On Human Communication

A Review, a Survey, and a Criticism by Colin Cherry

What are the differences between human and animal communication? How regular and universal are the rules of language? Can translation be mechanized? What does "meaning" mean?

On Human Communication discusses the opinions of internationally known authorities on such questions as these. It reviews the growth of modern interest in the relationships existing between such communication sciences as: linguistics, mathematics, cybernetics, psychology, semantics, and phonetics. Furthermore, it explores the extent to which these areas are becoming unified and warns about the difficulties which keep them separated. The author's critical approach cuts across a wide field of the literature, and his work emerges as a comprehensive source book of references, citations, and definitions. Invaluable to the specialist, *On Human Communication* also provides the novice with a fascinating introduction to the subject of information theory.

JOURNAL OF COMMUNICATION. "This is 'must' reading for anyone interested in the scientific study of human communication."

SCIENCE. "This volume fills a long-felt need for a synthesis of the theoretical and empirical knowledge of the field. . . . The design of the study is admirable, and the execution is superb. Moreover, the book itself is, indeed, a model of human communication."

Colin Cherry is Henry Mark Pease Reader in Telecommunication, Imperial College, University of London.

cover design: WARD & SAKS

SCIENCE EDITIONS, *440 Park Avenue South, New York 16, New York*

Science Editions

On Human Communication

Colin Cherry

$1.95, in Canada $2.25

E. Gabor, D., ' Theory of Communication," *J. Inst. Elec. Engrs. (London), 93,* Part III, 1946, p. 429.
F. Hartley, R. V. L., "Transmission of Information," *Bell System Tech. J., 7,* 1928, p. 535.
G. Stevens, S. S., and H. Davis, *Hearing—Its Psychology and Physiology,* John Wiley & Sons, Inc., New York, 1938.

CHAPTER FIVE

On the Statistical Theory of Communication

It is a very inconvenient habit of kittens (Alice had once made the remark) that, whatever you say to them, they always *purr. "If they would only purr for 'yes,' and mew for 'no' or any rule of that sort," she had said, "so that one could keep up a conversation! But how* can *you talk with a person if they* always *say the same thing?"*

Lewis Carroll (1832–1898)
Through the Looking Glass

1. DOUBT, INFORMATION, AND DISCRIMINATION

In this, as in other chapters, we shall make no attempt to compress a whole study within the compass of a few dozen pages, but rather try to convey to the reader some notion of the nature of the subject of statistical communication theory, which has aroused such widespread interest during recent years. We hope, too, to guide him through the literature and advise him on a preferred order of reading.

We shall be discussing the scientific concept of *information.* Now this is a word in everyday use; we speak of information as being *reliable, accurate, precise, timely, valuable,* et cetera. It is therefore not unnatural that the purely scientific use of the word should often be extrapolated into fields of discussion where it has doubtful place. Communication theory is a

scientific theory; it is not a vague descriptive treatment of everyday ideas of "information." It rests upon a solid foundation of mathematics, and cannot be understood by those who would avoid the mathematics; it cannot truly be "popularized." On the other hand, it is not at variance with commonsense views.

Communication theory first arose in telegraphy, with the need to specify precisely the *capacity* of various systems of telecommunication (to communicate information). The first attempt to formulate a measure mathematically was made by Hartley[A] in 1928, and his ideas are basic to the theory today. The newcomer to this subject can do no better than read his short classic paper first; it is easy reading. Engineers are concerned primarily with the *correct* transmission of signals, or (electric) representations of messages; they are not commonly interested, professionally, with the purposes of messages—whether they be trivial gossip, serious news, or racing tips. Provided the telegraph or telephone transmits the signals faithfully, the messages will have "meaning," value, truth, reliability, timeliness, and all their other properties. The signals must be correct; then all these human properties are inherent and consequential. Mathematical communication theory concerns the signals alone, and their information content, abstracted from all specific human uses. It concerns not the question "What sort of information?" but rather "How much information?"

This aspect of the theory was once described by Weaver as "bizarre," but now seems to be generally accepted as completely reasonable. The newcomer is referred to his discussion.[D,*] In this chapter we are concerned solely with this aspect—the information content of *signals*. In the following chapter we shall look more closely at the philosophical background, in an attempt to see relationships between the mathematical concept of information and other common and more human aspects.

Information can be received only where there is doubt; and doubt implies the existence of alternatives—where choice, selection, or discrimination is called for. We are continually making selections among alternatives, every moment of our lives, some consciously, but in the majority of cases unconsciously. It is a basic animal attribute; in the words of a psychologist: "discrimination is the simplest and most basic operation performable."†

But *selection* (or discrimination) can be carried out in non-human communication links. Perhaps the reader has seen that modern wonder, one teletype machine communicating with another. At the transmitting end,

* Read first Weaver's discussion on p. 95.

† Reference 314, with kind permission of the American Psychological Association.

the operator selects and presses keys one at a time; coded electric signals are thereby sent to the receiving machine, causing it to select and depress the correct keys automatically. We see the receiver keys going down, as though pressed by invisible fingers.

When we ourselves communicate one with another, we transmit signals, electric, acoustic, visual—physical embodiments of messages. Now it is customary to speak of signals as "conveying information," as though information were a kind of commodity. But signals do not convey information as railway trucks carry coal. Rather we should say: signals have an information content by virtue of their *potential for making selections*. Signals operate upon the alternatives forming the recipient's doubt; they give the power to discriminate amongst, or select from, these alternatives. And at present the "set of alternatives" with which we are concerned is a set of distinct signs which will be termed an *alphabet*.

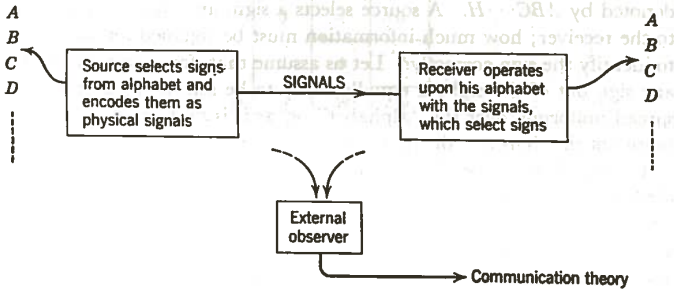

Fig. 5.1. "Information" as the selective potential of signals.

They may be the letters of a written language, numbers, printed words, the ordinates of wave forms (Chapter 4, Section 2.5, Fig. 4.7), semaphore or Morse code signs, or any discrete sign-types. But the alphabet must be specified, before the information content of messages can be discussed numerically; further, it must be assumed that the same alphabet exists at both the transmitting and receiving ends of the communication channel. It is then the function of the *source* of information to select the signs successively from this alphabet, thus constituting *messages*, and to transmit them in physical form as *signals*, through a channel, to the receiver. At the receiver, the signals operate upon an identical alphabet and select corresponding signs. Messages are then sent and received.

Note the distinction drawn here between *message* and *signal*. A message is regarded as the "selections from the alphabet," which is then put into physical form (signals) as sound, light, electricity, et cetera, for

transmission.* (A message might, for instance, be a thought, selected from an alphabet of thoughts.)

Perhaps such a naked description of this basic operation, illustrated by Fig. 5.1, emphasizes the dehumanized nature of the theory. But we shall breathe back the breath of life again in the next chapter.

Communication theory is written in the meta-language of an external observer; it is not a description of the process of communication as it appears to one of the participants. Figure 5.1 may thus be compared to Fig. 3.2(*a*) of Chapter 3.

2. HARTLEY'S THEORY: "INFORMATION" AS LOGICAL "INSTRUCTIONS TO SELECT"

Figure 5.2 shows, as an example, a simple alphabet of only eight signs, denoted by $ABC\cdots H$. A source selects a sign, and signals in some way to the receiver; how much information must be signaled for the receiver to identify the sign correctly? Let us assume that, from past observation, any sign out of the eight is equally likely to be selected. Doubt is then spread uniformly over the "alphabet" or, as it is said, the *a priori* probabilities of the signs are all equal (in this case, to 1/8).

The signals reaching the receiver represent instructions to select. Thus the first instruction answers the question: Is it in the first half of the alphabet, *yes* or *no*? (In Fig. 5.2, *yes* = 1, *no* = 0.) The range of doubt is *halved* by this. Then a second instruction divides each half into half again, and a third into half yet again. In this case then, three simple *yes*, *no* instructions (1, 0) serve to identify uniquely any one sign out of eight.

Such *yes*, *no* instructions are the simplest possible; each one successively halves the range of doubt. They are called *binary digits*, usually shortened to *bits* (or by some people, *binits*), and are used as the elementary *units of information capacity*. Notice that each sign in Fig. 5.2 is identified by a different sequence of 1, 0 digits. Thus *C* by 101, *G* by 001, et cetera. No two sequences differ by more than one digit; any single mistake therefore will cause ambiguity.

As we have already seen, all communicable messages (i.e., expressible by signs) *may* be coded into such binary 1, 0 sequences. The simplest illustration is provided by Morse code (dot, dash), which can code any written message in, at least, European languages.† We would remind

* The nomenclature of communication theory is still not universally established. However, the system adopted in this chapter in the greater part has been widely adopted in Britain and in the United States. A full list of definitions is given in the Appendix.

† Ignoring the letter- and word-space intervals; these can also be coded by a dot-dash sequence if required.

the reader too of the punched-card system of storaging information, illustrated by Fig. 2.2 (p. 34).

In our example, three bits of information are required for selection of each sign from among eight equally likely signs—because $2^3 = 8$ or

Sign	1st	2nd	3rd	Selections
A	1	1	1	
B	1	1	0	
C	1	0	1	
D	1	0	0	
E	0	1	1	
F	0	1	0	
G	0	0	1	
H	0	0	0	

Fig. 5.2. Binary coding of selections.

$\log_2 8 = 3$. A communication channel like this one, selecting signs at the rate of 100 per second, would have an *information rate* of 300 bits per second.

So much for the cases where the number of signs N in the alphabet is an exact power of 2. But suppose it is not? We shall show later that the information is still equal to $\log_2 N$ bits per sign selected, though this will involve an *averaging* process. But first, let us consider, as Hartley did, messages comprising wave forms, such as speech, rather than printed signs. Figure 5.3 shows (dotted) part of a continuous wave form $s(t)$, band-limited to F cycles per second, together with its representation by independent sample ordinates, spaced $1/2F$ second apart (see Section 2.5 of Chapter 4). These samples then define the wave form completely. Hartley appreciated that the amplitudes of such samples cannot be specified with absolute accuracy, in reality, although this is frequently done for the convenience of theoretical analysis. The amplitudes, being physical observations, must be quantized; in the figure, here, a comparatively coarse quantizing Δs of only eight levels has been assumed. (Such quantizing is in fact used practically in certain telecommunication systems, and the successive sample pulses are restricted to their nearest quantal

levels. The wave form then assumes a step-like character, which introduces a so-called quantization distortion.[252,*]) But such steps Δs may, in theory, be made as small as desired. The smaller Δs, the greater the number of levels, and the greater the *precision* of transmission; as we shall see, this implies also the greater the rate of transmission of information.

If we now label these ordinates arbitrarily, $ABC\cdots H$, then the successive selection of the sample ordinates may be regarded also as selection of these letter signs; such selections closely resemble our previous case,

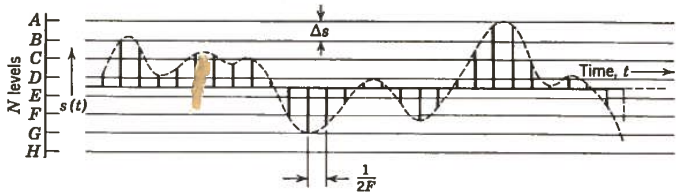

Fig. 5.3. Hartley's theory; band-limited wave-form source.

Fig. 5.2, with a source of discrete signs. However, it is advisable to distinguish between such (quantized) wave-form sources and sources of printed signs. For one thing, wave forms usually represent acoustic, electric, or other physical sources possessing *energy*, whereas we cannot readily associate energy with printed letters or other signs (excluding consideration of limiting physical light-quanta effects). The different levels, $ABC\cdots H$ represent possible *states* of this wave-form source; the successive sample ordinates select from these states. In general, if there are N such levels, or states, each sample ordinate contributes $\log_2 N$ bits of information (about the wave-form source) analogous to our previous case.

Consider a time interval of T seconds. This interval contains $2FT$ independent sample ordinates, each of which can have one of N levels. Thus in this interval there could be N^{2FT} different, distinct, wave forms. This set comprises all the different possible signals which such a quantized source is capable of transmitting each T seconds; it is called a band-limited *ensemble*, of duration T seconds. Hartley defined the information rate of such a source by the logarithm of this number of different signals (number of members of the *ensemble*), as H where, expressed to the base 2:

$$\begin{aligned} H &= 2FT \log_2 N \quad \text{bits per } T \text{ seconds} \\ &= 2F \log_2 N \quad \text{bits per second} \end{aligned} \tag{5.1}$$

* See Chapter 2, Section 2.

which is simply $\log_2 N$ (the information content per ordinate) times $2F$, the number of ordinates per second. This logarithmic measure is then one which permits addition of the information contents of successive independent ordinates.

And so, too, with any source of independent discrete signs, $ABC\cdots N$ (assuming for the moment they are equally likely). If a source selects from these at the rate of n per second, its information rate will be $n \log_2 N$ bits per second; and again there are N^n distinct alternative sequences of signs in an *ensemble* of one-second duration from such a source. When the term *independent* is applied to the successive signs selected by a discrete source, we mean, at present, that no one sign carries with it any information concerning its neighbors. We shall later refer to *statistical independence* in a more exact way.

