# AN INTRODUCTION TO INFORMATION THEORY

## Symbols, Signals & Noise

JOHN R. PIERCE
Professor of Engineering
California Institute of Technology

Second, Revised Edition

Dover Publications, Inc.
New York

extend beyond anything that we can establish with mathematical certainty.

As I have indicated earlier in this chapter, communication theory as Shannon has given it to us deals in a very broad and abstract way with certain important problems of communication and information, but it cannot be applied to all problems which we can phrase using the words *communication* and *information* in their many popular senses. Communication theory deals with certain aspects of communication which *can* be associated and organized in a useful and fruitful way, just as Newton's laws of motion deal with mechanical motion only, rather than with all the named and indeed different phenomena which Aristotle had in mind when he used the word *motion.*

To succeed, science must attempt the possible. We have no reason to believe that we can unify all the things and concepts for which we use a common word. Rather we must seek that part of experience which can be related. When we have succeeded in relating certain aspects of experience we have a theory. Newton's laws of motion are a theory which we can use in dealing with mechanical phenomena. Maxwell's equations are a theory which we can use in connection with electrical phenomena. Network theory we can use in connection with certain simple sorts of electrical *or* mechanical devices. We can use arithmetic very generally in connection with numbers of men, stones, or stars, and geometry in measuring land, sea, or galaxies.

Unlike Newton's laws of motion and Maxwell's equations, which are strongly physical in that they deal with certain classes of physical phenomena, communication theory is abstract in that it applies to many sorts of communication, written, acoustical, or electrical. Communication theory deals with certain important but abstract aspects of communication. Communication theory proceeds from clear and definite assumptions to theorems concerning information sources and communication channels. In this it is essentially mathematical, and in order to understand it we must understand the idea of a theorem as a statement which must be proved, that is, which must be shown to be the necessary consequence of a set of initial assumptions. This is an idea which is the very heart of mathematics as mathematicians understand it.

# CHAPTER II *The Origins of Information Theory*

MEN HAVE BEEN at odds concerning the value of history. Some have studied earlier times in order to find a universal system of the world, in whose inevitable unfolding we can see the future as well as the past. Others have sought in the past prescriptions for success in the present. Thus, some believe that by studying scientific discovery in another day we can learn how to make discoveries. On the other hand, one sage observed that we learn nothing from history except that we never learn anything from history, and Henry Ford asserted that history is bunk.

All of this is as far beyond me as it is beyond the scope of this book. I will, however, maintain that we can learn at least two things from the history of science.

One of these is that many of the most general and powerful discoveries of science have arisen, not through the study of phenomena as they occur in nature, but, rather, through the study of phenomena in man-made devices, in products of technology, if you will. This is because the phenomena in man's machines are simplified and ordered in comparison with those occurring naturally, and it is these simplified phenomena that man understands most easily.

Thus, the existence of the steam engine, in which phenomena involving heat, pressure, vaporization, and condensation occur in a simple and orderly fashion, gave tremendous impetus to the very powerful and general science of thermodynamics. We see this especially in the work of Carnot.[1] Our knowledge of aerodynamics and hydrodynamics exists chiefly because airplanes and ships exist, no because of the existence of birds and fishes. Our knowledge of electricity came mainly not from the study of lightning, but from the study of man's artifacts.

Similarly, we shall find the roots of Shannon's broad and elegant theory of communication in the simplified and seemingly easily intelligible phenomena of telegraphy.

The second thing that history can teach us is with what difficulty understanding is won. Today, Newton's laws of motion seem simple and almost inevitable, yet there was a day when they were undreamed of, a day when brilliant men had the oddest notions about motion. Even discoverers themselves sometimes seem incredibly dense as well as inexplicably wonderful. One might expect of Maxwell's treatise on electricity and magnetism a bold and simple pronouncement concerning the great step he had taken. Instead, it is cluttered with all sorts of such lesser matters as once seemed important, so that a naïve reader might search long to find the novel step and to restate it in the simple manner familiar to us. It is true, however, that Maxwell stated his case clearly elsewhere.

Thus, a study of the origins of scientific ideas can help us to value understanding more highly for its having been so dearly won. We can often see men of an earlier day stumbling along the edge of discovery but unable to take the final step. Sometimes we are tempted to take it for them and to say, because they stated many of the required concepts in juxtaposition, that they must really have reached the general conclusion. This, alas, is the same trap into which many an ungrateful fellow falls in his own life. When someone actually solves a problem that he merely has had ideas about, he believes that he understood the matter all along.

Properly understood, then, the origins of an idea can help to show what its real content is; what the degree of understanding was before the idea came along and how unity and clarity have been attained. But to attain such understanding we must trace the actual course of discovery, not some course which we feel discovery

should or could have taken, and we must see problems (if we can) as the men of the past saw them, not as we see them today.

In looking for the origin of communication theory one is apt to fall into an almost trackless morass. I would gladly avoid this entirely but cannot, for others continually urge their readers to enter it. I only hope that they will emerge unharmed with the help of the following grudgingly given guidance.

A particular quantity called *entropy* is used in thermodynamics and in statistical mechanics. A quantity called *entropy* is used in communication theory. After all, thermodynamics and statistical mechanics are older than communication theory. Further, in a paper published in 1929, L. Szilard, a physicist, used an idea of information in resolving a particular physical paradox. From these facts we might conclude that communication theory somehow grew out of statistical mechanics.

This easy but misleading idea has caused a great deal of confusion even among technical men. Actually, communication theory evolved from an effort to solve certain problems in the field of electrical communication. Its entropy was called entropy by mathematical analogy with the entropy of statistical mechanics. The chief relevance of this entropy is to problems quite different from those which statistical mechanics attacks.

In thermodynamics, the entropy of a body of gas depends on its temperature, volume, and mass—and on what gas it is—just as the energy of the body of gas does. If the gas is allowed to expand in a cylinder, pushing on a slowly moving piston, with no flow of heat to or from the gas, the gas will become cooler, losing some of its thermal energy. This energy appears as work done on the piston. The work may, for instance, lift a weight, which thus stores the energy lost by the gas.

This is a *reversible* process. By this we mean that if work is done in pushing the piston slowly back against the gas and so recompressing it to its original volume, the exact original energy, pressure, and temperature will be restored to the gas. In such a reversible process, the entropy of the gas remains constant, while its energy changes.

Thus, entropy is an indicator of reversibility; when there is no change of entropy, the process is reversible. In the example dis-

[1] N. L. S. Carnot (1796–1832) first proposed an ideal expansion of gas (the *Carnot cycle*) which will extract the maximum possible mechanical energy from the thermal energy of the steam.

cussed above, energy can be transferred repeatedly back and forth between thermal energy of the compressed gas and mechanical energy of a lifted weight.

Most physical phenomena are not reversible. Irreversible phenomena always involve an increase of entropy.

Imagine, for instance, that a cylinder which allows no heat flow in or out is divided into two parts by a partition, and suppose that there is gas on one side of the partition and none on the other. Imagine that the partition suddenly vanishes, so that the gas expands and fills the whole container. In this case, the thermal energy remains the same, but the entropy increases.

Before the partition vanished we could have obtained mechanical energy from the gas by letting it flow into the empty part of the cylinder through a little engine. After the removal of the partition and the subsequent increase in entropy, we cannot do this. The entropy can increase while the energy remains constant in other similar circumstances. For instance, this happens when heat flows from a hot object to a cold object. Before the temperatures were equalized, mechanical work could have been done by making use of the temperature difference. After the temperature difference has disappeared, we can no longer use it in changing part of the thermal energy into mechanical energy.

Thus, an increase in entropy means a decrease in our ability to change thermal energy, the energy of heat, into mechanical energy. An increase of entropy means a decrease of available energy.

While thermodynamics gave us the concept of entropy, it does not give a detailed physical picture of entropy, in terms of positions and velocities of molecules, for instance. *Statistical mechanics* does give a detailed mechanical meaning to entropy in particular cases. In general, the meaning is that an increase in entropy means a decrease in order. But, when we ask what order means, we must in some way equate it with knowledge. Even a very complex arrangement of molecules can scarcely be disordered if we know the position and velocity of every one. Disorder in the sense in which it is used in statistical mechanics involves unpredictability based on a lack of knowledge of the positions and velocities of molecules. Ordinarily we lack such knowledge when the arrangement of positions and velocities is "complicated."

Let us return to the example discussed above in which all the molecules of a gas are initially on one side of a partition in a cylinder. If the molecules are all on one side of the partition, and if we know this, the entropy is less than if they are distributed on both sides of the partition. Certainly, we know more about the positions of the molecules when we know that they are all on one side of the partition than if we merely know that they are somewhere within the whole container. The more detailed our knowledge is concerning a physical system, the less uncertainty we have concerning it (concerning the location of the molecules, for instance) and the less the entropy is. Conversely, more uncertainty means more entropy.

Thus, in physics, entropy is associated with the possibility of converting thermal energy into mechanical energy. If the entropy does not change during a process, the process is reversible. If the entropy increases, the available energy decreases. Statistical mechanics interprets an increase of entropy as a decrease in order or, if we wish, as a decrease in our knowledge.

The applications and details of entropy in physics are of course much broader than the examples I have given can illustrate, but I believe that I have indicated its nature and something of its importance. Let us now consider the quite different purpose and use of the entropy of communication theory.

In communication theory we consider a message source, such as a writer or a speaker, which may produce on a given occasion any one of many possible messages. The amount of information conveyed by the message increases as the amount of uncertainty as to what message actually will be produced becomes greater. A message which is one out of ten possible messages conveys a smaller amount of information than a message which is one out of a million possible messages. The entropy of communication theory is a measure of this uncertainty and the uncertainty, or entropy, is taken as the measure of the amount of information conveyed by a message from a source. The more we know about what message the source will produce, the less uncertainty, the less the entropy, and the less the information.

We see that the ideas which gave rise to the entropy of physics and the entropy of communication theory are quite different. One can be fully useful without any reference at all to the other. Nonetheless, both the entropy of statistical mechanics and that of communication theory can be described in terms of uncertainty, in similar mathematical terms. Can some significant and useful relation be established between the two different entropies and, indeed, between physics and the mathematical theory of communication?

Several physicists and mathematicians have been anxious to show that communication theory and its entropy are extremely important in connection with statistical mechanics. This is still a confused and confusing matter. The confusion is sometimes aggravated when more than one meaning of *information* creeps into a discussion. Thus, *information* is sometimes associated with the idea of *knowledge* through its popular use rather than with *uncertainty* and the resolution of uncertainty, as it is in communication theory.

We will consider the relation between communication theory and physics in Chapter X, after arriving at some understanding of communication theory. Here I will merely say that the efforts to marry communication theory and physics have been more interesting than fruitful. Certainly, such attempts have not produced important new results or understanding, as communication theory has in its own right.

Communication theory has its origins in the study of electrical communication, not in statistical mechanics, and some of the ideas important to communication theory go back to the very origins of electrical communication.

During a transatlantic voyage in 1832, Samuel F. B. Morse set to work on the first widely successful form of electrical telegraph. As Morse first worked it out, his telegraph was much more complicated than the one we know. It actually drew short and long lines on a strip of paper, and sequences of these represented, not the letters of a word, but numbers assigned to words in a dictionary or code book which Morse completed in 1837. This is (as we shall see) an efficient form of coding, but it is clumsy.

While Morse was working with Alfred Vail, the old coding was given up, and what we now know as the Morse code had been devised by 1838. In this code, letters of the alphabet are represented by spaces, dots, and dashes. The space is the absence of an electric current, the dot is an electric current of short duration, and the dash is an electric current of longer duration.

Various combinations of dots and dashes were cleverly assigned to the letters of the alphabet. E, the letter occurring most frequently in English text, was represented by the shortest possible code symbol, a single dot, and, in general, short combinations of dots and dashes were used for frequently used letters and long combinations for rarely used letters. Strangely enough, the choice was not guided by tables of the relative frequencies of various letters in English text nor were letters in text counted to get such data. Relative frequencies of occurrence of various letters were estimated by counting the number of types in the various compartments of a printer's type box!

We can ask, would some other assignment of dots, dashes, and spaces to letters than that used by Morse enable us to send English text faster by telegraph? Our modern theory tells us that we could only gain about 15 per cent in speed. Morse was very successful indeed in achieving his end, and he had the end clearly in mind. The lesson provided by Morse's code is that it matters profoundly how one translates a message into electrical signals. This matter is at the very heart of communication theory.

In 1843, Congress passed a bill appropriating money for the construction of a telegraph circuit between Washington and Baltimore. Morse started to lay the wire underground, but ran into difficulties which later plagued submarine cables even more severely. He solved his immediate problem by stringing the wire on poles.

The difficulty which Morse encountered with his underground wire remained an important problem. Different circuits which conduct a steady electric current equally well are not necessarily equally suited to electrical communication. If one sends dots and dashes too fast over an underground or undersea circuit, they are run together at the receiving end. As indicated in Figure II-1, when we send a short burst of current which turns abruptly on and off, we receive at the far end of the circuit a longer, smoothed-out rise and fall of current. This longer flow of current may overlap the current of another symbol sent, for instance, as an absence of current. Thus, as shown in Figure II-2, when a clear and distinct

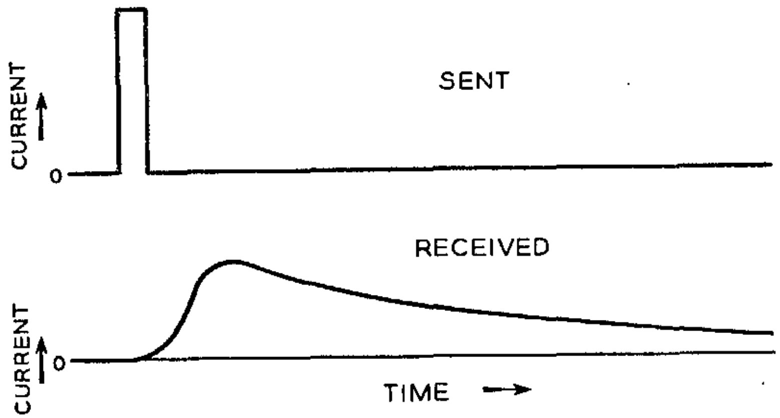


*Fig. II-1*

signal is transmitted it may be received as a vaguely wandering rise and fall of current which is difficult to interpret.

Of course, if we make our dots, spaces, and dashes long enough, the current at the far end will follow the current at the sending end better, but this slows the rate of transmission. It is clear that there is somehow associated with a given transmission circuit a limiting speed of transmission for dots and spaces. For submarine cables this speed is so slow as to trouble telegraphers; for wires on poles it is so fast as not to bother telegraphers. Early telegraphists were aware of this limitation, and it, too, lies at the heart of communication theory.

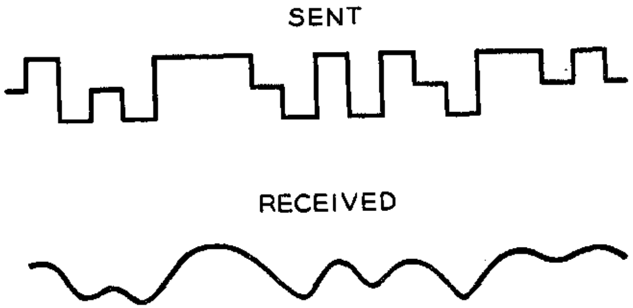


*Fig. II-2*

Even in the face of this limitation on speed, various things can be done to increase the number of letters which can be sent over a given circuit in a given period of time. A dash takes three times as long to send as a dot. It was soon appreciated that one could gain by means of double-current telegraphy. We can understand this by imagining that at the receiving end a galvanometer, a device which detects and indicates the direction of flow of small currents, is connected between the telegraph wire and the ground. To indicate a dot, the sender connects the positive terminal of his battery to the wire and the negative terminal to ground, and the needle of the galvanometer moves to the right. To send a dash, the sender connects the negative terminal of his battery to the wire and the positive terminal to the ground, and the needle of the galvanometer moves to the left. We say that an electric current in one direction (into the wire) represents a dot and an electric current in the other direction (out of the wire) represents a dash. No current at all (battery disconnected) represents a space. In actual double-current telegraphy, a different sort of receiving instrument is used.

In single-current telegraphy we have two elements out of which to construct our code: current and no current, which we might call 1 and 0. In double-current telegraphy we really have three elements, which we might characterize as forward current, or current into the wire; no current; backward current, or current out of the wire; or as $+1$, $0$, $-1$. Here the $+$ or $-$ sign indicates the direction of current flow and the number 1 gives the magnitude or strength of the current, which in this case is equal for current flow in either direction.

In 1874, Thomas Edison went further; in his quadruplex telegraph system he used two intensities of current as well as two directions of current. He used changes in intensity, regardless of changes in direction of current flow to send one message, and changes of direction of current flow regardless of changes in intensity, to send another message. If we assume the currents to differ equally one from the next, we might represent the four different conditions of current flow by means of which the two messages are conveyed over the one circuit simultaneously as $+3$, $+1$, $-1$, $-3$. The interpretation of these at the receiving end is shown in Table I.

TABLE I

| Current Transmitted | Meaning | |
|---|---|---|
| | *Message* 1 | *Message* 2 |
| +3 | on | on |
| +1 | off | on |
| −1 | off | off |
| −3 | on | off |

Figure II-3 shows how the dots, dashes, and spaces of two simultaneous, independent messages can be represented by a succession of the four different current values.

Clearly, how much information it is possible to send over a circuit depends not only on how fast one can send successive symbols (successive current values) over the circuit but also on how many different symbols (different current values) one has available to choose among. If we have as symbols only the two currents +1 or 0 or, which is just as effective, the two currents +1 and −1, we can convey to the receiver only one of two possibilities at a time. We have seen above, however, that if we can choose among any one of four current values (any one of four symbols) at a time, such as +3 or +1 or −1 or −3, we can convey by means of these current values (symbols) two independent pieces of information: whether we mean a 0 or 1 in message 1 and whether we mean a 0 or 1 in message 2. Thus, for a given rate of sending successive symbols, the use of four current values allows us to send two independent messages, each as fast as two current values allow us to send one message. We can send twice as many letters per minute by using four current values as we could using two current values.

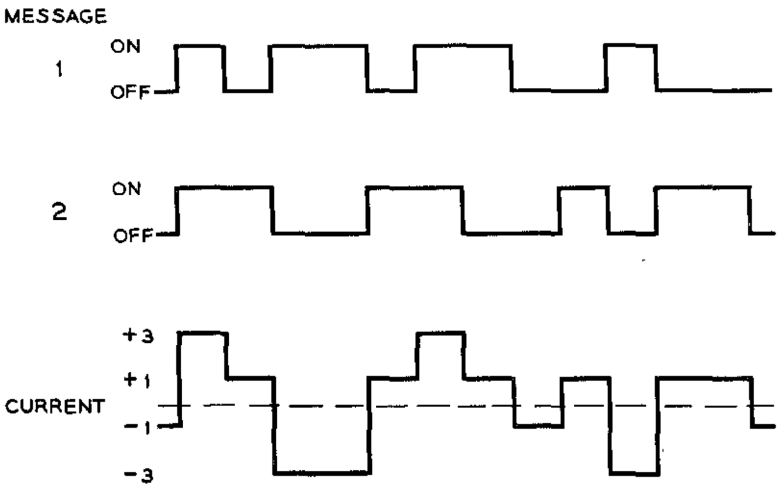


*Fig. II-3*

The use of multiplicity of symbols can lead to difficulties. We have noted that dots and dashes sent over a long submarine cable tend to spread out and overlap. Thus, when we look for one symbol at the far end we see, as Figure II-2 illustrates, a little of several others. Under these circumstances, a simple identification, as 1 or 0 or else +1 or −1, is easier and more certain than a more complicated indentification, as among +3, +1, −1, −3.

Further, other matters limit our ability to make complicated distinctions. During magnetic storms, extraneous signals appear on telegraph lines and submarine cables.[2] And if we look closely enough, as we can today with sensitive electronic amplifiers, we see that minute, undesired currents are always present. These are akin to the erratic Brownian motion of tiny particles observed under a microscope and to the agitation of air molecules and of all other matter which we associate with the idea of heat and temperature. Extraneous currents, which we call *noise*, are always present to interfere with the signals sent.

Thus, even if we avoid the overlapping of dots and spaces which is called *intersymbol interference*, noise tends to distort the received signal and to make difficult a distinction among many alternative symbols. Of course, increasing the current transmitted, which means increasing the power of the transmitted signal, helps to overcome the effect of noise. There are limits on the power that can be used, however. Driving a large current through a submarine cable takes a large voltage, and a large enough voltage can destroy the insulation of the cable—can in fact cause a short circuit. It is likely that the large transmitting voltage used caused the failure of the first transatlantic telegraph cable in 1858.

[2] The changing magnetic field of the earth induces currents in the cables. The changes in the earth's magnetic field are presumably caused by streams of charged particles due to solar storms.

Even the early telegraphists understood intuitively a good deal about the limitations associated with speed of signaling, interference, or noise, the difficulty in distinguishing among many alternative values of current, and the limitation on the power that one could use. More than an intuitive understanding was required, however. An exact mathematical analysis of such problems was needed.

Mathematics was early applied to such problems, though their complete elucidation has come only in recent years. In 1855, William Thomson, later Lord Kelvin, calculated precisely what the received current will be when a dot or space is transmitted over a submarine cable. A more powerful attack on such problems followed the invention of the telephone by Alexander Graham Bell in 1875. Telephony makes use, not of the slowly sent off-on signals of telegraphy, but rather of currents whose strength varies smoothly and subtly over a wide range of amplitudes with a rapidity several hundred times as great as encountered in manual telegraphy.

Many men helped to establish an adequate mathematical treatment of the phenomena of telephony: Henri Poincaré, the great French mathematician; Oliver Heaviside, an eccentric, English, minor genius; Michael Pupin, of *From Immigrant to Inventor* fame; and G. A. Campbell, of the American Telephone and Telegraph Company, are prominent among these.

The mathematical methods which these men used were an extension of work which the French mathematician and physicist, Joseph Fourier, had done early in the nineteenth century in connection with the flow of heat. This work had been applied to the study of vibration and was a natural tool for the analysis of the behavior of electric currents which change with time in a complicated fashion—as the electric currents of telephony and telegraphy do.

It is impossible to proceed further on our way without understanding something of Fourier's contribution, a contribution which is absolutely essential to all communication and communication theory. Fortunately, the basic ideas are simple; it is their proof and the intricacies of their application which we shall have to omit here.

Fourier based his mathematical attack on some of the problems of heat flow on a very particular mathematical function called a *sine wave.* Part of a sine wave is shown at the right of Figure II-4. The height of the wave *h* varies smoothly up and down as time passes, fluctuating so forever and ever. A sine wave has no beginning or end. A sine wave is not just any smoothly wiggling curve. The height of the wave (it may represent the strength of a current or voltage) varies in a particular way with time. We can describe this variation in terms of the motion of a crank connected to a shaft which revolves at a constant speed, as shown at the left of Figure II-4. The height *h* of the crank above the axle varies exactly sinusoidally with time.

A sine wave is a rather simple sort of variation with time. It can be characterized, or described, or differentiated completely from any other sine wave by means of just three quantities. One of these is the maximum height above zero, called the *amplitude.* Another is the time at which the maximum is reached, which is specified as the *phase.* The third is the time *T* between maxima, called the *period.* Usually, we use instead of the period the reciprocal of the period called the *frequency,* denoted by the letter *f.* If the period *T* of a sine wave is 1/100 second, the frequency *f* is 100 cycles per second, abbreviated cps. A *cycle* is a complete variation from crest, through trough, and back to crest again. The sine wave is *periodic* in that one variation from crest through trough to crest again is just like any other.

Fourier succeeded in proving a theorem concerning sine waves which astonished his, at first, incredulous contemporaries. He showed that any variation of a quantity with time can be accurately represented as the sum of a number of sinusoidal variations of

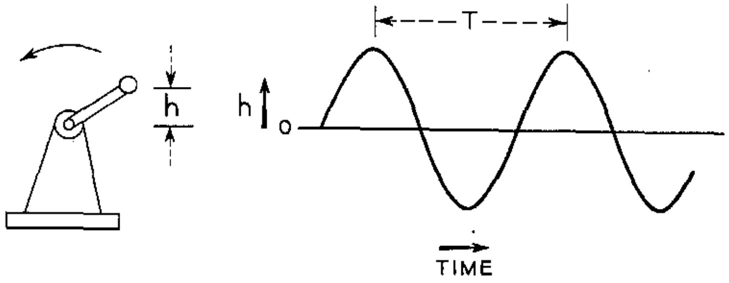


*Fig. II-4*

different amplitudes, phases, and frequencies. The quantity concerned might be the displacement of a vibrating string, the height of the surface of a rough ocean, the temperature of an electric iron, or the current or voltage in a telephone or telegraph wire. All are amenable to Fourier's analysis. Figure II-5 illustrates this in a simple case. The height of the periodic curve *a* above the centerline is the sum of the heights of the sinusoidal curves *b* and *c*.

The mere representation of a complicated variation of some physical quantity with time as a sum of a number of simple sinusoidal variations might seem a mere mathematician's trick. Its utility depends on two important physical facts. The circuits used in the transmission of electrical signals do not change with time, and they behave in what is called a *linear* fashion. Suppose, for instance, we send one signal, which we will call an *input signal,* over the line and draw a curve showing how the amplitude of the received signal varies with time. Suppose we send a second input signal and draw a curve showing how the corresponding received signal varies with time. Suppose we now send the sum of the two input signals, that is, a signal whose current is at every moment the simple sum of the currents of the two separate input signals. Then, the received output signal will be merely the sum of the two output signals corresponding to the input signals sent separately.

We can easily appreciate the fact that communication circuits don't change significantly with time. Linearity means simply that if we know the output signals corresponding to any number of input signals sent separately, we can calculate the output signal when several of the input signals are sent together merely by adding the output signals corresponding to the input signals. In a linear electrical circuit or transmission system, signals act as if they were present independently of one another; they do not interact. This is, indeed, the very criterion for a circuit being called a linear circuit.

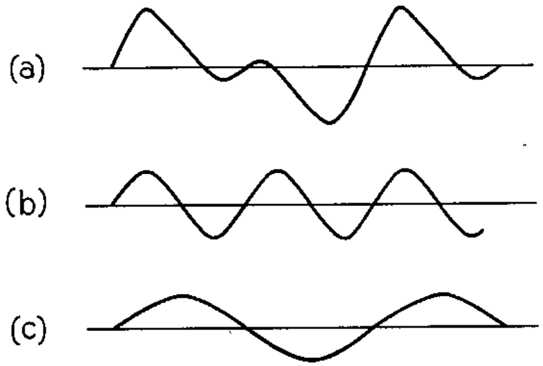


*Fig. II-5*

While linearity is a truly astonishing property of nature, it is by no means a rare one. All circuits made up of the resistors, capacitors, and inductors discussed in Chapter I in connection with network theory are linear, and so are telegraph lines and cables. Indeed, usually electrical circuits are linear, except when they include vacuum tubes, or transistors, or diodes, and sometimes even such circuits are substantially linear.

Because telegraph wires are linear, which is just to say because telegraph wires are such that electrical signals on them behave independently without interacting with one another, two telegraph signals can travel in opposite directions on the same wire at the same time without interfering with one another. However, while linearity is a fairly common phenomenon in electrical circuits, it is by no means a universal natural phenomenon. Two trains can't travel in opposite directions on the same track without interference. Presumably they could, though, if all the physical phenomena comprised in trains were linear. The reader might speculate on the unhappy lot of a truly linear race of beings.

With the very surprising property of linearity in mind, let us return to the transmission of signals over electrical circuits. We have noted that the output signal corresponding to most input signals has a different shape or variation with time from the input signal. Figures II-1 and II-2 illustrate this. However, it can be shown mathematically (but not here) that, if we use a sinusoidal signal, such as that of Figure II-4, as an input signal to a linear transmission path, we always get out a sine wave of the *same* period, or frequency. The amplitude of the output sine wave may be less than that of the input sine wave; we call this *attenuation* of the sinusoidal signal. The output sine wave may rise to a peak later than the input sine wave; we call this *phase shift,* or *delay* of the sinusoidal signal.

The amounts of the attenuation and delay depend on the frequency of the sine wave. In fact, the circuit may fail entirely to transmit sine waves of some frequencies. Thus, corresponding to an input signal made up of several sinusoidal *components,* there will be an output signal having components of the same frequencies but of different relative phases or delays and of different amplitudes. Thus, in general the shape of the output signal will be different from the shape of the input signal. However, the difference can be thought of as caused by the changes in the relative delays and amplitudes of the various components, differences associated with their different frequencies. If the attenuation and delay of a circuit is the same for all frequencies, the shape of the output wave will be the same as that of the input wave; such a circuit is *distortionless.*

Because this is a very important matter, I have illustrated it in Figure II-6. In *a* we have an input signal which can be expressed as the sum of the two sinusoidal components, *b* and *c*. In transmission, *b* is neither attenuated nor delayed, so the output *b′* of the same frequency as *b* is the same as *b*. However, the output *c′* due to the input *c* is attenuated and delayed. The total output *a′*, the sum of *b′* and *c′*, clearly has a different shape from the input *a*. Yet, the output is made up of two components having the same frequencies that are present in the input. The frequency components merely have different relative phases or delays and different relative amplitudes in the output than in the input.

The *Fourier analysis* of signals into components of various frequencies makes it possible to study the transmission properties of a linear circuit for all signals in terms of the attenuation and delay it imposes on sine waves of various frequencies as they pass through it.

Fourier analysis is a powerful tool for the analysis of transmission problems. It provided mathematicians and engineers with a bewildering variety of results which they did not at first clearly understand. Thus, early telegraphists invented all sorts of shapes and combinations of signals which were alleged to have desirable properties, but they were often inept in their mathematics and wrong in their arguments. There was much dispute concerning the efficacy of various signals in ameliorating the limitations imposed by circuit speed, intersymbol interference, noise, and limitations on transmitted power.

In 1917, Harry Nyquist came to the American Telephone and Telegraph Company immediately after receiving his Ph.D. at Yale (Ph.D.'s were considerably rarer in those days). Nyquist was a much better mathematician than most men who tackled the problems of telegraphy, and he always was a clear, original, and philosophical thinker concerning communication. He tackled the problems of telegraphy with powerful methods and with clear insight. In 1924, he published his results in an important paper, "Certain Factors Affecting Telegraph Speed."

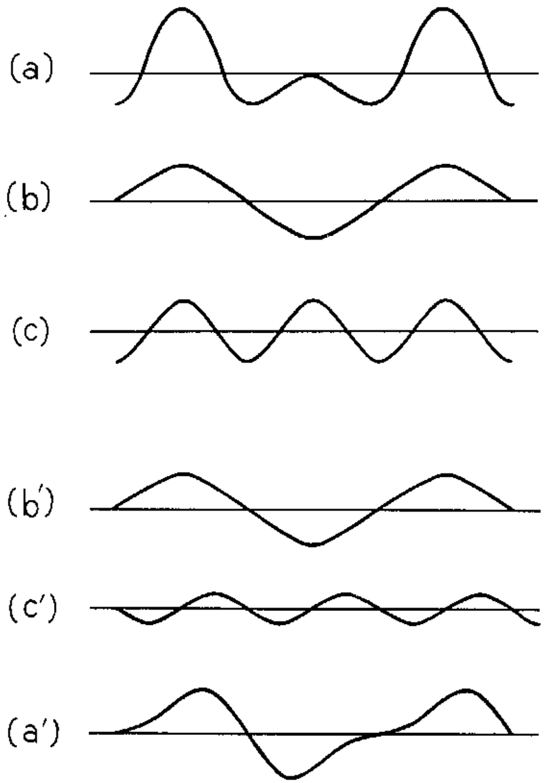


*Fig. II-6*

This paper deals with a number of problems of telegraphy. Among other things, it clarifies the relation between the speed of telegraphy and the number of current values such as $+1$, $-1$ (two current values) or $+3$, $+1$, $-1$, $-3$ (four current values). Nyquist says that if we send symbols (successive current values) at a constant rate, the speed of transmission, $W$, is related to $m$, the number of different symbols or current values available, by

$$W = \mathrm{K} \log m$$

Here K is a constant whose value depends on how many successive current values are sent each second. The quantity $\log m$ means logarithm of $m$. There are different *bases* for taking logarithms. If we choose 2 as a base, then the values of $\log m$ for various values of $m$ are given in Table II.

