius

$$r = \sqrt{2WTP}. \tag{13}$$

If noise is added to the signal in transmission, it means that the point corresponding to the signal has been moved a certain distance in the space proportional to the rms value of the noise. Thus noise produces a small region of uncertainty about each point in the space. A fixed distortion in the channel corresponds to a warping of the space, so that each point is moved, but in a definite fixed way.

In ordinary three-dimensional space it is possible to set up many different coordinate systems. This is also possible in the signal space of $2TW$ dimensions that we are considering. A different coordinate system corresponds to a different way of describing the same signal function. The various ways of specifying a function given above are special cases of this. One other way of particular importance in communication is in terms of frequency components. The function $f(t)$ can be expanded as a sum of sines and cosines of frequencies $1/T$ apart, and the coefficients used as a different set of coordinates. It can be shown that these coordinates are all perpendicular to each other and are obtained by what is essentially a rotation of the original coordinate system.

Passing a signal through an ideal filter corresponds to projecting the corresponding point onto a certain region in the space. In fact, in the frequency-coordinate system those components lying in the pass band of the filter are retained and those outside are eliminated, so that the projection is on one of the coordinate lines, planes, or hyperplanes. Any filter performs a linear operation on the vectors of the space, producing a new vector linearly related to the old one.

IV. Geometrical Representation of Messages

We have associated a space of $2TW$ dimensions with the set of possible signals. In a similar way one can associate a space with the set of possible messages. Suppose we are considering a speech system and that the messages consist

DOCKET ALARM
Find authenticated court documents without watermarks at docketalarm.com.

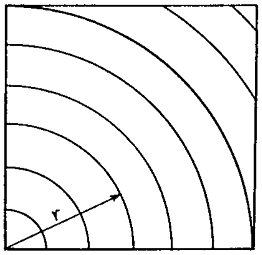

Fig. 2. Reduction of dimensionality through equivalence classes.

of all possible sounds which contain no frequencies over a certain limit W_1 and last for a time T_1.

Just as for the case of the signals, these messages can be represented in a one-to-one way in a space of $2T_1W_1$ dimensions. There are several points to be noted, however. In the first place, various different points may represent the same message, insofar as the final destination is concerned. For example, in the case of speech, the ear is insensitive to a certain amount of phase distortion. Messages differing only in the phases of their components (to a limited extent) sound the same. This may have the effect of reducing the number of essential dimensions in the message space. All the points which are equivalent for the destination can be grouped together and treated as one point. It may then require fewer numbers to specify one of these "equivalence classes" than to specify an arbitrary point. For example, in Fig. 2 we have a two-dimensional space, the set of points in a square. If all points on a circle are regarded as equivalent, it reduces to a one-dimensional space—a point can now be specified by one number, the radius of the circle. In the case of sounds, if the ear were completely insensitive to phase, then the number of dimensions would be reduced by one-half due to this cause alone. The sine and cosine components a_n and b_n for a given frequency would not need to be specified independently, but only $\sqrt{a_n^2 + b_n^2}$; that is, the total amplitude for this frequency. The reduction in frequency discrimination of the ear as frequency increases indicates that a further reduction in dimensionality occurs. The vocoder makes use to a considerable extent of these equivalences among speech sounds, in the first place by eliminating, to a large degree, phase information, and in the second place by lumping groups of frequencies together, particularly at the higher frequencies.

In other types of communication there may not be any equivalence classes of this type. The final destination is sensitive to any change in the message within the full message space of $2T_1W_1$ dimensions. This appears to be the case in television transmission.

A second point to be noted is that the information source may put certain restrictions on the actual messages. The space of $2T_1W_1$ dimensions contains a point for *every* function of time $f(t)$ limited to the band W_1 and of duration T_1. The class of messages we wish to transmit may be only a small subset of these functions. For example, speech sounds must be produced by the human vocal system. If we are willing to forego the transmission of any other sounds, the effective dimensionality may be considerably decreased. A similar effect can occur through probability considerations. Certain messages may be possible, but so improbable relative to the others that we can, in a certain sense, neglect them. In a television image, for example, successive frames are likely to be very nearly identical. There is a fair probability of a particular picture element having the same light intensity in successive frames. If this is analyzed mathematically, it results in an effective reduction of dimensionality of the message space when T_1 is large.

We will not go further into these two effects at present, but let us suppose that, when they are taken into account, the resulting message space has a dimensionality D, which will, of course, be less than or equal to $2T_1W_1$. In many cases, even though the effects are present, their utilization involves too much complication in the way of equipment. The system is then designed on the basis that all functions are different and that there are no limitations on the information source. In this case, the message space is considered to have the full $2T_1W_1$ dimensions.

V. Geometrical Representation of the Transmitter and Receiver

We now consider the function of the transmitter from this geometrical standpoint. The input to the transmitter is a message; that is, one point in the message space. Its output is a signal—one point in the signal space. Whatever form of encoding or modulation is performed, the transmitter must establish some correspondence between the points in the two spaces. Every point in the message space must correspond to a point in the signal space, and no two messages can correspond to the same signal. If they did, there would be no way to determine at the receiver which of the two messages was intended. The geometrical name for such a correspondence is a mapping. The transmitter maps the message space into the signal space.