This introduction to the information measure follows historical lines. Communication theory first arose in telegraphy[A] and we have used technical telegraphic terms, like *coding*. But the reader should appreciate the basic nature of the ideas. We are concerned not only with coding in the technical sense, but more broadly, with the *making of representations (of messages)*. The information received enables the recipient to add to his representations at his end, and the binary-digit measure tells by how much. The idea of "correspondence" is inherent in the concept of "communication"—the reproduction, or replication, of a representation.

2.1. Reversible and Irreversible Operations upon Signals

Each of the sample ordinates of a band-limited wave form (Fig. 5.3) selects a level (defines a state) of the source, $ABC\cdots H$. Each may be reduced to binary selections, as illustrated by Fig. 5.2. In Fig. 5.4, (*a*) a portion of a wave form is shown, together with (*b*) its binary-code representation, according to this coding scheme of Fig. 5.2. (A system of telecommunication coding, called *pulse-code modulation*, uses such representations practically, for transmitting speech and music;[82,252] for this purpose, *yes* (or 1) is coded as a sharp electrical impulse, whereas *no* (or 0) is coded by leaving a blank—no impulse. Figure 5.4(*c*) illustrates such impulse signals.)

Such codings, or representations, are clearly reversible; from (*c*) we may reconstruct the wave form (*a*) by setting up the ordinates and using the correct interpolation function (see Chapter 4, Section 2.5). Another, very familiar, reversible coding is the Morse code; with this, printed letters may be represented by dot-dash signals, but converted back into print without any loss or error.

The coded chain of impulses (*c*) may itself be regarded as sample ordinates of a wave form. Notice then that they are now three times as

closely spaced as in (*a*) ($\log_2 8 = 3$). In general, with quantization into N levels, the binary-coded signal will have samples with spacings reduced $\log_2 N$ times, requiring a bandwidth F' correspondingly increased. At the same time, the binary signal has only two levels ($N' = 2$). Thus, from Eq. 5.1, the information content of these signals has been unchanged by such coding.

Such a reversible coding represents a change of dimensionality; that is, a "trading" of bandwidth for numbers of levels, or alternative states.

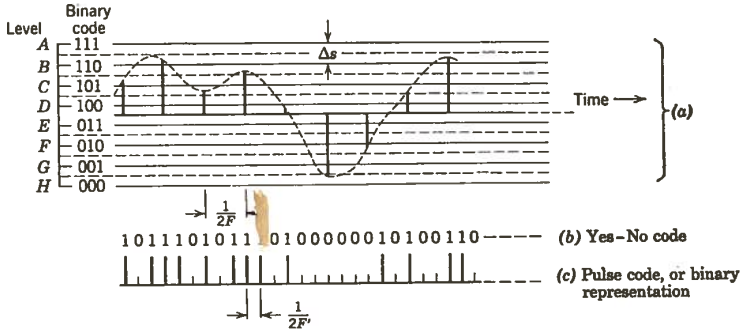

Fig. 5.4. Binary-pulse (reversible) code. Horizontal dotted lines represent *thresholds* of the quantization process.

The initial quantization itself represents an *irreversible* process—information content thrown away; each wave-form sample, assumed to be known at first with an unlimited precision, when quantized is reproduced with less precision. The original wave form then cannot be reconstructed with its original accuracy, since the necessary information has been destroyed; only the quantized wave form is recoverable.

But a more important cause of information loss (and so leading to an irreversible process) is *noise*. Noise is the destroyer of information and sets the ultimate upper limit to the information capacity of a channel, as we shall discuss later, in Section 6.1.

Hartley did not consider what it is that limits the fineness of quantization, in practical channels of the type so far considered; he did not refer to noise, nor did he consider the *probabilities* of the various states of a message source. It is these two aspects which have received so much attention recently. The statistical theory of communication is built up upon Hartley's foundations, but the idea of a determinate source of signals has become replaced by the concept of a *statistical ensemble*. Such a statistical

approach to telecommunication may be said to have originated with studies of the phenomenon of random electrical noise, in the 1920's. We shall return to such statistical aspects of our subject later, in Section 6.

2.2. When the number of alternative states is not a power of two

Selection of any one sign out of an alphabet of N signs can only be specified in whole numbers. We cannot speak of "fractions of a selection"; a choice is either made or not made—*yes* or *no*. If then N is not a power of 2, the selective information content of any one sign out of this alphabet cannot be specified as $\log_2 N$, since this will be fractional. But it is easily shown that this measure is still relevant if *averaged* over long sequences of selections.[c]

We are still assuming that all selections out of the N are equally likely. Consider an interval of time T, during which a wave-form source gives out a sequence of $2FT$ independent ordinates (or, analogously, nT selections from a discrete alphabet). During this interval one of $S = N^{2FT}$ possible different wave forms could be transmitted; then, as before:

$$\log_2 S = 2FT \log_2 N \qquad \text{but now this is fractional}$$

$$= r + \delta \qquad \text{where } r \text{ is whole number and } \delta \text{ a fraction}$$

To select this one wave form out of the S equally likely possibilities must require a *whole* number of elementary selections. The nearest whole number is r where

$$(\log_2 S) - \delta = r \text{ bits} \tag{5.2}$$

But if we speak of *average* number of selections, per sample ordinate (or sign) of the sequence, then as the interval T becomes large, this number of bits per *sample* becomes:

$$H_N = \lim_{2FT \to \infty} \frac{1}{2FT} [(\log_2 S) - \delta] = \log_2 N \text{ bits per sample} \tag{5.3}$$

Alternatively, the information *per second* from this source is H:

$$H = 2F \log_2 N \text{ bits per second} \tag{5.4}$$

exactly as for the case, Eq. 5.1, where N is a power of 2.

Notice that H is an information *rate;* so many binary selections (*yes, no*) *per second*. H may be fractional, but only by virtue of being taken on the *average*. This logarithmic measure of information rate can only be applied in this average sense. We can speak of a source possessing a certain "average rate of information." There are, however, certain cases in which it is convenient to regard the incremental contribution of single

signs (their information *content*), but such uses of the term information should be carefully distinguished.

The whole of the Wiener-Shannon theory is based upon average rates, and selections are always made as integral numbers of *yes*, *no* decisions.

2.3. Storage of information; capacity for information

Hartley's measurement of information rate, as we have approached it here, is seen to be in terms of the number of *yes*, *no* decisions required to specify the sample ordinates, or signs, emitted by a source, fractions arising only through averaging. One advantage of this is that it enables us to consider information storage and capacity.

Binary digits (*yes*, *no;* 1, 0; etc.) may readily be stored. Punched holes on paper cards were used in the Jacquard loom (for coding weaving patterns),* and the method remains in common use today in computing and accounting machines. Modern computing machines use relays, electronic tubes, magnetic storage drums, and other technical means.[25,33]† All such are used as two-state devices; they are either on or off.

The output signals from a source of information may be expressed as a chain, or *time series*, of binary pulses [Fig. 5.4(*c*)]. A source emitting H independent binary digits per second could fill a store of capacity Q binary elements in Q/H seconds on the average. Alternatively we may speak of the capacity of the source as H bits per second. But, as we have already seen, there need be no upper limit to the number of distinguishable signs, N, in an alphabet (or distinguishable amplitude levels of wave forms) were it not for noise. Consequently, a noise-free course can, in principle, have its capacity for transmitting information increased indefinitely, simply by increasing N.[328]

We define, then, the capacity of a communication channel as the number of independent *yes*, *no* digits which it may transmit per unit time. We shall return later to the question of an upper *limit* to capacity, in the presence of noise.

3. WHEN THE ALTERNATIVE SIGNS ARE NOT EQUALLY LIKELY TO OCCUR

With most practical sources of information, the signs are not equally likely to occur. A glance back at Fig. 2.4 (Chapter 2) for example, will show the relative frequencies of the letters in "English print" as they were assessed by Samuel Morse in his day. How does the Hartley logarithmic measure of information rate apply to such a source?

* See Section 3 of Chapter 2.

† See punched card, Fig. 2.2, p. 34.

3.1. Stationary and non-stationary sources

The relative frequencies p_i of the various signs may be estimated by an observer, if he watches the source for a long time; however, in practical cases, the possibility of making such an assessment with any pretence to accuracy depends upon the source being *statistically stationary*. This means that if the observer watches for a very long time T, the relative-frequency estimates he makes will not depend upon the actual moment of starting—the statistical parameters of a stationary source are invariant under a shift of the time origin. This assumption of stationariness is normally required in statistical communication theory, and is one of its present limitations. Many practical communication sources are, in fact, far from being stationary; thus spoken and written languages change their statistical (micro) structure continually (Chapter 3, Section 5); again, if the source possesses learning ability, it will change its behavior with the passage of time. In most fields of real *human* communication, the assumption of stationary sign behavior cannot be made, and this is one principal obstacle to the application of the mathematical theory to individual human communicative behavior.

3.2. Information rate of a stationary source of independent signs

Let $p_a p_b p_c \cdots p_i \cdots p_N$ be the relative frequencies of the N signs of an alphabet, $a, b, c, \cdots, N$, where $\sum_i p_i = 1$. Further, assume that the successive signs emitted by the source are independent, meaning that there are no rules (no "syntax"), determinate or statistical, by which any one sign is known to relate to another. Each selected sign is considered a separate event. In this case, the information rate of the source can be a function only of these relative frequencies p_i, and does not depend upon the *order* in which the signs are selected at the source.

This alphabet of signs, having certain relative frequencies, forms a *statistical ensemble*, upon which the source operates selectively. Figure 5.5(*a*) shows one way of illustrating such an ensemble; in this example there are eight signs, $abc \cdots h$, having the relative frequencies:

$$p = \tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{16}, \tfrac{1}{16}, \tfrac{1}{32}, \tfrac{1}{32}, \tfrac{1}{32}, \tfrac{1}{32},$$

respectively. A thick line of unit length (100 per cent) beneath this ensemble is shown divided up into segments of length proportional to these frequencies. This line, with the segments, represents a "range of doubt."

The source information rate is determined as before, in terms of equally likely, *yes*, *no* decisions, by successively halving the range of doubt. The "range of doubt" scale has been redrawn in Fig. 5.5(*b*), which may be

compared and contrasted with the equally likely case of Fig. 5.2. Thus, a first selection is made such that the ensemble is divided into two groups, of equal probability $p = \frac{1}{2}$. The transmitted sign is equally likely to come

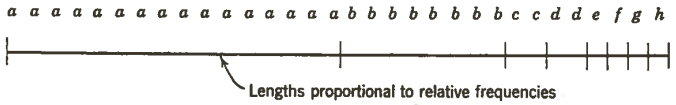

Fig. 5.5(a). An ensemble of eight signs, representing a "range of doubt."

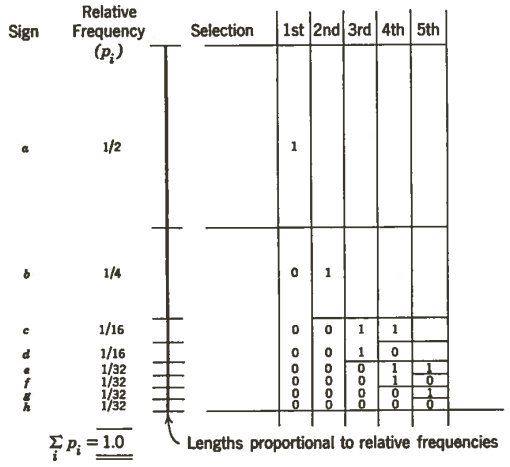

Fig. 5.5(b). Binary coding of selections of unequal probabilities.

from either group—on a long-term basis. Now the reader may object that such equal subdivision is only possible because we have chosen a most convenient set of probabilities in this example! True; this may not be possible in general, but let us assume for a moment it is, and return to this point later. A second subdivision, as shown, divides the ensemble into subgroups of equal probability $p = \frac{1}{4}$; a third, into sub-subgroups of $p = \frac{1}{8}$ and so on, until all signs are uniquely identified. The *yes*, *no* codes (1, 0) are shown in this figure, which illustrates also that the lower the probability of a sign in the ensemble, the more *yes*, *no* elementary selections are required; that is, the rarer the signs, the higher their information content.[C,*] Information content is then measured in terms of

* See Huffman under reference 166 for further treatment of such type of coding.

the *statistical rarity* of the signs (likened, by some people, to their "surprise value").

Each such division, into groups of equal probability, halves the range of *average* doubt; it therefore represents one *bit* of information. Let a particular sign be i, requiring say K_i successive binary subdivisions to identify it. Its probability is p_i; consequently, the final subdivision, which identifies it, divided a range $2p_i$ into equal parts; the subdivision before that divided the range 2^2p_i; the one before that 2^3p_i; and so on until we arrive at the initial division of the *whole* alphabet, having a probability $\sum p_i = 1$. Hence:

$$2^{K_i} p_i = 1$$

or

$$K_i = -\log_2 p_i \tag{5.5}$$

The average, or *expected* value of K_i, taken over the whole alphabet $a, b, c, \cdots, N$ (in the general case) is then:*

$$H(i) = \overline{-\log p_i} = -\sum_i p_i \log p_i \text{ bits per sign} \tag{5.6}$$

We shall return to this important formula, which represents the *average* number of *yes*, *no* digits required, per sign transmitted—the information rate of this source of independent discrete signs.

3.2.1. WHEN THE ALPHABET DOES NOT DIVIDE INTO EQUALLY LIKELY SUBGROUPS. The argument above, due to Fano, is very descriptive, but the following method is an alternative. Consider now those cases in which the alphabet does *not* divide consecutively, so conveniently, into equally likely subgroups. The argument is rather similar to that of Section 2.2; we cannot deal with *single* signs, but only with averages, over very long sequences given out by the source.