TABLE II

| *m* | *log m* |
|---|---|
| 1 | 0 |
| 2 | 1 |
| 3 | 1.6 |
| 4 | 2 |
| 8 | 3 |
| 16 | 4 |

To sum up the matter by means of an equation, $\log x$ is such a number that

$$2^{\log x} = x$$

We may see by taking the logarithm of each side that the following relation must be true:

$$\log 2^{\log x} = \log x$$

If we write $M$ in place of $\log x$, we see that

$$\log 2^M = M$$

All of this is consistent with Table II.

We can easily see by means of an example why the logarithm is the appropriate function in Nyquist's relation. Suppose that we wish to specify two independent choices of off-or-on, 0-or-1, simultaneously. There are four possible combinations of two independent 0-or-1 choices, as shown in Table III.

TABLE III

| *Number of Combination* | *First* 0-OR-1 *Choice* | *Second* 0-OR-1 *Choice* |
|---|---|---|
| 1 | 0 | 0 |
| 2 | 0 | 1 |
| 3 | 1 | 0 |
| 4 | 1 | 1 |

Further, if we wish to specify three independent choices of 0-or-1 at the same time, we find eight combinations, as shown in Table IV.

TABLE IV

| *Number of Combination* | *First* 0-OR-1 *Choice* | *Second* 0-OR-1 *Choice* | *Third* 0-OR-1 *Choice* |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 2 | 0 | 0 | 1 |
| 3 | 0 | 1 | 0 |
| 4 | 0 | 1 | 1 |
| 5 | 1 | 0 | 0 |
| 6 | 1 | 0 | 1 |
| 7 | 1 | 1 | 0 |
| 8 | 1 | 1 | 1 |

Similarly, if we wish to specify four independent 0-or-1 choices, we find sixteen different combinations, and, if we wish to specify $M$ different independent 0-or-1 choices, we find $2^M$ different combinations.

If we can specify $M$ independent 0-or-1 combinations at once, we can in effect send $M$ independent messages at once, so surely the speed should be proportional to $M$. But, in sending $M$ messages at once we have $2^M$ possible combinations of the $M$ independent 0-or-1 choices. Thus, to send $M$ messages at once, we need to be able to send $2^M$ different symbols or current values. Suppose that we can choose among $2^M$ different symbols. Nyquist tells us that we should take the logarithm of the number of symbols in order to get the line speed, and

$$\log 2^M = M$$

Thus, the logarithm of the number of symbols is just the number of independent 0-or-1 choices that can be represented simultaneously, the number of independent messages we can send at once, so to speak.

Nyquist's relation says that by going from off-on telegraphy to three-current $(+1, 0, -1)$ telegraphy we can increase the speed of sending letters or other symbols by 60 per cent, and if we use four current values $(+3, +1, -1, -3)$ we can double the speed. This is, of course, just what Edison did with his quadruplex telegraph, for he sent two messages instead of one. Further, Nyquist showed that the use of eight current values $(0, 1, 2, 3, 4, 5, 6, 7$, or $+7, +5, +3, +1, -1, -3, -5, -7)$ should enable us to send four times as fast as with two current values. However, he clearly realized that fluctuations in the attenuation of the circuit, interference or noise, and limitations on the power which can be used, make the use of many current values difficult.

Turning to the rate at which signal elements can be sent, Nyquist defined the *line speed* as one half of the number of signal elements (dots, spaces, current values) which can be transmitted in a second. We will find this definition particularly appropriate for reasons which Nyquist did not give in this early paper.

By the time that Nyquist wrote, it was common practice to send telegraph and telephone signals on the same wires. Telephony makes use of frequencies above 150 cps, while telegraphy can be carried out by means of lower frequency signals. Nyquist showed how telegraph signals could be so shaped as to have no sinusoidal components of high enough frequency to be heard as interference by telephones connected to the same line. He noted that the line speed, and hence also the speed of transmission, was proportional to the width or extent of the range or *band* (in the sense of strip) of frequencies used in telegraphy; we now call this range of frequencies the *band width* of a circuit or of a signal.

Finally, in analyzing one proposed sort of telegraph signal, Nyquist showed that it contained at all times a steady sinusoidal component of constant amplitude. While this component formed a part of the transmitter power used, it was useless at the receiver, for its eternal, regular fluctuations were perfectly predictable and could have been supplied at the receiver rather than transmitted thence over the circuit. Nyquist referred to this useless component of the signal, which, he said, conveyed no intelligence, as *redundant*, a word which we will encounter later.

Nyquist continued to study the problems of telegraphy, and in 1928 he published a second important paper, "Certain Topics in Telegraph Transmission Theory." In this he demonstrated a number of very important points. He showed that if one sends some number $2N$ of different current values per second, all the sinusoidal components of the signal with frequencies greater than $N$ are redundant, in the sense that they are not needed in deducing from the received signal the succession of current values which were sent. If all of these higher frequencies were removed, one could still deduce by studying the signal which current values had been transmitted. Further, he showed how a signal could be constructed which would contain no frequencies about $N$ cps and from which it would be very easy to deduce at the receiving point what current values had been sent. This second paper was more quantitative and exact than the first; together, they embrace much important material that is now embodied in communication theory.

R. V. L. Hartley, the inventor of the Hartley oscillator, was thinking philosophically about the transmission of information at about this time, and he summarized his reflections in a paper, "Transmission of Information," which he published in 1928.

Hartley had an interesting way of formulating the problem of communication, one of those ways of putting things which may seem obvious when stated but which can wait years for the insight that enables someone to make the statement. He regarded the sender of a message as equipped with a set of symbols (the letters of the alphabet for instance) from which he mentally selects symbol after symbol, thus generating a sequence of symbols. He observed that a chance event, such as the rolling of balls into pockets, might equally well generate such a sequence. He then defined $H$, the information of the message, as the logarithm of the number of possible sequences of symbols which might have been selected and showed that

$$H = n \log s$$

Here $n$ is the number of symbols selected, and $s$ is the number of different symbols in the set from which symbols are selected.

This is acceptable in the light of our present knowledge of information theory only if successive symbols are chosen independently and if any of the $s$ symbols is equally likely to be selected. In this case, we need merely note, as before, that the logarithm of $s$, the number of symbols, is the number of independent 0-or-1 choices that can be represented or sent simultaneously, and it is reasonable that the rate of transmission of information should be the rate of sending symbols per second $n$, times the number of independent 0-or-1 choices that can be conveyed per symbol.

Hartley goes on to the problem of encoding the primary symbols (letters of the alphabet, for instance) in terms of secondary symbols (e.g., the sequences of dots, spaces, and dashes of the Morse code). He observes that restrictions on the selection of symbols (the fact that E is selected more often than Z) should govern the lengths of the secondary symbols (Morse code representations) if we are to transmit messages most swiftly. As we have seen, Morse himself understood this, but Hartley stated the matter in a way which encouraged mathematical attack and inspired further work. Hartley also suggested a way of applying such considerations to continuous signals, such as telephone signals or picture signals.

Finally, Hartley stated, in accord with Nyquist, that the amount of information which can be transmitted is proportional to the band width times the time of transmission. But this makes us wonder about the number of allowable current values, which is also important to speed of transmission. How are we to enumerate them?

After the work of Nyquist and Hartley, communication theory appears to have taken a prolonged and comfortable rest. Workers busily built and studied particular communication systems. The art grew very complicated indeed during World War II. Much new understanding of particular new communication systems and devices was achieved, but no broad philosophical principles were laid down.

During the war it became important to predict from inaccurate or "noisy" radar data the courses of airplanes, so that the planes could be shot down. This raised an important question: Suppose that one has a varying electric current which represents data concerning the present position of an airplane but that there is added to it a second meaningless erratic current, that is, a noise. It may be that the frequencies most strongly present in the signal are different from the frequencies most strongly present in the noise. If this is so, it would seem desirable to pass the signal with the noise added through an electrical circuit or *filter* which attenuates the frequencies strongly present in the noise but does not attenuate very much the frequencies strongly present in the signal. Then, the resulting electric current can be passed through other circuits in an effort to estimate or predict what the value of the original signal, without noise, will be a few seconds from the present. But what sort of combination of electrical circuits will enable one best to predict from the present noisy signal the value of the true signal a few seconds in the future?

In essence, the problem is one in which we deal with not one but with a whole *ensemble* of possible signals (courses of the plane), so that we do not know in advance which signal we are dealing with. Further, we are troubled with an unpredictable noise.

This problem was solved in Russia by A. N. Kolmogoroff. In this country it was solved independently by Norbert Wiener. Wiener is a mathematician whose background ideally fitted him to deal with this sort of problem, and during the war he produced a yellow-bound document, affectionately called "the yellow peril" (because of the headaches it caused), in which he solved the difficult problem.

During and after the war another mathematician, Claude E. Shannon, interested himself in the general problem of communication. Shannon began by considering the relative advantages of many new and fanciful communication systems, and he sought some basic method of comparing their merits. In the same year (1948) that Wiener published his book, *Cybernetics,* which deals with communication and control, Shannon published in two parts a paper which is regarded as the foundation of modern communication theory.

Wiener and Shannon alike consider, not the problem of a single signal, but the problem of dealing adequately with *any* signal selected from a group or ensemble of possible signals. There was a free interchange among various workers before the publication of either Wiener's book or Shannon's paper, and similar ideas and expressions appear in both, although Shannon's interpretation appears to be unique.

Chiefly, Wiener's name has come to be associated with the field of extracting signals of a given ensemble from noise of a known type. An example of this has been given above. The enemy pilot follows a course which he choses, and our radar adds noise of natural origin to the signals which represent the position of the plane. We have a set of possible signals (possible courses of the airplane), not of our own choosing, mixed with noise, not of our own choosing, and we try to make the best estimate of the present or future value of the signal (the present or future position of the airplane) despite the noise.

Shannon's name has come to be associated with matters of so encoding messages chosen from a known ensemble that they can be transmitted accurately and swiftly in the presence of noise. As an example, we may have as a message source English text, not of our own choosing, and an electrical circuit, say, a noisy telegraph cable, not of our own choosing. But in the problem treated by Shannon, we are allowed to choose how we shall represent the message as an electrical signal—how many current values we shall allow, for instance, and how many we shall transmit per second. The problem, then, is not how to treat a signal plus noise so as to get a best estimate of the signal, but what sort of signal to send so as best to convey messages of a given type over a particular sort of noisy circuit.

This matter of efficient encoding and its consequences form the chief substance of information theory. In that an ensemble of messages is considered, the work reflects the spirit of the work of Kolmogoroff and Wiener and of the work of Morse and Hartley as well.

It would be useless to review here the content of Shannon's work, for that is what this book is about. We shall see, however, that it sheds further light on all the problems raised by Nyquist and Hartley and goes far beyond those problems.

In looking back on the origins of communication theory, two other names should perhaps be mentioned. In 1946, Dennis Gabor published an ingenious paper, "Theory of Communication." This, suggestive as it is, missed the inclusion of noise, which is at the heart of modern communication theory. Further, in 1949, W. G. Tuller published an interesting paper, "Theoretical Limits on the Rate of Transmission of Information," which in part parallels Shannon's work.

The gist of this chapter has been that the very general theory of communication which Shannon has given us grew out of the study of particular problems of electrical communication. Morse was faced with the problem of representing the letters of the alphabet by short or long pulses of current with intervening spaces of no current—that is, by the dots, dashes, and spaces of telegraphy. He wisely chose to represent common letters by short combinations of dots and dashes and uncommon letters by long combinations; this was a first step in efficient encoding of messages, a vital part of communication theory.

Ingenious inventors who followed Morse made use of different intensities and directions of current flow in order to give the sender a greater choice of signals than merely off-or-on. This made it possible to send more letters per unit time, but it made the signal more susceptible to disturbance by unwanted electrical disturbances called noise as well as by inability of circuits to transmit accurately rapid changes of current.

An evaluation of the relative advantages of many different sorts of telegraph signals was desirable. Mathematical tools were needed for such a study. One of the most important of these is Fourier analysis, which makes it possible to represent any signal as a sum of sine waves of various frequencies.

Most communication circuits are linear. This means that several signals present in the circuit do not interact or interfere. It can be shown that while even linear circuits change the shape of most signals, the effect of a linear circuit on a sine wave is merely to make it weaker and to delay its time of arrival. Hence, when a complicated signal is represented as a sum of sine waves of various frequencies, it is easy to calculate the effect of a linear circuit on each sinusoidal component separately and then to add up the weakened or attenuated sinusoidal components in order to obtain the over-all received signal.

Nyquist showed that the number of distinct, different current values which can be sent over a circuit per second is twice the total range or band width of frequencies used. Thus, the rate at which letters of text can be transmitted is proportional to band width. Nyquist and Hartley also showed that the rate at which letters of text can be transmitted is proportional to the logarithm of the number of current values used.

A complete theory of communication required other mathematical tools and new ideas. These are related to work done by Kolmogoroff and Wiener, who considered the problem of an unknown signal of a given type disturbed by the addition of noise. How does one best estimate what the signal is despite the presence of the interfering noise? Kolmogoroff and Wiener solved this problem.

The problem Shannon set himself is somewhat different. Suppose we have a message source which produces messages of a given type, such as English text. Suppose we have a noisy communication channel of specified characteristics. How can we represent or encode messages from the message source by means of electrical signals so as to attain the fastest possible transmission over the noisy channel? Indeed, how fast can we transmit a given type of message over a given channel without error? In a rough and general way, this is the problem that Shannon set himself and solved.

# CHAPTER III *A Mathematical Model*

A MATHEMATICAL THEORY which seeks to explain and to predict the events in the world about us always deals with a simplified model of the world, a mathematical model in which only things pertinent to the behavior under consideration enter.

Thus, planets are composed of various substances, solid, liquid, and gaseous, at various pressures and temperatures. The parts of their substances exposed to the rays of the sun reflect various fractions of the different colors of the light which falls upon them, so that when we observe planets we see on them various colored features. However, the mathematical astronomer in predicting the orbit of a planet about the sun need take into account only the total mass of the sun, the distance of the planet from the sun, and the speed and direction of the planet's motion at some initial instant. For a more refined calculation, the astronomer must also take into account the total mass of the planet and the motions and masses of other planets which exert gravitational forces on it.

This does not mean that astronomers are not concerned with other aspects of planets, and of stars and nebulae as well. The important point is that they need not take these other matters into consideration in computing planetary orbits. The great beauty and power of a mathematical theory or model lies in the separation of the relevant from the irrelevant, so that certain observable behavior can be related and understood without the need of comprehending the whole nature and behavior of the universe.

Mathematical models can have various degrees of accuracy or applicability. Thus, we can accurately predict the orbits of planets by regarding them as rigid bodies, despite the fact that no truly rigid body exists. On the other hand, the long-term motions of our moon can only be understood by taking into account the motion of the waters over the face of the earth, that is, the tides. Thus, in dealing very precisely with lunar motion we cannot regard the earth as a rigid body.

In a similar way, in network theory we study the electrical properties of interconnections of ideal inductors, capacitors, and resistors, which are assigned certain simple mathematical properties. The components of which the actual useful circuits in radio, TV, and telephone equipment are made only approximate the properties of the ideal inductors, capacitors, and resistors of network theory. Sometimes, the difference is trivial and can be disregarded. Sometimes it must be taken into account by more refined calculations.

Of course, a mathematical model may be a very crude or even an invalid representation of events in the real world. Thus, the self-interested, gain-motivated "economic man" of early economic theory has fallen into disfavor because the behavior of the economic man does not appear to correspond to or to usefully explain the actual behavior of our economic world and of the people in it.

In the orbits of the planets and the behavior of networks, we have examples of idealized *deterministic* systems which have the sort of predictable behavior we ordinarily expect of machines. Astronomers can compute the positions which the planets will occupy millennia in the future. Network theory tells us all the subsequent behavior of an electrical network when it is excited by a particular electrical signal.

Even the individual economic man is deterministic, for he will always act for his economic gain. But, if he at some time gambles on the honest throw of a die because the odds favor him, his economic fate becomes to a degree unpredictable, for he may lose even though the odds do favor him.

We can, however, make a mathematical model for purely chance

events, such as the drawing of some number, say three, of white or black balls from a container holding equal numbers of white and black balls. This model tells us, in fact, that after many trials we will have drawn all white about ⅛ of the time, two whites and a black about ⅜ of the time, two blacks and a white about ⅜ of the time, and all black about ⅛ of the time. It can also tell us how much of a deviation from these proportions we may reasonably expect after a given number of trials.

Our experience indicates that the behavior of actual human beings is neither as determined as that of the economic man nor as simply random as the throw of a die or as the drawing of balls from a mixture of black and white balls. It is clear, however, that a deterministic model will not get us far in the consideration of human behavior, such as human communication, while a random or statistical model might.

We all know that the actuarial tables used by insurance companies make fair predictions of the fraction of a large group of men in a given age group who will die in one year, despite the fact that we cannot predict when a particular man will die. Thus a statistical model may enable us to understand and even to make some sort of predictions concerning human behavior, even as we can predict how often, on the average, we will draw three black balls by chance from an equal mixture of white and black balls.

It might be objected that actuarial tables make predictions concerning groups of people, not predictions concerning individuals. However, experience teaches us that we can make predictions concerning the behavior of *individual* human beings as well as of groups of individuals. For instance, in counting the frequency of usage of the letter E in all English prose we will find that E constitutes about 0.13 of all the letters appearing, while W, for instance, constitutes only about 0.02 of all letters appearing. But, we also find almost the same proportions of E's and W's in the prose written by any one person. Thus, we can predict with some confidence that if you, or I, or Joe Doakes, or anyone else writes a long letter, or an article, or a book, about 0.13 of the letters he uses will be E's.

This predictability of behavior limits our freedom no more than does any other habit. We don't have to use in our writing the same fraction of E's, or of any other letter, that everyone else does. In fact, several untrammeled individuals have broken away from the common pattern. William F. Friedman, the eminent cryptanalyst and author of *The Shakesperian Cipher Examined,* has supplied me with the following examples.

Gottlob Burmann, a German poet who lived from 1737 to 1805, wrote 130 poems, including a total of 20,000 words, without once using the letter R. Further, during the last seventeen years of his life, Burmann even omitted the letter from his daily conversation.

In each of five stories published by Alonso Alcala y Herrera in Lisbon in 1641 a different vowel was suppressed. Francisco Navarrete y Ribera (1659), Fernando Jacinto de Zurita y Haro (1654), and Manuel Lorenzo de Lizarazu y Berbuizana (1654) provided other examples.

In 1939, Ernest Vincent Wright published a 267-page novel, *Gadsby,* in which no use is made of the letter E. I quote a paragraph below:

> Upon this basis I am going to show you how a bunch of bright young folks did find a champion; a man with boys and girls of his own; a man of so dominating and happy individuality that Youth is drawn to him as is a fly to a sugar bowl. It is a story about a small town. It is not a gossipy yarn; nor is it a dry, monotonous account, full of such customary "fill-ins" as "romantic moonlight casting murky shadows down a long, winding country road." Nor will it say anything about tinklings lulling distant folds; robins carolling at twilight, nor any "warm glow of lamplight" from a cabin window. No. It is an account of up-and-doing activity; a vivid portrayal of Youth as it is today; and a practical discarding of that worn-out notion that "a child don't know anything."

While such exercises of free will show that it is not impossible to break the chains of habit, we ordinarily write in a more conventional manner. When we are not going out of our way to demonstrate that we can do otherwise, we customarily use our due fraction of 0.13 E's with almost the consistency of a machine or a mathematical rule.

We cannot argue from this to the converse idea that a machine into which the same habits were built could write English text. However, Shannon has demonstrated how English words and text can be approximated by a mathematical process which could be carried out by a machine.

Suppose, for instance, that we merely produce a sequence of letters and spaces with equal probabilities. We might do this by putting equal numbers of cards marked with each letter and with the space into a hat, mixing them up, drawing a card, recording its symbol, returning it, remixing, drawing another card, and so on. This gives what Shannon calls the zero-order approximation to English text. His example, obtained by an equivalent process, goes:

1. Zero-order approximation (symbols independent and equiprobable)

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD.

Here there are far too many Zs and Ws, and not nearly enough E's and spaces. We can approach more nearly to English text by choosing letters independently of one another, but choosing E more often than W or Z. We could do this by putting many E's and few W's and Z's into the hat, mixing, and drawing out the letters. As the *probability* that a given letter is an E should be .13, out of every hundred letters we put into the hat, 13 should be E's. As the probability that a letter will be W should be .02, out of each hundred letters we put into the hat, 2 should be W's, and so on. Here is the result of an equivalent procedure, which gives what Shannon calls a first-order approximation of English text:

2. First-order approximation (symbols independent but with frequencies of English text).

OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL

In English text we almost never encounter any pair of letters beginning with Q except QU. The probability of encountering QX or QZ is essentially zero. While the probability of QU is not 0, it is so small as not to be listed in the tables I consulted. On the other hand, the probability of TH is .037, the probability of OR is .010 and the probability of WE is .006. These probabilities have the following meaning. In a stretch of text containing, say, 10,001 letters, there are 10,000 successive pairs of letters, i.e., the first and second, the second and third, and so on to the next to last and the last. Of the pairs a certain number are the letters TH. This might be 370 pairs. If we divide the total number of times we find TH, which we have assumed to be 370 times, by the total number of pairs of letters, which we have assumed to be 10,000, we get the probability that a randomly selected pair of letters in the text will be TH, that is, 370/10,000, or .037.

Diligent cryptanalysts have made tables of such *digram probabilities* for English text. To see how we might use these in constructing sequences of letters with the same digram probabilities as English text, let us assume that we use 27 hats, 26 for digrams beginning with each of the letters and one for digrams beginning with a space. We will then put a large number of digrams into the hats according to the probabilities of the digrams. Out of 1,000 digrams we would put in 37 TH's, 10 WE's, and so on.

Let us consider for a moment the meaning of these hats full of digrams in terms of the original counts which led to the evaluations of digram probabilities.

In going through the text letter by letter we will encounter every T in the text. Thus, the number of digrams *beginning* with T, all of which we put in one hat, will be the same as the number of T's. The fraction these represent of the total number of digrams counted is the probability of encountering T in the text; that is, .10. We might call this probability $p(T)$

$$p(T) = .10$$

We may note that this is also the fraction of digrams, distributed among the hats, which *end* in T as well as the fraction that *begin* with T.

Again, basing our total numbers on 1,001 letters of text, or 1,000 digrams, the number of times the digram TH is encountered is 37, and so the probability of encountering the digram TH, which we might call $p(T, H)$ is

$$p(T, H) = .037$$

Now we see that 0.10, or 100, of the digrams will begin with T and hence will be in the T hat and of these 37 will be TH. Thus, the fraction of the T digrams which are TH will be 37/100, or 0.37. Correspondingly, we say that the probability that a digram beginning with T is TH, which we might call $p_T(H)$, is

$$p_T(H) = .37$$

This is called the *conditional probability* that the letter following a T will be an H.

One can use these probabilities, which are adequately represented by the numbers of various digrams in the various hats, in the construction of text which has both the same *letter* frequencies and *digram* frequencies as does English text. To do this one draws the first digram at random from any hat and writes down its letters. He then draws a second digram from the hat indicated by the second letter of the first digram and writes down the second letter of this second digram. Then he draws a third digram from the hat indicated by the second letter of the second digram and writes down the second letter of this third digram, and so on. The space is treated just like a letter. There is a particular probability that a space will follow a particular letter (ending a "word") and a particular probability that a particular letter will follow a space (starting a new "word").

By an equivalent process, Shannon constructed what he calls a second-order approximation to English; it is:

3. Second-order approximation (digram structure as in English).

ON IE ANTSOUTINYS ARE T INCTORE ST BE S
DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE
FUSO TIZIN ANDY TOBE SEACE CTISBE

Cryptanalysts have even produced tables giving the probabilities of groups of three letters, called *trigram probabilities*. These can be used to construct what Shannon calls a third-order approximation to English. His example goes:

4. Third-order approximation (trigram structure as in English).

IN NO IST LAT WHEY CRATICT FROURE BIRS
GROCID PONDENOME OF DEMONSTURES OF THE
REPTAGIN IS REGOACTIONA OF CRE

When we examine Shannon's examples 1 through 4 we see an increasing resemblance to English text. Example 1, the zero-order approximation, has no wordlike combinations. In example 2, which takes letter frequencies into account, OCRO and NAH somewhat resemble English words. In example 3, which takes digram frequencies into account, all the "words" are pronounceable, and ON, ARE, BE, AT, and ANDY occur in English. In example 4, which takes trigram frequencies into account, we have eight English words and many English-sounding words, such as GROCID, PONDENOME, and DEMONSTURES.

G. T. Guilbaud has carried out a similar process using the statistics of Latin and has so produced a third-order approximation (one taking into account trigram frequencies) resembling Latin, which I quote below:

IBUS CENT IPITIA VETIS IPSE CUM VIVIVS
SE ACETITI DEDENTUR

The underlined words are genuine Latin words.

It is clear from such examples that by giving a machine certain statistics of a language, the probabilities of finding a particular letter or group of 1, or 2, or 3, or *n* letters, and by giving the machine an ability equivalent to picking a ball from a hat, flipping a coin, or choosing a random number, we could make the machine produce a close approximation to English text or to text in some other language. The more complete information we gave the machine, the more closely would its product resemble English or other text, both in its statistical structure and to the human eye.

If we allow the machine to choose groups of three letters on the basis of their probability, then any three-letter combination which it produces must be an English word or a part of an English word and any two letter "word" must be an English word. The machine is, however, less inhibited than a person, who ordinarily writes down only sequences of letters which do spell words. Thus, he misses ever writing down pompous PONDENOME, suspect ILONASIVE, somewhat vulgar GROCID, learned DEMONSTURES, and wacky but delightful DEAMY. Of course, a man in principle *could* write down such combinations of letters but ordinarily he doesn't.

We could cure the machine of this ability to produce un-English words by making it choose among groups of letters as long as the longest English word. But, it would be much simpler merely to supply the machine with words rather than letters and to let it produce these words according to certain probabilities.

Shannon has given an example in which words were selected independently, but with the probabilities of their occurring in English text, so that *the, and, man,* etc., occur in the same proportion as in English. This could be achieved by cutting text into words, scrambling the words in a hat, and then drawing out a succession of words. He calls this a first-order word approximation. It runs as follows:

5. First-order word approximation. Here words are chosen independently but with their appropriate frequencies.

REPRESENTING AND SPEEDILY IS AN GOOD APT
OR COME CAN DIFFERENT NATURAL HERE HE THE
A IN CAME THE TO OF TO EXPERT GRAY COME
TO FURNISHES THE LINE MESSAGE HAD BE THESE

There are no tables which give the probability of different pairs of words. However, Shannon constructed a random passage in which the probabilities of pairs of words were the same as in English text by the following expedient. He chose a first pair of words at random in a novel. He then looked through the novel for the next occurrence of the second word of the first pair and added the word which followed it in this new occurrence, and so on.

This process gave him the following second-order word approximation to English.

6. Second-order word approximation. The word transition probabilities are correct, but no further structure is included.

THE HEAD AND IN FRONTAL ATTACK ON AN
ENGLISH WRITER THAT THE CHARACTER OF THIS
POINT IS THEREFORE ANOTHER METHOD FOR THE
LETTERS THAT THE TIME OF WHO EVER TOLD THE
PROBLEM FOR AN UNEXPECTED.

We see that there are stretches of several words in this final passage which resemble and, indeed, might occur in English text.

Let us consider what we have found. In actual English text, in that text which we send by teletypewriter, for instance, particular letters occur with very nearly constant frequencies. Pairs of letters and triplets and quadruplets of letters occur with almost constant frequencies over long stretches of the text. Words and pairs of words occur with almost constant frequencies. Further, we can by means of a random mathematical process, carried out by a machine if you like, produce sequences of English words or letters exhibiting these same statistics.

Such a scheme, even if refined greatly, would not, however, produce all sequences of words that a person might utter. Carried to an extreme, it would be confined to combinations of words which *had* occurred; otherwise, there would be no statistical data available on them. Yet I may say, "The magenta typhoon whirled the farded bishop away," and this may well never have been said before.

The real rules of English text deal not with letters or words alone but with classes of words and their rules of association, that is, with grammar. Linguists and engineers who try to make machines for translating one language into another must find these rules, so that their machines can combine words to form grammatical utterances even when these exact combinations have not occurred before (and also so that the meaning of words in the text to be translated can be deduced from the context). This is a big problem. It is easy, however, to describe a "machine" which randomly produces endless, grammatical utterances of a limited sort.

Figure III-1 is a diagram of such a "machine." Each numbered box represents a *state* of the machine. Because there is only a finite number of boxes or states, this is called a *finite-state* machine.

From each box a number of arrows go to other boxes. In this particular machine, only two arrows go from each box to each of two other boxes. Also, in this case, each arrow is labeled ½. This indicates that the probability of the machine passing from, for instance, state 2 to state 3 is ½ and the probability of the machine passing from state 2 to state 4 is ½.

To make the machine run, we need a sequence of random choices, which we can obtain by flipping a coin repeatedly. We can let *heads* (*H*) mean *follow the top arrow* and *tails* (*T*), *follow the bottom arrow*. This will tell us to pass to a new state. When we do this we print out the word, words, or symbol written in that state box and flip again to get a new state.