In a similar way, the receiver maps the signal space back into the message space. Here, however, it is possible to have more than one point mapped into the same point. This means that several different signals are demodulated or decoded into the same message. In AM, for example, the phase of the carrier is lost in demodulation. Different signals which differ only in the phase of the carrier are demodulated into the same message. In FM the shape of the signal wave above the limiting value of the limiter does not affect the recovered message. In PCM considerable distortion of the received pulses is possible, with no effect on the output of the receiver.

We have so far established a correspondence between a communication system and certain geometrical ideas. The correspondence is summarized in Table 1.

VI. Mapping Considerations

It is possible to draw certain conclusions of a general nature regarding modulation methods from the geometrical picture alone. Mathematically, the simplest types of mappings are those in which the two spaces have the same

DOCKET ALARM Find authenticated court documents without watermarks at docketalarm.com.

Table 1

Communication System	*Geometrical Entity*
The set of possible signals	A space of $2TW$ dimensions
A particular signal	A point in the space
Distortion in the channel	A warping of the space
Noise in the channel	A region of uncertainty about each point
The average power of the signal	$(2TW)^{-1}$ times the square of the distance from the origin to the point
The set of signals of power P	The set of points in a sphere of radius $\sqrt{2TW\,P}$
The set of possible messages	A space of $2T_1W_1$ dimensions
The set of actual messages distinguishable by the destination	A space of D dimensions obtained by regarding all equivalent messages as one point, and deleting messages which the source could not produce
A message	A point in this space
The transmitter	A mapping of the message space into the signal space
The receiver	A mapping of the signal space into the message space

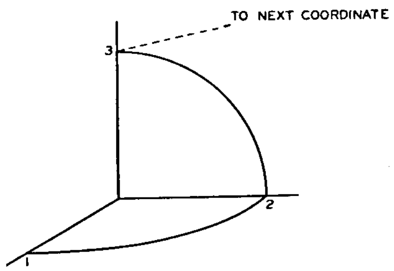

Fig. 3. Mapping similar to frequency modulation.

number of dimensions. Single-sideband amplitude modulation is an example of this type and an especially simple one, since the coordinates in the signal space are proportional to the corresponding coordinates in the message space. In double-sideband transmission the signal space has twice the number of coordinates, but they occur in pairs with equal values. If there were only one dimension in the message space and two in the signal space, it would correspond to mapping a line onto a square so that the point x on the line is represented by (x, x) in the square. Thus no significant use is made of the extra dimensions. All the messages go into a subspace having only $2T_1W_1$ dimensions.

In frequency modulation the mapping is more involved. The signal space has a much larger dimensionality than the message space. The type of mapping can be suggested by Fig. 3, where a line is mapped into a three-dimensional space. The line starts at unit distance from the origin on the first coordinate axis, stays at this distance from the origin on a circle to the next coordinate axis, and then goes to the third. It can be seen that the line is lengthened in this mapping in proportion to the total number of coordinates. It is not, however, nearly as long as it could be if it wound back and forth through the space, filling up the internal volume of the sphere it traverses.

This expansion of the line is related to the improved signal-to-noise ratio obtainable with increased bandwidth. Since the noise produces a small region of uncertainty about each point, the effect of this on the recovered message will be less if the map is in a large scale. To obtain as large a scale as possible requires that the line wander back and

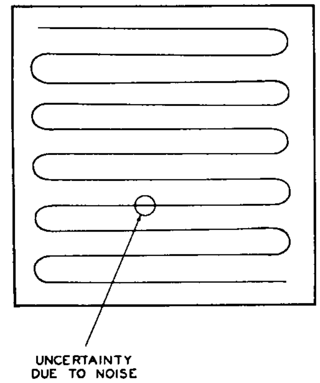

Fig. 4. Efficient mapping of a line into a square.

forth through the higher dimensional region as indicated in Fig. 4, where we have mapped a line into a square. It will be noticed that when this is done the effect of noise is small relative to the length of the line, provided the noise is less than a certain critical value. At this value it becomes uncertain at the receiver as to which portion of the line contains the message. This holds generally, and it shows that any system which attempts to use the capacities of a wider band to the full extent possible will suffer from a threshold effect when there is noise. If the noise is small, very little distortion will occur, but at some critical noise amplitude the message will become very badly distorted. This effect is well known in PCM.

Suppose, on the other hand, we wish to reduce dimensionality, i.e., to compress bandwidth or time or both. That is, we wish to send messages of band W_1 and duration T_1 over a channel with $TW < T_1W_1$. It has already been indicated that the effective dimensionality D of the message space may be less than $2T_1W_1$ due to the properties of the source and of the destination. Hence we certainly need no more than D dimension in the signal space for a good mapping. To make this saving it is necessary, of course, to isolate the effective coordinates in the message space, and to send these only. The reduced bandwidth transmission of speech by the vocoder is a case of this kind.

The question arises, however, as to whether further reduction is possible. In our geometrical analogy, is it possible to map a space of high dimensionality onto one of lower dimensionality? The answer is that it is possible, with certain reservations. For example, the points of a square can be described by their two coordinates which could be written in decimal notation

$$x = .a_1a_2a_3\cdots$$
$$y = .b_1b_2b_3\cdots. \quad (14)$$

From these two numbers we can construct one number by taking digits alternately from x and y

$$z = .a_1b_1a_2b_2a_3b_3\cdots. \quad (15)$$

A knowledge of x and y determines z, and z determines both x and y. Thus there is a one-to-one correspondence between the points of a square and the points of a line.

DOCKET ALARM
Find authenticated court documents without watermarks at docketalarm.com.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.