If we observe extremely long sequences, then the various signs $a, b, c, \cdots, N$ will in fact occur with almost their estimated probabilities $p_a p_b \cdots p_n$ (the source being statistically stationary); consider an ensemble of all the n possible different message sequences, each of S signs in length, distinguished only by different orders of occurrence. Then all such long sequences will have nearly equal probabilities $p(S)$ of occurring in the source, and the number of different messages in the ensemble will be

* *Expected value*: the expression 5.6 is a way of writing average values, as used particularly by statisticians. Suppose we have a chain of the numbers $a_1\ a_2\ a_3$ (perhaps $a_i = \log p_i$) from a source, of which the following is a sample of 12 successions: $a_3\ a_1\ a_3\ a_2\ a_3\ a_3\ a_1\ a_1\ a_2\ a_3\ a_1\ a_3$ (twelve numbers).

Average of this =

$$\frac{(4 \times a_1) + (2 \times a_2) + (6 \times a_3)}{12} = (p_1 \times a_1) + (p_2 \times a_2) + (p_3 \times a_3) = \sum p_i a_i.$$

$n = 1/p(S)$ where:

$$p(S) = {p_a}^{S \cdot p_a} \cdot {p_b}^{S \cdot p_b} \cdot {p_c}^{S \cdot p_c} \cdots {p_N}^{S \cdot p_N} \tag{5.7}$$

We see this, as follows: the probability of a sequence is the product of the probabilities of all the signs forming it. Then a occurs about $S \cdot p_a$ times in each long sequence; hence, since p_a is the probability of any *one* a occurring, the joint probability of the number Sp_a occurring is ${p_a}^{Sp_a}$. Similarly for b, c, d, et cetera.

Now all these n messages being so nearly equally likely, the information content of any one is obtained as for our first elementary case (Fig. 5.2). It is simply $\log_2 n$ bits per sequence S, or $\frac{1}{S} \log_2 n$ bits per sign. That is, from 5.6:

$$H(i) = \frac{1}{S} \log \frac{1}{p(S)} = -\sum_i p_i \log p_i \text{ bits per sign} \tag{5.8}$$

which is identical with Eq. 5.6.

4. THE USE OF PRIOR INFORMATION: REDUNDANCY

It is one of the merits of statistical communication theory that it takes into account the effect, upon communication, of *prior information*. Though a receiver may not know exactly what messages are coming to him next, he is not necessarily in a state of complete ignorance. We have already assumed that he knows the alphabet of signs and has had experience of their relative frequencies of occurrence. In Chapter 4 we considered his knowledge of the channel itself: of bandwidth, signal power, types of coding; of the *structure* of the signals, as dependent upon the channel properties. All this has been brought into consideration in measuring information rates. But other prior information may exist, by virtue of known constraints between the signs; that is, from syntactical rules. If such rules are known, determinate or statistical, then the signals reaching the receiver bear less information than they would if the successive signs were independent. The information conveyed by signals is always relative; it depends upon the *difference* in the receiver's doubt before and after their receipt.

4.1. Syntactical redundancy: its measurement

The rules of syntax of human languages are complicated and varied; such rules introduce *redundancy* into the messages, thereby making their correct reception more certain. We have already discussed this question, in a purely descriptive way, in Section 6.3 of Chapter 3. In communication theory, redundancy is treated mathematically, the syntax being

described, not necessarily as a linguist would commonly view it, but as a set of conditional probabilities.[D]

A source of information which selects signs according to probabilities is called a *stochastic source*, and the message sequences *stochastic series*. We may consider also *transition probabilities*, or the relative frequencies with which different signs, say, follow a given sign or, alternatively, precede it. In printed English, for instance, the rule of spelling, "I before E except after C," with a very few exceptions, suggests that

$$p_e(\text{EI}) \gg p_e(\text{IE})$$

We read a transition probability $p_x(y)$ as "the probability of y given x." An alternative notation is $p(y|x)$.

Other conditional probabilities may be known, referring not only to adjacent signs of a sequence, but to any specified spacings or groupings such as "letter bridges" or "word bridges."[357]

The existence of constraints, in terms of transition or other conditional probabilities will, if known *a priori*, introduce redundancy into the messages received from a source—being something known statistically about the messages beforehand (prior statistical information).

In English texts, or those of other human languages, the various transition probabilities governing the appearance of the successive letters are very unequal. As an illustration, suppose a teletype machine gives out the following sequence:

$$\cdots\cdots\cdots\cdots \textit{with the arrival of t}|$$

where the bar represents the instant "now." The next letter is governed by a whole set of conditional probabilities, and depends, in the limit, upon *all* that has gone before. However, the influence of the letters and words several lines, paragraphs, or pages removed in the past will be very slight. It is the few letters immediately preceding "now" which have the greatest control, with certain exceptions owing to rigid grammatical rules. But, as regards *numerical* measurement of redundancy, we have available only those conditional probabilities which have to be gathered by the patient labor of cryptographers and language students.[D,85,96,273,294,367] The task of assessing monogram, digram, and trigram frequencies is formidable, let alone going beyond this. The fact that we ourselves can guess successive letters of a text, with fair accuracy, implies that we possess immense mental stores of the *rank orderings* of letters and words; but we do not know the various transitions as numerical relative frequencies.[294]

With the help of statistical tables of letter or word frequencies, together with digrams, trigrams, or other grouping frequencies, it is possible to construct texts which resemble, say, English passages (though they may

continually "wander off the point!"). But this experiment need cause no surprise, and has no philosophical interest whatever; it merely shows the correctness of the tables used. Jonathan Swift made biting comment upon this experiment (Chapter 2, Section 1).

Rather than an English message, such as that cited above, let us consider a Teletype machine operating in code* and, for simplicity, using only the letters A, B, C, D. A typical sequence might be:

$$\cdots\cdots\cdots\cdots\; B\ A\ A\ C\ D\ B\ A\ D\ C\ D\ A\ B\ A \Big| \begin{matrix} A \\ B \\ C \\ D \end{matrix}$$

where the bar represents "now," to be followed by one of A, B, C, D, according to a whole set of conditional probabilities. To assess any one, say the frequency with which B follows A, $p_A(B)$, we pick out all the A's, in a *very long* sequence, and observe what fraction are followed by B. Since *some* letter must follow any given one

$$\sum_j p_i(j) = \sum_j p(j) = 1 \tag{5.9}$$

That is to say, the summation is obviously independent of the preceding sign, i.

There is a simple, yet very important, theorem concerning statistical constraints and redundancy; it should be clear from the illustration above:

> If all the various transition probabilities $p_i(j)$ are equal, then the individual signs, or letters, become statistically independent and equally probable. In such a case there are absolutely no preferred guesses as to what letters will be given out by the source; redundancy is provided by the existence of *unequal* transition probabilities.

Such a source of equi-probable, statistically independent letters or other signs has a maximum information rate (other factors being fixed). Equation 5.8 gives the information rate for a source of independent signs, and this expression is maximized when all p_i are equal.[D]

But notice that the converse argument does not hold; it is easy to arrange that all letters should be equi-probable, yet have unequal transition probabilities. An example will suffice; suppose this is a typical sequence:

$$\cdots\ B\ B\ B\ B\ A\ A\ A\ A\ C\ C\ C\ C\ A\ A\ A\ A\ D\ D\ D\ D\ C\ C\ C\ C\ B\ B\ B\ B\ \cdots$$

* It can be helpful, to the beginner, to consider examples in code, rather than in plain language, because the mind is so easily side-tracked by the "meaningfulness" of the latter. *Meaning* is quite irrelevant to our present context, but we shall consider its place in relation to communication theory later, in Chapter 6.

Then although $p(A) = p(B) = p(C) = p(D)$, it is possible that, say, $p_A(C) < p_C(C)$. Given any one letter of the sequence, our best guess here, for the next, would be the same letter.

Sequences for which only pairs of adjacent signs are considered, as we have done so far, are called Markoff chains,[225] though the term is frequently used for series with known trigram or higher-order (finite) structure. Quantitatively speaking, the redundancy of a source is assessable only *relative* to the known set of probabilities. Thus we can quote the redundancy of a source on a monogram basis [knowing only the various $p(i)$], or a digram basis [knowing also $p(i, j)$], or a trigram basis, et cetera. But we cannot simply give "its redundancy," on an unspecified basis.

Suppose that we have assessed the relative frequencies with which a source emits different alternative sequences of S letters; let us write such S-gram joint probabilities as $p(a\ b\ c \cdots S)$. For example, in our four-letter source used above, we may know the values of $p(A\ B\ C)$, $p(A\ C\ B)$, $p(B\ A\ C)$, $p(B\ C\ A)$, et cetera—all the trigrams. Then these may readily be interpreted in terms of successive *transitions* since:

$$\begin{aligned} P(a\ b\ c \cdots S) &= P(a) \cdot P_a(b\ c \cdots S) \\ &= P(a) \cdot P_a(b) \cdot P_{a\,b}(c \cdots S) \\ &= \cdots\cdots \text{etc.} \end{aligned} \tag{5.10}$$

Probability constraints between successive letters may then be specified either in terms of joint probabilities $p(a\ b\ c \cdots S)$ or as different transition probabilities $P_a(b)$, $P_{ab}(c \cdots S)$, et cetera.

Knowing such conditional probabilities, we may then assess the corresponding redundancy—which is still to be defined.

The redundancy of a source may be quoted as a percentage:

$$\text{Redundancy}^{\text{D}} = \frac{H_{\max} - H}{H_{\max}} \times 100 \text{ per cent} \tag{5.11}$$

where H = information rate (bits per sign, or second) of the source

$H_{\max}$ = maximum information rate which it could possess if recoded into the same alphabet of signs by equalizing all transition probabilities, and hence equalizing all sign probabilities, thus rendering them independent.

For illustration, Fig. 5.5 shows the encoding of a redundant source; the signs of the alphabet, $a, b, \cdots, h$, having unequal probabilities are shown encoded into 1, 0 signs (digits). But there are clearly many alternative ways of doing this. The alphabet might have been divided successively into two parts, represented by a 1 and a 0, in different ways. One way,

however, will give H_{max}, and will render the frequencies of the 1 and 0 signs equal, on an average, from the source.

We have defined the rate of information of a source of signs, as $H(i)$ in Eq. 5.8; this was given as minus the expected value (average, over alphabet) of the log probability of the various signs. Its maximum value $H_{max}(i)$ would be reached if all $p(i)$ were made equal by recoding. But we have not yet defined the information rate of a source of signs having known transition probabilities. This may be done on the same basis as before: as minus the expected value of the log probabilities of the various signs of the alphabet. Suppose, for example, that we know not only all the sign probabilities $p(i)$ but also all the transition probabilities with which any sign j may follow a given sign i; that is, we know all $p(i)$ and all $p_i(j)$. Then at any given instant the last sign i, emitted by the transmitter, is known at the receiver (the channel being noiseless); consequently doubt about the next sign j depends upon the probability $p_i(j)$, not upon $p(j)$. Consequently the relevant "doubt measure" is $-\log p_i(j)$, which must be averaged over all the *digrams* (ij). Thus the information rate of such a redundant source $H_i(j)$ is

$$H_i(j) = -\sum_i \sum_j p(i,j) \log p_i(j)$$
$$= -\sum_i \sum_j p(i) p_i(j) \log p_i(j) \text{ bits per sign} \qquad (5.12)$$

Similarly the information rate $H_{ij}(k)$ may be calculated for a source having a known trigram structure; and so on.

Shannon has estimated the redundancy of English,[294] on a letter basis, from the published data[273] on letter frequencies, and digram and trigram transitions $p_i(j)$, $p_{ij}(k)$ [tables of higher n-grams are not available]. He gives the following figures: $H(i) = 4.14$, $H_i(j) = 3.56$, and $H_{ij}(k) = 3.3$ bits per letter. A 26-letter alphabet is used, with the word space ignored. He gives figures also for a 27-letter alphabet and for the information rate on a *word* basis,[85] together with an interesting experimental method of estimating rates with higher-order transition constraints (see Chapter 3, Section 6.3), showing that the information rate tends toward a limit of roughly 1.5 bits per letter.

On the other hand, suppose we know not the transition but the *joint* probabilities of adjacent pairs of signs $p(i,j)$, ranging over all signs of the source alphabet. Then the receiver's doubt about each arriving digram (ij) depends upon $\log p(i,j)$. It is as though the alphabet was considered to be rewritten as a digram alphabet, from which the source selects digrams. The information rate, relative to *a priori* knowledge of this kind, is:

$$H(i,j) = -\sum_i \sum_j p(i,j) \log p(i,j) \text{ bits per sign} \qquad (5.13)$$

4.2. Redundancy: its function in correcting errors

"Redundancy" may be said to be due to an additional set of rules, whereby it becomes increasingly difficult to make an undetectable mistake. The term therefore is rather a misnomer, for it may be a valuable property of a source of information. If a source has zero redundancy, then any errors in transmission and reception, owing to disturbances or noise, will cause the receiver to make an uncorrectable and unidentifiable mistake.

Redundancy may be contributed in many ways; different kinds of determinate or statistical rules may be used. In human languages, such rules constitute *syntax*, where "rules" may better be called "habits," for none are inviolate (see Chapter 3, Section 6). But with codes or invented sign systems, regular rules may be introduced. All redundancy is, in effect, a form of addition; a larger number of instructions are sent than are barely necessary. The simplest form of addition is plain repetition of each sign n times, as with the sequence given above, though this is not very efficient. From a non-redundant source of, say, independent and equiprobable letters, any specified sequence of letters must be capable of occurring; none are "forbidden." But in the English language, for example, with its 26 letters, there are many sequences which virtually never occur. If you were to receive the following telegram, you would have no difficulty in correcting the "obvious" mistakes:

BEST WISHES FOR VERY HAPPP BIRTFDAY

because sequences such as HAPPP do not occur in the language. By virtue of redundancy, messages may become changed by errors into something *more* improbable. Similarly with speech; speech sounds appear only in certain sequences, in language, so that extraneous noises superpose and convert the sequences into something the listener knows to be most improbable. He detects a mistake and asks the speaker to repeat; if the extraneous noise, by chance, converts a sequence into something resembling a true speech sequence, the listener may mishear. But speech perception raises other problems far beyond such simple illustrations, which we shall discuss in Chapter 7.