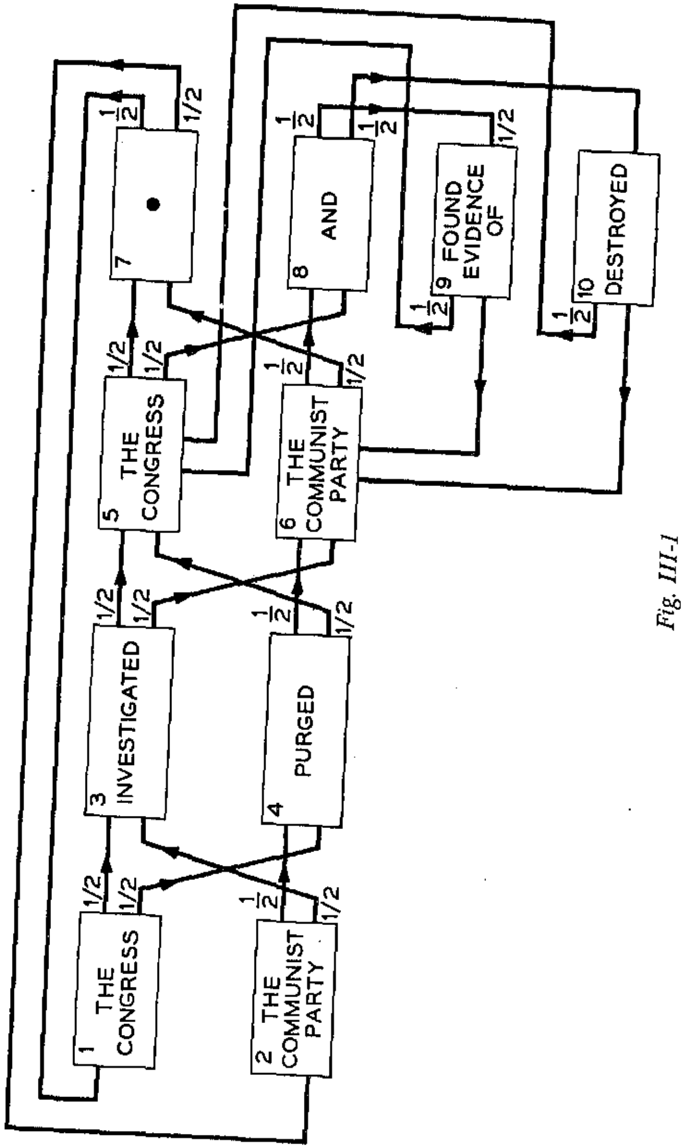


*Fig. III-1*

As an example, if we started in state 7 and flipped the following sequence of heads and tails: T H H H T T H T T T H H H H, the "machine would print out"

THE COMMUNIST PARTY INVESTIGATED THE CONGRESS. THE COMMUNIST PARTY PURGED THE CONGRESS AND DESTROYED THE COMMUNIST PARTY AND FOUND EVIDENCE OF THE CONGRESS.

This can go on and on, never retracing its whole course and producing "sentences" of unlimited length.

Random choice according to a table of probabilities of sequences of symbols (letters and space) or words can produce material resembling English text. A finite-state machine with a random choice among allowed transitions from state to state can produce material resembling English text. Either process is called a *stochastic* process, because of the random element involved in it.

We have examined a number of properties of English text. We have seen that the average frequency of E's is commonly constant for both the English text produced by one writer and, also, for the text produced by all writers. Other more complicated statistics, such as the frequency of digrams (TH, WE, and other letter pairs), are also essentially constant. Further, we have shown that English-like text can be produced by a sequence of random choices, such as drawings of slips of paper from hats, or flips of a coin, if the proper probabilities are in some way built into the process. One way of producing such text is through the use of a finite-state machine, such as that of Figure III-1.

We have been seeking a mathematical model of a source of English text. Such a model should be capable of producing text which corresponds closely to actual English text, closely enough so that the problem of encoding and transmitting such text is essentially equivalent to the problem of encoding and transmitting actual English text. The mathematical properties of the model must be mathematically defined so that useful theorems can be proved concerning the encoding and transmission of the text is produces, theorems which are applicable to a high degree of approximation to the encoding of actual English text. It would, however, be asking too much to insist that the production of actual English text conform with mathematical exactitude to the operation of the model.

The mathematical model which Shannon adopted to represent the production of text (and of spoken and visual messages as well) is the *ergodic source.* To understand what an ergodic source is, we must first understand what a *stationary source* is, and to explain this is our next order of business.

The general idea of a stationary source is well conveyed by the name. Imagine, for instance, a process, i.e., an imaginary machine, that produces forever after it is started the sequences of characters

A E A E A E A E A E, etc.

Clearly, what comes later is like what has gone before, and *stationary* seems an apt designation of such a source of characters. We might contrast this with a source of characters which, after starting, produced

A E A A E E A A A E E E, etc.

Here the strings of A's and E's get longer and longer without end; certainly this is not a stationary source.

Similarly, a sequence of characters chosen at random with some assigned probabilities (the first-order letter approximation of example 1 above) constitutes a stationary source and so do the digram and trigram sources of examples 2 and 3. The general idea of a stationary source is clear enough. An adequate mathematical definition is a little more difficult.

The idea of stationarity of a source demands no change with time. Yet, consider a digram source, in which the probability of the second character depends on what the previous character is. If we start such a source out on the letter A, several different letters can follow, while if we start such a source out on the letter Q, the second letter must be U. In general, the manner of starting the source will influence the statistics of the sequence of characters produced, at least for some distance from the start.

To get around this, the mathematician says, let us not consider just one sequence of characters produced by the source. After all, our source is an imaginary machine, and we can quite well imagine that it has been started an infinite number of times, so as to produce an infinite number of sequences of characters. Such an infinite number of sequences is called an *ensemble* of sequences.

These sequences could be started in any specified manner. Thus, in the case of a digram source, we can if we wish start a fraction, 0.13, of the sequences with E (this is just the probability of E in English text), a fraction, 0.02, with W (the probability of W), and so on. *If we do this,* we will find that the fraction of E's is the same, averaging over all the *first* letters of the ensemble of sequences, as it is averaging over all the *second* letters of the ensemble, as it is averaging over all the *third* letters of the ensemble, and so on. No matter what position from the beginning we choose, the fraction of E's or of any other letter occurring in that position, taken over all the sequences in the ensemble, is the same. This independence with respect to position will be true also for the probability with which TH or WE occurs among the first, second, third, and subsequent *pairs* of letters in the sequences of the ensemble.

This is what we mean by stationarity. If we can find a way of assigning probabilities to the various starting conditions used in forming the ensemble of sequences of characters which we allow the source to produce, probabilities such that any statistic obtained by averaging over the ensemble doesn't depend on the distance from the start at which we take an average, then the source is said to be stationary. This may seem difficult or obscure to the reader, but the difficulty arises in giving a useful and exact mathematical form to an idea which would otherwise be mathematically useless.

In the argument above we have, in discussing the infinite ensemble of sequences produced by a source, considered averaging over-all *first* characters or over-all *second* or *third* characters (or pairs, or triples of characters, as other examples). Such an average is called an *ensemble* average. It is different from a sort of average we talked about earlier in this chapter, in which we lumped together all the characters in *one* sequence and took the average over them. Such an average is called a *time* average.

The time average and the ensemble average can be different. For instance, consider a source which starts a third of the time with A and produces alternately A and B, a third of the time with B and produces alternately B and A, and a third of the time with E and produces a string of E's. The possible sequences are

1. A B A B A B A B, etc.
2. B A B A B A B A, etc.
3. E E E E E E E E, etc.

We can see that this is a stationary source, yet we have the probabilities shown in Table V.

TABLE V

| *Probability of* | *Time Average Sequence (1)* | *Time Average Sequence (2)* | *Time Average Sequence (3)* | *Ensemble Average* |
|---|---|---|---|---|
| A | ½ | ½ | 0 | ⅓ |
| B | ½ | ½ | 0 | ⅓ |
| E | 0 | 0 | 1 | ⅓ |

When a source is stationary, and when every possible ensemble average (of letters, digrams, trigrams, etc.) is equal to the corresponding time average, the source is said to be ergodic. The theorems of information theory which are discussed in subsequent chapters apply to ergodic sources, and their proofs rest on the assumption that the message source is ergodic.[1]

While we have here discussed *discrete* sources which produce sequences of characters, information theory also deals with continuous sources, which generate smoothly varying signals, such as the acoustic waves of speech or the fluctuating electric currents which correspond to these in telephony. The sources of such signals are also assumed to be ergodic.

Why is an ergodic message source an appropriate and profitable mathematical model for study? For one thing, we see by examining the definition of an ergodic source as given above that for an ergodic source the statistics of a message, for instance, the frequency of occurrence of a letter, such as E, or of a digram, such as TH, do not vary along the length of the message. As we analyze a longer and longer stretch of a message, we get a better and better estimate of the probabilities of occurrence of various letters and letter groups. In other words, by examining a longer and longer stretch of a message we are able to arrive at and refine a mathematical description of the source.

Further, the probabilities, the description of the source arrived at through such an examination of one message, apply equally well to *all* messages generated by the source and not just to the

[1] Some work has been done on the encoding of nonstationary sources, but it is not discussed in this book.

particular message examined. This is assured by the fact that the time and ensemble averages are the same.

Thus, an ergodic source is a particularly simple kind of probabilistic or stochastic source of messages, and simple processes are easier to deal with mathematically than are complicated processes. However, simplicity in itself is not enough. The ergodic source would not be of interest in communication theory if it were not reasonably realistic as well as simple.

Communication theory has two sides. It has a mathematically exact side, which deals rigorously with hypothetical, exactly ergodic sources, sources which we can imagine to produce infinite ensembles of infinite sequences of symbols. Mathematically, we are free to investigate rigorously either such a source itself or the infinite ensemble of messages which it can produce.

We *use* the theorems of communication theory in connection with the transmission of actual English text. A human being is not a hypothetical, mathematically defined machine. He cannot produce even one infinite sequence of characters, let alone an infinite ensemble of sequences.

A man does, however, produce many long sequences of characters, and all the writers of English together collectively produce a great many such long sequences of characters. In fact, part of this huge output of very long sequences of characters constitutes the messages actually sent by teletypewriter.

We will, thus, think of all the different Americans who write out telegrams in English as being, approximately at least, an ergodic source of telegraph messages and of all Americans speaking over telephones as being, approximately at least, an ergodic source of telephone signals. Clearly, however, all men writing French plus all men writing English could not constitute an ergodic source. The output of each would have certain time-average probabilities for letters, digrams, trigrams, words, and so on, but the probabilities for the English text would be different from the probabilities for the French text, and the ensemble average would resemble neither.

We will not assert that all writers of English (and all speakers of English) constitute a strictly ergodic message source. The statistics of the English we produce change somewhat as we change subject or purpose, and different people write somewhat differently. Too, in producing telephone signals by speaking, some people speak softly, some bellow, and some bellow only when they are angry. What we do assert is that we find a remarkable uniformity in many statistics of messages, as in the case of the probability of E for different samples of English text. Speech and writing as ergodic sources are not quite true to the real world, but they are far truer than is the economic man. They are true enough to be useful.

This difference between the exactly ergodic source of the mathematical theory of communication and the approximately ergodic message sources of the real world should be kept in mind. We must exercise a reasonable caution in applying the conclusions of the mathematical theory of communication to actual problems. We are used to this in other fields. For instance, mathematics tells us that we can deduce the diameter of a circle from the coordinates or locations of any three points on the circle, and this is true for absolutely exact coordinates. Yet no sensible man would try to determine the diameter of a somewhat fuzzy real circle drawn on a sheet of paper by trying to measure very exactly the positions of three points a thousandth of an inch apart on its circumference. Rather, he would draw a line through the center and measure the diameter directly as the distance between diametrically opposite points. This is just the sort of judgment and caution one must always use in applying an exact mathematical theory to an inexact practical case.

Whatever caution we invoke, the fact that we have used a random, probabilistic, stochastic process as a model of man in his role of a message source raises philosophical questions. Does this mean that we imply that man acts at random? There is no such implication. Perhaps if we knew enough about a man, his environment, and his history, we could always predict just what word he would write or speak next.

In communication theory, however, we assume that our only knowledge of the message source is obtained either from the messages that the source produces or perhaps from some less-than-complete study of man himself. On the basis of information so obtained, we can derive certain statistical data which, as we have seen, help to narrow the probability as to what the next word or letter of a message will be. There remains an element of uncertainty. For us who have incomplete knowledge of it, the message source behaves *as if* certain choices were made at random, insofar as we cannot predict what the choices will be. If we could predict them, we should incorporate the knowledge which enables us to make the predictions into our statistics of the source. If we had more knowledge, however, we might see that the choices which *we* cannot predict are not really random, in that they are (on the basis of knowledge that we do not have) predictable.

We can see that the view we have taken of finite-state machines, such as that of Figure III-1, has been limited. Finite-state machines can have inputs as well as outputs. The transition from a particular state to one among several others need not be chosen randomly; it could be determined or influenced by various inputs to the machine. For instance, the operation of an electronic digital computer, which is a finite-state machine, is determined by the program and data fed to it by the programmer.

It is, in fact, natural to think that man may be a finite-state machine, not only in his function as a message source which produces words, but in all his other behavior as well. We can think if we like of all possible conditions and configurations of the cells of the nervous system as constituting states (states of mind, perhaps). We can think of one state passing to another, sometimes with the production of a letter, word, sound, or a part thereof, and sometimes with the production of some other action or of some part of an action. We can think of sight, hearing, touch, and other senses as supplying inputs which determine or influence what state the machine passes into next. If man is a finite-state machine, the number of states must be fantastic and beyond any detailed mathematical treatment. But, so are the configurations of the molecules in a gas, and yet we can explain much of the significant behavior of a gas in terms of pressure and temperature merely.

Can we someday say valid, simple, and important things about the working of the mind in producing written text and other things as well? As we have seen, we can already predict a good deal concerning the statistical nature of what a man will write down on paper, unless he is deliberately trying to behave eccentrically, and, even then, he cannot help conforming to habits of his own.

Such broad considerations are not, of course, the real purpose or meat of this chapter. We set out to find a mathematical model adequate to represent some aspects of the human being in his role as a source of messages and adequate to represent some aspects of the messages he produces. Taking English text as an example, we noted that the frequencies of occurrence of various letters are remarkably constant, unless the writer deliberately avoids certain letters. Likewise, frequencies of occurrence of particular pairs, triplets, and so on, of letters are very nearly constant, as are frequencies of various words.

We also saw that we could generate sequences of letters with frequencies corresponding to those of English text by various random or stochastic processes, such as, cutting a lot of text into letters (or words), scrambling the bits of paper in a hat, and drawing them out one at a time. More elaborate stochastic processes, including finite-state machines, can produce an even closer approximation to English text.

Thus, we take a generalized stochastic process as a model of a message source, such as, a source producing English text. But, how must we mathematically define or limit the stochastic sources we deal with so that we can prove theorems concerning the encoding of messages generated by the sources? Of course, we must choose a definition consistent with the character of real English text.

The sort of stochastic source chosen as a model of actual message sources is the ergodic source. An ergodic source can be regarded as a hypothetical machine which produces an infinite number of or ensemble of infinite sequences of characters. Roughly, the nature or statistics of the sequences of characters or messages produced by an ergodic source do not change with time; that is, the source is stationary. Further, for an ergodic source the statistics based on one message apply equally well to all messages that the source generates.

The theorems of communication theory are proved exactly for truly ergodic sources. All writers writing English text together constitute an *approximately* ergodic source of text. The mathematical model—the truly ergodic source—is close enough to the actual situation so that the mathematics we base on it is very useful. But we must be wise and careful in applying the theorems and results of communication theory, which are exact for a mathematical ergodic source, to actual communication problems.

# CHAPTER IV *Encoding and Binary Digits*

A SOURCE OF INFORMATION may be English text, a man speaking, the sound of an orchestra, photographs, motion picture films, or scenes at which a television camera may be pointed. We have seen that in information theory such sources are regarded as having the properties of ergodic sources of letters, numbers, characters, or electrical signals. A chief aim of information theory is to study how such sequences of characters and such signals can be most effectively encoded for transmission, commonly by electrical means.

Everyone has heard of codes and the encoding of messages. Romantic spies use secret codes. Edgar Allan Poe popularized cryptography in *The Gold Bug.* The country is full of amateur cryptanalysts who delight in trying to read encoded messages that others have devised.

In this historical sense of cryptography or secret writing, codes are used to conceal the content of an important message from these for whom it is not intended. This may be done by substituting for the words of the message other words which are listed in a code book. Or, in a type of code called a cipher, letters or numbers may be substituted for the letters in the message according to some previously agreed upon secret scheme.

The idea of encoding, of the accurate representation of one thing by another, occurs in other contexts as well. Geneticists believe that the whole plan for a human body is written out in the chromosomes of the germ cell. Some assert that the "text" consists of an orderly linear arrangement of four different units, or "bases," in the DNA (desoxyribonucleic acid) forming the chromosome. This text in turn produces an equivalent text in RNA (ribonucleic acid), and by means of this RNA text proteins made up of sequences of twenty amino acids are synthesized. Some cryptanalytic effort has been spent in an effort to determine how the four-character message of RNA is reencoded into the twenty-character code of the protein.

Actually, geneticists have been led to such considerations by the existence of information theory. The study of the transmission of information has brought about a new general understanding of the problems of encoding, an understanding which is important to any sort of encoding, whether it be the encoding of cryptography or the encoding of genetic information.

We have already noted in Chapter II that English text can be encoded into the symbols of Morse code and represented by short and long pulses of current separated by short and long spaces. This is one simple form of encoding. From the point of view of information theory, the electromagnetic waves which travel from an FM transmitter to the receiver in your home are an encoding of the music which is transmitted. The electric currents in telephone circuits are an encoding of speech. And the sound waves of speech are themselves an encoding of the motions of the vocal tract which produce them.

Nature has specified the encoding of the motions of the vocal tract into the sounds of speech. The communication engineer, however, can choose the form of encoding by means of which he will represent the sounds of speech by electric currents, just as he can choose the code of dots, dashes, and spaces by means of which he represents the letters of English text in telegraphy. He wants to perform this encoding well, not poorly. To do this he must have some standard which distinguishes good encoding from bad encoding, and he must have some insight into means for achieving good encoding. We learned something of these matters in Chapter II.

It is the study of this problem, a study that might in itself seem limited, which has provided through information theory new ideas important to all encoding, whether cryptographic or genetic. These new ideas include a measure of amount of information, called *entropy*, and a unit of measurement, called the *bit.*

I would like to believe that at this point the reader is clamoring to know the meaning of "amount of information" as measured in bits, and if so I hope that this enthusiasm will carry him over a considerable amount of intervening material about the encoding of messages.

It seems to me that one can't understand and appreciate the solution to a problem unless he has some idea of what the problem is. You can't explain music meaningfully to a man who has never heard any. A story about your neighbor may be full of insight, but it would be wasted on a Hottentot. I think it is only by considering in some detail how a message can be encoded for transmission that we can come to appreciate the need for and the meaning of a measure of amount of information.

It is easiest to gain some understanding of the important problems of coding by considering simple and concrete examples. Of course, in doing this we want to learn something of broad value, and here we may foresee a difficulty.

Some important messages consist of sequences of discrete characters, such as the successive letters of English text or the successive digits of the output of an electronic computer. We have seen, however, that other messages seem inherently different.

Speech and music are variations with time of the pressure of air at the ear. This pressure we can accurately represent in telephony by the voltage of a signal traveling along a wire or by some other quantity. Such a variation of a signal with time is illustrated in *a* of Figure IV-1. Here we assume the signal to be a voltage which varies with time, as shown by the wavy line.

Information theory would be of limited value if it were not applicable to such *continuous* signals or messages as well as to discrete messages, such as English text.

In dealing with continuous signals, information theory first invokes a mathematical theorem called the *sampling theorem,* which we will use but not prove. This theorem states that a continuous signal can be represented completely by and reconstructed perfectly from a set of measurements or *samples* of its amplitude which are made at equally spaced times. The interval between such

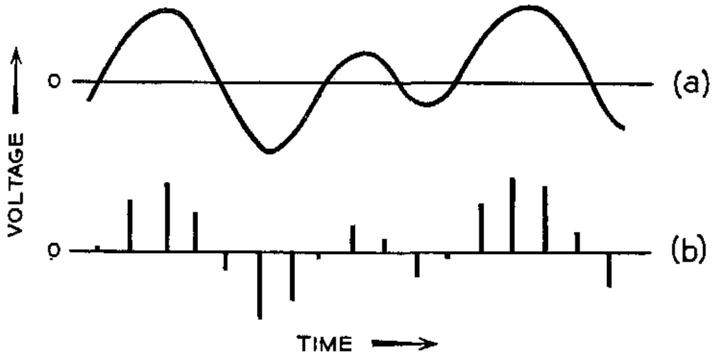


*Fig. IV-1*

samples must be equal to or less than one-half of the period of the highest frequency present in the signal. A set of such measurements or samples of the amplitude of the signal *a*, Figure IV-1, is represented by a sequence of vertical lines of various heights in *b* of Figure IV-1.

We should particularly note that for such samples of the signal to represent a signal perfectly they must be taken frequently enough. For a voice signal including frequencies from 0 to 4,000 cycles per second we must use 8,000 samples per second. For a television signal including frequencies from 0 to 4 million cycles per second we must use 8 million samples per second. In general, if the frequency range of the signal is *f* cycles per second we must use at least 2*f* samples per second in order to describe it perfectly.

Thus, the sampling theorem enables us to represent a smoothly varying signal by a sequence of samples which have different amplitudes one from another. This sequence of samples is, however, still inherently different from a sequence of letters or digits. There are only ten digits and there are only twenty-six letters, but a sample can have any of an infinite number of amplitudes. The amplitude of a sample can lie anywhere in a *continuous* range of values, while a character or a digit has only a limited number of *discrete* values.

The manner in which information theory copes with samples having a continuous range of amplitudes is a topic all in itself, to which we will return later. Here we will merely note that a signal need not be described or reproduced perfectly. Indeed, with real physical apparatus a signal *cannot* be reproduced perfectly. In the transmission of speech, for instance, it is sufficient to represent the amplitude of a sample to an accuracy of about 1 per cent. Thus, we can, if we wish, restrict ourselves to the numbers 0 to 99 in describing the amplitudes of successive speech samples and represent the amplitude of a given sample by that one of these hundred integers which is closest to the actual amplitude. By so *quantizing* the signal samples, we achieve a representation comparable to the discrete case of English text.

We can, then, by sampling and quantizing, convert the problem of coding a continuous signal, such as speech, into the seemingly simpler problem of coding a sequence of discrete characters, such as the letters of English text.

We noted in Chapter II that English text can be sent, letter by letter, by means of the Morse code. In a similar manner, such messages can be sent by teletypewriter. Pressing a particular key on the transmitting machine sends a particular sequence of electrical pulses and spaces out on the circuit. When these pulses and spaces reach the receiving machine, they activate the corresponding type bar, and the machine prints out the character that was transmitted.

Patterns of pulses and spaces indeed form a particularly useful and general way of describing or encoding messages. Although Morse code and teletypewriter codes make use of pulses and spaces of different lengths, it is possible to transmit messages by means of a sequence of pulses and spaces of equal length, transmitted at perfectly regular intervals. Figure IV-2 shows how the electric current sent out on the line varies with time for two different patterns, each six intervals long, of such equal pulses and spaces. Sequence *a* is a pulse-space-space-pulse-space-pulse. Sequence *b* is pulse-pulse-pulse-space-pulse-pulse.

The presence of a pulse or a space in a given interval specifies one of two different possibilities. We could use any pair of symbols to represent such patterns of pulses or spaces as those of Figure IV-2: yes, no; +, −; 1, 0. Thus we could represent pattern *a* as follows:

| | | | | | |
|---|---|---|---|---|---|
| *pulse* | *space* | *space* | *pulse* | *space* | *pulse* |
| *Yes* | *No* | *No* | *Yes* | *No* | *Yes* |
| + | − | − | + | − | + |
| 1 | 0 | 0 | 1 | 0 | 1 |

The representation by 1 or 0 is particularly convenient and important. It can be used to relate patterns of pulses to numbers expressed in the *binary system* of notation.

When we write 315 we mean

$$\begin{aligned} &3 \times 10^2 + 1 \times 10^1 + 5 \times 1 \\ &= 3 \times 100 + 1 \times 10 + 5 \times 1 \\ &= 315 \end{aligned}$$

In this ordinary *decimal* system of representing numbers we make use of the ten different digits: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. In the binary system we use only two digits, 0 and 1. When we write 1 0 0 1 0 1 we mean

$$\begin{aligned} &1 \times 2^5 + 0 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 0 \times 2 + 1 \times 1 \\ &= 1 \times 32 + 0 \times 16 + 0 \times 8 + 1 \times 4 + 0 \times 2 + 1 \times 1 \\ &= 37 \text{ in decimal notation} \end{aligned}$$

It is often convenient to let zeros precede a number; this does not change its value. Thus, in decimal notation we can say,

$$0016 = 16$$

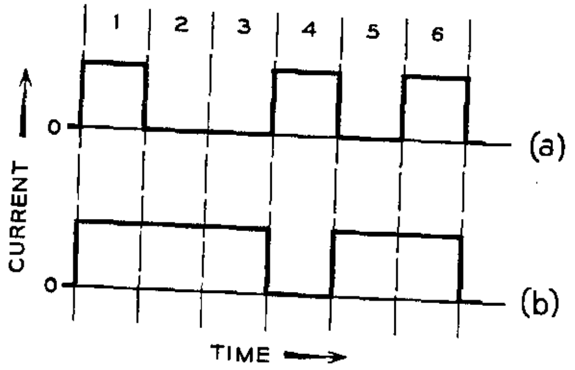


*Fig. IV-2*

Or in binary notation

$$001010 = 1010$$

In binary numbers, each 0 or 1 is a binary digit. To describe the pulses or spaces occurring in six successive intervals, we can use a sequence of six binary digits. As a pulse or space in one interval is equivalent to a binary digit, we can also refer to a pulse group of six binary digits, or we can refer to the pulse or space occurring in one interval as one binary digit.

Let us consider how many patterns of pulses and spaces there are which are three intervals long. In other words, how many three-digit binary numbers are there? These are all shown in Table VI.

Table VI

| | |
|---|---|
| 000 | (0) |
| 001 | (1) |
| 010 | (2) |
| 011 | (3) |
| 100 | (4) |
| 101 | (5) |
| 110 | (6) |
| 111 | (7) |

The decimal numbers corresponding to these sequences of 1's and 0's regarded as binary numbers are shown in parentheses to the right.

We see that there are 8 (0 and 1 through 7) three-digit binary numbers. We may note that 8 is $2^3$. We can, in fact, regard an orderly listing of binary digits $n$ intervals long as simply setting down $2^n$ successive binary numbers, starting with 0. As examples, in Table VII the numbers of different patterns corresponding to different numbers $n$ of binary digits are tabulated.

We see that the number of different patterns increases very rapidly with the number of binary digits. This is because we double the number of possible patterns each time we add one digit. When we add one digit, we get all the old sequences preceded by a 0 plus all the old sequences preceded by a 1.

Table VII

| *n (Number of Binary Digits)* | *Number of Patterns ($2^n$)* |
|---|---|
| 1 | 2 |
| 2 | 4 |
| 3 | 8 |
| 4 | 16 |
| 5 | 32 |
| 10 | 1,024 |
| 20 | 1,048,576 |

The binary system of notation is not the only alternative to the decimal system. The octal system is very important to people who use computers. We can regard the octal system as made up of the eight digits 0, 1, 2, 3, 4, 5, 6, 7.

When we write 356 in the octal system we mean

$$\begin{aligned} &3 \times 8^2 + 5 \times 8 + 6 \times 1 \\ &= 3 \times 64 + 5 \times 8 + 6 \times 1 \\ &= 238 \text{ in decimal notation} \end{aligned}$$

We can convert back and forth between the octal and the binary systems very simply. We need merely replace each successive block of three binary digits by the appropriate octal digit, as, for instance,

| | | | | |
|---|---|---|---|---|
| *binary* | 0 1 0 | 1 1 1 | 0 1 1 | 1 1 0 |
| *octal* | 2 | 7 | 3 | 6 |

People who work with binary notation in connection with computers find it easier to remember and transcribe a short sequence of octal digits than a long group of binary digits. They learn to regard patterns of three successive binary digits as an entity, so that they will think of a sequence of twelve binary digits as a succession of four patterns of three, that is, as a sequence of four octal digits.

It is interesting to note, too, that, just as a pattern of pulses and spaces can correspond to a sequence of binary digits, so a sequence of pulses of various amplitudes (0, 1, 2, 3, 4, 5, 6, 7) can correspond to a sequence of octal digits. This is illustrated in Figure IV-3. In *a,* we have the sequence of off-on, 0-1 pulses corresponding to the binary number 010111011110. The corresponding octal number is 2736, and in *b* this is represented by a sequence of four pulses of current having amplitudes 2, 7, 3, 6.

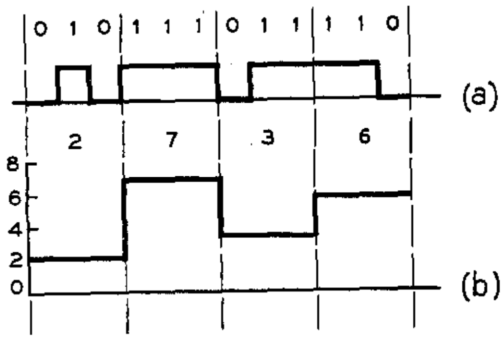


*Fig. IV-3*

Conversion from binary to decimal numbers is not so easy. On the average, it takes about 3.32 binary digits to represent one decimal digit. Of course we can assign four binary digits to each decimal digit, as shown in Table VIII, but this means that some patterns are wasted; there are more patterns than we use.

It is convenient to think of sequences of 0's and 1's or sequences of pulses and spaces as binary numbers. This helps us to understand how many sequences of a different length there are and how numbers written in the binary system correspond to numbers written in the octal or in the decimal system. In the transmission of information, however, the particular number assigned to a sequence of binary digits is irrelevent. For instance, if we wish merely to *transmit* representations of octal digits, we could make the assignments shown in Table IX rather than those in Table VI.

Table VIII

| *Binary Number* | *Decimal Digit* |
|---|---|
| 0000 | 0 |
| 0001 | 1 |
| 0010 | 2 |
| 0011 | 3 |
| 0100 | 4 |
| 0101 | 5 |
| 0110 | 6 |
| 0111 | 7 |
| 1000 | 8 |
| 1001 | 9 |
| 1010 | not used |
| 1011 | not used |
| 1100 | not used |
| 1101 | not used |
| 1110 | not used |
| 1111 | not used |

Table IX

| *Sequence of Binary Digits* | *Octal Digit Represented* |
|---|---|
| 000 | 5 |
| 001 | 7 |
| 010 | 1 |
| 011 | 6 |
| 100 | 0 |
| 101 | 4 |
| 110 | 2 |
| 111 | 3 |

Here the "binary numbers" in the left column designate octal numbers of different numerical value.

In fact, there is another way of looking at such a correspondence between binary digits and other symbols, such as octal digits, a way in which we do not regard the sequence of binary digits as part of a binary number but rather as means of choosing or designating a particular symbol.

We can regard each 0 or 1 as expressing an elementary choice between two possibilities. Consider, for instance, the "tree of choice" shown in Figure IV-4. As we proceed upward from the root to the twigs, let 0 signify that we take the left branch and let 1 signify that we take the right branch. Then 0 1 1 means left, right, right and takes us to the octal digit 6, just as in Table IX.

Just as three binary digits give us enough information to determine one among eight alternatives, four binary digits can determine one among sixteen alternatives, and twenty binary digits can determine one among 1,048,576 alternatives. We can do this by assigning the required binary numbers to the alternatives in any order we wish.

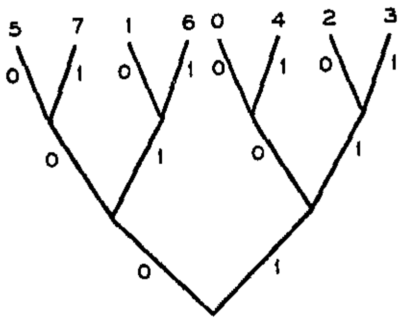


*Fig. IV-4*

The alternatives which we wish to specify by successions of binary digits need not of course be numbers at all. In fact, we began by considering how we might encode English text so as to transmit it electrically by sequences of pulses and spaces, which can be represented by sequences of binary digits.

A bare essential in transmitting English text letter by letter is twenty-six letters plus a space, or twenty-seven symbols in all. This of course allows us no punctuation and no Arabic numbers. We can write out the numbers (three, not 3) if we wish and use words for punctuation, (stop, comma, colon, etc.).

Mathematics says that a choice among 27 symbols corresponds to about 4.75 binary digits. If we are not too concerned with efficiency, we can assign a different 5-digit binary number to each character, which will leave five 5-digit binary numbers unused.

My typewriter has 48 keys, including shift and shift lock. We might add two more "symbols" representing carriage return and line advance, making a total of 50. I could encode my actions in typing, capitalization, punctuation, and all (but not insertion of the paper) by a succession of choices among 50 symbols, each choice corresponding to about 5.62 binary digits. We could use 6 binary digits per character and waste some sequences of binary digits.

This waste arises because there are only thirty-two 5-digit binary numbers, which is too few, while there are sixty-four 6-digit binary numbers, which is too many. How can we avoid this waste? If we have 50 characters, we have 125,000 possible different groups of 3 ordered characters. There are 131,072 different combinations of 17 binary digits. Thus, if we divide our text into *blocks* of 3 successive characters, we can specify any possible block by a 17-digit binary number and have a few left over. If we had represented each separate character by 6 binary digits, we would have needed 18 binary digits to represent 3 successive characters. Thus, by this *block coding,* we have cut down the number of binary digits we use in encoding a given length of text by a factor 17/18.

Of course, we might encode English text in quite a different way. We can say a good deal with 16,384 English words. That's quite a large vocabulary. There are just 16,384 fourteen-digit binary numbers. We might assign 16,357 of these to different useful words and 27 to the letters of the alphabet and the space, so that we could spell out any word or sequence of words we failed to include in our word vocabulary. We won't need to put a space between words to which numbers have been assigned; it can be assumed that a space goes with each word.

If we have to spell out words very infrequently, we will use about 14 binary digits per word in this sort of encoding. In ordinary English text there are on the average about 4.5 letters per word. As we must separate words by a space, when we send the message character by character, even if we disregard capitalization and punctuation, we will require on the average 5.5 characters per word. If we encode these using 5 binary digits per character, we will use on the average 27.5 binary digits per word, while in encoding the message word by word we need only 14 binary digits per word.

How can this be so? It is because, in spelling out the message letter by letter, we have provided means for sending with equal facility all sequences of English letters, while, in sending word by word, we restrict ourselves to English words.

Clearly, the average number of binary digits per word required to represent English text depends strongly on how we encode the text.

Now, English text is just one sort of message we might want to transmit. Other messages might be strings of numbers, the human voice, a motion picture, or a photograph. If there are efficient and inefficient ways of encoding English text, we may expect that there will be efficient and inefficient ways of encoding other signals as well.

Indeed, we may be led to believe that there exists in principle some *best* way of encoding the signals from a given message source, a way which will on the average require fewer binary digits per character or per unit time than any other way.

If there is such a best way of encoding a signal, then we might use the average number of binary digits required to encode the signal as a measure of the amount of information per character or the amount of information per second of the message source which produced the signal.

This is just what is done in information theory. How it is done and further reasons for so doing will be considered in the next chapter.

Let us first, however, review very briefly what we have covered in this chapter. In communication theory, we regard coding very broadly, as representing one signal by another. Thus a radio wave can represent the sounds of speech and so form an encoding of these sounds. Encoding is, however, most simply explained and explored in the case of discrete message sources, which produce messages consisting of sequences of characters or numbers. Fortunately, we can represent a continuous signal, such as the current in a telephone line, by a number of samples of its amplitude, using, each second, twice as many samples as the highest frequency present in the signal. Further we can if we wish represent the amplitude of each of these samples approximately by a whole number.

The representation of letters or numbers by sequences of off-or-on signals, which can in turn be represented directly by sequences of the binary digits 0 and 1, is of particular interest in communication theory. For instance, by using sequences of 4 binary digits we can form 16 binary numbers, and we can use 10 of these to represent the 10 decimal digits. Or, by using sequences of 5 binary digits we can form 32 binary numbers, and we can use 27 of these to represent the letters of the English alphabet plus the space. Thus, we can transmit decimal numbers or English text by sending sequences of off-or-on signals.

We should note that while it may be convenient to regard the sequences of binary digits so used as binary numbers, the numerical value of the binary number has no particular significance; we can choose any binary number to represent a particular decimal digit.

If we use 10 of the 16 possible 5-digit binary numbers to encode the 10 decimal digits, we never use (we waste) 6 binary numbers. We could, but never do, transmit these sequences as sequences of off-or-on signals. We can avoid such waste by means of block coding, in which we encode sequences of 2, 3, or more decimal digits or other characters by means of binary digits. For instance, all sequences of 3 decimal digits can be represented by 10 binary digits, while it takes a total of 12 binary digits to represent separately each of 3 decimal digits.

Any sequence of decimal digits may occur, but only certain sequences of English letters ever occur, that is, the words of the English language. Thus, it is more efficient to encode English words as sequences of binary digits rather than to encode the letters of the words individually. This again emphasizes the gain to be made by encoding sequences of characters, rather than encoding each character separately.

All of this leads us to the idea that there may be a best way of encoding the messages from a message source, a way which calls for the least number of binary digits.

# CHAPTER V Entropy

IN THE LAST CHAPTER, we have considered various ways in which messages can be encoded for transmission. Indeed, all communication involves some sort of encoding of messages. In the electrical case, letters may be encoded in terms of dots or dashes of electric current or in terms of several different strengths of current and directions of current flow, as in Edison's quadruplex telegraph. Or we can encode a message in the binary language of zeros and ones and transmit it electrically as a sequence of pulses or absences of pulses.

Indeed, we have shown that by periodically sampling a continuous signal such as a speech wave and by representing the amplitudes of each sample approximately by the nearest of a set of discrete values, we can represent or encode even such a continuous wave as a sequence of binary digits.

We have also seen that the number of digits required in encoding a given message depends on how it is encoded. Thus, it takes fewer binary digits per character when we encode a group or block of English letters than when we encode the letters one at a time. More important, because only a few combinations of letters form words, it takes considerably fewer digits to encode English text word by word than it does to encode the same text letter by letter.

Surely, there are still other ways of encoding the messages produced by a particular ergodic source, such as a source of English text. How many binary digits per letter or per word are *really* needed? Must we try all possible sorts of encoding in order to find

out? But, if we did try all forms of encoding we could think of, we would still not be sure we had found the best form of encoding, for the best form might be one which had not occurred to us.

Is there not, in principle at least, some statistical measurement we can make on the messages produced by the source, a measure which will tell us the minimum average number of binary digits per symbol which will serve to encode the messages produced by the source?

In considering this matter, let us return to the model of a message source which we discussed in Chapter III. There we regarded the message source as an ergodic source of symbols, such as letters or words. Such an ergodic source has certain unvarying statistical properties: the relative frequencies of symbols; the probability that one symbol will follow a particular other symbol, or pair of symbols, or triplet of symbols; and so on.

In the case of English text, we can speak in the same terms of the relative frequencies of words and of the probability that one word will follow a particular word or a particular pair, triplet, or other combination of words.

In illustrating the statistical properties of sequences of letters or words, we showed how material resembling English text can be produced by a sequence of random choices among letters and words, provided that the letters or words are chosen with due regard for their probabilities or their probabilities of following a preceding sequence of letters or words. In these examples, the throw of a die or the picking of a letter out of a hat can serve to "choose" the next symbol.

In writing or speaking, we exercise a similar choice as to what we shall set down or say next. Sometimes we have no choice; Q must be followed by U. We have more choice as to the next symbol in beginning a word than in the middle of a word. However, in any message source, living or mechanical, choice is continually exercised. Otherwise, the messages produced by the source would be predetermined and completely predictable.

Corresponding to the choice exercised by the message source in producing the message, there is an uncertainty on the part of the recipient of the message. This uncertainty is resolved when the recipient examines the message. It is this resolution of uncertainty which is the aim and outcome of communication.

If the message source involved no choice, if, for instance, it could produce only an endless string of ones or an endless string of zeros, the recipient would not need to receive or examine the message to know what it was; he could predict it in advance. Thus, if we are to measure information in a rational way, we must have a measure that increases with the amount of choice of the source and, thus, with the uncertainty of the recipient as to what message the source may produce and transmit.

Certainly, for any message source there are more long messages than there are short messages. For instance, there are 2 possible messages consisting of 1 binary digit, 4 consisting of 2 binary digits, 16 consisting of 4 binary digits, 256 consisting of 8 binary digits, and so on. Should we perhaps say that amount of information should be measured by the number of such messages? Let us consider the case of four telegraph lines used simultaneously in transmitting binary digits between two points, all operating at the same speed. Using the four lines, we can send 4 times as many digits in a given period of time as we could using one line. It also seems reasonable that we should be able to send 4 times as much information by using four lines. If this is so, we should measure information in terms of the number of binary digits rather than in terms of the number of different messages that the binary digits can form. This would mean that amount of information should be measured, not by the number of possible messages, but by the logarithm of this number.

The measure of amount of information which communication theory provides does this and is reasonable in other ways as well. This measure of amount of information is called *entropy*. If we want to understand this entropy of communication theory, it is best first to clear our minds of any ideas associated with the entropy of physics. Once we understand entropy as it is used in communication theory thoroughly, there is no harm in trying to relate it to the entropy of physics, but the literature indicates that some workers have never recovered from the confusion engendered by an early admixture of ideas concerning the entropies of physics and communication theory.

The entropy of communication theory is measured in *bits*. We may say that the entropy of a message source is so many bits per letter, or per word, or per message. If the source produces symbols at a constant rate, we can say that the source has an entropy of so many bits per second.

Entropy increases as the number of messages among which the source may choose increases. It also increases as the freedom of choice (or the uncertainty to the recipient) increases and decreases as the freedom of choice and the uncertainty are restricted. For instance, a restriction that certain messages must be sent either very frequently or very infrequently decreases choice at the source and uncertainty for the recipient, and thus such a restriction must decrease entropy.

It is best to illustrate entropy first in a simple case. The mathematical theory of communication treats the message source as an ergodic process, a process which produces a string of symbols that are to a degree unpredictable. We must imagine the message source as selecting a given message by some random, i.e., unpredictable means, which, however, must be ergodic. Perhaps the simplest case we can imagine is that in which there are only two possible symbols, say, $X$ and $Y$, between which the message source chooses repeatedly, each choice uninfluenced by any previous choices. In this case we can know only that $X$ will be chosen with some probability $p_0$ and $Y$ with some probability $p_1$, as in the outcomes of the toss of a biased coin. The recipient can determine these probabilities by examining a long string of characters ($X$'s, $Y$'s) produced by the source. The probabilities $p_0$ and $p_1$ must not change with time if the source is to be ergodic.

For this simplest of cases, the entropy $H$ of the message source is defined as

$$H = -(p_0 \log p_0 + p_1 \log p_1) \quad \text{bits per symbol}$$

Thus, the entropy is the negative of the sum of the probability $p_0$ that $X$ will be chosen (or will be received) times the logarithm of $p_0$ and the probability $p_1$ that $Y$ will be chosen (or will be received) times the logarithm of this probability.

Whatever plausible arguments one may give for the use of entropy as defined in this and in more complicated cases, the real and true reason is one that will become apparent only as we proceed, and the justification of this formula for entropy will therefore be deferred. It is, however, well to note again that there are different kinds of logarithms and that, in information theory, we use logarithms to the base 2. Some facts about logarithms to the base 2 are noted in Table X.

TABLE X

| *Fraction p* | *Another Way of Writing p* | *Still Another Way of Writing p* | *Log p* |
|---|---|---|---|
| $\frac{3}{4}$ | $\frac{1}{2^{.415}}$ | $2^{-.415}$ | $-.415$ |
| $\frac{1}{2}$ | $\frac{1}{2^{1}}$ | $2^{-1}$ | $-1$ |
| $\frac{3}{8}$ | $\frac{1}{2^{1.415}}$ | $2^{-1.415}$ | $-1.415$ |
| $\frac{1}{4}$ | $\frac{1}{2^{2}}$ | $2^{-2}$ | $-2$ |
| $\frac{1}{8}$ | $\frac{1}{2^{3}}$ | $2^{-3}$ | $-3$ |
| $\frac{1}{16}$ | $\frac{1}{2^{4}}$ | $2^{-4}$ | $-4$ |
| $\frac{1}{64}$ | $\frac{1}{2^{6}}$ | $2^{-6}$ | $-6$ |
| $\frac{1}{256}$ | $\frac{1}{2^{8}}$ | $2^{-8}$ | $-8$ |

The logarithm to the base 2 of a number is the power to which 2 must be raised to give the number.

Let us consider, for instance, a "message source" which consists of the tossing of an honest coin. We can let $X$ represent heads and $Y$ represent tails. The probability $p_1$ that the coin will turn up heads is ½ and the probability $p_0$ that the coin will turn up tails is also ½. Accordingly, from our expression for entropy and from Table X we find that

$$H = -(\tfrac{1}{2} \log \tfrac{1}{2} + \tfrac{1}{2} \log \tfrac{1}{2})$$
$$H = -[(\tfrac{1}{2})(-1) + (\tfrac{1}{2})(-1)]$$
$$H = 1 \text{ bit per toss}$$

If the message source is the sequence of heads and tails obtained by tossing a coin, it takes one bit of information to convey whether heads or tails has turned up.

Let us notice, now, that we can represent the outcome of successively tossing a coin by a number of binary digits equal to the number of tosses, letting 1 stand for heads and 0 stand for tails. Hence, in this case at least, the entropy, one bit per toss, and the number of binary digits which can represent the outcome, one binary digit per toss, are equal. In this case at least, the number of binary digits necessary to transmit the message generated by the source (the succession of heads and tails) is equal to the entropy of the source.

Suppose the message source produces a string of 1's and 0's by tossing a coin so weighted that it turns up heads ¾ of the time and tails only ¼ of the time. Then

$$p_1 = \tfrac{3}{4}$$
$$p_0 = \tfrac{1}{4}$$
$$H = -(\tfrac{1}{4} \log \tfrac{1}{4} + \tfrac{3}{4} \log \tfrac{3}{4})$$
$$H = -[(\tfrac{1}{4})(-2) + (\tfrac{3}{4})(-.415)]$$
$$H = .811 \text{ bit per toss}$$

We feel that, in the case of a coin which turns up heads more often than tails, we know more about the outcome than if heads or tails were equally likely. Further, if we were constrained to choose heads more often than tails we would have less choice than if we could choose either with equal probability. We feel that this must be so, for if the probability for heads were 1 and for tails 0, we would have no choice at all. And, we see that the entropy for the case above is only .811 bit per toss. We feel somehow that we ought to be able to represent the outcome of a sequence of such biased tosses by fewer than one binary digit per toss, but it is not immediately clear how many binary digits we must use.

If we choose heads over tails with probability $p_1$, the probability $p_0$ of choosing tails must of course be $1 - p_1$. Thus, if we know $p_1$ we know $p_0$ as well. We can compute $H$ for various values of $p_1$ and plot a graph of $H$ vs. $p_1$. Such a curve is shown in Figure V-1. $H$ has a maximum value of 1 when $p_1$ is 0.5 and is 0 when $p_1$ is 0 or 1, that is, when it is certain that the message source always produces either one symbol or the other.

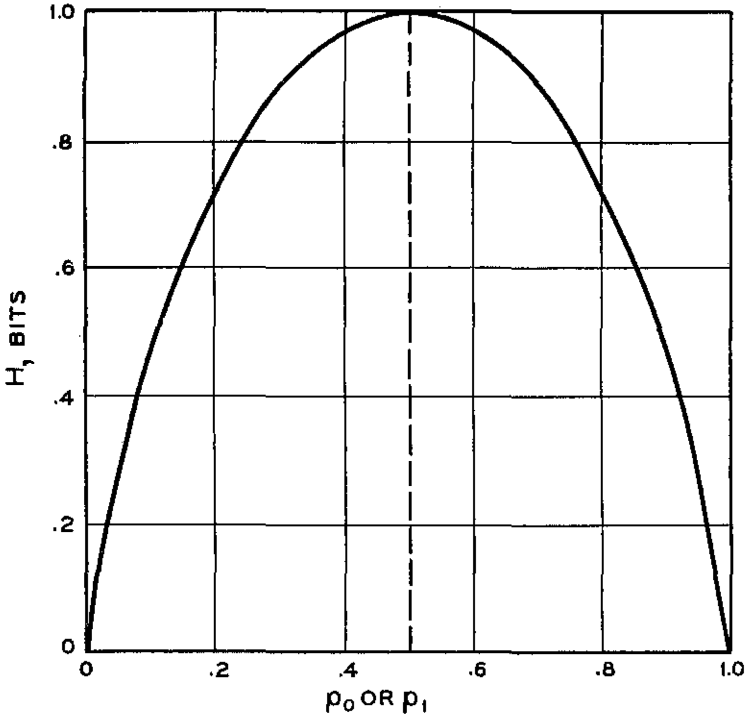


*Fig. V-1*

Really, whether we call heads $X$ and tails $Y$ or heads $Y$ and tails $X$ is immaterial, so the curve of $H$ vs. $p_1$ must be the same as $H$ vs. $p_0$. Thus, the curve of Figure V-1 is symmetrical about the dashed center line at $p_1$ and $p_0$ equal to 0.5.

A message source may produce successive choices among the ten decimal digits, or among the twenty-six letters of the alphabet, or among the many thousands of words of the English language. Let us consider the case in which the message source produces one among $n$ symbols or words, with probabilites which are independent of previous choices. In this case the entropy is defined as

$$H = -\sum_{i=1}^{n} p_i \log p_i \text{ bits per symbol} \qquad (5.1)$$

Here the sign $\Sigma$ (sigma) means to sum or to add up various terms.

$p_i$ is the probability of the $i^{\text{th}}$ symbol being chosen. The $i = 1$ below and $n$ above the $\Sigma$ mean to let $i$ be 1, 2, 3, etc. up to $n$, so the equation says that the entropy will be given by adding $p_1 \log p_1$ and $p_2 \log p_2$ and so on, including all symbols. We see that when $n = 2$ we have the simple case which we considered earlier.

Let us take an example. Suppose, for instance, that we toss two coins simultaneously. Then there are four possible outcomes, which we can label with the numbers 1 through 4:

$H$ $H$ or 1
$H$ $T$ or 2
$T$ $H$ or 3
$T$ $T$ or 4

If the coins are honest, the probability of each outcome is ¼ and the entropy is

$$H = -(\tfrac{1}{4}\log\tfrac{1}{4} + \tfrac{1}{4}\log\tfrac{1}{4} + \tfrac{1}{4}\log\tfrac{1}{4} + \tfrac{1}{4}\log\tfrac{1}{4})$$
$$H = -(-\tfrac{1}{2} - \tfrac{1}{2} - \tfrac{1}{2} - \tfrac{1}{2})$$
$$H = 2 \text{ bits per pair tossed}$$

It takes 2 bits of information to describe or convey the outcome of tossing a pair of honest coins simultaneously. As in the case of tossing one coin which has equal probabilities of landing heads or tails, we can in this case see that we can use 2 binary digits to describe the outcome of a toss: we can use 1 binary digit for each coin. Thus, in this case too, we can transmit the message generated by the process (of tossing two coins) by using a number of binary digits equal to the entropy.

If we have some number $n$ of symbols all of which are equally probable, the probability of any particular one turning up is $1/n$, so we have $n$ terms, each of which is $1/n \log 1/n$. Thus, the entropy is in this case

$$H = -\log 1/n \quad \text{bits per symbol}$$

For instance, an honest die when rolled has equal probabilities of turning up any number from 1 to 6. Hence, the entropy of the sequence of numbers so produced must be $-\log \tfrac{1}{6}$, or 2.58 bits per throw.

More generally, suppose that we choose each time with equal likelihood among all binary numbers with $N$ digits. There are $2^N$ such numbers, so

$$n = 2^N$$

From Table X we easily see that

$$\log 1/n = \log 2^{-N} = -N$$

Thus, for a source which produces at each choice with equal likelihood some $N$-digit binary number, the entropy is $N$ bits per number. Here the message produced by the source *is* a binary number which can certainly be represented by binary digits. And, again, the message can be represented by a number of binary digits equal to the entropy of the message, measured in bits. This example illustrates graphically how the logarithm *must* be the correct mathematical function in the entropy.

Ordinarily the probability that the message source will produce a particular symbol is different for different symbols. Let us take as an example a message source which produces English words independently of what has gone before but with the probabilities characteristic of English prose. This corresponds to the first-order word approximation given in Chapter III.

In the case of English prose, we find as an empirical fact that if we order the words according to frequency of usage, so that the most frequently used, the most probable word (*the,* in fact) is word number 1, the next most probable word (*of*) is number 2, and so on, then the probability for the $r^{\text{th}}$ word is very nearly (if $r$ is not too large)

$$p_r = .1/r \qquad (5.2)$$

If equation 5.2 were strictly true, the points in Figure V-2, in which word probability or frequency $p_r$ is plotted against word order or rank $r,$ would fall on the solid line which extends from upper left to lower right. We see that this is very nearly so. This empirical inverse relation between word probability and word rank is known as Zipf's law. We will discuss Zipf's law in Chapter XII; here, we propose merely to use it.

We can show that this equation (5.2) cannot hold for all words. To see this, let us consider tossing a coin. If the probability of heads

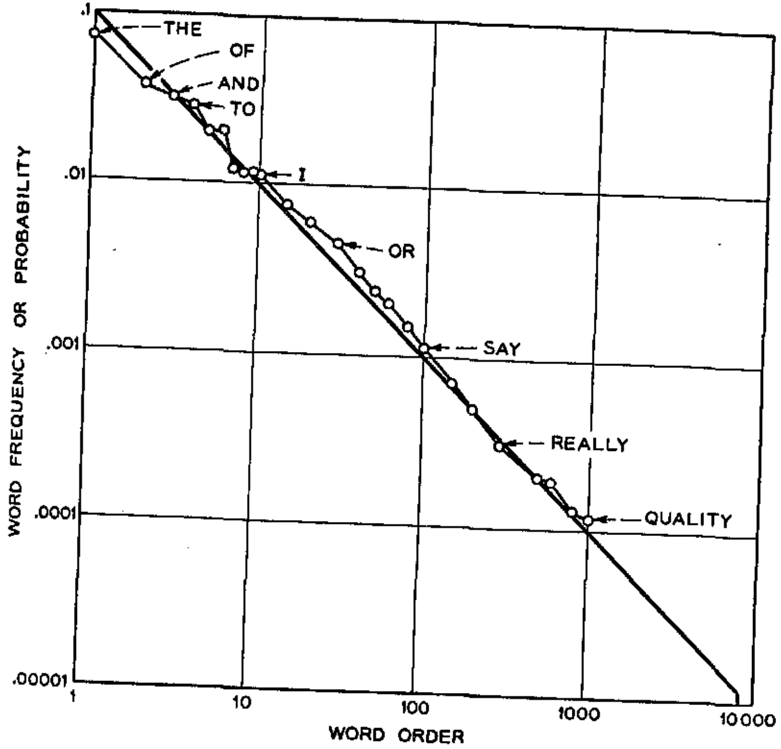


*Fig. V-2*

turning up is ½ and the probability of tails turning up is ½, then there is no other possible outcome: ½ + ½ = 1. If there were an additional probability of ⅒ that the coin would stand on edge, we would have to conclude that in a hundred tosses we would expect 110 outcomes: heads 50 times, tails 50 times, and standing on edge 10 times. This is patently absurd. The probabilities of all outcomes must add up to unity. Now, let us note that if we add up successively $p_1$ plus $p_2$, etc., as given by equation 5.2, we find that by the time we came to $p_{8727}$ the sum of the successive probabilities has become unity. If we took this literally, we would conclude that no additional word could ever occur. Equation 5.1 must be a little in error.

Nonetheless, the error is not great, and Shannon used equation

5.2 in computing the entropy of a message source which produces words independently but with the probability of their occurring in English text. In order to make the sum of the probabilities of all words unity, he included only the 8,727 most frequently used words. He found the entropy to be 9.14 bits per word.

In Chapter IV, we saw that English text can be encoded letter by letter by using 5 binary digits per character or 27.5 binary digits per word. We also saw that by providing different sequences of binary digits for each of 16,357 words and 27 characters, we could encode English text by using about 14 binary digits per word. We are now beginning to suspect that the number of binary digits actually required is given by the entropy, and, as we have seen, Shannon's estimate, based on the relative probabilities of English words, would be 9.14 binary digits per word.

As a next step in exploring this matter of the number of binary digits required to encode the message produced by a message source, we will consider a startling theorem which Shannon proved concerning the "messages" produced by an ergodic source which selects a sequence of letters or words independently with certain probabilities.

Let us consider all of the messages the source can produce which consist of some particular large number of characters. For example, we might consider all messages which are 100,000 symbols (letters, words, characters) long. More generally, let us consider messages having a number $M$ of characters. Some of these messages are more probable than others. In the probable messages, symbol 1 occurs about $Mp_1$ times, symbol 2 occurs about $Mp_2$ times, etc. Thus, in these probable messages each symbol occurs with about the frequency characteristic of the source. The source *might* produce other sorts of messages, for instance, a message consisting of one symbol endlessly repeated or merely a message in which the numbers of the various symbols differed markedly from $M$ times their probabilities, but it seldom does.

The remarkable fact is that, if $H$ is the entropy of the source per symbol, there are just about $2^{MH}$ probable messages, and the rest of the messages all have vanishingly small probabilities of ever occurring. In other words, if we ranked the messages from most probable to least probable, and assigned binary numbers of $MH$ digits to the $2^{MH}$ most probable messages, we would be almost certain to have a number corresponding to any $M$-symbol message that the source actually produced.

Let us illustrate this in particular simple cases. Suppose that the symbols produced are 1 or 0. If these are produced with equal probabilities, a probability ½ that for 1 and a probability ½ that for 0 the entropy $H$ is, as we have seen, 1 bit per symbol. Let us let the source produce messages $M$ digits long. Then $MH = 1{,}000$, and, according to Shannon's theorem, there must be $2^{1000}$ different probable messages.

Now, by using 1,000 binary digits we can write just $2^{1000}$ different binary numbers. Thus, in order to assign a different binary number to each probable message, we must use binary numbers 1,000 digits long. This is just what we would expect. In order to designate to the message destination which 1,000 digit binary number the message source produces, we must send a message 1,000 binary digits long.

But, suppose that the digits constituting the messages produced by the message source are obtained by tossing a coin which turns up heads, designating 1, ¾ of the time and tails, designating 0, ¼ of the time. The typical messages so produced will contain more 1's than 0's, but that is not all. We have seen that in this case the entropy $H$ is only .811 bit per toss. If $M$, the length of the message, is again taken as 1,000 binary digits, $MH$ is only 811. Thus, while as before there are $2^{1000}$ *possible* messages, there are only $2^{811}$ *probable* messages.

Now, by using 811 binary digits we can write $2^{811}$ different binary numbers, and we can assign one of these to each of the 1,000-digit probable messages, leaving the other improbable 1,000-digit messages unnumbered. Thus, we can send word to a message destination which *probable* 1,000-digit message our message source produces by sending only 811 binary digits. And the chance that the message source will produce an improbable 1,000-digit message, to which we have assigned no number, is negligible. Of course, the scheme is not quite foolproof. The message source may still very occasionally turn up a message for which we have no label among all $2^{811}$ of our 811-digit binary labels. In this case we cannot transmit the message—at least, not by using 811 binary digits.

We see that again we have a strong indication that the number of binary digits required to transmit a message is just the entropy in bits per symbol times the number of symbols. And, we might note that in this last illustration we achieved such an economical transmission by block encoding—that is, by lumping 1,000 (or some other large number) message digits together and representing each probable combination of digits by its individual code (of 811 binary digits).

How firmly and generally can this supposition be established? So far we have considered only cases in which the message source produces each symbol (number, letter, word) independently of the symbols it has produced before. We know this is not true for English text. Besides the constraints of word frequency, there are constraints of word order, so that the writer has less choice as to what the next word will be than he would if he could choose it independently of what has gone before.

How are we to handle this situation? We have a clue in the block coding which we discussed in Chapter IV, and which has been brought to our mind again in the last example. In an ergodic process the probability of the next letter may depend only on the preceding 1, 2, 3, 4, 5, or more letters but not on earlier letters. The second and third order approximations to English given in Chapter III illustrate text produced by such a process. Indeed, in any ergodic process of which we are to make mathematical sense the effect of the past on what symbol will be produced next must decrease as the remoteness of that past is greater. This is reasonably valid in the case of real English as well. While we can imagine examples to the contrary (the consistent use of the same name for a character in a novel), in general the word I write next does not depend on just what word I wrote 10,000 words back.

Now, suppose that before we encode a message we divide it up into very long blocks of symbols. If the blocks are long enough, only the symbols near the beginning will depend on symbols in the previous block, and, if we make the block long enough, these symbols that do depend on symbols in the previous block will form a negligible part of all the symbols in the block. This makes it possible for us to compute the entropy *per block* of symbols by means of equation 5.1. To keep matters straight, let us call the

probability of a particular one of the multitudinous long blocks of symbols, which we will call the $i^{\text{th}}$ block, $P(B_i)$. Then the entropy per block will be

$$H = -\sum_i P(B_i) \log P(B_i) \quad \text{bits per block}$$

Any mathematician would object to calling this the entropy. He would say, the quantity $H$ given by the above equation *approaches* the entropy as we make the block longer and longer, so that it includes more and more symbols. Thus, we must assume that we make the blocks very long indeed and get a very close approximation to the entropy. With this proviso, we can obtain the entropy per symbol by dividing the entropy per block by the number $N$ of symbols per block

$$H = -(1/N)\sum_i P(B_i) \log P(B_i) \quad \text{bits per symbol} \qquad (5.3)$$

In general, an estimate of entropy is always high *if* it fails to take into account some relations between symbols. Thus, as we make $N$, the number of symbols per block, greater and greater, $H$ as given by 5.3 will decrease and approach the true entropy.

We have insisted from the start that amount of information must be so defined that if separate messages are sent over several telegraph wires, the total amount of information must be the sum of the amounts of information sent over the separate wires. Thus, to get the entropy of several message sources operating simultaneously, we add the entropies of the separate sources. We can go further and say that if a source operates intermittently we must multiply its information rate or entropy by the fraction of the time that it operates in order to get its average information rate.

Now, let us say that we have one message source when we have just sent a particular sequence of letters such as TH. In this case the probability that the next letter will be E is very high. We have another particular message source when we have just sent NQ. In this case the probability that the next symbol will be U is unity. We calculate the entropy for each of these message sources. We multiply the entropy of a source which we label $B_i$ by the probability $p(B_i)$ that this source will occur (that is, by the fraction of instances in which this source is in operation). We multiply the entropy of each other source by the probability that that source will occur, and so on. Then we add all the numbers we get in this way in order to get the average entropy or rate of the over-all source, which is a combination of the many different sources, each of which operates only part time. As an example, consider a source involving digram probabilities only, so that the whole effect of the past is summed up in the letter last produced. One source will be the source we have when this letter is E; this will occur in .13 of the total instances. Another source will be the source we have when the letter just produced is W; this will occur in .02 of the total instances.