It will suffice here to give one elementary method of adding redundancy to coded signals, and to refer the reader to more advanced treatments of the subject of *noise-combating* codes. Depending upon the type of noise, and the type of channel, redundancy is best added in different ways; but the whole subject is very difficult.[D,131,142,*] Shannon has indicated a general technique of *coding* messages in advantageous ways, for combating

* See also Laemmel, under reference 166.

noise, that is more subtle than mere repetition of every transmitted sign.[D]

When messages, originally expressed by some form of signs (such as the letters of printed texts), are transformed into another set of signs, in a way agreed upon between the transmitter and the receiver, and such that they may be unambiguously transformed back again, they are said to be *coded*. When transformed into code groups containing only two distinct signs, they are said to be in *binary code*. This code, which we have seen is of basic interest, is illustrated by a simple example in Fig. 5.2. Here, the alphabet of eight letters $ABC\cdots H$ can be expressed alternatively by *yes*, *no* or 1, 0 digits before being signaled to a receiver, who may recover the letters unambiguously. In this example, the various 1, 0 groups, corresponding to each letter, differ from one another by only one digit; thus, any error, resulting in the conversion of a 1 into a 0, or vice versa, causes an undetectable mistake in decoding of the received letter. But suppose we add one redundant digit to each group, as follows:

Letters		Code Groups	Letters		Code Groups
A	=	1111	E	=	0110
B	=	1100	F	=	0101
C	=	1010	G	=	0011
D	=	1001	H	=	0000

(5.14)

On inspection, it will be seen that such code groups enable *one* single mistake, in any 1, 0 digit, to be detected (but not corrected). For instance, the group 1111, for A, might be converted to any of the following, by noise: 1110, 1101, 1011, 0111, none of which appears in the code. With this redundancy, one digit error per letter is detectable, but not correctable; thus, the group 1110 could be produced either by a single error in the code for A, for B, for C, or for E.

To give greater safeguard against error, further redundant digits could be added, making the set of code groups differ, one from another, by as many 1, 0 digits as possible. We may regard this as seeking to place the code groups as "far apart" from one another as can be, where "far apart" means a distance in a code hyperspace. To visualize this, we must reduce the space to two or three dimensions, so that we can draw it. Taking an alphabet of four letters only, we can code this as follows:

$$A = 11 \qquad C = 01$$
$$B = 10 \qquad D = 00 \tag{5.15}$$

Each code group here has only two degrees of freedom and so may be represented by a diagram in two dimensions. Figure 5.6(a) shows two axes, representing the first and second digit, so that the four code groups may be placed at the four corners of a square. Moving parallel to the vertical axis changes the first digit, or parallel to the horizontal axis, the second digit. Only four distinct code groups are possible, given by Eq. 5.15, and, of these, those at the ends of either diagonal of the square are "farthest apart." Figure 5.6(b) shows the similar case with three degrees of freedom; here eight code groups exist (corresponding to Fig

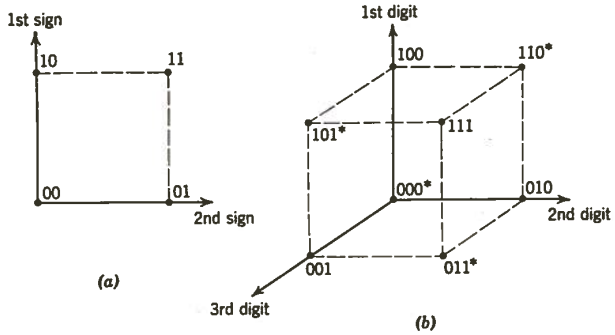

Fig. 5.6. Binary coding: (a) with two and (b) with three degrees of freedom.

5.2), and four which are mutually "farthest" apart are shown with asterisks, lying at the corners of a tetrahedron. Alternatively, the four without asterisks have the identical property.

This process may be carried into spaces of m dimensions, in which the code groups each have m binary (1, 0) digits. The complete set of distinct code groups would then possess 2^m members, which might be used to encode an alphabet of 2^m signs (e.g., letters), but with no chance of detecting errors. Out of this set, a number N could be selected so as to differ from one another by at least d digits. The problem is then to choose these in such a way as to maximize the number N for use as code groups.[121,142,*]

(*Postscript:* The question is often asked during student lectures on communication theory: If a million copies of a newspaper are printed, is the information content increased a millionfold? The answer is that, should any one person (a "receiver") read them all, a millionfold *redundancy* would exist!)

* See also Laemmel under reference 166.

5. MESSAGES REPRESENTED AS WAVE FORMS: "CONTINUOUS" INFORMATION

Let us now glance at a few aspects of signal wave forms, such as those of speech, rather than sequences of discrete signs, such as letters. Notice there are two ways of regarding a source of speech; we could imagine, say, speech reduced to phonetic symbols and these treated as a finite alphabet of signs or, rather more naturally, we could treat the raw speech wave forms as the communication medium. Then, in such cases, what are the "signs"? It is here that the Sampling Theorem comes to our aid (as discussed in Section 2.5 of Chapter 4). If the bandwidth of the wave-form source is restricted to any value, F cycles per second, chosen arbitrarily or by practical considerations, then the wave forms are specified completely by the values of their ordinates spaced apart along the time scale by intervals of $1/2F$ seconds (the time origin may be chosen arbitrarily). Figure 5.3 illustrates a sampled wave form. Strictly speaking, *there is no need to consider "continuous" wave forms at all in signal analysis.*[122] "Continuous" functions are the creation of mathematicians,[266] and enable methods of analysis of great elegance to be used. But such analysis may well be done algebraically.* Against this, it may be argued that algebraic methods must necessarily introduce approximations;† this may be true, but it should be remembered that signal analysis concerns the use of mathematical methods for describing *physical signals* and their properties. Mathematicians deal with mental *constructs*, not with description of physical situations. Approximations can be reduced as much as we wish, at the price of increased algebraic labor. A "continuous" function is not a physical idea but a mathematical one; when solving problems in physics (or applied mathematics), such an idea need not be regarded as holy, as sometimes seems to be the case.

Communication sources, emitting wave forms, are sometimes referred to as *continuous sources*. This, however, is not because wave forms are "continuous functions of time," $s(t)$, but rather because the successive independent sample ordinates $s(\tau_1)$, $s(\tau_2)$, et cetera, may have a continuous range of amplitudes; an ensemble of such wave forms (or their sample ordinate sequences) may have a continuous *amplitude distribution.*‡

* For example, see reference 333. Tustin denotes a sequence of wave-form ordinates by a sequence of numbers, representing their amplitudes; he then determines the rules for addition and multiplication of such *time series*.

† All applied mathematics is necessarily approximate, of course, because we cannot describe a physical situation in its entirety. But whether it is the mathematics or the physics which is approximate is not a real question. Rather, we should say that the two can never fit one another perfectly.

‡ However, in practice, such distributions can only be estimated from a *finite* set of observations.

Against this, it could be objected that amplitude quantization is a necessity, since the wave forms represent physical observations of signals; but to take refuge in this idea, and so make wave-form sources similar to sources of discrete signs (Fig. 5.3) is, although quite justifiable, rather distasteful to people whose interest is primarily mathematical. For the smaller we make the amplitude quantum Δs, the greater the number of alternatives in the "alphabet" of ordinate amplitudes, and so the greater the information content contributed by the selection of any one of them; then, as $\Delta s \to 0$, in this limit does the information rate of such a source become infinite? This is an interesting theoretical point, which we shall discuss shortly (Section 6).

5.1. The idea of "statistical matching"

Wave-form analysis concerns signal wave forms, their properties, and relations between them. It is really, then, a "syntactic" study.* But there are certain distinctions between sources of wave forms and sources of, say, printed signs, apart from the question of "continuity." One distinction is this: an alphabet of printed signs may be listed in arbitrary order; but the ordinates of wave forms are rank-ordered along a scale of amplitude, or energy. An ordinate having an amplitude $s(t) \pm \Delta s/2$ specifies a wave-form sample having an energy proportional to the square of this amplitude, and so the selection of this ordinate, by the source, requires that this energy be supplied. Sources of information emitting wave forms require supplies of power, and any limitation set to the value of this power imposes a constraint upon the source. Such limitation may be set in several ways; frequently it is set as a fixed mean value (Chapter 4, Section 2) and sometimes as a peak value or as a maximum wave-form ordinate magnitude. Different types of telecommunication channel use different systems of modulation, and these, in turn, impose different types of power constraint. The power of wave-form transmitters must always be limited to a finite value.

We have now mentioned a few constraints which practical telecommunication channels impose upon the signals they transmit. In particular, they restrict the *bandwidth* (and hence the number of independent ordinates per second) and the *power;* again, the source itself, prior to encoding, possesses a certain statistical structure. Such constraints demand that, for efficient transmission, a source of information should be *statistically matched* to the physical channel, for transmission.[D]

This concept of statistical matching is extremely important because, in communication theory, it gives an exact mathematical formulation of a universal principle of human behavior. When carrying out any goal-

* We shall enlarge upon this notion in the next chapter.

seeking task, the way in which this task is organized will depend upon the constraints imposed—that is, upon the individual's freedom of action. The achieving of some optimum result depends upon organization of the task, whilst keeping within the limits imposed. The key word here is *organize*. The encoding of messages is a process of organization, converting or transforming messages from one sign representation into another, possibly more suited to the type of communication channel employed; and the channel may impose limitations of bandwidth, or of power, (and, as we see later, noise) which determine how this encoding should best be done.

To give a simple human illustration, when I send a telegram in Britain, I am charged so many pence per word; therefore I express ("represent") my messages in certain preferred ways, omitting prepositions, et cetera, and choosing subtle words. The statistics of my language become changed. On the other hand, when I talk to young children I am constrained to use words of one syllable, though perhaps many more of them than I would use for an adult. So the statistics are again altered. All such constraints of the channel, then, determine a preferred statistical structure for the transmitted signals. But such examples are very vague. In communication theory, this idea is given exact mathematical expression, in terms of the encoding of messages so as to *match* the physical constraints of the channel of transmission. For example, suppose a source selects messages which are represented by an alphabet of printed letters; then, as we saw before, the greatest rate of transmission (in this medium of *print*) is achieved when the letters are statistically independent and equally probable. But suppose we wish to transmit these printed messages over a telegraph channel; then the letters should be encoded into electrical signals, such that they use the limited available electric power of the telegraph channel in a most efficient manner. "Most efficient" here means that, with the given power, the electric signals shall be able to convey information at the greatest possible rate, or at least possess a capacity greater than that required by the message source itself. Now the coded messages may be represented as electric wave forms in many ways; two in particular we have already illustrated, namely (*a*) simple amplitude variation (Fig. 5.3), in which the amplitude of any ordinate represents a sign, and (*b*) pulse-code modulation (Fig. 5.4), in which all the electric pulses are identical in amplitude. And it may be shown that in the former case, with the assumption that the *mean* signal power is fixed, the greatest information rate is achieved if the messages be so coded that the transmitted wave-form ordinates are statistically independent and approximate to a *Gaussian* amplitude probability distribution.[D] Briefly, in the case of a source of *printed* letters (no power consideration),

the information rate is greatest when the letters are equi-probable and independent; but in the case of messages represented as bandwidth-limited wave forms, with the *mean* power limited to P_{avg}, then the maximum rate is reached with a Gaussian* amplitude distribution:

$$\sigma \cdot P\left(\frac{s}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \epsilon^{-s^2/2\sigma^2} \tag{5.16}$$

where $\sigma^2 = P_{avg}$, the mean power. This equation gives the probability densities of s, the wave-form ordinate amplitudes, relative to their root-

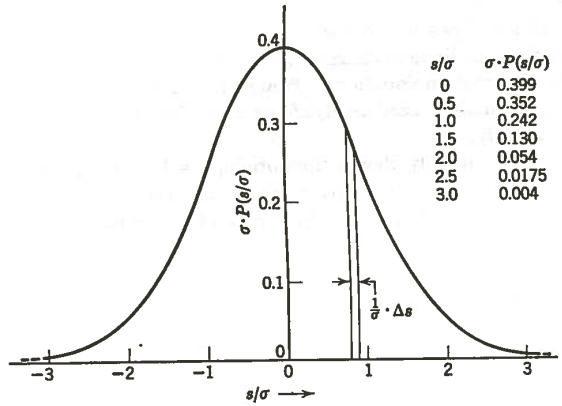

Fig. 5.7. The Gaussian, or Normal Density Function $P\left(\frac{s}{\sigma}\right)$ (area lying under this curve is unity).

mean-square value σ. Here σ^2 is also called the *variance* of this bell-shaped distribution (Fig. 5.7), and it is a normalizing factor of the curve.

But this way of discussing sources of messages represented as wave forms is not wholly satisfactory. We have imagined the wave-form ordinates to be quantized into a finite number of states, possibly quite large. But we have so far avoided this question of a *continuous* range of amplitudes, which would seem to result in the possibility of an infinite rate of communication of information. There are certain essential distinctions between such continuous sources and discrete-sign sources. In particular, information rate can only be considered to be relative, not absolute; again, continuous sources cannot readily be discussed in a practical way

* Sometimes called Normal Density Function (see reference E for tables), as illustrated by Fig. 5.7.

the problem of continuous sources in a slightly different way.