Putting this in formal mathematical terms, we say that if a particular block of $N$ symbols, which we designate by $B_i$, has just occurred, the probability that the next symbol will be symbol $S_j$ is

$$p_{B_i}(S_j)$$

The entropy of this "source" which operates only when a particular block of $N$ symbols designated by $B_i$ has just been produced is

$$-\sum_j p_{B_i}(S_j) \log p_{B_i}(S_j)$$

But, in what fraction of instances does this particular message source operate? The fraction of instances in which this source operates is the fraction of instances in which we encounter block $B_i$ rather than some other block of symbols; we call this fraction

$$p(B_i)$$

Thus, taking into account all blocks of $N$ symbols, we write the sum of the entropies of all the separate sources (each separate source defined by what particular block $B_i$ of $N$ symbols has preceded the choice of the symbol $S_j$) as

$$H_N = -\sum_{i,j} p(B_i) p_{B_i}(S_j) \log p_{B_i}(S_j) \qquad (5.4)$$

The $i,j$ under the summation sign mean to let $i$ and $j$ assume all possible values and to add all the numbers we get in this way.

As we let the number $N$ of symbols preceding symbol $S_j$ become very large, $H_N$ approaches the entropy of the source. If there are no statistical influences extending over more than $N$ symbols (this will be true for a digram source for $N = 1$ and for a trigram source for $N = 2$), then $H_N$ is the entropy.

Shannon writes equation 5.4 a little differently. The probability $p(B_i, S_j)$ of encountering the block $B_i$ followed by the symbol $S_j$ is the probability $p(B_i)$ of encountering the block $B_i$ times the probability $p_{B_i}(S_j)$ that symbol $S_j$ will follow block $B_i$. Hence, we can write 5.4 as follows:

$$H_N = -\sum_{i,j} p(B_i, S_j) \log p_{B_i}(S_j)$$

In Chapter III we consider a finite-state machine, such as that shown in Figure III-3, as a source of text. We can, if we wish, base our computation of entropy on such a machine. In this case, we regard each state of the machine as a message source and compute the entropy for that state. Then we multiply the entropy for that state by the probability that the machine will be in that state and sum (add up) all states in order to get the entropy.

Putting the matter symbolically, suppose that when the machine is in a particular state $i$ it has a probability $p_i(j)$ of producing a particular symbol which we designate by $j$. For instance, in a state labeled $i = 10$ it might have a probability of 0.03 of producing the third letter of the alphabet, which we label $j = 3$. Then

$$p_{10}(3) = .03$$

The entropy $H_i$ of state $i$ is computed in accord with 5.1:

$$H_i = -\sum_j p_i(j) \log p_i(j)$$

Now, we say that the machine has a probability $P_i$ of being in the $i^{\text{th}}$ state. The entropy per symbol for the machine as a source of symbols is then

$$H = \sum_i P_i H_i \quad \text{bits per symbol}$$

We can write this as

$$H = -\sum_{i,j} P_i p_i(j) \log p_i(j) \quad \text{bits per symbol} \qquad (5.5)$$

$P_i$ is the probability that the finite-state machine is in the $i^{\text{th}}$ state, and $p_i(j)$ is the probability that it produces the $j^{\text{th}}$ symbol when it is in the $i^{\text{th}}$ state. The $i$ and $j$ under the $\Sigma$ mean to allow both $i$ and $j$ to assume all possible values and to add all the numbers so obtained.

Thus, we have gone easily and reasonably from the entropy of a source which produces symbols independently and to which equation 5.1 applies to the more difficult case in which the probability of a symbol occurring depends on what has gone before. And, we have three alternative methods for computing or defining the entropy of the message source. These three methods are equivalent and rigorously correct for true ergodic sources. We should remember, of course, that the source of English text is only approximately ergodic.

Once having defined entropy per symbol in a perfectly general way, the problem is to relate it unequivocally to the average number of binary digits per symbol necessary to encode a message.

We have seen that if we divide the message into a block of letters or words and treat each possible block as a symbol, we can compute the entropy per block by the same formula we used per independent symbol and get as close as we like to the source entropy merely by making the blocks very long.

Thus, the problem is to find out how to encode efficiently in binary digits a sequence of symbols chosen from a very large group of symbols, each of which has a certain probability of being chosen. Shannon and Fano both showed ways of doing this, and Huffman found an even better way, which we shall consider here.

Let us for convenience list all the symbols vertically in order of decreasing probability. Suppose the symbols are the eight words *the, man, to, runs, house, likes, horse, sells,* which occur independently with probabilities of their being chosen, or appearing, as listed in Table XI.

We can compute the entropy per word by means of 5.1; it is 2.21 bits per word. However, if we merely assigned one of the eight 3-digit binary numbers to each word, we would need 3 digits to transmit each word. How can we encode the words more efficiently?

Figure V-3 shows how to construct the most efficient code for encoding such a message *word by word.* The words are listed to the left, and the probabilities are shown in parentheses. In constructing the code, we first find the two lowest probabilities, .02 (*sells*) and .03 (*horse*), and draw lines to the point marked .05, the probability of either *horse* or *sells.* We then disregard the individual probabilities connected by the lines and look for the two lowest probabilities, which are .04 (like) and .04 (house). We draw lines to the right to a point marked .08, which is the sum of .04 and .04. The two lowest remaining probabilities are now .05 and .08, so we draw a line to the right connecting them, to give a point marked

TABLE XI

| *Word* | *Probability* |
|---|---|
| the | .50 |
| man | .15 |
| to | .12 |
| runs | .10 |
| house | .04 |
| likes | .04 |
| horse | .03 |
| sells | .02 |

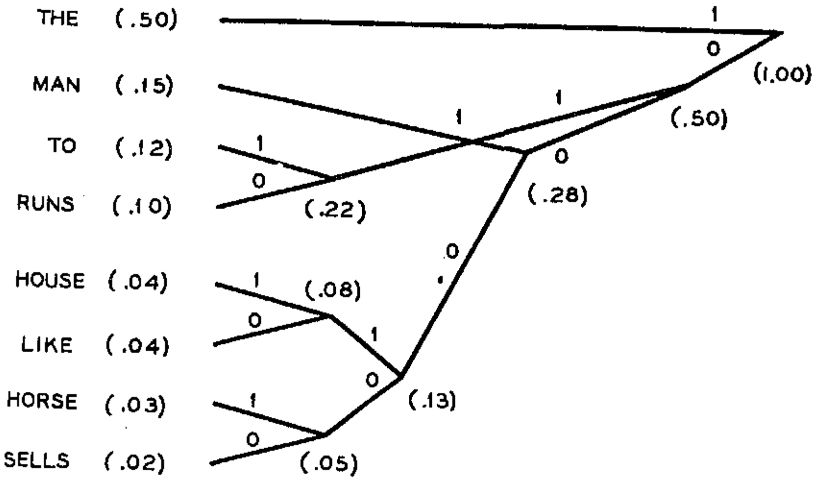


*Fig. V-3*

.13. We proceed thus until paths run from each word to a common point to the right, the point marked 1.00. We then label each upper path going to the left from a point 1 and each lower path 0. The code for a given word is then the sequence of digits encountered going left from the common point 1.00 to the word in question. The codes are listed in Table XII.

TABLE XII

| *Word* | *Probability p* | *Code* | *Number of Digits in Code, N* | *Np* |
|---|---|---|---|---|
| the | .50 | 1 | 1 | .50 |
| man | .15 | 001 | 3 | .45 |
| to | .12 | 011 | 3 | .36 |
| runs | .10 | 010 | 3 | .30 |
| house | .04 | 00011 | 5 | .20 |
| likes | .04 | 00010 | 5 | .20 |
| horse | .03 | 00001 | 5 | .15 |
| sells | .02 | 00000 | 5 | .10 |
| | | | | 2.26 |

In Table XII we have shown not only each word and its code but also the probability of each code and the number of digits in each code. The probability of a word times the number of digits in the code gives the average number of digits per word in a long message due to the use of that particular word. If we add the products of the probabilities and the numbers of digits for all the words, we get the average number of digits per word, which is 2.26. This is a little larger than the entropy per word, which we found to be 2.21 bits per word, but it is a smaller number of digits than the 3 digits per word we would have used if we had merely assigned a different 3-digit code to each word.

Not only can it be proved that this Huffman code is the most efficient code for encoding a set of symbols having different probabilities, it can be proved that it always calls for less than one binary digit per symbol more than the entropy (in the above example, it calls for only 0.05 extra binary digits per symbol).

Now suppose that we combine our symbols into blocks of 1, 2, 3, or more symbols before encoding. Each of these blocks will have a probability (in the case of symbols chosen independently, the probability of a sequence of symbols will be the product of the probabilities of the symbols). We can find a Huffman code for these blocks of symbols. As we make the blocks longer and longer, the number of binary digits in the code for each block will increase. Yet, our Huffman code will take less than one extra digit per block above the entropy in bits per block! Thus, as the blocks and their codes become very long, the less-than-one extra digit of the Huffman code will become a negligible fraction of the total number of digits, and, as closely as we like (by making the blocks longer), the number of binary digits per block will equal the entropy in bits per block.

Suppose we have a communication channel which can transmit a number $C$ of off-or-on pulses per second. Such a channel can transmit $C$ binary digits per second. Each binary digit is capable of transmitting one bit of information. Hence we can say that the information *capacity* of this communication channel is $C$ bits per second. If the entropy $H$ of a message source, measured in bits per second, is less than $C$, then, by encoding with a Huffman code, the signals from the source can be transmitted over the channel.

Not all channels transmit binary digits. A channel, for instance, might allow three amplitudes of pulses, or it might transmit different pulses of different lengths, as in Morse code. We can imagine connecting various different message sources to such a channel. Each source will have some entropy or information rate. Some source will give the highest entropy that can be transmitted over the channel, and this highest possible entropy is called the *channel capacity* $C$ of the channel and is measured in bits per second.

By means of the Huffman code, the output of the channel when it is transmitting a message of this greatest possible entropy can be coded into some least number of binary digits per second, and, when long stretches of message are encoded into long stretches of binary digits, it must take very close to $C$ binary digits per second to represent the signals passing over the channel.

This encoding can, of course, be used in the reverse sense, and $C$ independent binary digits per second can be so encoded as to be transmitted over the channel. Thus, a source of entropy $H$ can be encoded into $H$ binary digits per second, and a general discrete channel of capacity $C$ can be used to transmit $C$ bits per second.

We are now in a position to appreciate one of the fundamental theorems of information theory. Shannon calls this the fundamental theorem of the noiseless channel. He states it as follows:

> Let a source have entropy H (bits per symbol) and a channel have a capacity [to transmit] C bits per second. Then it is possible to encode the ousput of the source in such a way as to transmit at the average rate $(C/H) - \varepsilon$ symbols per second over the channel, where $\varepsilon$ is arbitrarily small. It is not possible to transmit at an average rate greater than C/H.

Let us restate this without mathematical niceties. Any discrete channel that we may specify, whether it transmits binary digits, letters and numbers, or dots, dashes, and spaces of certain distinct lengths has some particular unique channel capacity $C$. Any ergodic message source has some particular entropy $H$. If $H$ is less than or equal to $C$, we can transmit the messages generated by the source over the channel. If $H$ is greater than $C$, we had better not try to do so, because we just plain can't.

We have indicated above how the first part of this theorem can be proved. We have not shown that a source of entropy $H$ cannot be encoded in less than $H$ binary digits per symbol, but this also can be proved.

We have now firmly arrived at the fact that the entropy of a message source measured in bits tells us how many binary digits (or off-or-on pulses, or yeses-or-noes) are required, per character, or per letter, or per word, or per second in order to transmit messages produced by the source. This identification goes right back to Shannon's original paper. In fact, the word *bit* is merely a contraction of *binary digit* and is generally used in place of *binary digit.*

Here I have used *bit* in a particular sense, as a measure of amount of information, and in other contexts I have used a different expression, binary digit. I have done this in order to avoid a confusion which might easily have arisen had I started out by using *bit* to mean two different things.

After all, in practical situations the entropy in bits is usually different from the number of binary digits involved. Suppose, for instance, that a message source randomly produces the symbol 1 with a probability ¼ and the symbol 0 with the probability ¾ and that it produces 10 symbols per second. Certainly such a source produces binary digits at a rate of 10 per second, but the information rate or entropy of the source is .811 bit per binary digit and 8.11 bits per second. We could encode the sequence of binary digits produced by this source by using on the average only 8.11 binary digits per second.

Similarly, suppose we have a communication channel which is capable of transmitting 10,000 arbitrarily chosen off-or-on pulses per second. Certainly, such a channel has a channel capacity of 10,000 bits per second. However, if the channel is used to transmit a completely repetitive pattern of pulses, we must say that the actual rate of transmission of information is 0 bits per second, despite the fact that the channel is certainly transmitting 10,000 binary digits per second.

Here we have used bit only in the sense of a binary measure of amount of information, as a measure of the entropy or information rate of a message source in bits per symbol or in bits per second or as a measure of the information transmission capabilities of a channel in bits per symbol or bits per second. We can describe it as an elementary binary choice or decision among two possibilities which have equal probabilities. At the message source a bit represents a certain amount of choice as to the message which will be generated; in writing grammatical English we have on the average a choice of about one bit per letter. At the destination a bit of information resolves a certain amount of uncertainty; in receiving English text there is on the average, about one bit of uncertainty as to what the next letter will be.

When we are transmitting messages generated by an information source by means of off-or-on pulses, we know how many binary digits we are transmitting per second even when (as in most cases) we don't know the entropy of the source. (If we know the entropy of the source in bits per second to be less than the binary digits used per second, we would know that we could get along in principle with fewer binary digits per second.) We know how to use the binary digits to specify or determine one out of several possibilities, either by means of a tree such as that of Figure IV-4 or by means of a Huffman code such as that of Figure V-3. It is common in such a case to speak of the rate of transmission of binary digits as a bit rate, but there is a certain danger that the inexperienced may muddy their thinking if they do this.

All that I really ask of the reader is to remember that we have used *bit* in one sense only, as a measure of information and have called 0 or 1 a binary digit. If we can transmit 1,000 freely chosen binary digits per second, we can transmit 1,000 bits of information a second. It may be convenient to use *bit* to mean *binary digit,* but when we do so we should be sure that we understand what we are doing.

Let us now return for a moment to an entirely different matter, the Huffman code given in Table XII and Figure V-3. When we encode a message by using this code and get an uninterrupted string of symbols, how do we tell whether we should take a particular 1 in the string of symbols as indicating the word *the* or as part of the code for some other word?

We should note that of the codes in Table XII, none forms the first part of another. This is called the *prefix property.* It has important and, indeed, astonishing consequences, which are easily illustrated. Suppose, for instance, that we encode the message: the man sells the house to the man the horse runs to the man. The encoded message is as follows:

```
|the|  man  |     sells     |the|    house     |
  1   0 0 1   0 0  0 0 0 1 0   0 0 1   1
             ¦     likes     ¦   man    ¦the¦

   to   |the|  man  |the|    horse    |
 0 1 1    1   0 0 1   1   0 0 0 0 1
   to   ¦the¦  man  ¦the¦    horse    ¦

  runs  |   to   |the|  man  |
 0 1 0    0 1 1    1   0 0 1
  runs  ¦   to   ¦the¦  man  ¦
```

Here the message words are written above the code groups.

Now suppose we receive only the digits following the first vertical dashed line below the digits. We start to decode by looking for the shortest sequence of digits which constitutes a word in our code. This is 00010, which corresponds to *likes.* We go on in this fashion. The "decoded" words are written under the code, separated by dashed lines.

We see that after a few errors the dashed lines correspond to the solid lines, and from that point on the deciphered message is correct. We don't need to know where the message starts in order to decode it as correctly as possible (unless all code words are of equal length).

When we look back we can see that we have fulfilled the purpose of this chapter. We have arrived at a measure of the amount of information per symbol or per unit time of an ergodic source, and we have shown how this is equal to the average number of binary digits per symbol necessary to transmit the messages produced by the source. We have noted that to attain transmission with negligibly more bits than the entropy, we must encode the messages produced by the source in long blocks, not symbol by symbol.

We might ask, however, how long do the blocks have to be? Here we come back to another consideration. There are two reasons for encoding in long blocks. One is, in order to make the average number of binary digits per symbol used in the Huffman code negligibly larger than the entropy per symbol. The other is, that to encode such material as English text efficiently we must take into account the influence of preceding symbols on the probability that a given symbol will appear next. We have seen that we can do this using equation 5.3 and taking very long blocks.

We return, then, to the question: how many symbols $N$ must the block of characters have so that (1) the Huffman code is very efficient, (2) the entropy per block, disregarding interrelations outside of the block, is very close to $N$ times the entropy per symbol? In the case of English text, condition 2 is governing.

Shannon has estimated the entropy per letter for English text by measuring a person's ability to guess the next letter of a message after seeing 1, 2, 3, etc., preceding letters. In these texts the "alphabet" used consisted of 26 letters plus the space.

Figure V-4 shows the upper and lower bounds on the entropy of English plotted vs. the number of letters the person saw in

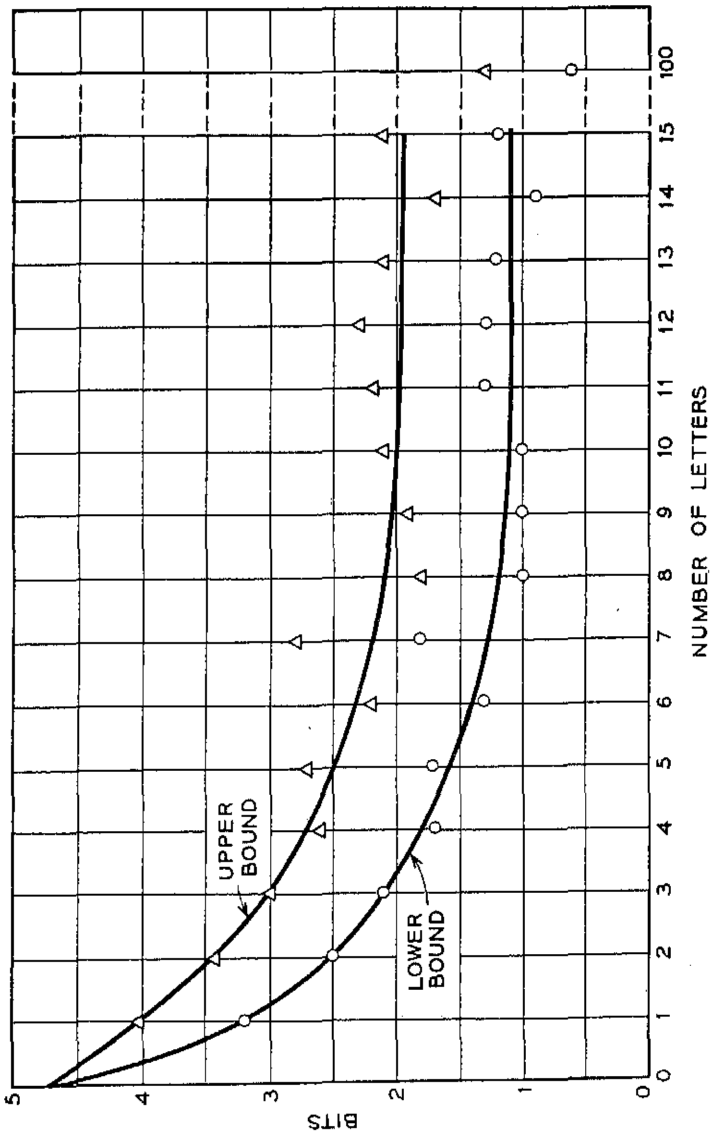


*Fig. V-4*

making his prediction. While the curve seems to drop slowly as the number of letters is increased from 10 to 15, it drops substantially between 15 and 100. This would appear to indicate that we might have to encode in blocks as large as 100 letters long in order to encode English really efficiently.

From Figure V-4 it appears that the entropy of English text lies somewhere between 0.6 and 1.3 bits per letter. Let us assume a value of 1 bit per letter. Then it will take on the average 100 binary digits to encode a block of 100 letters. This means that there are $2^{100}$ probable English sequences of 100 letters. In our usual decimal notation, $2^{100}$ can be written as 1 followed by 30 zeroes, a fantastically large number.

In endeavoring to find the probability in English text of all meaningful blocks of letters 100 letters long, we would have to count the relative frequency of occurrence of each such block. Since there are $10^{30}$ highly likely blocks, this would be physically impossible.

Further, this is impossible in principle. Most of these $10^{30}$ sequences of letters and spaces (which do not include *all* meaningful sequences) have never been written down! Thus, it is impossible to speak of their relative frequencies or probabilities of such long blocks of letters as derived from English text.

Here we are really confronted with two questions: the accuracy of the description of English text as the product of an ergodic source and the most appropriate statistical description of that source. One may believe that appropriate probabilities do exist in some form in the human being even if they cannot be evaluated by the examination of existing text. Or one may believe that the probabilities exist and that they can be derived from data taken in some way more appropriate than a naïve computation of the probabilities of sequences of letters. We may note, for instance, that equations 5.4 and 5.5 also give the entropy of an ergodic source. Equation 5.5 applies to a finite-state machine. We have noted at the close of Chapter III that the idea of a human being being in some particular state and in that state producing some particular symbol or word is an appealing one.

Some linguists hold, however, that English grammar is inconsistent with the output of a finite-state machine. Clearly, in trying to understand the structure and the entropy of actual English text we would have to consider such text much more deeply than we have up to this point.

It is safe if not subtle to apply an exact mathematical theory blindly and mechanically to the ideal abstraction for which it holds. We must be clever and wise in using even a good and appropriate mathematical theory in connection with actual, nonideal problems. We should seek a simple and realistic description of the laws governing English text if we are to relate it with communication theory as successfully as possible. Such a description must certainly involve the grammar of the language, which we will discuss in the next chapter.

In any event, we know that there are some valid statistics of English text, such as letter and word frequencies, and the coding theorems enable us to take advantage of such known statistics.

If we encode English letter by letter, disregarding the relative frequencies of the letters, we require 4.76 binary digits per character (including space). If we encode letter by letter, taking into account the relative probabilities of various letters, we require 4.03 binary digits per character. If we encode word by word, taking into account relative frequencies of words, we require 1.66 binary digits per character. And, by using an ingenious and appropriate means, Shannon has estimated the entropy of English text to be between .6 and 1.3 bits per letter, so that we may hope for even more efficient encoding.

If, however, we mechanically push some particular procedure for finding the entropy of English text to the limit, we can easily engender not only difficulties but nonsense. Perhaps we can ascribe this nonsense partly to differences between man as a source of English text and our model of an ideal ergodic source, but partly we should ascribe it to the use of an inappropriate approach. We can surely say that the model of man as an ergodic source of text is good and useful if not perfect, and we should regard it highly for these qualities.

This chapter has been long and heavy going, and a summary seems in order. Clearly, it is impossible to recapitulate briefly all those matters which took so many pages to expound. We can only re-emphasize the most vital points.

In communication theory the entropy of a signal source in bits per symbol or per second gives the average number of binary digits, per symbol or per second, necessary to encode the messages produced by the source.

We think of the message source as randomly, that is, unpredictably, choosing one among many possible messages for transmission. Thus, in connection with the message source we think of entropy as a measure of choice, the amount of choice the source excercises in selecting the one particular message that is actually transmitted.

We think of the recipient of the message, prior to the receipt of the message, as being uncertain as to which among the many possible messages the message source will actually generate and transmit to him. Thus, we think of the entropy of the message source as measuring the uncertainty of the recipient as to which message will be received, an uncertainty which is resolved on receipt of the message.

If the message is one among $n$ equally probable symbols or messages, the entropy is log $n$. This is perfectly natural, for if we have log $n$ binary digits, we can use them to write out

$$2^{\log n} = n$$

different binary numbers, and one of these numbers can be used as a label for each of the $n$ messages.

More generally, if the symbols are not equally probable, the entropy is given by equation 5.1. By regarding a very long block of symbols, whose content is little dependent on preceding symbols, as a sort of super symbol, equation 5.1 can be modified to give the entropy per symbol for information sources in which the probability that a symbol is chosen depends on what symbols have been chosen previously. This gives us equation 5.3. Other general expressions for entropy are given by equations 5.4 and 5.5.

By assuming that the symbols or blocks of symbols which a source produces are encoded by a most efficient binary code called a Huffman code, it is possible to prove that the entropy of an ergodic source measured in bits is equal to the average number of binary digits necessary to encode it.

An error-free communication channel may not transmit binary digits; it may transmit letters or other symbols. We can imagine attaching different message sources to such a channel and seeking (usually mathematically) the message source that causes the entropy of the message transmitted over the channel to be as large as possible. This largest possible entropy of a message transmitted over an error-free channel is called the channel capacity. It can be proved that, if the entropy of a source is less than the channel capacity of the channel, messages from the source can be encoded so that they can be transmitted over the channel. This is Shannon's fundamental theorem for the noiseless channel.

In principle, expressions such as equations 5.1, 5.3, 5.4, and 5.5 enable us to compute the entropy of a message source by statistical analysis of messages produced by the source. Even for an ideal ergodic source, this would often call for impractically long computations. In the case of an actual source, such as English text, some naïve prescriptions for computing entropy can be meaningless.

An approximation to the entropy can be obtained by disregarding the effect of some past symbols on the probability of the source producing a particular symbol next. Such an approximation to the entropy is always too large and calls for encoding by means of more binary digits than are absolutely necessary. Thus, if we encode English text letter by letter, disregarding even the relative probabilities of letters, we require 4.76 binary digits per letter, while if we encode word by word, taking into account the relative probability of words, we require 1.66 binary digits per letter.

If we wanted to do even better we would have to take into account other features of English such as the effect of the constraints imposed by grammar on the probability that a message source will produce a particular word.

While we do not know how to encode English text in a highly efficient way, Shannon made an ingenious experiment which shows that the entropy of English text must lie between .6 and 1.3 bits per character. In this experiment a person guessed what letter would follow the letters of a passage of text many letters long.

# CHAPTER VI *Language and Meaning*

THE TWO GREAT TRIUMPHS of information theory are establishing the channel capacity and, in particular, the number of binary digits required to transmit information from a particular source and showing that a noisy communication channel has an information rate in bits per character or bits per second up to which errorless transmission is possible despite the noise. In each case, the results must be demonstrated for discrete and for continuous sources and channels.

After four chapters of by no means easy preparation, we were finally ready to essay in the previous chapter the problem of the number of binary digits required to transmit the information generated by a truly ergodic discrete source. Were this book a text on information theory, we would proceed to the next logical step, the noisy discrete channel, and then on to the ergodic continuous channel.

At the end of such a logical progress, however, our thoughts would necessarily be drawn back to a consideration of the message sources of the real world, which are only approximately ergodic, and to the estimation of their entropy and the efficient encoding of the messages they produce.

Rather than proceeding further with the strictly mathematical aspects of communication theory at this point, is it not more attractive to pause and consider that chief form of communication, language, in the light of communication theory? And, in doing so, why should we not let our thoughts stray a little in viewing an important part of our world from the small eminence we have attained? Why should we not see whether even the broad problems of language and meaning seem different to us in the light of what we have learned?

In following such a course the reader should heed a word of caution. So far the main emphasis has been on what we *know.* What we know is the hard core of science. However, scientists find it very difficult to share the things that they know with laymen. To understand the sure and the reasonably sure knowledge of science takes the sort of hard thought which I am afraid was required of the reader in the last few chapters.

There is, however, another and easier though not entirely frivolous side to science. This is a peculiar type of informed ignorance. The scientist's ignorance is rather different from the layman's ignorance, because the background of established fact and theory on which the scientist bases his peculiar brand of ignorance excludes a wide range of nonsense from his speculations. In the higher and hazier reaches of the scientist's ignorance, we have scientifically informed ignorance about the origin of the universe, the ultimate basis of knowledge, and the relation of our present scientific knowledge to politics, free will, and morality. In this particular chapter we will dabble in what I hope to be scientifically informed ignorance about language.

The warning is, of course, that much of what will be put forward here about language is no more than informed ignorance. The warning seems necessary because it is very hard for laymen to tell scientific ignorance from scientific fact. Because the ignorance is necessarily expressed in broader, sketchier, and less qualified terms than is the fact, it is easier to assimilate. Because it deals with grand and unsolved problems, it is more romantic. Generally, it has a wider currency and is held in higher esteem than is scientific fact.

However hazardous such ignorance may be to the layman, it is valuable to the scientist. It is this vision of unattained lands, of unscaled heights, which rescues him from complacency and spurs him beyond mere plodding. But when the scientist is airing his ignorance he usually knows what he is doing, while the unwarned layman apparently often does not and is left scrambling about on cloud mountains without ever having set foot on the continents of knowledge.

With this caution in mind, let us return to what we have already encountered concerning language and proceed thence.

In what follows we will confine ourselves to a discussion of grammatical English. We all know (and especially those who have had the misfortune of listening to a transcription of a seemingly intelligible conversation or technical talk) that much spoken English appears to be agrammatical, as, indeed, much of Gertrude Stein is. So are many conventions and clichés. "Me heap big chief" is perfectly intelligible anywhere in the country, yet it is certainly not grammatical. Purists do not consider the inverted word order which is so characteristic of second-rate poetry as being grammatical.

Thus, a discussion of grammatical English by no means covers the field of spoken and written communication, but it charts a course which we can follow with some sense of order and interest.

We have noted before that, if we are to write what will be accepted as English text, certain constraints must be obeyed. We cannot simply set down any word following any other. A complete grammar of a language would have to express all of these constraints fully. It should allow within its rules the construction of any sequence of English words which will be accepted, at some particular time and according to some particular standard, as grammatical.

The matter of acceptance of constructions as grammatical is a difficult and hazy one. The translators who produced the King James Bible were free to say "fear not," "sin not," and "speak not" as well as "think not," "do not," or "have not," and we frequently repeat the aphorism "want not, waste not." Yet in our everyday speech or writing we would be constrained to say "do not fear," "do not sin," or "do not speak," and we might perhaps say, "If you are not to want, you should not waste." What is grammatical certainly changes with time. Here we can merely notice this and pass on to other matters.

Certainly, a satisfactory grammar must prescribe certain rules which allow the construction of all possible grammatical utterances and of grammatical utterances only. Besides doing this, satisfactory rules of grammar should allow us to analyze a sentence so as to distinguish the features which were determined merely by the rules of grammar from any other features.

If we once had such rules, we would be able to make a new estimate of the entropy of English text, for we could see what part of sentence structure is a mere mechanical following of rules and what part involves choice or uncertainty and hence contributes to entropy. Further, we could transmit English efficiently by transmitting as a message only data concerning the choices exercised in constructing sentences; at the receiver, we could let a grammar machine build grammatical sentences embodying the choices specified by the received message.

Even grammar, of course, is not the whole of language, for a sentence can be very odd even if it is grammatical. We can imagine that, if a machine capable of producing only grammatical sentences made its choices at random, it might perhaps produce such a sentence as "The chartreuse semiquaver skinned the feelings of the manifold." A man presumably makes his choices in some other way if he says, "The blue note flayed the emotions of the multitude." The difference lies in what choices one makes while following grammatical rules, not in the rules themselves. An understanding of grammar would not unlock to us all of the secrets of language, but it would take us a long step forward.

What sort of rules will result in the production of grammatical sentences only and of all grammatical sentences, even when choices are made at random? In Chapter III we saw that English-like sequences of words can be produced by choosing a word at random according to its probability of succeeding a preceding sequence of words some $M$ words long. An example of a second-order word approximation, in which a word is chosen on the basis of its succeeding the previous word, was given.