5.2. Sources of wave forms: time averages and ensemble averages

We shall now consider "continuous" sources of signals as *wave forms* having a bandwidth F cycles per second. Such wave forms may be represented completely by a series of ordinates spaced apart by $1/2F$ seconds (as in Fig. 5.3). It should be appreciated that only this *spacing* is of consequence, and the time origin of the samples is empirical. True, if a different set of points be chosen, also spaced by $1/2F$ seconds, then a different set of ordinates will result; but these will be related to the first set by transformation equations. However, if any sequence of ordinates be chosen, equally spaced by $1/2F$ seconds, they will specify the wave form completely.

Given the arbitrarily chosen time origin, $t = 0$ at the position of any one ordinate, the nth ordinate from this in the positive time direction will mark the instant $t = n/2F$ or, in the negative direction $t = -n/2F$. The wave form $s(t)$ is then represented by the summation of the sequence of *interpolation functions* of $\sin x/x$ form (Eq. 4.16), having amplitudes given by these sample ordinates, as illustrated by Fig. 4.7. That is:

$$s(t) = \sum_{-\infty}^{+\infty} s\left(\frac{n}{2F}\right) \frac{\sin 2\pi F\left(t - \frac{n}{2F}\right)}{2\pi F\left(t - \frac{n}{2F}\right)} \tag{5.17}$$

Here we are imagining the wave form of the source output to have unlimited duration. If this duration is limited to a time T, then n will range over the values $1, 2, \cdots, 2FT$. This equation, 5.17, represents the set of all the possible wave forms which can be emitted by this band-limited source.

Since a set of discrete ordinates completely defines the signal wave form $s(t)$, we should expect to be able to express all the various statistical properties of the signal in terms only of these ordinates. "Statistics" are "averages"; and there are two distinct ways whereby such statistical parameters may be specified. The two ways, which, in certain important cases, become equivalent, are illustrated by Fig. 5.8; let us take them in turn.

5.2.1. "Time-average" source statistics. Figure 5.8 illustrates typical wave-form segments (each of duration T seconds) from a number of different sources—source 1, source 2, et cetera. We shall be regarding so that each of these signals will have a large number of degrees of freedom $2FT$. Any one wave form is uniquely specified by the values of the $2FT$

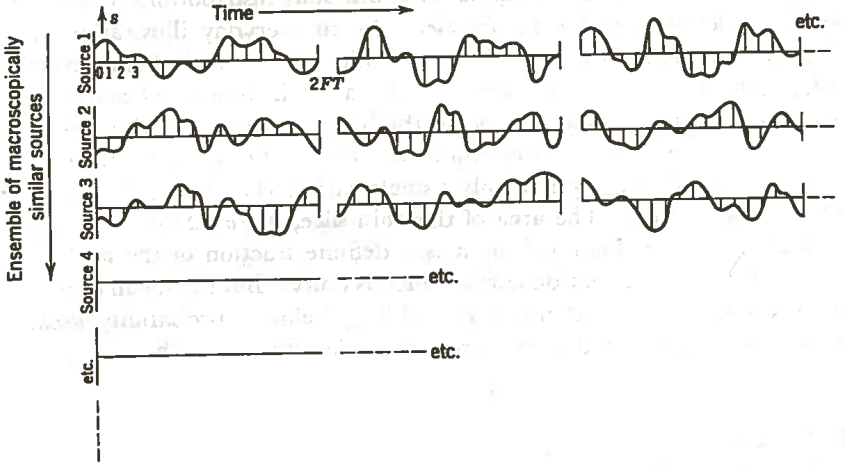

Fig. 5.8. Time-average and (source) ensemble-average statistics. For simplicity, we show relatively short durations T here, representing few degrees of freedom $2FT$ for the wave forms.

equally spaced ordinates, so that, starting with the first ordinate of each or any of these wave forms, we may label the successive ordinates, as before, $0, 1, 2, \cdots, n, \cdots, 2FT$ and refer to their amplitudes as $s_1 s_2 \cdots s_n \cdots s_{2FT}$ (rather than the $s(n/2F)$ notation used in Eq. 5.17).

This set of band-limited wave forms, all of duration T, may be considered to represent alternative messages which the source may select, just as we earlier spoke of a source as selecting from alternative long sequences of printed signs (Section 3.2.1). Again, analogous to the discrete case, we may speak of this set of wave forms as a band-limited *ensemble*, defined by a probability distribution $p(s_n)$ where

$$p(s_n) = p(s_1 s_2 \cdots s_n \cdots s_{2FT}) \tag{5.18}$$

As distinct from the discrete case, this is here assumed a continuous distribution since the various s_n may have *any* values. The source of information may now be said to exert its selective action upon this continuous ensemble of wave forms. The total probability must be unity;

hence the constraint, or *normalizing* condition:

$$\iint \cdots \int p(s_1 s_2 \cdots s_n \cdots s_{2FT}) ds_1 ds_2 \cdots ds_{2FT} = 1 \qquad (5.19)$$

In the case of discrete letters, a definite numerical value may be estimated for the probability (relative frequency) of any letter in the alphabet or set. But now we are speaking of continuous distributions and so we can consider only *probability densities*. As an everyday illustration of a density, we cannot speak of the probability of "a man's being exactly h feet tall in Britain"—we must consider a small interval of height, Δh, and speak of the probability of height lying between h and $(h + \Delta h)$. Figure 5.7 shows the very important type of density function—the Gaussian, or Normal (having here only a single variate s) and a typical interval $(1/\sigma)\Delta s$ is marked. The area of this thin slice, $p(s/\sigma) \cdot \Delta s$ has a definite probability value, inasmuch as it is a definite fraction of the *total* area lying under the distribution curve, which is unity. But the mean ordinate of this slice $\sigma \cdot p(s/\sigma)$ is not a probability, being a probability *density*. Similarly $p(s_n)$ in Eq. 5.18 is a probability density, whilst the

$$p(s_1 s_2 \cdots s_{2FT}) ds_1 ds_2 \cdots ds_{2FT}$$

in Eq. 5.19 is a probability.

Statistics relating to such ensembles are *time averages;* we have taken the set of all possible wave forms, having duration T and hence $2FT$ degrees of freedom, emitted from one particular source at different times.

Consequently, such a method of averaging is suited only to stationary sources; for only if the statistics remain unchanging with time can we assess them usefully from wave forms emitted at different times.

As we saw to be true of the case of a source of discrete signs (Section 2.2), the information rate of a continuous source should also be regarded as an *average* rate—averaged over long sequences of ordinates. On such a basis, the information rate may be expressed as the minimum number of *yes, no* instructions required to select the wave forms from the ensemble. In this case of a continuous source, the wave forms constituting the ensemble must have a large number of degrees of freedom $2FT$; that is, their duration T must be long. The root reason for the requirement arises from the Law of Large Numbers,[E] which concerns a deceptive point about our intuitive notions of a probability as a relative frequency. Briefly, it is this. Imagine a source of wave forms, quantized in amplitude into intervals Δs which may be made very small (Fig. 5.3, for example, though Δs is a coarse quantizing there). Then, over a very long time T, the fractions of the total number of ordinates $2FT$ which fall into these various quantum levels constitute an estimate of the amplitude probability

distribution. The "true probabilities" are never attainable by real-life experiments, however small the quantum intervals Δs, but represent tendencies, or mathematical limits.* For consider what wave forms *might* occur from a sequence of $2FT$ ordinates, on the assumption that there is no mutual influence between successive ordinates (that is, if they are independent events). From a sequence of $2FT$ ordinates, quantized into N levels, we can generate N^{2FT} different possible wave forms, as was emphasized in Hartley's theory (Section 2). Any sequence of ordinate amplitudes *might* occur, to constitute a wave form. It is conceivable, for instance, that a wave form might occur for which the whole sequence of $2FT$ ordinates had equal amplitudes, or even zero amplitudes; then we should say that these are not "typical wave forms" of the source. (Again, when playing cards, you have no reason to be surprised if, one day, you draw a complete hand of spades! Such a hand is just as possible as any other *stated* hand; but it is "not typical"; a "typical" hand would contain some hearts, clubs, diamonds, and spades.) In the case of our source of wave forms, suppose we actually observe it for a long time T and make an estimate of the amplitude distribution; if a second sample, also of duration T, be observed, another estimate may be made, and a third, fourth, and so on. These different estimates, made from successive wave forms of duration T, will *fluctuate* about a mean distribution. The Law of Large Numbers states the mathematical fact that the longer the sample duration T (i.e., the greater $2FT$), the greater will be the fraction of these wave forms having amplitude distributions lying very close to the "true" probability values. That is to say, non-typical wave forms will become relatively rarer. But it is important to appreciate that non-typical ones *can* occur; they merely have, by chance, fluctuations very wide of the statistical mark.[E]

5.2.2. "ENSEMBLE AVERAGES." The classical theory of communication, as developed mainly by Shannon, was concerned with *stationary* sources.[D] It was intended for application to problems arising in the telecommunication engineer's field—to telephone systems, telegraphs, television, and other systems—together with certain analogous problems in cryptography.[293] In such systems the assumption of stationariness is not a severe limitation.

But there are certain problems (some of which arise in the engineering field too) in which the changes of the signal statistics, as time passes, are of particular interest. The communication theory of learning sources would be one case, for example, but so far as your author knows, little such theory has yet been presented.[122] Various social studies, too, such

* It is legitimate to question whether in fact these limits exist, or whether they are merely assumed to, as a postulate. See reference 206 for a popular discussion.

brought about by macroscopic changes in the physical controlling factors, in the social field, the distribution of wealth may suddenly be changed by a war, a revolution, or a new system of taxation; in physics, the velocity distribution of the particles of a body of gas will be changed by application of a source of heat. But always, when dealing with the question of the relative stationariness of statistics, the time scale should be borne in mind, for all fluctuations may be smoothed out if a sufficiently long averaging time be taken. The longer this time, the more detail will be lost concerning shifts and changes taking place as the controlling factors vary.

In cases of non-stationary sources of information, time averaging cannot be used, because the estimates of the source statistics, made from successive sequences of $2FT$ wave-form ordinates, would show a steady change, the origins of these successive sequences being at different instants 0, T, $2T$, $\cdots$, et cetera, on the time axis. However, it can be appropriate and often very useful to replace this concept by that of an *ensemble average*.[B,79] For this purpose we regard Fig. 5.8 vertically, and imagine a large number of similar sources, all operating under identical macroscopic physical controlling conditions. The sources are not microscopically *identical*, but each emits its own wave forms or time sequences of ordinates. These sources all experience the same changes in the physical controlling conditions as time passes if, in fact, such changes occur to cause non-stationariness. If we label the successive sample ordinates as the 1st, 2nd, $\cdots$, nth, $\cdots$, et cetera, then an *ensemble average* may be taken over *each* of these; for example, taking the nth ordinates of the (simultaneous) wave forms of all these sources, various statistical parameters may be estimated from that collection of data. The sources being non-stationary in time, the statistics relating to the 1st, 2nd, $\cdots$, nth, $\cdots$, ordinates will in general change. Ensemble averaging is extremely useful in non-stationary system study.

It should be clear that, in stationary examples, time averaging and ensemble averaging give like results; for the successive sequences of duration T, emitted by a particular source, might well have been emitted by a succession of sources, if operating under identical macroscopic controlling conditions. But in non-stationary cases the results will, in general, differ. These ideas are equally relevant to sources possessing redundancy, which show definite probability constraints between successive ordinates, provided that such interordinate influences extend over relatively short sequences only.

The remainder of this chapter will be devoted to a barest sketch of the main concepts of the statistical theory of communication when noise is present. This condition more closely approaches reality than the ideal "noiseless" conditions assumed hitherto. We shall discuss in particular the concepts of "information rate," "channel capacity," and "equivocation." These concepts are not easy to acquire, or simple to apply correctly. They are essentially mathematical and, what is most important, they are primarily of application to certain technical problems (mainly in telecommunication) under clearly defined conditions. It is only too easy and tempting to use these terms vaguely and descriptively, especially in relation to human communication—"by analogy." The concepts and the methods of communication theory demand strict discipline in their use.

6.1. Noise, disturbances, cross-talk: the ultimate limitations to communication

In real life, all communication signals are subject to disturbances, usually beyond the control of the transmitter or of the receiver. The theory as treated so far has assumed that no disturbances are present; the source selects messages, and transmits signals, which are received without error, enabling the receiver to make an identical set of selections from his ensemble. No question of mistakes in reception arises, for no causes have yet been cited.

Disturbances may take on many forms, in practical channels. In radio reception there may be the sporadic impulsive noise of "atmospherics"; on the telephone, there may be similar crackling and hissing noises, owing to electric disturbances; a television picture may occasionally be spoiled by a splash of white dots, caused by motor-car ignition systems. There is another kind of noise of a somewhat different nature, often called "cross-talk," which can arise on faulty telephone lines, resulting in a third voice's breaking in upon the conversation. In a sense, conversation with a friend at a noisy party provides an example of a speech channel subjected to disturbance by the cross-talk of other people's speech. Cross-talk is one type of noise of particular importance; it may be specified statistically by a set of parameters in a manner similar to that for a wanted speech source.