One can construct higher-order word approximations by using the knowledge of English which is stored in our heads. One can, for instance, obtain a fourth-order word approximation by simply showing a sequence of three connected words to a person and asking him to think up a sentence in which the sequence of words occurs and to add the next word. By going from person to person a long string of words can be constructed, for instance:

1. When morning broke after an orgy of wild abandon he said here head shook vertically aligned in a sequence of words signifying what.

2. It happened one frosty look of trees waving gracefully against the wall.

3. When cooked asparagus has a delicious flavor suggesting apples.

4. The last time I saw him when he lived.

These "sentences" are as sensible as they are because selections of words were not made at random but by thinking beings. The point to be noted is how astonishingly grammatical the sentences are, despite the fact that rules of grammar (and sense) were applied to only four words at a time (the three shown to each person and the one he added). Still, example 4 is perhaps dubiously grammatical.

If Shannon is right and there is in English text a choice of about 1 bit per symbol, then choosing among a group of 4 words could involve about 22 binary choices, or a choice among some 10 million 4-word combinations. In principle, a computer could be made to add words by using such a list of combinations, but the result would not be assuredly grammatical, nor could we be sure that this cumbersome procedure would produce all possible grammatical sequences of words. There probably are sequences of words which could form a part of a grammatical sentence in one case and could not in another case. If we included such a sequence, we would produce some nongrammatical sentences, and, if we excluded it, we would fail to produce all grammatical sentences.

If we go to combinations of more than four words, we will favor grammar over completeness. If we go to fewer than four words, we will favor completeness over grammar. We can't have both.

The idea of a finite-state machine recurs at this point. Perhaps at each point in a sentence a sentence-producing machine should be in a particular state, which allows it certain choices as to what state it will go to next. Moreover, perhaps such a machine can deal with certain classes or subclasses of words, such as singular nouns, plural nouns, adjectives, adverbs, verbs of various tense and number, and so on, so as to produce grammatical structures into which words can be fitted rather than sequences of particular words.

The idea of grammar as a finite-state machine is particularly appealing because a mechanist would assert that man must be a finite-state machine, because he consists of only a finite number of cells, or of atoms if we push the matter further.

Noam Chomsky, a brilliant and highly regarded modern linguist, rejects the finite-state machine as either a possible or a proper model of grammatical structure. Chomsky points out that there are many rules for constructing sequences of characters which cannot be embodied in a finite-state machine. For instance, the rule might be, choose letters at random and write them down until the letter Z shows up, then repeat all the letters since the preceding Z in reverse order, and then go on with a new set of letters, and so on. This process will produce a sequence of letters showing clear evidence of long-range order. Further, there is no limit to the possible length of the sequence between Z's. No finite-state machine can simulate this process and this result.

Chomsky points out that there is no limit to the possible length of grammatical sentences in English and argues that English sentences are organized in such a way that this is sufficient to rule out a finite-state machine as a source of all possible English text. But, can we really regard a sentence miles long as grammatical when we know darned well that no one ever has or will produce such a sentence and that no one could understand it if it existed?

To decide such a question, we must have a standard of being grammatical. While Chomsky seems to refer being or not being grammatical, and some questions of punctuation and meaning as well, to spoken English, I think that his real criterion is: a sentence is grammatical if, in reading or saying it aloud with a natural expression and thoughtfully but ingenuously, it is deemed grammatical by a person who speaks it, or perhaps by a person who hears it. Some problems which might plague others may not bother Chomsky because he speaks remarkably well-connected and grammatical English.

Whether or not the rules of grammar can be embodied in a finite-state machine, Chomsky offers persuasive evidence that it is wrong and cumbersome to try to generate a sentence by basing the choice of the next word entirely and solely on words already written down. Rather, Chomsky considers the course of sentence generation to be something of this sort:

We start with one or another of several general forms the sentence might take; for example, a noun phrase followed by a verb phrase. Chomsky calls such a particular form of sentence a *kernel sentence.* We then invoke rules for expanding each of the parts of the kernel sentence. In the case of a noun phrase we may first describe it as an article plus a noun and finally as "the man." In the case of a verb phrase we may describe it as a verb plus an object, the object as an article plus a noun, and, in choosing particular words, as "hit the ball." Proceeding in this way from the kernel sentence, noun phrase plus verb phrase, we arrive at the sentence, "The man hit the ball." At any stage we could have made other choices. By making other choices at the final stages we might have arrived at "A girl caught a cat."

Here we see that the element of choice is not exercised sequentially along the sentence from beginning to end. Rather, we choose an over-all skeletal plan or scheme for the whole final sentence at the start. That scheme or plan is the kernel sentence. Once the kernel sentence has been chosen, we pass on to parts of the kernel sentence. From each part we proceed to the constituent elements of that part and from the constituent elements to the choice of particular words. At each branch of this treelike structure growing from the kernel sentence, we exercise choice in arriving at the particular final sentence, and, of course, we chose the kernel sentence to start with.

Here I have indicated Chomsky's ideas very incompletely and very sketchily. For instance, in dealing with irregular forms of words Chomsky will first indicate the root word and its particular grammatical form, and then he will apply certain obligatory rules in arriving at the correct English form. Thus, in the branching construction of a sentence, use is made both of optional rules, which allow choice, and of purely mechanical, deterministic obligatory rules, which do not.

To understand this approach further and to judge its merit, one must refer to Chomsky's book,[1] and to the references he gives.

Chomsky must, of course, deal with the problem of ambiguous sentences, such as, "The lady scientist made the robot fast while she ate." The author of this sentence, a learned information theo-

[1] Noam Chomsky, *Syntactic Structures,* Mouton and Co., 's-Gravenhage, 1957.

rist, tells me that, allowing for the vernacular, it has at least four different meanings. It is perhaps too complicated to serve as an example for detailed analysis.

We might think that ambiguity arises only when one or more words can assume different meanings in what is essentially the same grammatical structure. This is the case in "he was mad" (either angry or insane) or "the pilot was high" (in the sky or in his cups). Chomsky, however, gives a simple example of a phrase in which the confusion is clearly grammatical. In "the shooting of the hunters," the noun hunters may be either the subject, as in "the growling of lions" or the object, as in "the growing of flowers."

Chomsky points out that different rules of transformation applied to different kernel sentences can lead to the same sequence of grammatical elements. Thus, "the picture was painted by a real artist" and "the picture was painted by a new technique" seem to correspond grammatically word for word, yet the first sentence could have arisen as a transformation of "a real artist painted the picture" while the second could not have arisen as a transformation of a sentence having this form. When the final words as well as the final grammatical elements are the same, the sentence is ambiguous.

Chomsky also faces the problem that the distinction between the provinces of grammar and meaning is not clear. Shall we say that grammar allows adjectives but not adverbs to modify nouns? This allows "colorless green." Or should grammar forbid the association of some adjectives with some nouns, of some nouns with some verbs, and so on? With one choice, certain constructions are grammatical but meaningless; with the other they are ungrammatical.

We see that Chomsky has laid out a plan for a grammar of English which involves at each point in the synthesis of a sentence certain steps which are either obligatory or optional. The processes allowed in this grammar cannot be carried out by a finite-state machine, but they can be carried out by a more general machine called a *Turing machine,* which is a finite-state machine plus an infinitely long tape on which symbols can be written and from which symbols can be read or erased. The relation of Chomsky's grammar to such machines is a proper study for those interested in automata.

We should note, however, that if we arbitrarily impose some bound on the length of a sentence, even if we limit the length to 1,000 or 1 million words, then Chomsky's grammar *does* correspond to a finite-state machine. The imposition of such a limit on sentence length seems very reasonable in a practical way.

Once a general specification or model of a grammar of the sort Chomsky proposes is set up, we may ask under what circumstances and how can an entropy be derived which will measure the choice or uncertainty of a message source that produces text according to the rules of the grammar? This is a question for the mathematically skilled information theorist.

Much more important is the production of a plausible and workable grammar. This might be a *phrase-structure* grammar, as Chomsky proposes, or it might take some other form. Such a grammar might be incomplete in that it failed to produce or analyze some constructions to be found in grammatical English. It seems more important that its operation should correspond to what we know of the production of English by human beings. Further, it should be simple enough to allow the generation and analysis of text by means of an electronic computer. I believe that computers must be used in attacking problems of the structure and statistics of English text.

While a great many people are convinced that Chomsky's phrase-structure approach is a very important aspect of grammar, some feel that his picture of the generation of sentences should be modified or narrowed if it is to be used to describe the actual generation of sentences by human beings. Subjectively, in speaking or listening to a speaker one has a strong impression that sentences are generated largely from beginning to end. One also gets the impression that the person generating a sentence doesn't have a very elaborate pattern in his head at any one time but that he elaborates the pattern as he goes along.

I suspect that studies of the form of grammars and of the statistics of their use as revealed by language will in the not distant future tell us many new things about the nature of language and about the nature of men as well. But, to say something more particular than this, I would have to outreach present knowledge—mine and others.

A grammar must specify not only rules for putting different types of words together to make grammatical structures; it must divide the actual words of English into classes on the basis of the places in which they can appear in grammatical structures. Linguists make such a division purely on the basis of grammatical function without invoking any idea of meaning. Thus, all we can expect of a grammar is the generation of grammatical sentences, and this includes the example given earlier: "The chartreuse semiquaver skinned the feelings of the manifold." Certainly the division of words into grammatical categories such as nouns, adjectives, and verbs is not our sole guide concerning the use of words in producing English text.

What does influence the choice among words when the words used in constructing grammatical sentences are chosen, not at random by a machine, but rather by a live human being who, through long training, speaks or writes English according to the rules of the grammar? This question is not to be answered by a vague appeal to the word *meaning.* Our criteria in producing English sentences can be very complicated indeed. Philosophers and psychologists have speculated about and studied the use of words and language for generations, and it is as hard to say anything entirely new about this as it is to say anything entirely true. In particular, what Bishop Berkeley wrote in the eighteenth century concerning the use of language is so sensible that one can scarcely make a reasonable comment without owing him credit.

Let us suppose that a poet of the scanning, rhyming school sets out to write a grammatical poem. Much of his choice will be exercised in selecting words which fit into the chosen rhythmic pattern, which rhyme, and which have alliteration and certain consistent or agreeable sound values. This is particularly notable in Poe's "The Bells," "Ulalume," and "The Raven."

Further, the poet will wish to bring together words which through their sound as well as their sense arouse related emotions or impressions in the reader or hearer. The different sections of Poe's "The Bells" illustrate this admirably. There is a marked contrast between:

> How they tinkle, tinkle, tinkle,
> In the icy air of night!
> While the stars that oversprinkle
> All the heavens, seem to twinkle
> In a crystalline delight; . . .

and

> Through the balmy air of night
> How they ring out their delight!
> From the molten-golden notes,
> And all in tune,
> What a liquid ditty floats . . .

Sometimes, the picture may be harmonious, congruous, and moving without even the trivial literal meaning of this verse of Poe's, as in Blake's two lines:

> Tyger, Tyger, burning bright
> In the forests of the night . . .

In instances other than poetry, words may be chosen for euphony, but they are perhaps more often chosen for their associations with and ability to excite passions such as those listed by Berkeley: fear, love, hatred, admiration, disdain. Particular words or expressions move each of us to such feelings. In a given culture, certain words and phrases will have a strong and common effect on the majority of hearers, just as the sights, sounds or events with which they are associated do. The words of a hymn or psalm can induce a strong religious emotion; political or racial epithets, a sense of alarm or contempt, and the words and phrases of dirty jokes, sexual excitement.

One emotion which Berkeley does not mention is a sense of understanding. By mouthing commonplace and familiar patterns of words in connection with ill-understood matters, we can associate some of our emotions of familiarity and insight with our perplexity about history, life, the nature of knowledge, consciousness, death, and Providence. Perhaps such philosophy as makes use of common words should be considered in terms of assertion of a reassurance concerning the importance of man's feelings rather than in terms of meaning.

One could spend days on end examining examples of motivation in the choice of words, but we do continually get back to the matter of meaning. Whatever meaning may be, all else seems lost without it. A Chinese poem, hymn, deprecation, or joke will have little effect on me unless I understand Chinese in whatever sense those who know a language understand it.

Though Colin Cherry, a well-known information theorist, appears to object, I think that it is fair to regard meaningful language as a sort of code of communication. It certainly isn't a simple code in which one mechanically substitutes a word for a deed. It's more like those elaborate codes of early cryptography, in which many alternative code words were listed for each common letter or word (in order to suppress frequencies). But in language, the listings may overlap. And one person's code book may have different entries from another's, which is sure to cause confusion.

If we regard language as an imperfect code of communication, we must ultimately refer meaning back to the intent of the user. It is for this reason that I ask, "What do you mean?" even when I have heard your words. Scholars seek the intent of authors long dead, and the Supreme Court seeks to establish the intent of Congress in applying the letter of the law.

Further, if I become convinced that a man is lying, I interpret his words as meaning that he intends to flatter or deceive me. If I find that a sentence has been produced by a computer, I interpret it to mean that the computer is functioning very cleverly.

I don't think that such matters are quibbles; it seems that we are driven to such considerations in connection with meaning if we do regard language as an imperfect code of communication, and as one which is sometimes exploited in devious ways. We are certainly far from any adequate treatment of such problems.

Grammatical sentences do, however, have what might be called a formal meaning, regardless of intent. If we had a satisfactory grammar, a machine should be able to establish the relations between the words of a sentence, indicating subject, verb, object, and what modifying phrases or clauses apply to what other words. The next problem beyond this in seeking such formal meaning in sentences is the problem of associating words with objects, qualities, actions, or relations in the world about us, including the world of man's society and of his organized knowledge.

In the simple communications of everyday life, we don't have much trouble in associating the words that are used with the proper objects, qualities, actions, and relations. No one has trouble with "close the east window" or "Henry is dead," when he hears such a simple sentence in simple, unambiguous surroundings. In a familiar American room, anyone can point out the window; we have closed windows repeatedly, and we know what direction east is. Also, we know Henry (if we don't get Henry Smith mixed up with Henry Jones), and we have seen dead people. If the sentence is misheard or misunderstood, a second try is almost sure to succeed.

Think, however, how puzzling the sentence about the window would be, even in translation, to a shelterless savage. And we can get pretty puzzled ourselves concerning such a question as, is a virus living or dead?

It appears that much of the confusion and puzzlement about the associations of words with things of the world arose through an effort by philosophers from Plato to Locke to give meaning to such ideas as window, cat, or dead by associating them with general ideas or ideal examples. Thus, we are presumed to identify a window by its resemblance to a general idea of a window, to an ideal window, in fact, and a cat by its resemblance to an ideal cat which embodies all the attributes of cattiness. As Berkeley points out, the abstract idea of a (or the ideal) triangle must at once be "neither oblique, rectangle, equilateral, equicrural nor scaleron, but all and none of these at once."

Actually, when a doctor pronounces a man dead he does so on the basis of certain observed *signs* which he would be at a loss to identify in a virus. Further, when a doctor makes a diagnosis, he does not start out by making an over-all comparison of the patient's condition with an ideal picture of a disease. He first looks for such signs as appearance, temperature, pulse, lesions of the skin, inflammation of the throat, and so on, and he also notes such *symptoms* as the patient can describe to him. Particular combinations of signs and symptoms indicate certain diseases, and in differential diagnoses further tests may be used to distinguish among diseases producing similar signs and symptoms.

In a similar manner, a botanist identifies a plant, familiar or unfamiliar, by the presence or absence of certain qualities of size, color, leaf shape and disposition, and so on. Some of these qualities, such as the distinction between the leaves of monocotyledonous and dicotyledonous plants, can be decisive; others, such as size, can be merely indicative. In the end, one is either sure he is right or perhaps willing to believe that he is right; or the plant may be a new species.

Thus, in the workaday worlds of medicine and botany, the ideal disease or plant is conspicuous by its absence as any actual useful criterion. Instead, we have lists of qualities, some decisive and some merely indicative.

The value of this observation has been confirmed strongly in recent work toward enabling machines to carry out tasks of recognition or classification. Early workers, perhaps misled by early philosophers, conceived the idea of matching a letter to an ideal pattern of a letter or the spectrogram of a sound to an ideal spectrogram of the sound. The results were terrible. Audrey, a pattern-matching machine with the bulk of a hippo and brains beneath contempt, could recognize digits spoken by one voice or a selected group of voices, but Audrey was sadly fallible. We should, I think, conclude that human recognition works this way in very simple cases only, if at all.

Later and more sophisticated workers in the field of recognition look for significant features. Thus, as a very simple example, rather than having an ideal pattern of a capital Q, one might describe Q as a closed curve without corners or reversals of curvature and with something attached between four and six o'clock.

In 1959, L. D. Harmon built at the Bell Laboratories a simple device weighing a few pounds which almost infallibly recognizes the digits from one to zero written out as words in longhand. Does this gadget match the handwriting against patterns? You bet it doesn't! Instead, it asks such questions as, how many times did the stylus go above or below certain lines? Were I's dotted or T's crossed?

Certainly, no one doubts that words refer to classes of objects, actions, and so on. We are surrounded by and involved with a large number of classes and subclasses of objects and actions which we can usefully associate with words. These include such objects as plants (peas, sunflowers . . .), animals (cats, dogs . . .), machines (autos, radios . . .), buildings (houses, towers . . .), clothing (skirts, socks . . .), and so on. They include such very complicated sequences of actions as dressing and undressing (the absent-minded, including myself, repeatedly demonstrate that they can do this unconsciously); tying one's shoes (an act which children have considerable difficulty in learning), eating, driving a car, reading, writing, adding figures, playing golf or tennis (activities involving a host of distinct subsidiary skills), listening to music, making love, and so on and on and on.

It seems to me that what delimits a particular class of objects, qualities, actions, or relations is not some sort of ideal example. Rather, it is a list of qualities. Further, the list of qualities cannot be expected to enable us to divide experience up into a set of logical, sharply delimited, and all-embracing categories. The language of science may approach this in dealing with a narrow range of experience, but the language of everyday life makes arbitrary, overlapping, and less than all-inclusive divisions of experience. Yet, I believe that it is by means of such lists of qualities that we identify doors, windows, cats, dogs, men, monkeys, and other objects of daily life. I feel also that this is the way in which we identify common actions such as running, skipping, jumping, and tying, and such symbols as words, written and spoken, as well.

I think that it is only through such an approach that we can hope to make a machine classify objects and experience in terms of language, or recognize and interpret language in terms of other language or of action. Further, I believe that when a word cannot offer a table of qualities or signs whose elements can be traced back to common and familiar experiences, we have a right to be wary of the word.

If we are to understand language in such a way that we can hope some day to make a machine which will use language successfully, we must have a grammar and we must have a way of relating words to the world about us, but this is of course not enough. If we are to regard sentences as meaningful, they must in some way correspond to life as we live it.

Our lives do not present fresh objects and fresh actions each day. They are made up of familiar objects and familiar though complicated sequences of actions presented in different groupings and orders. Sometimes we learn by adding new objects, or actions, or combinations of objects or sequences of actions to our stock, and so we enrich or change our lives. Sometimes we forget objects and actions.

Our particular actions depend on the objects and events about us. We dodge a car (a complicated sequence of actions). When thirsty, we stop at the fountain and drink (another complicated but recurrent sequence). In a packed crowd we may shoulder someone out of the way as we have done before. But our information about the world does not all come from direct observation, and our influence on others is happily not confined to pushing and shoving. We have a powerful tool for such purposes: language and words.

We use words to learn about relations among objects and activities and to remember them, to instruct others or to receive instruction from them, to influence people in one way or another. For the words to be useful, the hearer must understand them in the same sense that the speaker means them, that is, insofar as he associates them with nearly enough the same objects or skills. It's no use, however, to tell a man to read or to add a column of figures if he has never carried out these actions before, so that he doesn't have these skills. It is no use to tell him to shoot the aardvark and not the gnu if he has never seen either.

Further, for the sequences of words to be useful, they must refer to real or possible sequences of events. It's of no use to advise a man to walk from London to New York in the forenoon immediately after having eaten a seven o'clock dinner.

Thus, in some way the meaningfulness of language depends not only on grammatical order and on a workable way of associating words with collections of objects, qualities, and so on; it also depends on the structure of the world around us. Here we encounter a real and an extremely serious difficulty with the idea that we can in some way translate sentences from one language into another and accurately preserve the "meaning."

One obvious difficulty in trying to do this arises from differences in classification. We can refer to either the foot or the lower leg; the Russians have one word for the foot plus the lower leg. Hungarians have twenty fingers (or toes), for the word is the same for either appendage. To most of us today, a dog is a dog, male or female, but men of an earlier era distinguished sharply between a dog and a bitch. Eskimos make, it is said, many distinctions among snow which in our language would call for descriptions, and for us even these descriptions would have little real content of importance or feeling, because in our lives the distinctions have not been important. Thus, the parts of the world which are common and meaningful to those speaking different languages are often divided into somewhat different classes. It may be impossible to write down in different languages words or simple sentences that specify exactly the same range of experience.

There is a graver problem than this, however. The range of experience to which various words refer is not common among all cultures. What is one to do when faced with the problem of translating a novel containing the phrase, "tying one's shoelace," which as we have noted describes a complicated action, into the language of a shoeless people? An elaborate description wouldn't call up the right thing at all. Perhaps some cultural equivalent (?) could be found. And how should one deal with the fact that "he built a house" means personal tree cutting and adzing in a pioneer novel, while it refers to the employment of an architect and a contractor in a contemporary story?

It is possible to make some sort of translation between closely related languages on a word-for-word or at least phrase-for-phrase basis, though this is said to have led from "out of sight, out of mind" to "blind idiot." When the languages and cultures differ in major respects, the translator has to think what the words mean in terms of objects, actions, or emotions and then express this meaning in the other language. It may be, of course, that the culture with which the language is associated has no close equivalents to the objects or actions described in the passage to be translated. Then the translator is really stuck.

How, oh how is the man who sets out to build a translating machine to cope with a problem such as this? He certainly cannot do so without in some way enabling the machine to deal effectively with what we refer to as understanding. In fact, we see understanding at work even in situations which do not involve translation from one language into another. A screen writer who can quite accurately transfer the essentials of a scene involving a dying uncle in Omsk to one involving a dying father in Dubuque will repeatedly make complete nonsense in trying to rephrase a simple technical statement. This is clearly because he understands grief but not science.

Having grappled painfully with the word *meaning*, we are now faced with the word *understanding*. This seems to have two sides. If we understand algebra or calculus, we can use their manipulations to solve problems we haven't encountered before or to supply proofs of theorems we haven't seen proved. In this sense, understanding is manifested by a power to do, to create, not merely to repeat. To some degree, an electronic computer which proves theorems in mathematical logic which it has not encountered before (as computers can be programmed to do) could perhaps be said to understand the subject. But there is an emotional side to understanding, too. When we can prove a theorem in several ways and fit it together with other theorems or facts in various manners, when we can view a field from many aspects and see how it all fits together, we say that we understand the subject deeply. We attain a warm and confident feeling about our ability to cope with it. Of course, at one time or another most of us have felt the warmth without manifesting the ability. And how disillusioned we were at the critical test!

In discussing language from the point of view of information theory, we have drifted along a tide of words, through the imperfectly charted channels of grammar and on into the obscurities of meaning and understanding. This shows us how far ignorance can take one. It would be absurd to assert that information theory, or anything else, has enabled us to solve the problems of linguistics, of meaning, of understanding, of philosophy, of life. At best, we can perhaps say that we are pushing a little beyond the mechanical constraints of language and getting at the amount of choice that language affords. This idea suggests views concerning the use and function of language, but it does not establish them. The reader may share my freely offered ignorance concerning these matters, or he may prefer his own sort of ignorance.

# CHAPTER VII *Efficient Encoding*

We will never again understand nature as well as Greek philosophers did. A general explanation of common phenomena in terms of a few all-embracing principles no longer satisfies us. We know too much. We must explain many things of which the Greeks were unaware. And, we require that our theories harmonize in detail with the very wide range of phenomena which they seek to explain. We insist that they provide us with useful guidance rather than with rationalizations. The glory of Newtonian mechanics is that it has enabled men to predict the positions of planets and satellites and to understand many other natural phenomena as well; it is surely not that Newtonian mechanics once inspired and supported a simple mechanistic view of the universe at large, including life.

Present-day physicists are gratified by the conviction that all (non-nuclear) physical, chemical, and biological properties of matter can in principle be completely and precisely explained in all their detail by known quantum laws, assuming only the existence of electrons and of atomic nuclei of various masses and charges. It is somewhat embarrassing, however, that the only physical system all of whose properties actually have been calculated exactly is the isolated hydrogen atom.

Physicists are able to predict and explain some other physical phenomena quite accurately and many more semiquantitatively. However, a basic and accurate theoretical treatment, founded on electrons, nuclei, and quantum laws only, without recourse to other experimental data, is lacking for most common thermal, mechanical, electrical, magnetic, and chemical phenomena. Tracing complicated biological phenomena directly back to quantum first principles seems so difficult as to be scarcely relevant to the real problems of biology. It is almost as if we knew the axioms of an important field of mathematics but could prove only a few simple theorems.

Thus, we are surrounded in our world by a host of intriguing problems and phenomena which we cannot hope to relate through one universal theory, however true that theory may be in principle. Until recently the problems of science which we commonly associate with the field of physics have seemed to many to be the most interesting of all the aspects of nature which still puzzle us. Today, it is hard to find problems more exciting than those of biochemistry and physiology.

I believe, however, that many of the problems raised by recent advances in our technology are as challenging as any that face us. What could be more exciting than to explore the potentialities of electronic computers in proving theorems or in simulating other behavior we have always thought of as "human"? The problems raised by electrical communication are just as challenging. Accurate measurements made by electrical means have revolutionized physical acoustics. Studies carried out in connection with telephone transmission have inaugurated a new era in the study of speech and hearing, in which previously accepted ideas of physiology, phonetics, and liguistics have proved to be inadequate. And, it is this chaotic and intriguing field of much new ignorance and of a little new knowledge to which communication theory most directly applies.

If communication theory, like Newton's laws of motion, is to be taken seriously, it must give us useful guidance in connection with problems of communication. It must demonstrate that it has a real and enduring substance of understanding and power. As the name implies, this substance should be sought in the efficient and accurate transmission of information. The substance indeed exists. As we have seen, it existed in an incompletely understood form even before Shannon's work unified it and made it intelligible.

To deal with the matter of accurate transmission of information we need new basic understanding, and this matter will be tackled in the next chapter. The foregoing chapters have, however, put us in a position to discuss some challenging aspects of the efficient transmission of information.

We have seen that in the entropy of an information source measured in bits per symbol or per second we have a measure of the number of binary digits, of off-or-on pulses, per symbol or per second which are necessary to transmit a message. Knowing this number of binary digits required for encoding and transmission, we naturally want a means of actually encoding messages with, at the most, not many more binary digits than this minimum number.

Novices in mathematics, science, or engineering are forever demanding infallible, universal, mechanical methods for solving problems. Such methods are valuable in proving that problems can be solved, but in the case of difficult problems they are seldom practical, and they may sometimes be completely unfeasible. As an example, we may note that an explicit solution of the general cubic equation exists, but no one ever uses it in a practical problem. Instead, some approximate method suited to the type or class of cubics actually to be solved is resorted to.

The person who isn't a novice thinks hard about a specific problem in order to see if there isn't some better approach than a machine-like application of what he has been taught. Let us see how this applies in the case of information theory. We will first consider the case of a discrete source which produces a string of symbols or characters.

In Chapter V, we saw that the entropy of a source can be computed by examining the relative probabilities of occurrence of various long blocks of characters. As the length of the block is increased, the approximation to the entropy gets closer and closer. In a particular case, perhaps blocks 5, or 10, or 100 characters in length might be required to give a very good approximation to the entropy.

We also saw that by dividing the message into successive blocks of characters, to each of which a probability of occurrence can be attached, and by encoding these blocks into binary digits by means of the Huffman code, the number of digits used per character approaches the entropy as the blocks of characters are made longer and longer.

Here indeed is our foolproof mechanical scheme. Why don't we simply use it in all cases?

To see one reason, let us examine a very simple case. Suppose that an information source produces a binary digit, a 1 or a 0, randomly and with equal probability and then follows it with the same digit twice again before producing independently another digit. The message produced by such a source might be:

0 0 0 1 1 1 0 0 0 1 1 1 1 1 1 0 0 0 0 0 0 1 1 1

Would anyone be foolish enough to divide such a message successively into blocks of 1, 2, 3, 4, 5, etc., characters, compute the probabilities of the blocks, encode them with a Huffman code, and note the improvement in the number of binary digits required for transmission? I don't know; it sometimes seems to me that there are no limits to human folly.

Clearly, a much simpler procedure is not only adequate but absolutely perfect. Because of the repetition, the entropy is clearly the same as for a succession of a third as many binary digits chosen randomly and independently with equal probability of 1 or 0. That is, it is ⅓ binary digit per character of the repetitious message. And, we can transmit the message perfectly efficiently simply by sending every third character and telling the recipient to write down each received character three times.

This example is simple but important. It illustrates the fact that we should look for natural structure in a message source, for salient features of which we can take advantage.

The discussion of English text in Chapter IV illustrates this. We might, for instance, transmit text merely as a picture by television or facsimile. This would take many binary digits per character. We would be providing a transmission system capable of sending not only English text, but Cyrillic, Greek, Sanskrit, Chinese, and other text, and pictures of landscapes, storms, earthquakes, and Marilyn Monroe as well. We would not be taking advantage of the elementary and all-important fact that English text is made up of letters.

If we encode English text letter by letter, taking no account of the different probabilities of various letters (and excluding the space), we need 4.7 binary digits per letter. If we take into account the relative probabilities of letters, as Morse did, we need 4.14 binary digits per letter.

If we proceeded mechanically to encode English text more efficiently, we might go on to encoding pairs of letters, sequences of three letters, and so on. This, however, would provide for encoding many sequences of letters which aren't English words. It seems much more sensible to go on to the next larger unit of English text, the word. We have seen in Chapter IV that we would expect to use only about 9 binary digits per word or 1.7 binary digits per character in so encoding English text.

If we want to proceed further, the next logical step would be to consider the structure of phrases or sentences; that is, to take advantage of the rules of grammar. The trouble is that we don't know the rules of grammar completely enough to help us, and if we did, a communication system which made use of these rules would probably be impractically complicated. Indeed, in practical cases it still seems best to encode the letters of English text independently, using at least 5 binary digits per character.

It is, however, important to get some idea of what *could* be accomplished in transmitting English text. To this end, Shannon considered the following communication situation. Suppose we ask a man, using all his knowledge of English, to guess what the next character in some English text is. If he is right we tell him so, and he writes the character down. If he is wrong, we may either tell him what the character actually is or let him make further guesses until he guesses the right character.

Now, suppose that we regard this process as taking place at the transmitter, and say that we have an absolutely identical twin to guess for us at the receiver, a twin who makes just the same mistakes that the man at the transmitter does. Then, to transmit the text, we let the man at the receiver guess. When the man at the transmitter guesses right, so will the man at the receiver. Thus, we need send information to the man at the receiver only when the man at the transmitter guesses wrong and then only enough information to enable the men at the transmitter and the receiver to write down the right character.

Shannon has drawn a diagram of such a communication system, which is shown in Figure VII-1. A predictor acts on the original text. The prediction of the next letter is compared with the actual letter. If an error is noted, some information is transmitted. At the receiver, a prediction of the next character is made from the already reconstructed text. A comparison involving the received signal is carried out. If no error has been made, the predicted character is used; if an error has been made, the "reduced text" information coming in will make it possible to correct the error.