But there is one other class of noise of outstanding interest, which has received great attention from mathematicians and physicists, often called

Gaussian noise; it is produced by the random superposition of a great number of independent causes. Historically, the first random source of this kind to be studied was the so-called "Brownian motion." In 1827 Robert Brown,[76] an English botanist, saw through his microscope the rapid and apparently random motions of minute colloidal particles suspended in a liquid—haphazard movements due to chance collisions with the liquid molecules. Figure 5.9 illustrates a part of a typical path taken by one particle, a path such as Brown himself and others since

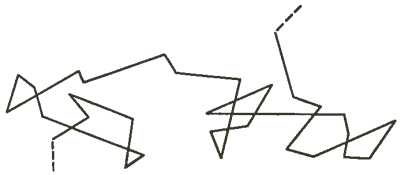

Fig. 5.9. "Brownian" (random) motion.

have tried to trace. If such a path be observed for a long time (i.e., many collisions), it is found that *all* directions are equally probable. We cannot predict and control such movements in detail, mainly because we can never know the exact positions, directions, and speeds of all the molecules at any instant of time—for there are far too many. But, fortunately, it is possible to describe and predict the motions statistically—that is, on a long-term average.[133,360] The appropriate mathematical method to be applied to such problems, involving enormous numbers of variables which can never be known in detail (microscopically) but only statistically (macroscopically), is not simple mechanics but statistical mechanics.[B]

Similar random motions arise among the electrons in all electrical conductors, in telephones, in radio receivers, and in all telecommunication apparatus, and give rise to the phenomenon of random Gaussian noise. Such random disturbing signals always exist, in varying degrees of magnitude, and are microscopically unpredictable and so cannot be allowed for or annulled. Such noise is the ultimate limiter of the fineness with which wave-form ordinates may be effectively quantized, Δs, and is the ultimate limiter of the information capacity of a telecommunication channel—the ultimate limit set by Nature.

A source of such Gaussian noise may be observed and its statistical parameters specified, like any other source of noise, or source of information. The noise disturbing a wanted source of information may either depend upon this source itself or not. Thus, statistical dependency

might be the consequence of some physical control exerted by the signal transmitter upon the source of noise. The theory of communication has so far been applied, almost entirely, to cases in which the information source and the noise source are completely independent, and the source signals and noise are simply added; any knowledge of the information source, or signals received from it, can give no information about the moment-by-moment noise values. However, communication theory

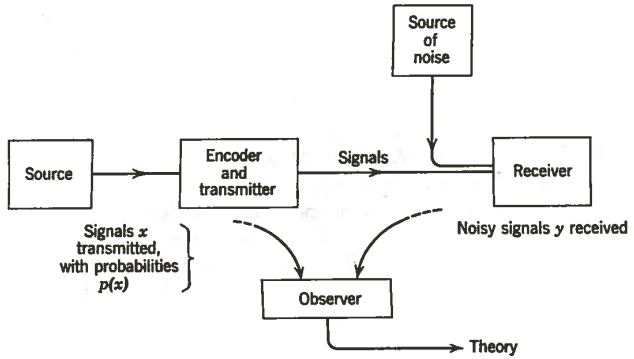

Fig. 5.10. Communication of information, when noise is present.

demonstrates the surprising fact that, solely from knowledge of the statistical parameters of the noise source, the *average* rate of loss of information may be determined.[D]

Figure 5.10 illustrates a source of information selecting messages, which are encoded and transmitted as physical signals, perhaps as wave forms. To these signals noise disturbances are directly added, before they reach the receiver. The receiver has no means of knowing by how much the true signals are perturbed, moment by moment, by this noise. The received noisy signals will consist then of two parts: first, that part representing the (wanted) *yes, no* instructions from the selective actions of the message source; and, second, that part embodying *bogus* instructions from the noise source which is making its own selections from its ensemble of random functions. These bogus instructions interfere with those from the message source and destroy information at a definite rate. The noise source thus *increases* the receiver's doubt, and we may regard it as possessing a certain rate of destruction of information ("negative information").

But we have not, as yet, considered how to specify the information rate

of a continuous source; let us do this now and show that, if noise be included, this rate cannot be infinite as seemed to be the case from our earlier arguments (Section 5).

6.2. The weighing of evidence and formation of verdicts

When noise disturbs the signals, the instructions which they embody to the receiver, to select messages from his ensemble, are not complete, perfect, or definite. The situation is then not one of precise cause and effect, but rather one of effect and *probable* cause. The received noisy signals do not completely represent the messages from the source but constitute only evidence of those messages. The receiver can, at best, weigh this evidence in the light of all the past (*a priori*) knowledge he possesses and make a verdict—his verdict or *decision* being the "best guess" about the transmitted message. And, as with all verdicts based upon limited evidence, this "guess" may be wrong.

That is the logic of the situation, and it may be described mathematically. The process of communication in the presence of noise is essentially one of inductive inference and the appropriate description of the situation is given by Bayes's theorem, which we briefly discussed earlier.*

Call the transmitted signal x and the corresponding received signal y. Then y differs from x, for it has noise in addition, or in combination in some way. The receiver's problem is to extract, from his received signal y, all the possible information about the transmitted signal x (and hence about the message represented by x), and to reject the inherent "bogus information" about the noise source.

Imagine y to be a noisy signal, received on some one specific occasion; before that moment the receiver's doubt about what signal *might* be sent depends upon the transmitter ensemble probabilities $p(x)$ (so-called *a priori* probabilities). On receiving y he possesses this as *evidence* concerning the actual transmitted x; his doubt is now represented by a new distribution $p(x|y)$, being the probability that any x was sent, when the particular y is received† (so called *a posteriori* probabilities). Then if $p(x|y)$ can be determined by the receiver, the whole of the information about x, contained in the noisy signal y, will be extracted.‡

* The suggestion that this approach might be appropriate and useful seems to have been made independently by Woodward and Davies, 1950 (reference 361), and by Cherry (see under reference 167).

† We use a different notation now for conditional probability because $p_y(x)$, etc., was used before for the special case of *transition* probabilities.

‡ See reference 136, Chapter 6, on rational decisions, for general mathematical treatment of Bayes's theorem and of its use for the weighing of evidence. Dr. Good discusses the general problem in a way immediately interpretable in terms of our message extraction problem here.

This process represents the "weighing of the evidence," but does not touch upon the verdict. That is, the process of finding the *a posteriori* distribution $p(x|y)$ does not extract the actual message. The verdict, or "best judgment" as to the actual message, is arrived at after consideration of $p(x|y)$, but, as we shall see later, the receiver does not necessarily choose the maximum value of this function (the most likely message x).

If the logarithmic measure be used, as before, then the gain in information, on receiving y, may be expressed:

$$\left.\begin{matrix}\text{Information content}\\ \text{of a received signal } y\end{matrix}\right\} = I_y = \log \frac{p(x|y)}{p(x)} \tag{5.20}$$

The following calculation is carried out in the meta-language of our *external* observer (Fig. 5.10) and not in that of a human transmitter or receiver (participant). Let $p(x, y)$ be the probability (or density if x and y are continuous) of the joint event: x transmitted, y received. From the product law:

$$p(x, y) = p(x)p(y|x) = p(y)p(x|y) \tag{5.21}$$

so that the required distribution:

$$p(x|y) = \frac{p(x)}{p(y)} \cdot p(y|x) \tag{5.22}$$

However, since y is some one definite received signal, $p(y)$ is known numerically, as a constant $1/K$, which is given by the condition that $\sum_x p(x|y) = 1$ as we shall see by example. Then

$$p(x|y) = K \cdot p(x) \cdot p(y|x) \tag{5.23}$$

As a simple illustration, there is no better example than that given by Woodward.* Suppose it rains four days out of seven, and that when it rains the barometer is low three times out of four, whilst when it is fine, the barometer is high two times in three. One day the barometer is high; what will the weather be?

Here the barometer is giving *evidence* of the weather, not an absolute indication. If F = Fine, R = Rain, whilst H = High, L = Low, we

The reader may ask: "How does the receiver assess the transmitter ensemble probabilities $p(x)$ if he never has (noise-free) access to the transmitter? Surely, his prior doubt can depend only upon the probabilities of his own, received message ensemble as gathered from his own past experience and decisions concerning the messages?" The answer is that the theory is expressed in the meta-language of an external observer [Fig. 3.2(*a*)], and it assumes the transmitted ensemble to be known at both ends.

* By kind permission. See under reference 167, p. 167.

may represent the problem as a set of equally likely possibilities, thus:

$$\begin{array}{ll} R\ R\ R\ R\ F\ F\ F & [p(R) = \frac{4}{7};\ p(F) = \frac{3}{7}] \\ L\ L\ \underbrace{L\downarrow H\ H\ H\downarrow L} & [p(L|R) = \frac{3}{4};\ p(H|F) = \frac{2}{3}] \end{array} \tag{5.24}$$

From inspection we see that $p(F|H) = \frac{2}{3}$, $p(R|H) = \frac{1}{3}$ is the required answer—the chances of fine or rainy weather when the barometer is high.

This method of enumeration is a much more self-evident demonstration of inverse probability than is direct appeal to the Eq. 5.23. However, we might instead have substituted there, giving:

$$\left.\begin{array}{l} p(R|H) = K\cdot p(R)\cdot p(H|R) = K\cdot\frac{4}{7}\cdot\frac{1}{4} \\ p(F|H) = K\cdot p(F)\cdot p(H|F) = K\cdot\frac{3}{7}\cdot\frac{2}{3} \end{array}\right\} \tag{5.25}$$

where $K = 1/p(H)$ and is given by the condition $p(R|H) + p(F|H) = 1$, so that $K = \frac{7}{3}$, which is obvious also from inspection of Eq. 5.24.

This simple example illustrates one further important point, namely that $p(y|x)$ is not really a probability density (or relative frequency) at all, because the y is one received signal (or evidence) on this one particular occasion. It has a definite value. In our example the barometer was reading high (H) on some occasion. Then $p(H|F)$ and $p(H|R)$ are really *likelihoods* of fine or rain on that specific occasion. Then, in general, $p(y|x)$ is a likelihood function of x, written $L(x)$:

$$p(y|x) = L(x) \quad \text{a likelihood function} \tag{5.26}$$

The method of enumeration, represented by Eq. 5.24, clearly shows the relations between the *a priori* probabilities $p(x)$, the *a posteriori* probabilities $p(x|y)$, and the likelihood function $L(x)$, as in Eq. 5.23. In words, we may describe these functions thus:

- $p(x)$ is the probability of message x being sent, assessed from past observations of the transmitter.
- $p(x|y)$ is the probability of an x being sent, on those occasions when y is received.
- $L(x)$ is the likelihood that, if any particular x *had* been sent, the specific y would be received.

Then Eq. 5.23 expresses the fact that the probability that a message x has been sent, in the face of some received signal evidence y, is proportional to the likelihood of x, weighted by its prior probability.

6.3. The average information rate of a continuous source, when noise is present

So much for the "information content" of a particular received signal y. Let us now consider the regular flow of signals between a transmitter and receiver and, furthermore, go straight to the case of *continuous* signals, having any of a continuous but bounded range of values.

For example, the signals might be transmitted and received wave forms, having a continuous range of amplitudes between zero and some peak value. The reader will recall that such continuous cases previously led us into difficulties (Section 5), for we saw that if the $\sum$ expression for the rate of information of a discrete source be interpreted as an integral, for a continuous source, the answer was infinity. But we have now included noise, and two statistical sources are at work, one supplying information to the receiver, one destroying it, at different rates.

Rather than write $p(y) = 1/K$, we shall now retain it as $p(y)$ because *all possible* received signal y values must now be considered; it will also be appropriate to retain the form $p(y|x)$ rather than $L(x)$. Putting Equation 5.22 in logarithmic form:[136]

$$-\log p(x) + \log p(x|y) = -\log p(y) + \log p(y|x) \tag{5.27}$$

Equation 5.22 has expressed the information content of one particular received noisy signal y; to determine the mean rate of information, we must average over all possible x and y. To do this, multiply by the joint-probability density $p(x, y)\,dx\,dy$ and integrate* over the ranges of x and y values.

$$\begin{aligned} &-\iint p(x, y)\log p(x)\,dx\,dy + \iint p(x, y)\log p(x|y)\,dx\,dy \\ &\quad = -\iint p(x, y)\log p(y)\,dx\,dy + \iint p(x, y)\log p(y|x)\,dx\,dy \end{aligned} \tag{5.28}$$

Using the product rules, Eq. 5.21, this equation simplifies; thus we may rewrite the different terms in Eq. 5.28 as follows:

$$\begin{aligned} (a)\quad -\iint p(x, y)\log p(x)\,dx\,dy &= -\int p(y|x)\int p(x)\log p(x)\,dx\,dy \\ &= -\int p(x)\log p(x)\,dx = H(x) \end{aligned}$$

the information rate of the *ideal*, noiseless source. This information rate can never be realized through our practical noisy channel; for notice the second term in Eq. 5.28:

$$(b)\qquad +\iint p(x, y)\log p(x|y)\,dx\,dy = -H(x|y)$$

which represents the average ambiguity, produced by the noise source, in the received signals y; that is, the average rate of production of doubt

* For note on this averaging process, see footnote on p. 179.

("negative information") about what actual x values are transmitted, even when the received signals y are known.

We may write the left-hand side of Equation 5.28 now:

$$H(x) - H(x|y) = R \tag{5.29}$$

the true rate of transmission of information over the noisy channel. It is the difference of two rates: $H(x)$ is the rate of production of information at the source itself, all of which is not accessible to the receiver because of the inherent effects of noise; it represents the receiver's *a priori* (average) doubt. Even after receiving the signals y, the *a posteriori* doubt $H(x|y)$ remains, because the noise renders the signals ambiguous. Thus $H(x|y)$ represents a rate of loss of information, caused by the noise, and it has been termed the channel *equivocation* by Shannon.[D] Notice that it is distinct from $H(y|x)$, which represents the rate of production of "bogus information" by the noise source.

All these rates have the units of bits per degree of freedom (as was the case for discrete sources), for we may regard the continuous signals as being defined by the values of $2F$ sample ordinates per second. Thus $2FR$ represents the channel rate, in bits per second. Once again, this measure of information rate is equivalent to a specification of the minimum number of *yes, no* instructions *about the source* messages conveyed by the noisy signals.