Of course, we don't have such identical twins or any other highly effective identical predictors. Nonetheless, a much simpler but purely mechanical system based on this diagram has been used in transmitting pictures. Shannon's purpose was different, however. By using just one person, and not twins, he was able to find what transmission rate would be required in such a system merely by examining the errors made by the one man in the transmitter situation. The results are summed up in Figure V-4 of Chapter V. A better prediction is made on the basis of the 100 preceding letters than on the basis of the preceding 10 or 15. To correct the errors in prediction, something between 0.6 and 1.3 binary digits per character is required. This tells us that, insofar as this result is correct, the entropy of English text must lie between .6 and 1.3 bits per letter.

A discrete source of information provides a good example for discussion but not an example of much practical importance in communication. The reason is that, by modern standards of electrical communication, it takes very few binary digits or off-or-on pulses to send English text. We have to hurry to speak a few hundred words a minute, yet it is easy to send over a thousand words of text over a telephone connection in a minute or to send 10 million words a minute over a TV channel, and, in principle if not in practice, we could transmit some 50,000 words a minute over a telephone channel and some 50 million words a minute over a TV channel. As a matter of fact, in practical cases we have even retreated from Morse's ingenious code which sends an E faster than a Z. A teletype system uses the same length of signal for any letter.

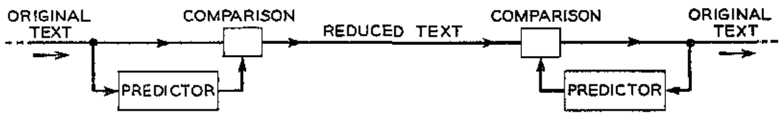


*Fig. VII-1*

Efficient encoding is thus potentially more important for voice transmission than for transmission of text, for voice takes more binary digits per word than does text. Further, efficient encoding is potentially more important for TV than for voice.

Now, a voice or a TV signal is inherently continuous as opposed to English text, numbers, or binary digits, which are discrete. Disregarding capitalization and punctuation, an English character may be any one of the letters or the space. At a given moment, the sound wave or the human voice may have any pressure at all lying within some range of pressures. We have noted in Chapter IV that if the frequencies of such a continuous signal are limited to some bandwidth *B*, the signal can be accurately represented by 2*B* samples or measurements of amplitude per second.

We remember, however, that the entropy per character depends on how many values the character can assume. Since a continuous signal can assume an infinite number of different values at a sample point, we are led to assume that a continuous signal must have an entropy of an infinite number of bits per sample.

This would be true if we required an absolutely accurate reproduction of the continuous signal. However, signals are transmitted to be heard or seen. Only a certain degree of fidelity of reproduction is required. Thus, in dealing with the samples which specify continuous signals, Shannon introduces a *fidelity criterion.* To reproduce the signal in a way meeting the fidelity criterion requires only a finite number of binary digits per sample or per second, and hence we can say that, within the accuracy imposed by a particular fidelity criterion, the entropy of a continuous source has a particular value in bits per sample or bits per second.

It is extremely important to realize that the fidelity criterion should be associated with long stretches of the signal, not with individual samples. For instance, in transmitting a sound, if we make each sample 10 per cent larger, we will merely make the sound louder, and no damage will be done to its quality. If we make a random error of 10 per cent in each sample, the recovered signal will be very noisy. Similarly, in picture transmission an error in brightness or contrast which changes smoothly and gradually across the picture will pass unnoticed, but an equal but random error differing from point to point will be intolerable.

We have seen that we can send a continuous signal by quantizing each sample, that is, by allowing it to assume only certain preassigned values. It appears that 128 values are sufficient for the transmission of telephone-quality speech or of pictures. We must realize, however, that, in quantizing a speech signal or a picture signal sample by sample, we are proceeding in a very unsophisticated manner, just as we are if we encode text letter by letter rather than word by word.

The name *hyperquantization* has been given to the quantization of continuous signals of more than one sample at a time. This is undoubtedly the true road to efficient encoding of continuous signals. One can easily ruin his chances of efficient encoding completely by quantizing the samples at the start. Yet, to hyperquantize a continuous signal is not easy. Samples are quantized independently in present pulse code modulation systems that carry telephone conversations from telephone office to telephone office and from town to town, and in the digital switching systems that provide much long distance switching. Samples are quantized independently in sending pictures back from Mars, Jupiter and farther planets.

In pulse code modulation, the nearest of one of a number of standard levels or amplitudes is assigned to each sample. As an example, if eight levels were used, they might be equally spaced as in a of Figure VII-2. The level representing the sample is then transmitted by sending the binary number written to the right of it.

Some subtlety of encoding can be used even in such a system. Instead of the equally spaced amplitudes of Figure VII-2*a,* we can use quantization levels which are close together for small signals and farther apart for large signals, as shown in Figure VII-2*b.* The reason for doing this is, of course, that our ears are sensitive to a fractional error in signal amplitude rather than to an error of so many dynes below or above average pressure or so many volts positive or negative, in the signal. By such *companding* (*compressing* the high amplitudes at the transmitter and *expanding* them again at the receiver), 7 binary digits per sample can give a signal almost as good as 11 binary digits would if the signal levels transmitted were separated by equal differences in amplitude.

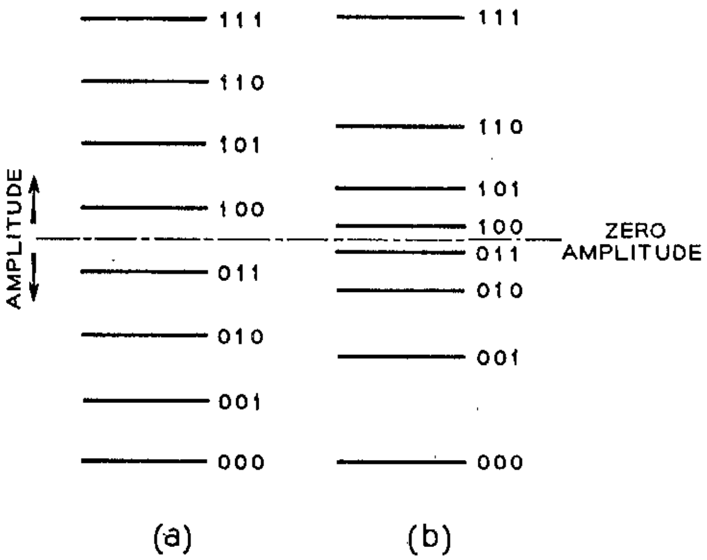


*Fig. VII-2*

To send speech more efficiently than this, we need to examine the characteristics both of speech and of hearing. After all, we require only enough accuracy of transmission to convince the hearer that transmission is good enough.

Efficiency is not everything. A vocoder can transmit only one voice, not two or more at a time. Also, vocoders behave badly when one speaks in the presence of loud noise. Trying to transmit the actual speech waveform more efficiently, or *waveform decoding,* avoids these problems, but 15,000–20,000 binary digits per second are required for acceptable speech.

Figure VII-3 shows the wave forms of several speech sounds, that is, how the pressure of the sound wave or the voltage representing it in a communication system varies with time. We see that many of the wave forms, and especially those for the vowels (*a* through *d*), repeat over and over almost exactly. Couldn't we perhaps transmit just one complete period of variation and use it

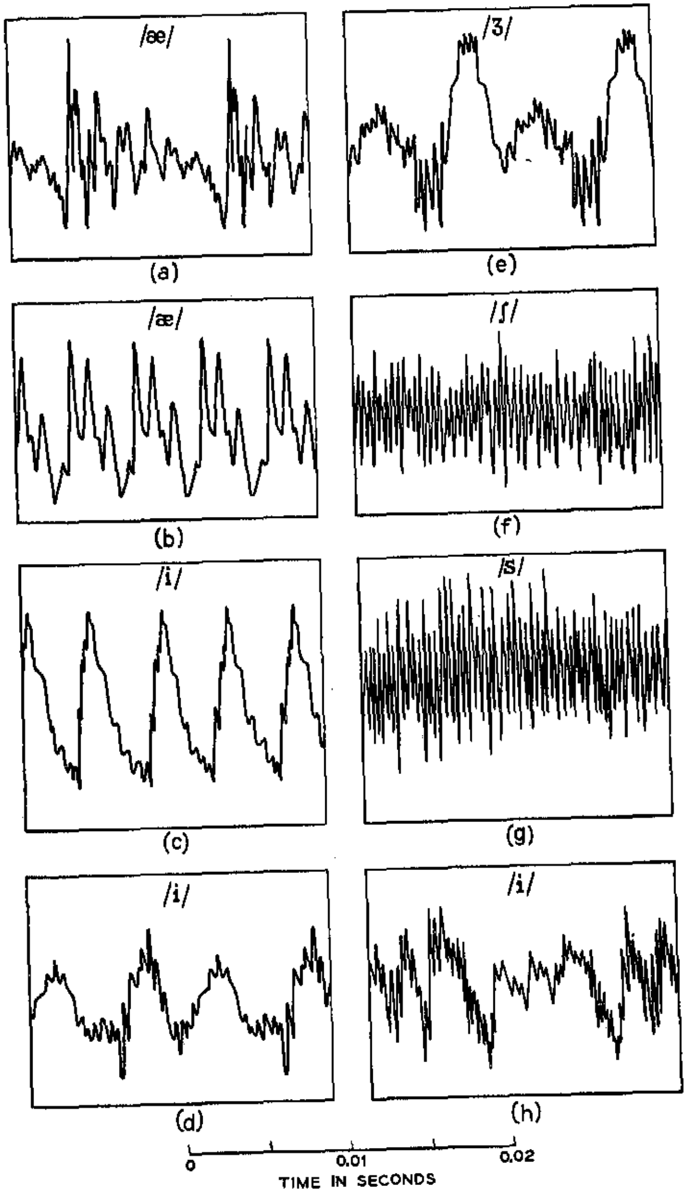


*Fig. VII-3*

to replace several succeeding periods? This is very difficult, for it is hard for a machine to determine just how long a period is in actual speech. It has been tried. The speech reproduced is intelligible but seriously distorted.

If speech is to be encoded efficiently, a much more fundamental approach is required. We must know how great a variety of speech sounds must be transmitted and how effective our sense of hearing is in distinguishing among speech sounds.

The fluctuations of air pressure which constitute the sounds of speech are very rapid indeed, of the order of thousands per second. Our voluntary control over our vocal tracts is exercised at a much lower rate. At the most, we change the manner of production of sounds a few tens of times a second. Thus, speech may well be (and is) simpler than we might conclude by examining the rapidly fluctuating sound waves of speech.

What control do we exercise over our vocal organs? First of all, we control the production of *voiced* sounds by our control over our vocal cords. These are two lips or folds of muscular tissue attached to a cartilaginous box called the *larynx,* which is prominent in man as the Adam's apple. When we are not giving voice to sound, these are wide open. They can be drawn together more or less tightly, so that when air from the lungs is forced through them they emit a sound something like a Bronx cheer. If they are held very tight, the sound has a high pitch; if they are more relaxed, the sound has a lower pitch.

The pulses of air passing the vocal cords contain many frequencies. The mouth and lips act as a complex resonator which emphasizes certain frequencies more than others. What frequencies are emphasized depends on how much and at what position the tongue is raised or humped in the mouth, on whether the soft palate opens the nasal cavities to the mouth and throat, and on the opening of the jaws and the position of the lips.

Particular sounds of voiced speech, which includes vowels and other *continuants,* such as m and r, are formed by exciting the vocal cords and giving particular characteristic shapes to the mouth.

*Stop consonants,* or *plosives,* such as p, b, g, t, are formed by stopping off the vocal passage at various points with the tongue or lips, creating an air pressure, and suddenly releasing it. The vocal cords are used in producing some of these sounds (b, for instance) and not in producing others (p, for instance).

*Fricatives,* such as s and sh, are produced by the passage of air through various constrictions. Sometimes the vocal cords are used as well (in a zh sound, as in azure).

A specification of the movements of the vocal organs would be much more slowly changing than a description of the sound produced. May this not be a clue to efficient encoding of speech?

In the early thirties, long before Shannon's work on information theory, Homer Dudley of the Bell Laboratories invented such a form of speech transmission, which he called the vocoder (from voice coder). The transmitting (analyzer) and receiving (synthesizer) units of a vocoder are illustrated in Figure VII-4.

In the analyzer, an electrical replica of the speech is fed to 16 filters, each of which determines the strength of the speech signal in a particular band of frequencies and transmits a signal to the synthesizer which gives this information. In addition, an analysis is made to determine whether the sound is voiceless (s, f) or voiced (o, u) and, if voiced, what the pitch is.

At the synthesizer, if the sound is voiceless, a hissing noise is produced; if the sound is voiced a sequence of electrical pulses is produced at the proper rate, corresponding to the puffs of air passing the vocal cords of the speaker.

The hiss or pulses are fed to an array of filters, each passing a band of frequencies corresponding to a particular filter in the analyzer. The amount of sound passing through a particular filter in the synthesizer is controlled by the output of the corresponding analyzer filter so as to be the same as that which the analyzer filter indicates to be present in the voice in that frequency range.

This process results in the reproduction of intelligible speech. In effect, the analyzer listens to and analyzes speech, and then instructs the synthesizer, which is an artificial speaking machine, how to say the words all over again with the very pitch and accent of the speaker.

Most vocoders have a strong and unpleasant electrical accent. The study of this has led to new and important ideas concerning what determines and influences speech quality; we cannot afford time to go into this matter here. Even imperfect vocoders can be

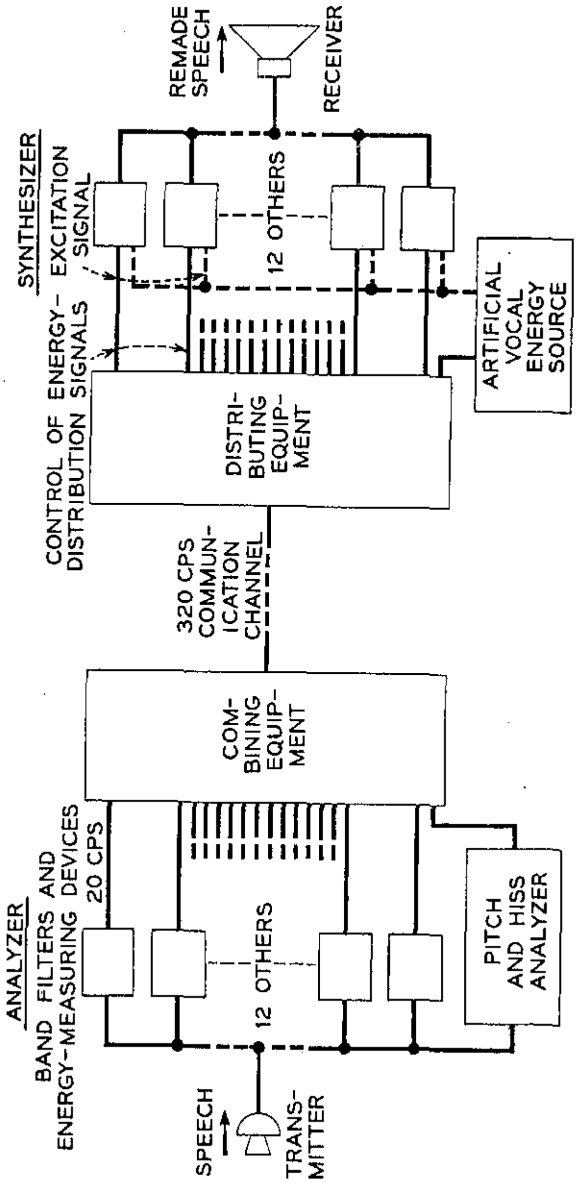


*Fig. VII-4*

very useful. For instance, it is sometimes necessary to resort to enciphered speech transmission. If one merely directly reduces speech to binary digits by pulse code modulation, 30,000 to 60,000 binary digits per second must be sent. By using a vocoder, speech can be sent with about 2,400 binary digits per second.

The channel vocoder of Figure VII-4 is only one example of a large class of devices (we may call them all vocoders, if we wish) that analyze speech and transmit signals which drive a speaking machine. In linear predictive encoding the analysis finds slowly varying coefficients that predict the next speech sample as a weighted sum of several past samples. An error signal can be sent as well, which is used to correct the output of the speaking machine. Linear predictive coding gives very good speech if 9,600 binary digits per second are transmitted, intelligible speech at 2,400 binary digits per second, and barely intelligible speech at 600 binary digits per second.

Various other parameters of speech can be derived from the linear predictive coefficients. The channel signals characteristic of the channel vocoder of Figure VII-4 can be derived from the linear predictive coefficients. So can the resonant frequencies of the vocal tract characteristic of various speech sounds. These resonant frequencies are called *formants*. When we transmit these resonant frequencies and use them to reconstruct speech we say we have a *formant tracking vocoder*. It has been proposed to derive parameters describing the shape of the vocal tract and to transmit these. If, only if, we could use the coefficients to recognize speech sounds, or *phonemes*, and merely transmit their labels, we would have a *phoneme vocoder* that would transmit speech with the efficiency of text.

Let us consider the vocoder for a moment before leaving it.

We note that transmission of voice using even the most economical of vocoders takes many more binary digits per word than transmission of English text. Partly, this is because of the technical difficulties of analyzing and encoding speech as opposed to print. Partly, it is because, in the case of speech, we are actually transmitting information about speech quality, pitch, and stress, and accent as well as such information as there is in text. In other words, the entropy of speech is somewhat greater per word than the entropy of text.

That the vocoder does encode speech more efficiently than other methods depends on the fact that the configuration of the vocal tract changes less rapidly than the fluctuations of the sound waves which the vocal tract produces. Its effectiveness also depends on limitations of the human sense of hearing.

From an electrical point of view, the most complicated speech sounds are the hissing fricatives, such as sh (*f* of Figure VII-3) and s (*g* of Figure VII-3). Furthermore, the wave forms of two s's uttered successively may have quite a different sequence of ups and downs. It would take many binary digits per second to transmit each in full detail. But, to the ear, one s sounds just like another if it has in a broad way the same frequency content. Thus, the vocoder doesn't have to reproduce the s sound the speaker uttered; it has merely to reproduce an s sound that has roughly the same frequency content and hence sounds the same.

We see that, in transmitting speech, the royal road to efficient encoding appears to be the detection of certain simple and important patterns and their recreation at the receiving end. Because of the greater channel capacity required, efficient encoding is even more important in TV transmission than in speech transmission. Can we perhaps apply a similar principle in TV?

The TV problem is much more difficult than the speech transmission problem. Partly, this is because the sense of sight is inherently more detailed and discriminating than the sense of hearing. Partly, though, it is because many sorts of pictures from many sources are transmitted by TV, while speech is all produced by the same sort of vocal apparatus.

In the face of these facts, is some vocoder-like way of transmitting pictures possible if we confine ourselves to one sort of picture source, for instance, the human face?

One can conceive of such a thing. Imagine that we had at the receiver a sort of rubbery model of a human face. Or we might have a description of such a model stored in the memory of a huge electronic computer. First, the transmitter would have to look at the face to be transmitted and "make up" the model at the receiver in shape and tint. The transmitter would also have to note the sources of light and reproduce these in intensity and direction at the receiver. Then, as the person before the transmitter talked, the transmitter would have to follow the movements of his eyes, lips and jaws, and other muscular movements and transmit these so that the model at the receiver could do likewise. Such a scheme might be very effective, and it could become an important invention if anyone could specify a useful way of carrying out the operations I have described. Alas, how much easier it is to say what one would like to do (whether it be making such an invention, composing Beethoven's tenth symphony, or painting a masterpiece on an assigned subject) than it is to do it.

In our day of unlimited science and technology, people's unfulfilled aspirations have become so important to them that a special word, popular in the press, has been coined to denote such dreams. That word is *breakthrough*. More rarely, it may also be used to describe something, usually trivial, which has actually been accomplished.

If we turn from such dreams of the future, we find that all actual picture-transmission systems follow a common pattern. The picture or image to be transmitted is *scanned* to discover the brightness at successive points. The scanning is carried out along a sequence of closely spaced lines. In color TV, three images of different colors are scanned simultaneously. Then, at the receiver, a point of light whose intensity varies in accord with the signal from the transmitter paints out the picture in light and shade, following the same line pattern. So far all practical attempts at efficient encoding have started out with the signal generated by such a scanning process.

The outstanding efficient encoding scheme is that used in color TV. The brightness of a color TV picture has very fine detail; the pattern of color has very much less detail. Thus, color TV of almost the same detail as monochrome TV can be sent over the same channel as is used for monochrome. Of course, color TV uses an analog signal; the picture is not reduced to discrete on-or-off pulses.

Increasingly, pulse code modulation will be used to transmit all sorts of signals, including television signals. The picture to be transmitted will be scanned in a conventional way, but its brightness will be encoded as a succession of binary numbers that specify the brightnesses of a succession of discrete picture elements or pixels that lie along each scanning line. This is how pictures were sent back from Mars by the Mariner lander, and from Jupiter and its moons by the Voyager spacecraft.

All recent work aimed at encoding television efficiently is digital. It deals with successions of binary numbers that represent successive pixel brightnesses.

In large parts of a TV picture the brightness changes gradually and smoothly from pixel to pixel. In such areas of the picture, a good prediction can be made of the brightness of the next pixel from the brightness of preceding pixels in the same line, and perhaps in the preceding line. At the receiver we need know only the error in such a prediction, so we need transmit only the small difference between the true brightness and a brightness which we predict at the receiver as well as at the transmitter. Of course, in "busy" portions of the picture, prediction will be poor, and the brightness difference that must be sent will be great.

We can transmit brightness differences most efficiently by using a Huffman code, with short code words for more frequently occurring small brightness differences and long code words for less frequently occurring, large brightness differences. If we do this, the binary digits of the coded differences will be generated at an uneven rate, at a slow rate when smooth portions of the picture are scanned and at a faster rate when busy portions of the picture are scanned. In order to transmit the binary digits at a constant rate, the digits must be fed into a *buffer*, which stores the incoming digits and feeds them out at a constant rate equal to the average rate at which they come in. A similar buffer must be used at the receiving end.

By means of such *intraframe* encoding, the number of binary digits per second needed to transmit a good TV picture can be reduced to ½ to ⅓ of the number of binary digits used in initially encoding the pixel brightnesses.

Much greater gains can be made through *interframe* encoding, in which the pixel brightnesses of the whole previous TV picture are stored and used in predicting the brightness of the next pixel to be sent. This is particularly effective in transmitting pictures of people against a fixed background, for the brightnesses of pixels in the background don't change from frame to frame.

Even more elaborate experimental schemes make use of the fact that when a figure in front of a background moves, it moves as a whole. Thus, the brightnesses of the pixels in the moving figure can be predicted from the brightnesses of pixels which are a constant distance away in the previous frame.

If each pixel of a TV picture is represented by 8 binary digits (a very good picture), the picture can be transmitted by sending around 100 million binary digits per second. By intraframe encoding this can be reduced to perhaps 32 million. With interframe coding this has been reduced to as little as 6 million. A reduction to 1.5 million seems conceivable for such pictures as the head of a person against a fixed background.

The *transform* method is another approach to the efficient transmission of TV pictures. In the transform method, the pattern of pixel brightnesses that make up the TV picture, or some portion of it, is represented as the sum of a chosen set of standardized patterns whose amplitudes are transmitted with chosen accuracies.

Reviewing what has been said, we see that there are three important principles in encoding signals efficiently: (1) Don't encode the signal one sample or one character at a time; encode a considerable stretch of a signal at a time (hyperquantization); (2) take into account the limitations on the source of the signal; (3) take into account any inabilities of the eye or the ear to detect errors in a reconstruction of the signal.

The vocoder illustrates these principles excellently. The fine temporal structure of the speech wave is not examined in detail. Instead, a description specifying the average intensities over certain ranges of frequencies is transmitted, together with a signal which tells whether the speech is voiced or unvoiced and, if it is voiced, what its pitch is. This description of a signal is efficient because the vocal organs don't change position rapidly in producing speech. At the receiver, the vocoder generates a speech signal which doesn't resemble the original speech signal in fine detail but sounds like the original speech signal, because of the natural limitations of our hearing.

The vocoder is a sort of paragon of efficient transmission devices. Next perhaps comes color TV, in which the variations of color over the picture are defined much less sharply than variations of intensity are. This takes advantage of the eyes' inability to see fine detail in color patterns.

Beyond this, the present art of communication has had to make use of means which, because they do not encode long stretches of signal at a time, must, according to communication theory, be rather inefficient.

Still, efficient encoding is potentially important. This is especially so in the case of the transmission of relatively broad-band signals (TV or even voice signals) over very expensive circuits, such as transoceanic telephone cables.

No doubt much ingenuity will be spent in efficient encoding in the future, and many startling results will be attained. But we should perhaps beware of going too far.

Imagine, for instance, that we send English text letter by letter. If we make an error in sending a few letters we can still make some sense out of the text:

Hore I hove reploced a few vowols by o.

We can even replace the vowels by x's and read with some facility:

Hxrx X hxvx rxplxcxd thx vxwxls bx x.

It is more efficient to encode English text word by word. In this case, if an error is made in transmission, we are not tipped off by finding a misspelled word. Instead, one word is replaced by another. This might have embarrassing results. Suppose it changed "The President is a good Republican" to "The President is a good Communist" (or donkey, or poltroon, or many other nouns).

We might still detect an error by the fact that the word was inappropriate. But suppose we used a more refined encoding scheme that could reproduce grammatical utterances only. Then we would have little chance of detecting an error in transmission.

English text, and most other information sources are *redundant* in that the messages they produce give many clues to the recipient. A few errors caused by replacing one letter by another don't destroy the message because we can infer it from other letters which are transmitted correctly. Indeed, it is only because of this redundancy that anyone can read my handwriting. When a continuous signal is sent a sample at a time, a few errors in sample amplitude result in a few clicks in sound transmission or in a few specks in picture transmission.

Our ideal so far has been to remove this redundancy, so that we transmit the absolutely minimum number of clues by means of which the message can be reconstructed. But we see that if we do this with perfect success, any error in transmission will send, not a distorted message, but a false and misleading message. If we fall a little short of the ideal, an error may produce merely a terrible garble.

We all know that there is some noise in electrical communication —a hiss in the background on radio and a little snow at least in TV. That such noise is an inevitable fact of nature we must accept. Is this going to vitiate in principle our grand plan to encode the messages from a signal source into scarcely more binary digits than the entropy of the source?

This is the subject that we will consider in the next chapter.

# CHAPTER VIII *The Noisy Channel*

It is hard to put oneself in the place of another, and, especially, it is hard to put oneself in the place of a person of an earlier day. What would a Victorian have thought of present-day dress? Were Newton's laws of motion and of gravitation as astonishing and disturbing to his contemporaries as Einstein's theory of relativity appears to have been to his? And what is disturbing about relativity? Present-day students accept it, not only without a murmur, but with a feeling of inevitability, as if any other idea must be very odd, surprising, and inexplicable.

Partly, this is because our attitudes are bred of our times and surroundings. Partly, in the case of science at least, it is because ideas come into being as a response to new or better-phrased questions. We remember that according to Plato, Socrates drew a geometrical proof from a slave simply by means of an ingenious sequence of questions. Those who have not seriously asked themselves a particular question are not likely to have come upon the proper answer, and, sometimes, when the question is phrased with the answer in mind, the answer appears to be obvious.

Those interested in communication have been aware from the very beginning that communication circuits or channels are imperfect. In telephony and radio, we hear the desired signal against a background of noise, which may be strong or faint and which may vary in quality from the crackling of static to a steady hiss.

In TV, the picture is overlaid faintly or strongly with an ever-changing granular "snow." In teletypewriter transmission, the received character may occasionally differ from that transmitted.

Suppose that one had questioned a communication engineer about this general problem of "noise" in 1945. One might have asked, "What can one *do* about noise?" The engineer might have answered, "You can increase the transmitter power or make the receiver less noisy. And be sure that the receiver is insensitive to disturbances with frequencies other than the signal frequencies."

One might have persisted, "Can't one do anything else?" The engineer might have answered, "Well, by using frequency modulation, which takes a very large band width, one can reduce the effect of noise."

Suppose, however, that one had asked, "In teletypewriter systems, noise may cause some received characters to be wrong; how can one guard against this?" The engineer could and might perhaps have answered, "I know that if I use five off-or-on pulses to represent a decimal digit and assign to the decimal digits only such sequences as all have two ons and three offs, I can often tell when an error has been made in transmission, for when errors are made the received sequence may have other than 2 ons."

One might have pursued the matter further with, "If the teletypewriter circuit does cause errors is there any way that one can get the correct message to the destination?" The engineer might have answered, "I suppose you can if you repeat it enough times, but that's very wasteful. You'd better fix the circuit."

Here we are getting pretty close to questions that just hadn't been asked before Shannon asked them. Nonetheless, let us go on and imagine that one had said, "Suppose that I told you that by properly encoding my message, I can send it over even a noisy channel with a completely negligible fraction of errors, a fraction smaller than any assignable value. Suppose that I told you that, if the sort of noise in the channel is known and if its magnitude is known, I can calculate just how many characters I can send over the channel per second and that, if I send any number fewer than this, I can do so virtually without error, while if I try to send more, I will be bound to make errors."

The engineer might well have answered, "You'd sure have to show me. I never thought of things in quite that way before, but what you say seems extremely improbable. Why, every time the noise increases, the error rate increases. Of course, repeating a message several times does work better when there aren't too many errors. But, it is always very costly. Maybe there's something in what you say, but I'd be awfully surprised if there was. Still, the way you put it . . ."

Whatever we may imagine concerning an engineer benighted in the days of error, mathematicians and engineers who have survived the transition all feel that Shannon's results concerning the transmission of information over a noisy channel were and still are very surprising. Yet I have known an intelligent layman to see nothing remarkable in Shannon's results. What is one to think of this?

Perhaps the best course is merely to describe and explain the problem of the noisy channel as we *now* understand it, raising and answering questions that, however natural and inevitable they now seem, belong in their trend and content to the post-Shannon era. The reader can be surprised or not as he chooses.

So far we have discussed both simple and complex means for encoding text and numbers for efficient transmission. We have noted further that any electrical signal of limited band width $W$ can be represented by $2W$ amplitudes or samples per second, measured or taken at intervals $1/2W$ seconds apart. We have seen that, by means of pulse code modulation, we can use some number, around 7, of binary digits to represent adequately the amplitude of any sample. Thus, by using pulse code modulation or some more complicated and more efficient scheme, we can transmit speech or picture signals by means of a sequence of binary digits or off-or-on or positive-or-negative pulses of current.

All of this works perfectly if the recipient of the message receives the same signal that the sender transmits. The actual facts are different. Sometimes he receives a 0 when a 1 is transmitted, and sometimes he receives a 1 when a 0 is transmitted. This can happen through the malfunction of electrical relays in a slow-speed telegraph circuit or through the malfunction of vacuum tubes or transistors in a higher speed circuit. It can also happen because of interfering signals or noise, either noise from man-made apparatus, or noise from magnetic storms.

We can easily see in a simple case how errors can occur because of the admixture of noise with a signal. Imagine that we want to send a large number of binary digits, 0 or 1, per second over a wire by means of an electrical signal. We may represent the signal conveying these digits by the succession of samples *s* of Figure VIII-1, each of which will be +1 or −1. Here we have a succession of positive and negative voltages which represent the digits 1 0 1 1 1 0 0 1 0.