Take now the right-hand side of Equation 5.28.

$$(c) \quad -\iint p(x, y) \log p(y)\, dx\, dy = -\int p(x|y) \int p(y) \log p(y)\, dx\, dy$$

$$= -\int p(y) \log p(y)\, dy = H(y)$$

which, by analogy with $H(x)$, represents the "information rate" of the received signals y. But some of this is *bogus* (information about the noise source itself). The rate of bogus information is:

$$(d) \qquad +\iint p(x, y) \log p(y|x)\, dx\, dy = -H(y|x)$$

representing, on an average, the doubt about what y will be received, even when the transmitted signals are known. It should be remembered that it is the external observer who assesses these quantities, not the receiver himself.

Now we have another, and alternative, expression for the true rate of information:

$$H(y) - H(y|x) = R \tag{5.30}$$

which is similar to Equation 5.29, but with x and y reversed. Again this is

the difference of two rates; the rate corresponding to the received signals y, less the "negative" or bogus information rate of the noise source.

The true information rate R is thus, in both forms, given by the *difference* of two integral expressions. It is this fact which renders R finite, although each of the integrals might become infinite. We have not in fact proved here that this difference is finite, but would refer the reader to the original work,[D,*] because our purpose is not to present a condensed version of the theory, but rather to survey and discuss its basis, its objects, and its restrictions.

7. THE ULTIMATE CAPACITY OF A NOISY CHANNEL

Shannon's most important contribution to statistical communication theory is undoubtedly his Capacity Theorem;[D,328] this gives a result which would certainly not be suspected intuitively. It is this: *It is possible to encode a source of messages, having an information rate H, so that information can be transmitted through a noisy channel with an arbitrarily small frequency of errors, up to a certain limiting rate C, called the limiting capacity*, which depends upon the channel constraints (e.g., bandwidth, power restrictions, noise statistics, etc.), provided that $H \leqslant C$.

It might at first be thought that, since noise is present, errors are inevitable; or that perhaps redundancy could be added so as to combat the noise to some extent, but never to remove errors entirely, for this implies that information would be sent with absolute *certainty*, in spite of the unpredictable noise! In fact, any attempt to transmit at a higher rate than C will cause errors; but at any rate below C the errors can be made, in theory, vanishingly few. It is emphasized: *in theory*. For the practical accomplishment of such ideal codes has proved to be of extraordinary difficulty,[D,131,142,328,†] and is somewhat discouraged by the fact that the types of modulation and coding which have been invented already by telecommunication engineers have proved to be remarkably efficient.[252,296,‡] But the fact that the engineer has "got there first" does not detract one iota from the value of this theorem. Practical accomplishment so frequently precedes theory. The value here lies in the establishment of a *limit* to the capacity; anyone who tries to beat this limit is wasting his time! In this light, the Capacity Theorem is similar to the concept of Conservation of Energy.

* See also reference 133 for a very full discussion of this question. The basic reason why R is finite is that, although both $H(x)$ and $H(x|y)$ have magnitudes which depend upon the co-ordinates of x and y, their *difference* R is invariant under a transformation of these co-ordinates.

† See also Laemmel under reference 166 and Shannon under reference 167.

‡ See also Jelonek under reference 166.

7.1. Received information and the extraction of messages

One most significant point about the formulae for the rate of transmission of true information through noisy channels (Eqs. 5.29 and 5.30) is that they are expressed entirely in terms of probability distributions, $\log p(x)$, $\log p(y|x)$, et cetera, or their ensemble averages. The information content of a received signal y has been regarded as the logarithm of the ratio of the posterior to the prior probabilities (Eq. 5.20) of the different possible transmitted signals $x_1 x_2 \cdots x_n \cdots$. But, in practice, communication cannot be said to be established, between a transmitter and a receiver, if the receiver gets nothing but probabilities! A Teletype machine prints definite letters, not probability functions. Nevertheless the production of the posterior function $p(x|y)$ represents the extraction of the information content of the noisy signal y; but, at some stage, one definite value of x must be selected, based on the $p(x|y)$ evidence, as the "best choice" determining the received message.

Curiously enough, the "best choice" need not be the most probable value of x, although in fact it usually is. As I. J. Good has emphasized,* this choice may depend upon the future consequences or upon the purposes of the message; more generally, to borrow a term from the economists, the choice depends upon the *utilities* involved.† Good quotes a most convincing example, drawn from radar (a form of telecommunication very thoroughly treated, from the present point of view, by Woodward[360,362] and by Davies[78]), illustrated by Fig. 5.11. Suppose a radar station is given advance information whenever an enemy aircraft is approaching at a range lying between 100 and 400 miles. The radar receiver problem is to determine the correct range as accurately as possible. The various "possible ranges" now represent messages, x. Before a radar signal is received, the prior range probability $p(x)$ is assumed uniform between the 100 and 400 mile limits. Now suppose a noisy radar signal y is received and the complete posterior probability $p(x|y)$ determined, having the form shown in the figure, with a maximum value at range $x = 270$ miles but with a smaller peak at $x = 150$ miles. The radar operator might nevertheless decide to take action on the basis of the smaller peak at 150 miles, because this represents a more immediate danger.

The *whole* of the posterior distribution $p(x|y)$ represents information; it represents the receiver's "degree of belief" that any particular range x

* See discussion by Good under reference 166, p. 180.

† *Utility* is defined as "reasonable measure of value" (e.g., of money). The concept goes back to Bernoulli, in the early history of probability theory and its application to gambling. See reference 136, p. 52.

is the true one. If one point be chosen as the assumed "true signal x" and the remainder of the curve rejected, then information is thrown away.* This may be illustrated by the following argument. Suppose the choice be deferred and a second signal received, for example in this radar case. Then $p(x|y)$ now becomes the *prior* probability of x, for this

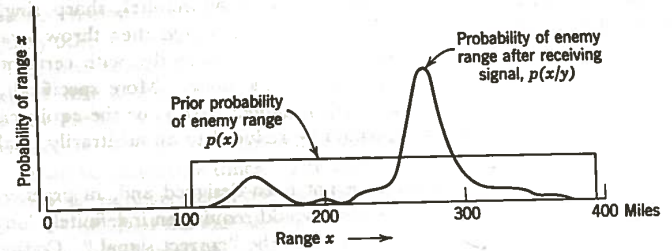

Fig. 5.11. Measurement of a target range by radar.

second observation. Suppose the process is continued and a series of consecutive signals are received, $y_1 y_2 \cdots y_r \cdots$, so rapidly that the true x (enemy range) remains substantially constant. Then, from Eqs. 5.23 and 5.26:

$$\left.\begin{aligned} &\text{After 1st observation} && p(x|y_1) = K_1 p(x) L_1(x) \\ &\text{After 2nd observation} && p(x|y_1 y_2) = K_2 p(x) L_1(x) L_2(x) \\ &\text{After 3rd observation} && p(x|y_1 y_2 y_3) = K_3 p(x) L_1(x) L_2(x) L_3(x) \end{aligned}\right\} \quad (5.31)$$

and so on.

It will normally happen that the probability curve will become sharper and sharper, centered upon the true x, though this is not inevitable,[360] because the successive true signals will be related whilst the successive noise contributions will be random.

This is similar to adding redundancy at the source by simple repetition of x. However, a radar target is an example of a particularly *unco-operative* source; the enemy does not obligingly code his radar echoes, adding redundancy as required, so as to overcome the noise disturbing the receiver!

* This whole question of the determination of the "best" signal x, when noisy signals are received, may be regarded as the testing of statistical hypotheses. The alternative "hypotheses" are the possible signals $x_1 x_2 \cdots x_r \cdots$ and the choice of any one carries with it some probability of error. For discussion of the various types of test, in relation to this problem of signal detection, see Middleton under reference 166. See also reference 78.

In more usual, friendly telecommunication systems, the transmitter and receiver co-operate. Coding may be designed to include redundancy in the best possible way (as limited in practice by ingenuity and economy) so as to overcome the noise and make the receiver's final selection of the "assumed correct signals x" from the posterior distribution $p(x|y)$ easier, and his chances of error fewer. Clearly, then, if ideal coding could be found, it should be such as to reduce $p(x|y)$ to an infinitely sharp, single peak. The selection of the "assumed correct x" would then throw away no information; the signal would be received correctly, with certainty and with no chance of error, in spite of the noise. More specifically, it is the ensemble average of $\log p(x|y)$, namely $H(x|y)$ or the equivocation, given by Eq. 5.29, which would be reduced to an arbitrarily small value by ideal coding.[D]

Such ideal coding methods have not been designed and, in practice, they would be unusable, because they would require an indefinitely long postponement of the final identification of the "correct signal." Coding which involves indefinitely long delay is impracticable, and some compromise must be sought.

7.2. Statistical matching of a source to a noisy channel

We have already made some preliminary discussion of *statistical matching* of a source to a channel of transmission, in Section 5.1. A channel, such as a telephone or telegraph channel, for example, exerts certain constraints upon the signals it transmits; in particular it restricts the electrical power available, and the bandwidth. In Section 5.1 we referred to the problem of coding the messages from the source in the best way, for transmission, subject to these constraints, where "best way" implied transmission of information at the maximum possible rate. However, we abandoned our discussion there, when it became clear that factors other than available bandwidth and power determine this maximum rate. We now see that this new factor is the noise. The noise also exerts a constraint upon the channel, and the manner of adding redundancy to the source messages, so as to change their statistical structure in the "best way," depends upon the structure of the noise.

The problem of statistical matching is to find a suitable code for the source such that the ensemble of transmitted signals is given a statistical structure which maximizes R, the rate of transmission of information through the noisy channel. From Eq. 5.29 and the integral expressions given there for $H(x)$ and $H(x|y)$, this ultimate capacity of a noisy channel, attained by such statistical matching, may be expressed thus:[D]

$$C = \lim_{T\to\infty}\left[\max_{p(x)} \frac{1}{T}\iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\,dx\,dy\right] \text{ bits per sec} \quad (5.32)$$

When we speak of transmitted signals x, these may be taken to be wave forms of duration T and bandwidth F; consequently, such signals are specified by the values $x_1x_2\cdots x_{2FT}$ at $2FT$ equi-spaced instants, so that the transmitted ensemble probability distribution has a finite dimensionality $2FT$. That is $p(x) = p(x_1x_2\cdots x_{2FT})$. This problem of maximizing the rate of information, as the integral expression, Eq. 5.32, over all possible ensembles $p(x)$ and subject to fixed power, bandwidth, and possibly other constraints, is an exercise in the calculus of variations.[D,122,*]

Repetitive redundancy, to which we referred in the last section, is the simplest way of combating noise and reducing the equivocation at the receiver. It involves prior agreement between the communicating parties that each transmitted sign (letter; binary-code 1, 0; wave-form ordinate, etc.) shall be repeated n times. The receiver then has a better chance of assessing the signs correctly, but the price he pays is a delay in the process; he must wait until the end of each sequence before making his decision. The same price is always paid; statistical coding involves delay, and this delay becomes longer and longer as better coding is employed, for transmission and errorless reception, at a rate approaching the ultimate capacity C of the channel. This rate can then in practice never be attained, but only approached asymptotically. We may infer that this is so from our earlier argument in Section 4. All forms of redundancy operate by calling upon past experience; perhaps by the inclusion of known digram, or trigram, constraints; perhaps by including the statistical influence of signs extending even farther back into the past. But to extract the ultimate information out of any sign, we should require to know all the statistical constraints upon it, involving knowledge of the preceding signs extending indefinitely far back into the past. Ideal coding involves taking into account, in the transmitter, indefinitely long blocks, or run-lengths of messages.

8. MANDELBROT'S EXPLICATION OF ZIPF'S LAW —CONTINUED

We are now able to take up again the threads of an earlier discussion (Chapter 3, Section 5.2) concerning Zipf's experimental "law," illustrated by Fig. 3.5 and Mandelbrot's theoretical treatment of this.† In this earlier chapter we were discussing within the field of linguistics; let us now treat messages strictly as sequences of words, each a sequence of

* See also Jelonek under reference 166. These authors have calculated a number of channel capacities for different signaling systems and noise conditions.

† See Chapter 3, Section 5.2 for references.

letters,* and regard written language as a "code." Difficulties concerning "the word" as a linguistic concept will not be raised again here.

In our earlier section, we referred to Mandelbrot's concept of the "cost" of letters and words (signs). Let c_n represent the cost of a word (assumed to be given) of rank order† n in the language, and let p_n be its frequency of occurrence (Fig. 3.5). Then the average cost, per word, of messages will be:

$$\text{Average cost per word} = \sum_n p_n c_n \tag{5.33}$$

Mandelbrot proceeds first to minimize this average cost, by carrying out a variation of the distribution of p_n over the different words; that is, he finds the optimum, "cheapest" word ensemble. This minimization process is carried out with the information rate (per word, average) held invariant; but the term "information rate" as used here needs a little clarification.

Shannon has shown that messages may be coded most efficiently if the process is carried out over long blocks of words, although such coding inevitably requires correspondingly long time delays.[D] But Mandelbrot points out that his own problem is different, since human language is uttered or written under conditions which cannot permit such very long time delays. Shannon's ideal coding would be very *efficient* (in information per sign) but not very *practical*. Mandelbrot makes the assumption of a constraint upon the tolerable delay, equal to the word length; that is *words* are considered to be coded one at a time. Again, every word is considered to end with a certain sign, "space," which never occurs inside a word. If the maximum message information rate be taken as Shannon's H_n, we have

$$H_n = -\sum_n p_n \log p_n \text{ bits per word} \qquad [(5.6)](5.34)$$

with coding carried out using very long blocks. With this rate held invariant, the "cheapest" ensemble of words is shown to have the distribution:

$$p_n = Q\epsilon^{-Kc_n} \tag{5.35}$$

where Q and K are constants.