Now suppose a random noise voltage, which may be either positive or negative, is added to the signal. We can represent this also by a number of noise samples *n* of Figure VIII-1 taken simultaneously with the signal samples. The signal plus the noise is obtained by adding the signal and the noise samples and is shown as $s + n$ in Figure VIII-1.

If we interpret a positive signal-plus-noise in the received message as a 1 and a negative signal-plus-noise as a 0, then the received message will be represented by the digits *r* of Figure VIII-1. Thus, errors in transmission, as indicated, occur in positions 2, 3, and 7.

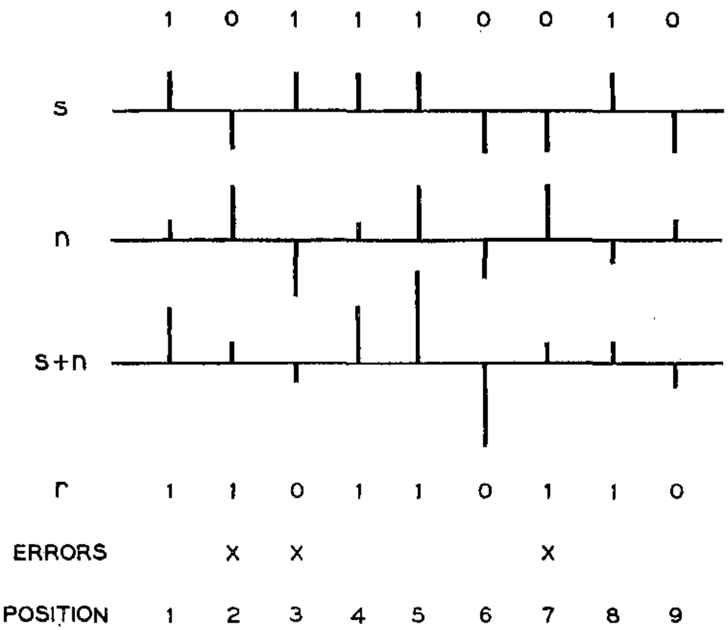


*Fig. VIII-1*

The effect of such errors in transmission can range from annoying to dangerous. In speech or picture transmission by means of simple coding schemes, they result in clicks, hissing noises, or "snow." If more efficient, block encoding schemes are used (hyperquantization) the effects of errors will be more pronounced. In general, however, we may expect the most dangerous effects of errors in the transmission of text.

In the transmission of English text by conventional means, errors merely put a wrong letter in here and there. The text is so redundant that we catch such errors by eye. However, when type is set remotely by teletypewriter signals, as it is, for instance, in the simultaneous printing of news magazines in several parts of the country, even errors of this sort can be costly.

When numbers are sent errors are much more serious. An error might change $1,000 into $9,000. If the error occurred in a program intended to make an electronic computer carry out a complicated calculation, the error could easily cause the whole calculation to be meaningless.

Further, we have seen that, if we encode English text or any other signal very efficiently, so as largely to remove the redundancy, an error can cause a gross change in the meaning of the received signal.

When errors are very important to us, how indeed may we guard against them? One way would be to send every letter twice or to send every binary digit used in transmitting a letter or a number twice. Thus, in transmitting the binary sequence 1 0 1 0 0 1 1 0 1, we might send and receive as follows:

| | |
|---|---|
| *sent* | 1 1 0 0 1 1 0 0 0 0 1 1 1 1 0 0 1 1 |
| *received* | 1 1 0 0 1 1 0 0 0 1 1 1 1 1 0 0 1 1 |
| | × <br> error |

For a given rate of sending binary digits, this will cut our rate of transmitting information in half, for we have to pause and retransmit every digit. However, we can now see from the received signal than an error has occurred at the marked point, because instead of a pair of like digits, 0 0 or 1 1, we have received a pair of unlike digits, 0 1. We don't know whether the correct, transmitted pair was 0 0 or 1 1. We have *detected* the error, but we have not *corrected* it.

If errors aren't too frequent, that is, if the chance of two errors occurring in the transmission of three successive digits is negligible, we can correct as well as detect an error by transmitting each digit three times, as follows:

```
sent      1 1 1 0 0 0 1 1 1 0 0 0 0 0 0 1 1 1 1 1 1
received  1 1 1 0 0 0 1 0 1 0 0 0 0 0 0 1 1 1 1 1 1
                         ^
                       error
```

We have now cut our rate of transmission to one-third, because we have to pause and retransmit each digit twice. However, we can now correct the error indicated by the fact that the digits in the indicated group 1 0 1 are not all the same. If we assume that there was only one error in the transmission of this group of digits, then the transmitted group must have been 1 1 1, representing 1, rather than 0 0 0, representing 0.

We see that a very simple scheme of repeating transmitted digits can detect or even correct infrequent errors of transmission. But how costly it is! If we use this means of error correction or detection, even when almost all of the transmitted digits are correct we have to cut our rate of transmission in half by repeating digits in order just to detect errors, and we have to cut our rate of transmission to one-third by transmitting each digit three times in order to get error correction. Moreover, these schemes won't work if errors are frequent enough so that more than one will sometimes occur in the transmission of two or three digits.

Clearly, this simple approach will never lead to a sound understanding of the possibility of error correction. What is required is a deep and powerful mathematical attack. This is just what Shannon provided in discovering and proving his fundamental theorem for the noisy channel. It is the course of his reasoning that we are about to follow.

In formulating an abstract and general model of noise or errors, we will deal with the case of a discrete communication system which transmits some group of characters, such as the digits from 0 to 9 or the letters of the alphabet. For convenience, let us consider a system for transmitting the digits 0 through 9. This is illustrated in Figure VIII-2. At the left we have a number of little circles labeled with the digits; we may regard these little circles as push-buttons. To ‛he right we have a number of little circles, again labeled with the digits. We may regard these as lights. When we push a digit button at the transmitter to the left, some digit light lights up at the receiver to the right.

If our communication system were noiseless, pushing the 0 button would always light the 0 light, pushing the 1 button would always light the 1 light, and so on. However, in an imperfect or noisy communication system, pushing the 4 button, for instance, may light the 0 light, or the 1 light, or the 2 light, or any other light, as shown by the lines radiating from the 4 button in Figure VIII-2. In a simple, noisy communication system, we can say that when we press a button the light which lights is a matter of chance,

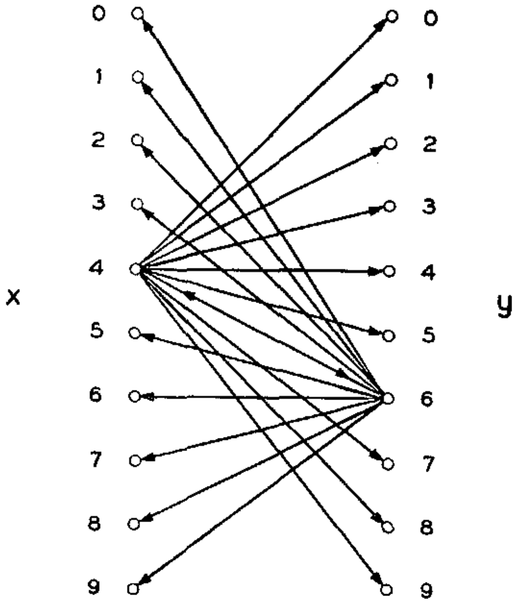


*Fig. VIII-2*

independent of what has gone before and that, if the 4 button is pressed, there is some probability $p_4(6)$ that the 6 light will light, and so on.

If the sender can't be sure which light will light when he presses a particular button, then the recipient of the message can't be sure which button was pressed when a particular light lights. This is indicated by the arrows from light 6 to various buttons on the left. If, for instance, light 6 lights, there is some probability $p_6$ (4) that button 4 was pressed, and so on. Only for a noiseless system will $p_6(6)$ be unity and $p_6(4)$, $p_6(9)$, etc., be zero.

The diagram of Figure VIII-2 would be too complicated if all possible arrows were put in, and the number of probabilities is too great to list, but I believe that the general idea of the degree and nature of uncertainty of the character received when the sender tries to send a particular character and the uncertainty of the character sent when the recipient receives a particular character, have been illustrated. Let us now consider this noisy communication channel in a rather general way. In doing so we will represent by $x$ all of the characters sent and by $y$ all of the characters received.

The characters $x$ are just the characters generated by the message source from which the message comes. If there are $m$ of these characters and if they occur independently with probabilities $p(x)$, then we know from Chapter V that the entropy $H(x)$ of the message source, the rate at which the message source generates information, must be

$$H(x) = \sum_{x=1}^{m} - p(x) \log p(x) \tag{8.1}$$

We can regard the output of the device, which we designate by $y$, as another message source. The number of lights need not be equal to the number of buttons, but we will assume that it is, so that there are $m$ lights. The entropy of the output will be

$$H(y) = \sum_{y=1}^{m} - p(y) \log p(y) \tag{8.2}$$

We note that while $H(x)$ depends only on the input to the communication channel, $H(y)$ depends both on the input to the channel and on the errors made in transmission. Thus, the probability of

receiving a 4 if nothing but a 4 is ever sent is different from the probability of receiving a 4 if transmitting buttons are pressed at random.

If we imagine that we can see both the transmitter and the receiver, we can observe how often certain combinations of $x$ and $y$ occur; say, how often 4 is sent and 6 is received. Or, knowing the statistics of the message source and the statistics of the noisy channel, we can compute such probabilities. From these we can compute another entropy.

$$H(x, y) = \sum_{x=1}^{m} \sum_{x=1}^{m} - p(x, y) \log p(x, y) \tag{8.3}$$

This is the uncertainty of the combination of $x$ and $y$.

Further, we can say, suppose that we know $x$ (that is, we know what key was pressed). What are the probabilities of various lights lighting (as illustrated by the arrows to the right in Figure VIII-2)? This leads to an entropy,

$$H_x(y) = \sum_{x=1}^{m} \sum_{y=1}^{m} - p(x) p_x(y) \log p_x(y) \tag{8.4}$$

This is a *conditional* entropy of uncertainty. Its form is reminiscent of the entropy of a finite-state machine. As in that case, we multiply the uncertainty for a given condition (state, value of $x$) by the probability that that condition (state, value of $x$) will occur and sum over all conditions (states, values of $x$).

Finally, suppose we know what light lights. We can say what the probabilities are that various buttons were pressed. This leads to another conditional entropy

$$H_y(x) = \sum_{y=1}^{m} \sum_{x=1}^{m} - p(y) p_y(x) \log p_y(x) \tag{8.5}$$

This is the sum over $y$ of the probability that $y$ is received times the uncertainty that $x$ is sent when $y$ is received.

These conditional entropies depend on the statistics of the message source, because they depend on how often $x$ is transmitted or how often $y$ is received, as well as on the errors made in transmission.

The entropies listed above are best interpreted as uncertainties involving the characters generated by the message source and the characters received by the recipient. Thus:

$H(x)$ is the uncertainty as to $x$, that is, as to which character will be transmitted.

$H(y)$ is the uncertainty as to which character will be received in the case of a given message source and a given communication channel.

$H(x, y)$ is the uncertainty as to when $x$ will be transmitted and $y$ received.

$H_x(y)$ is the uncertainty of receiving $y$ when $x$ is transmitted. It is the average uncertainty of the sender as to what will be received.

$H_y(x)$ is the uncertainty that $x$ was transmitted when $y$ is received. It is the average uncertainty of the message recipient as to what was actually sent.

There are relations among these quantities:

$$H(x, y) = H(x) + H_x(y) \qquad (8.6)$$

That is, the uncertainty of sending $x$ and receiving $y$ is the uncertainty of sending $x$ plus the uncertainty of receiving $y$ when $x$ is sent.

$$H(x, y) = H(y) + H_y(x) \qquad (8.7)$$

That is, the uncertainty of receiving $y$ and sending $x$ is the uncertainty of receiving $y$ plus the uncertainty that $x$ was sent when $y$ was received.

We see that when $H_x(y)$ is zero, $H_y(x)$ must be zero, and $H(y)$ is then just $H(x)$. This is the case of the noiseless channel, for which the entropy of the received signal is just the same as the entropy of the transmitted signal. The sender knows just what will be received, and the recipient of the message knows just what was sent.

The uncertainty as to which symbol was transmitted when a given symbol is received, that is, $H_y(x)$ seems a natural measure of the information lost in transmission. Indeed, this proves to be the case, and the quantity $H_y(x)$ has been given a special name; it is called the *equivocation* of the communication channel. If we take $H(x)$ and $H_y(x)$ as entropies in bits per second, the rate $R$ of transmission of information over the channel can be shown to be, in bits per second,

$$R = H(x) - H_y(x) \qquad (8.8)$$

That is, the rate of transmission of information is the source rate or entropy less the equivocation. It is the entropy of the message as sent less the uncertainty of the recipient as to what message was sent.

The rate is also given by

$$R = H(y) - H_x(y) \qquad (8.9)$$

That is, the rate is the entropy of the received signal $y$ less the uncertainty that $y$ was received when $x$ was sent. It is the entropy of the message as received less the sender's uncertainty as to what will be received.

The rate is also given by

$$R = H(x) + H(y) - H(x, y) \qquad (8.10)$$

The rate is the entropy of $x$ plus the entropy of $y$ less the uncertainty of occurrence of the combination $x$ and $y$. We will note from 8.3 that for a noiseless channel, since $p(x, y)$ is zero except when $x = y$, and $H(x, y) = H(x) = H(y)$. The information rate is just the entropy of the information source, $H(x)$.

Shannon makes expression 8.8 for the rate plausible by means of the sketch shown in Figure VIII-3. Here we assume a system in which an observer compares transmitted and received signals and then sends correction data by means of which the erroneous received signal is corrected. Shannon is able to show that in order to correct the message, the entropy of the correction signal must be equal to the equivocation.

We see that the rate $R$ of relation 8.8 depends both on the channel and on the message source. How can we describe the *capacity* of a noisy or imperfect channel for transmitting information? We can choose the message source so as to make the rate $R$ *as large as possible* for a given channel. This maximum possible rate of transmission for the channel is called the *channel capacity*

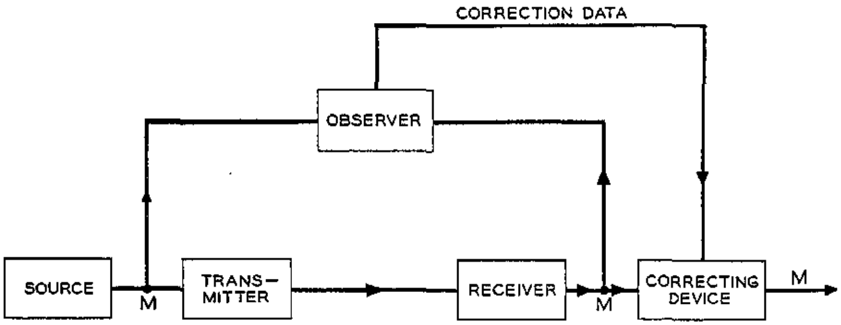


*Fig. VIII-3*

*C.* Shannon's fundamental theorem for a noisy channel involves the channel capacity *C*. It says:

> Let a discrete channel have a capacity C and a discrete source the entropy per second H. If H $<$ C there exists a coding system such that the output of the source can be transmitted over the channel with an arbitrarily small frequency of errors (or an arbitrarily small equivocation). If H $>$ C it is possible to encode the source so that the equivocation is less than H $-$ C $+$ $\varepsilon$, where $\varepsilon$ is arbitrarily small. There is no method of encoding which gives an equivocation less than H $-$ C.

This is a precise statement of the result which so astonished engineers and mathematicians. As errors in transmission become more probable, that is, as they occur more frequently, the channel capacity as defined by Shannon gradually goes down. For instance, if our system transmits binary digits and if some are in error, the channel capacity *C*, that is, number of bits of information we can send per binary digit transmitted, decreases. But the channel capacity decreases *gradually* as the errors in transmission of digits become more frequent. To achieve transmission with as few errors as we may care to specify, we have to reduce our rate of transmission so that it is equal to or less than the channel capacity.

How are we to achieve this result? We remember that in efficiently encoding an information source, it is necessary to lump many characters together and so to encode the message a long block of characters at a time. In making very efficient use of a noisy channel, it is also necessary to deal with sequences of received

characters, each many characters long. Among such blocks, only certain transmitted and received sequences of characters will occur with other than a vanishing probability.

In proving the fundamental theorem for a noisy channel, Shannon finds the average frequency of error for all possible codes (for all associations of particular input blocks of characters with particular output blocks of characters), when the codes are chosen at random, and he then shows that when the channel capacity is greater than the entropy of the source, the error rate averaged over all of these encoding schemes goes to zero as the block length is made very long. If we get this good a result by averaging over all codes chosen at random, then there must be some one of the codes which gives this good a result. One information theorist has characterized this mode of proof as weird. It is certainly not the sort of attack that would occur to an uninspired mathematician. The problem isn't one which would have occurred to an uninspired mathematician, either.

The foregoing work is entirely general, and hence it applies to all problems. I think it is illuminating, however, to return to the example of the binary channel with errors, which we discussed early in this chapter and which is illustrated in Figure VIII-1, and see what Shannon's theorem has to say about this simple and common case.

Suppose that the probability that over this noisy channel a 0 will be received as a 0 is equal to the probability $p$ that a 1 will be received as a 1. Then the probability that a 1 will be received as a 0 or a 0 as a 1 must be $(1 - p)$. Suppose further that these probabilities do not depend on past history and do not change with time. Then, the proper abstract representation of this situation is a symmetric binary channel (in the manner of Figure VIII-2) as shown in Figure VIII-4.

Because of the symmetry of this channel, the maximum information rate, that is, the channel capacity, will be attained for a message source such that the probability of sending a 1 is equal to the probability of sending a zero. Thus, in the case of $x$ (and, because the channel is symmetrical, in the case of $y$ also)

$$p(1) = p(0) = ½$$

We already know that under these circumstances

$$\begin{aligned} H(x) &= H(y) \\ &= -\left(\tfrac{1}{2}\log\tfrac{1}{2} + \tfrac{1}{2}\log\tfrac{1}{2}\right) \\ &= 1 \text{ bit per symbol} \end{aligned}$$

What about the conditional probabilities? What about the equivocation, for instance, as given by 8.5? Four terms will contribute to this conditional entropy. The sources and contributions are:

The probability that 1 is received is ½. When 1 is received, the probability that 1 was sent is $p$ and the probability that 0 was sent is $(1 - p)$. The contribution to the equivocation from these events is:

$$\tfrac{1}{2}(-p \log p - (1-p)\log(1-p))$$

There is a probability of ½ that 0 is received. When 0 is received, the probability that 0 was sent is $p$ and the probability that 1 was sent is $(1 - p)$. The contribution to the equivocation from these events is:

$$\tfrac{1}{2}(-p \log p - (1-p)\log(1-p))$$

Accordingly, we see that, for the symmetrical binary channel, the equivocation, the sum of these terms, is

$$H_y(x) = -p \log p - (1-p)\log(1-p)$$

Thus the channel capacity $C$ of the symmetrical binary channel is, from 8.8,

$$C = 1 + p \log p + (1-p)\log(1-p)$$

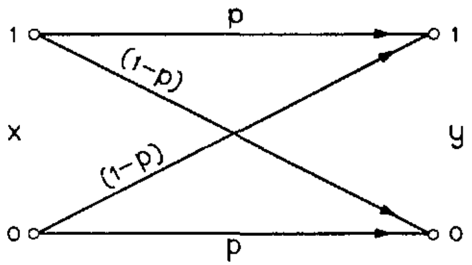


*Fig. VIII-4*

We should note that this channel capacity $C$ is just unity less the function plotted against $p$ in Figure V-1. We see that if $p$ is ½, the channel capacity is 0. This is natural, for in this case, if we receive a 1, it is equally likely that a 1 or a 0 was transmitted, and the received message does nothing to resolve our uncertainty as to what digit the sender sent. We should also note that the channel capacity is the same for $p = 0$ as for $p = 1$. If we consistently receive a 0 when we transmit a 1 and a 1 when we transmit a 0, we are just as sure of the sender's intentions as if we always get a 1 for a 1 and a 0 for a 0.

If, on the average, 1 digit in 10 is in error, the channel capacity is reduced to .53 of its value for errorless transmission, and for one error in 100 digits, the channel capacity is reduced to .92 merely.

The writer would like to testify at this point that the simplicity of the result we have obtained for the symmetrical binary channel is in a sense misleading (it was misleading to the writer at least). The expression for the optimum rate (channel capacity) of an unsymmetrical binary channel in which the probability that a 1 is received as a 1 is $p$ and the probability that a 0 is received as a 0 is a different number $q$ is a mess, and more complicated channels must offer almost intractable problems.

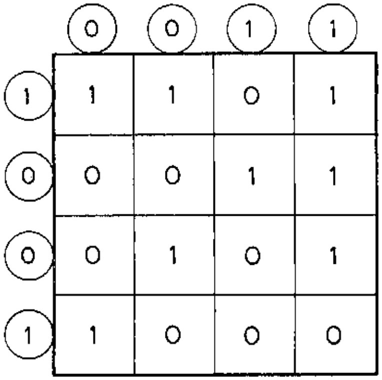
Perhaps for this reason as well as for its practical importance, much consideration has been given to transmission over the symmetrical binary channel. What sort of codes are we to use in order to attain errorless transmission over such a channel? Examples devised by R. W. Hamming were mentioned by Shannon in his original paper. Later, Marcel J. E. Golay published concerning error-correcting codes in 1949, and Hamming published his work in 1950. We should note that these codes were devised subsequent to Shannon's work. They *might,* I suppose, have been devised before, but it was only when Shannon showed error-free transmission to be possible that people asked, "How can we achieve it?"

We have noted that to get an efficient correction of errors, the encoder must deal with a long sequence of message digits. As a simple example, suppose we encode our message digits in blocks of 16 and add after each block a sequence of *check digits* which enable us to detect a single error in *any one* of the digits, message digits or check digits. As a particular example, consider the sequence of message digits 1 1 0 1 0 0 1 1 0 1 0 1 1 0 0 0. To find the appropriate check digits, we write the 0's and 1's constituting the message digits in the 4 by 4 grid shown in Figure VIII-5. Associated with each row and each column is a circle. In each circle is a 0 or a 1 chosen so as to make the total number of 1's in the column or row (including the circle as well as the squares) even. Such added digits are called *check digits.* For the particular assortment of message digits used as an example, together with the appropriately chosen check digits, the numbers of 1's in successive columns (left to right) and 2, 2, 2, 4, all being even numbers, and the numbers of 1's in successive rows (top to bottom) are 4, 2, 2, 2, which are again all even.

What happens if a single error is made in the transmission of a message digit among the 16? There will be an odd number of ones *in a row and in a column.* This tells us to change the message digit where the row and column intersect.

What happens if a single error is made in a check digit? In this case there will be an odd number of ones *in a row or in a column.* We have detected an error, but we see that it was not among the message digits.

The total number of digits transmitted for 16 message digits is 16 + 8, or 24; we have increased the number of digits needed in the ratio 24/16, or 1.5. If we had started out with 400 message digits, we would have needed 40 check digits and we would have increased the number of digits needed only in the ratio of 440/400,



|   | 0 | 0 | 1 | 1 |
|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 1 |
| 0 | 0 | 0 | 1 | 1 |
| 0 | 0 | 1 | 0 | 1 |
| 1 | 1 | 0 | 0 | 0 |

*Fig. VIII-5*

or 1.1. Of course, we would have been able to correct only one error in 440 rather than one error in 24.

Codes can be devised which can be used to correct larger numbers of errors in a block of transmitted characters. Of course, more check digits are needed to correct more errors. A final code, however we may devise it, will consist of some set of $2^M$ blocks of 0's and 1's representing all of the blocks of digits $M$ digits long which we wish to transmit. If the code were not error correcting, we could use a block just $M$ digits long to represent each block of $M$ digits which we wish to transmit. We will need more digits per block because of the error-correcting feature.

When we receive a given block of digits, we must be able to deduce from it which block was sent despite some number $n$ of errors in transmission (changes of 0 to 1 or 1 to 0). A mathematician would say that this is possible if the distance between any two blocks of the code is at least $2n + 1$.

Here *distance* is used in a queer sense indeed, as defined by the mathematician for his particular purpose. In this sense, the distance between two sequences of binary digits is the number of 0's or 1's that must be changed in order to convert one sequence into the other. For instance, the distance between 0 0 1 0 and 1 1 1 1 is 3, because we can convert one sequence into the other only by changing three digits *in one sequence or in the other.*

When we make $n$ errors in transmission, the block of digits we receive is a distance $n$ from the code word we sent. It *may be* a distance $n$ digits closer to some other code word. If we want to be sure that the received block will always be nearer to the correct code word, the one that was sent, than to any other code word, then the distance from any code word to any other code word must be at least $2n + 1$.

Thus, one problem of block coding is to find $2^M$ equal length code words (longer than $M$ binary digits) that are all at least a distance $2n + 1$ from one another. The code words must be as short as possible. The codes of Hamming and Golay are efficient, and other efficient codes have been found.

Another problem of block coding is to provide a feasible scheme for encoding and, especially, for decoding. Simply listing code words won't do. The list would be too long. Encoding blocks of 20 binary digits ($M = 20$) requires around a million code words. And, finding the code word nearest to some received block of digits would take far too long.

Algebraic coding theory provides means for coding and decoding with the correction of many errors. Slepian was a pioneer in this field and important contributors can be identified by the names of types of algebraic codes: Reed-Solomon codes and Bose-Chaudhuri-Hocquenghem codes provide examples. Elwin Berlekamp contributed greatly to mathematical techniques for calculating the nearest code word more simply.

*Convolutional codes* are another means of error correction. In convolution coding, the latest $M$ digits of the binary stream to be sent are stored in what is called a *shift register*. Every time a new binary digit comes in, 2 (or 3, or 4) are sent out by the coder. The digits sent out are produced by what is called modulo 2 addition of various digits stored in the shift register. (In modulo 2 addition of binary numbers one doesn't "carry.")

Convolutional encoding has been traced to early ideas of Elias, but the earliest coding and decoding scheme published is that in a patent of D. W. Hagelbarger, filed in 1958. Convolutional decoding really took off in 1967 when Andrew J. Viterbi invented an optimum and simple decoding scheme called maximum likelihood decoding.

Today, convolutional decoding is used in such valuable, noisy communication channels as in sending pictures of Jupiter and its satellites back from the Voyager spacecraft. Convolutional coding is particularly valuable in such applications because Viterbi's maximum likelihood decoding can take advantage of the *strength* as well as the *sign* of a received pulse.

If we receive a very small positive pulse, it is almost as likely to be a negative pulse plus noise as it is to be a positive pulse plus noise. But, if we receive a large positive pulse, it is much likelier to be a positive pulse plus noise than a negative pulse plus noise. Viterbi decoding can take advantage of this.

Block coding is used in protecting the computer storage of vital information. It can also be used in the transmission of binary information over inherently low-noise data circuits.

Many existing circuits that are used to transmit data are subject to long bursts of noise. When this is so, the most effective form of error correction is to divide the message up into long blocks of digits and to provide foolproof error detection. If an error is detected in a received block, retransmission of the block is requested.

Mathematicians are fascinated by the intricacies and challenges of block coding. In the eyes of some, information theory has become essentially algebraic coding theory. Coding theory is important to information theory. But, in its inception, in Shannon's work, information theory was, as we have seen, much broader. And even in coding itself, we must consider source coding as well as channel coding.

In Chapter VII, we discussed ways of removing redundancy from a message so that it could be transmitted by means of fewer binary digits. In this chapter, we have considered the matter of adding redundancy to a nonredundant message in order to attain virtually error-free transmission over a noisy channel. The fact that such error-free transmission *can* be attained using a noisy channel was and is surprising to communication engineers and mathematicians, but Shannon has proved that it is necessarily so.

Prior to receiving a message over an error-free channel, the recipient is uncertain as to what particular message out of many possible messages the sender will actually transmit. The amount of the recipient's uncertainty is the entropy or information rate of the message source, measured in bits per symbol or per second. The recipient's uncertainty as to what message the message source will send is completely resolved if he receives an exact replica of the message transmitted.

A message may be transmitted by means of positive and negative pulses of current. If a strong enough noise consisting of random positive and negative pulses is added to the signal, a positive signal pulse may be changed into a negative pulses or a negative signal pulse may be changed into a positive pulse. When such a noisy channel is used to transmit the message, if the sender sends any particular symbol there is some uncertainty as to what symbol will be received by the recipient of the message.

When the recipient receives a message over a noisy channel, he knows what message he has received, but he cannot ordinarily be sure what message was transmitted. Thus, his uncertainty as to what message the sender chose is not completely resolved even on the receipt of a message. The remaining uncertainty depends on the probability that a received symbol will be other than the symbol transmitted.

From the sender's point of view, the uncertainty of the recipient as to the true message is the uncertainty, or entropy, of the message source plus the uncertainty of the recipient as to what message was transmitted *when he knows what message was received.* The measure which Shannon provides of this latter uncertainty is the *equivocation,* and he defines the rate of transmission of information as the entropy of the message source less the equivocation.

The rate of transmission of information depends both on the amount of noise or uncertainty in the channel and on what message source is connected to the channel at the transmitting end. Let us suppose that we choose a message source such that this rate of transmission which we have defined is as great as it is possible to make it. This greatest possible rate of transmission is called the *channel capacity* for a noisy channel. The channel capacity is measured in bits per symbol or per second.

So far, the channel capacity is merely a mathematically defined quantity which we can compute if we know the probabilities of various sorts of errors in the transmission of symbols. The channel capacity is important, because Shannon proves, as his fundamental theorem for the noisy channel, that when the entropy or information rate of a message source is less than this channel capacity, the messages produced by the source can be so encoded that they can be transmitted over the noisy channel with an error less than any specified amount.

In order to encode messages for error-free transmission over noisy channels, long sequences of symbols must be lumped together and encoded as one supersymbol. This is the sort of block encoding that we have encountered earlier. Here we are using it for a new purpose. We are not using it to remove the redundancy of the messages produced by a message source. Instead, we are using it to add redundancy to nonredundant messages so that they can be transmitted without error over a noisy channel. Indeed, the whole problem of efficient and error-free communication turns out to be that of removing from messages the somewhat inefficient redundancy which they have and then adding redundancy of the right sort in order to allow correction of errors made in transmission.

The redundant digits we must use in encoding messages for error-free transmission, of course, slow the speed of transmission. We have seen that in using a binary symmetric channel in which 1 transmitted digit in 100 is erroneously received, we can send only 92 correct nonredundant message digits for each 100 digits we feed into the noisy channel. This means that on the average, we must use a redundant code in which, for each 92 nonredundant message digits, we must include in some way 8 extra check digits thus making the over-all stream of digits redundant.

Shannon's very general work tells us in principle how to proceed. But, the mathematical difficulties of treating complicated channels are great. Even in the case of the simple, symmetric, off-on binary channel, the problem of finding efficient codes is formidable, although mathematicians have found a large number of best codes. Alas, even these seem to be too complicated to use!

Is this a discouraging picture? How much wiser we are than in the days before information theory! We know what the problem is. We know in principle how well we can do, and the result has astonished engineers and mathematicians. Further, we do have effective error-correcting codes that are used in a variety of applications, including the transmission back to earth of glamorous pictures of far planets.