This result accords more or less with intuition, since it requires the most frequent words to be the cheapest. (We have already observed that Morse's code was based upon a similar assumption, applied to letters whilst Fano's code represents a more formalized version,[C,D,‡] if in both

* Or similarly with phonemes and transcribed texts.

† See footnote, p. 102.

‡ See also Huffman under reference 166. See our Fig. 5.5(*b*)

cases we take the length of the code sequences as a measure of their "cost.")

Notice that Eq. 5.35 implies that the rank order of the words is the same, whether quoted with respect to increasing cost c_n or decreasing probability p_n, since the exponential function is monotonic.

Another step in the theory leads to a relation between the cost of a word and its rank order. Mandelbrot considers words, in a first approximation, as random sequence of letters and spaces, and all conceivable sequences of letters of the alphabet are admitted as possible "words." The question, how to assign costs to the various letters of the alphabet, is answered by assuming, initially, that all letters are equally costly; subsequently it is shown that *any* distribution of costs will suffice, for, surprisingly, the choice makes no appreciable difference to the main conclusions. (From Eq. 5.35 we see that assignment of equal costs to letters implies also that all the letters, but not spaces, are equally probable.) Thus the cost of a word is equal to the sum of the costs of its letters, so that if letters be assumed to be equally costly, the cost of a word is proportional to the number of letters contained. Further, the longer any sequence of letters, the more the words that may be constructed having this length. Then, in an alphabet of M letters:

There are M possible, equally probable, equally costly 1-letter words.
There are M^2 possible, equally probable, equally costly 2-letter words.
There are M^3 possible, equally probable, equally costly 3-letter words.
. . . . . , etc.
There are M^l possible, equally probable, equally costly l-letter words.

In this table the word groups are ranked from top to bottom, as $1 \cdots l \cdots M$-letter sequences. They are therefore ranked in groups of increasing cost, on a linear scale, so that from Eq. 5.35 they are also ranked in groups of decreasing probability.

The various l-letter words, within any one group, may be regarded as ranked in arbitrary order. But by rank order, in Zipf's law, it is the order of every *word*, not word group, in the language which is meant. Thus we can say that, approximately, the rank order of any *word* of length l letters is n_l, being equal to the sum of all words of length equal to, or less than, l:

$$\left.\begin{aligned} n_l &\simeq 1 + \sum_{\lambda=1}^{l} M^\lambda \\ &= M^l \cdot \frac{M}{M-1} - \frac{1}{M-1} \end{aligned}\right\} \tag{5.36}$$

If now we write $M/(M-1)$ as M^{-l_0}, then:

$$M^{l-l_0} = n_l + \frac{1}{M-1}$$

or
$$l = l_0 + \log_M \left(n_l + \frac{1}{M-1} \right) \tag{5.37}$$

This shows that, to a first approximation, the length of any word is proportional to the logarithm of its rank order, with a correction which is serious only when $n_l \ll 1/(M-1)$. But, cost being proportional to word length, we may rewrite Eq. 5.37, dropping the subscript l, as:

$$c_n \simeq c_0 + \log_M n \tag{5.38}$$

and substituting this in Eq. 5.35 eliminates the costs c_n:

$$p_n = Pn^{-B}$$

where P and B are constants that depend upon the K and the Q in Eq. 5.35, and through K and Q, upon the information which we wish to transmit per word, or upon the average cost of transmission per word.

This, of course, is Zipf's law. But in this form, we see the role played by the index B as a measure of the *variety* of our available vocabulary. The smaller B is, the greater the variety.

As illustrated here, Mandelbrot's arguments have been reduced to their simplest terms. He has shown, however, by slightly more involved reasoning, that the relationship in Eq. 5.37 still holds, if *any* costs be assigned to the various letters of the alphabet—or even if the cost of any letter in a word depends upon the preceding letter*—so that this work may bear more relation to real-life printed language (and perhaps other human social constructs) than at first appears to be the case, with this simplest model discussed here.

Mandelbrot proceeds to develop analogous relations between his whole theory and certain results of thermodynamics, and we should refer the reader to his original texts. This question of the relationship between statistical communication theory and statistical thermodynamics has been deliberately avoided in this chapter, until now, for it is the writer's opinion that there is little necessity to make such comparisons, for the newer theory may well stand upon its own rights. However, this has frequently been done, especially invoking the concept of entropy; a few words on the subject may not be out of place at this point.

9. COMMENTS UPON INFORMATION INTERPRETED AS ENTROPY

Communication provides an example of a process which we regard as proceeding from the past into the future; time, we say, "has a direction." Phonograph records played backward sound as senseless gibberish.

* See Mandelbrot under references 26, 41.

A movie, in reverse, produces comic results—a diver rising from the water, landing on tiptoe; torn scrap paper coming together into folded news sheets; a drinker regurgitating a pint of beer into a glass. The world, run backward, looks ludicrous.

Yet Newton's laws of motion—the backbone of physical science—are reversible; time can have a positive or negative sign. We appear then to regard time in two distinct ways, reversibly and irreversibly. On one hand, if we study, say, the properties of some simple frictionless machine containing relatively few moving parts, we can calculate its precise motions, in detail; we may learn all about it and predict its future behavior with accuracy. In the equations of such mechanical motions, the sign of time may everywhere be reversed, with complete consistency. On the other hand there are whole realms wherein the "direction" of time is of major importance—in studies of life processes, of meteorology, of thermodynamics, or again in philosophical questions concerning "creative thinking," "intelligent beings," and many others.[26,289]

This concept of the apparent irreversibility of time has received its most elaborate mathematical formulation in thermodynamics, and is expressed in terms of the so-called Second Law, which holds that a certain quantity called *entropy* can never decrease. Thermodynamics was originally concerned with the properties of gases—that is, enormous assemblies of particles in violent motion. Of such assemblies we can have only partial knowledge; although Newton's laws apply to every individual particle, we cannot observe them all, or distinguish one from another. Their properties cannot be calculated precisely, like those of a simple machine, but may be discussed only in terms of probabilities, stochastically. We may measure and so learn about their macroscopic properties—their number of degrees of freedom, or dimensionality; their pressures, volumes, temperatures, energies. We may represent certain properties by statistical distributions, such as the particle velocities for example. We may, with great difficulty, observe some microscopic motions, but we can never have *complete* knowledge of every particle of the system.

Likewise with other systems, of which communication is an important example; it is not surprising that the same mathematical methods should be considered as applicable. We can have only partial knowledge of a communication source. We may know the ensemble properties, the coding system, and various constraints upon the messages or the signals; but we (as recipient or participant-observer) cannot know, *a priori*, the moment-by-moment states of the source, the exact messages it will give out next, in microscopic detail, or we should have foreknowledge and receive no information from the signals.

But it was the later formulation of the laws of thermodynamics in

terms of probabilities, in the classic work of Boltzmann and Gibbs in particular,[B] as a statistical-mechanical interpretation of the properties of gases which showed the great generality of the laws and concepts. The existence of a relationship between "entropy" and "information" is, in fact, inherently shown in their work, though the explicit relation was first shown, it appears, by Szilard, in a discussion upon the old problem of "Maxwell's demon."[318,*] This problem, and the entropy-information relation, has subsequently been discussed by Wiener,[349] and by Brillouin in particular.[35–37]

Entropy, in statistical thermodynamics, is a function of the probabilities of the states of the particles comprising a gas; information rate, in statistical communication theory, is a similar function of the probabilities of the states of a source. In both cases we have an *ensemble*—in the case of the gas, an enormous collection of particles, the states of which (i.e., the energies) are distributed according to some probability function; in the communication problem, a collection of messages, or states of a source, again described by a probability function.

The relationship between information and entropy is brought out most objectively by the Wiener-Shannon formula, Eq. 5.8:

$$H(i) = -\sum p_i \log p_i \qquad [(5.8)]$$

which (with a positive sign) bears resemblance to Boltzmann's formula for the entropy of a perfect gas. Now, when such an important relationship between two branches of science has been exhibited, there are two ways in which it may become exploited; precisely and mathematically, taking due care about the validity of applying the methods; or vaguely and descriptively. Since this relationship has been pointed out, we have heard of "entropies" of languages, social systems, and economic systems and of its use in various method-starved studies. It is the kind of sweeping generality which people will clutch like a straw. Some part of these interpretations has indeed been valid and useful, but the concept of entropy is one of considerable difficulty and of a deceptively apparent simplicity. It is essentially a mathematical concept and the rules of its application are clearly laid down.

In a descriptive sense, entropy is often referred to as a "measure of disorder" and the Second Law of thermodynamics as stating that "systems can only proceed to a state of increased disorder"; as time passes, "entropy can never decrease." The properties of a gas can change only in such a way that our knowledge of the positions and energies of the particles lessens; randomness always increases. In a similar descriptive

* The paper by the same author quoted by Weaver (see reference B, p. 95) appears to be a wrong reference.

way, information is contrasted, as bringing increasing order out of chaos. Information, then, is said to be "like" negative entropy. But any likeness that exists, exists between the mathematical descriptions which have been set up; between formulae and method.

Shannon refers to $H(i)$, as given by Eq. 5.8 above, as the "entropy" of a discrete source of information, having a finite number of states with known probabilities $p_1 p_2 \cdots p_n$. Wiener, earlier, has referred to "negative entropy" in a similar context, and there is a certain difference of point of view. Both physical entropy and information can be only relative, never absolute; we can only have changes. The reader will remember this point was brought out earlier, since the corresponding $H(x)$ for a *continuous* noiseless source appeared to be infinity (Section 6.3). In this case $H(x)$ becomes:

$$H(x) = -\int p(x) \log p(x)\, dx \qquad (5.39)$$

This represents the receiver's prior average doubt, or uncertainty; that is, it represents the "entropy" of the source ensemble. If a signal y is received from this source, perturbed by noise (the noise source itself having a certain "entropy"), the receiver's uncertainty concerning the message state of the source becomes changed—usually lessened—by the quantity I_y, given by Eq. 5.20, the information content of that signal y. If now signals are steadily received, the receiver's uncertainty reduces at an average rate R, given by the averaged contents of all the received signals, which was expressed by Eq. 5.29. This rate R is then the rate of received information, or the *negative* "entropy" (per sign, per degree of freedom, or per second, as required).

This aspect of communication is one special view of a general situation in physics—that of an observer "receiving information" from a physical system under observation. Physical (thermodynamic) entropy is defined for a *closed* system, a system which is considered utterly isolated and incapable of exchanging energy in any way with its surroundings. Again, the term is usually applied to systems which are in a state of near-randomness, and which consist of *truly enormous* systems or assemblies of elements.

In Szilard's discussion of the Maxwell demon problem, the demon was regarded as "receiving information" about the particle motions of a gas, this information enabling him to operate a heat engine and set up a *perpetuum mobile;* the demon was making use of his information, not simply receiving it and passing it into storage. This suggests a violation of the Second Law. But the demon is essentially a participant-observer and must receive energy, in order to make his observations, and so he himself must be regarded as part of the system.[35] As Szilard had shown, in his

1929 paper, the selective action represented by the demon's observations must give rise to an increase of entropy *at least* equal to the reduction he can effect by virtue of this information. The system and the demon exchange entropy, but no overall reduction is necessitated.

But these more general questions take us off our track. Questions of extracting information from Nature and of using this information to change our models or representations lie outside communication theory—for an observer looking down a microscope, or reading instruments, is not to be equated with a listener on a telephone receiving spoken messages. Mother Nature does not communicate to us with signs or language. A *communication channel* should be distinguished from a *channel of observation* and, without wishing to seem too assertive, the writer would suggest that in true communication problems the concept of entropy need not be evoked at all. And again, physical entropy is capable of a number of interpretations, albeit related, and its similarity with (selective, syntactic) information is not as straightforward as the simplicity and apparent similarity of the formulae suggests. This wider field, which has been studied in particular by MacKay,[218] Gabor,[123] and Brillouin,[35] as an aspect of scientific method, is referred to, at least in Britain, as *information theory*, a term which is unfortunately used elsewhere synonymously with communication theory. Again, the French sometimes refer to communication theory as *cybernetics*.[40] It is all very confusing!

BIBLIOGRAPHY

A. Hartley, R. V. L., "Transmission of Information," *Bell System Tech. J.*, 7, 1928, p. 535.

B. Tolman, R. C., *Principles of Statistical Mechanics*, Clarendon Press, Oxford, 1938.

C. Fano, R. M., "The Transmission of Information," *M.I.T., Research Lab. Electronics, Tech. Rept., 65*, 1949.

D. Shannon, C. E., and W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, 1949.

E. Feller, W., *An Introduction to Probability Theory and its Applications*, Vol. I., John Wiley & Sons, Inc., New York, 1950.

F. Woodward, P. M., *Probability and Information Theory, with Applications to Radar*, Pergamon Press, Ltd., London, 1953.

CHAPTER SIX

On the Logic of Communication (Syntactics, Semantics, and Pragmatics)

He was . . . 40 years old before he looked upon geometry; which happened accidentally. Being in a gentleman's library . . . Euclid's Elements *lay open and 'twas the 47 El, libri I. He read the Proposition. "By g*——" say'd he "this is impossible!" So he reads the demonstration of it, which referred him back to such a proposition; which proposition he read. That referred him back to another, which he also read.* Et sic deinceps, *that at last he was demonstratively convinced of that trueth. This made him in love with geometry.*

* (He would now and then sweare, by way of emphasis.)

John Aubrey (1626–1697), concerning
Thomas Hobbes
Brief Lives, Volume I, 1680

1. "SIGNIFICS"—OR MENTAL HYGIENE

The Honorable Lady Welby, who was Lady-in-Waiting to Queen Victoria, pioneered a movement, at the turn of the century, to tighten discipline of thought and expression in many fields of human interest, in