INFORMATION THEORY AND SOME OF ITS APPLICATIONS

B. S. RAMAKRISHNA

Department of Electrical Communication Engineering,
Indian Institute of Science, Bangalore

INTRODUCTION

IT is difficult to trace the precise origins of many scientific theories. With information theory, sometimes called communication theory, one need not go back farther than the twenties when Nyquist and Hartley tried to develop a quantitative measure of information to assess the capacities of telecommunication systems. It is only during the last decade or so, however, that a theory of information has been developed and its concepts have found widespread use outside telecommunication engineering. Norbert Wiener, to whom the basic philosophy of modern information theory is due, was the first to recognize the universal character of the communication problem encountered not only in telecommunication systems but also in living beings and social organizations. We read in his book on Cybernetics a panoramic description of the growth of these ideas against the background of the problems of the last war. A little later, in 1948, Claude E. Shannon of the Bell Telephone Laboratories published his classic papers on the mathematical theory of communication. Although Shannon was concerned in his work primarily with the telecommunication porblem, the mathematical model he set up for a communication system has been found to be useful in many different disciplines and the concepts of communication theory have penetrated into fields as far away as linguistics, psychology, neurophysiology and others. Today, ten years later, the domain of information theory extends far beyond telecommunication engineering and some of its most interesting problems lie just on the boundaries of telecommunication and other sciences. Nevertheless, in an expository account it is desirable to develop the fundamental concepts of information theory within the confines of telecommunication engineering before their use in other fields is considered. We shall lean heavily on Shannon's work in introducing the current notions of communication theory.

ANATOMY OF TELECOMMUNICATION SYSTEMS

In the interests of a general theory of communication we must abstract from the wide variety of communication systems the essential features which they all have in common. Every communication system is primarily a device for transmitting messages from their sources to their destinations. These messages may be spoken words with an acoustic pressure-time pattern as in telephone conversation, written characters as in telegraphy or the colour and intensity patterns of light from an object being televised or any other set of symbolic patterns. They may even be numerical data relating to some physical quantity such as temperature or density under observation. A very successful model of a general communication system, due to Shannon, may be represented schematically by the block diagram of Fig. 1.

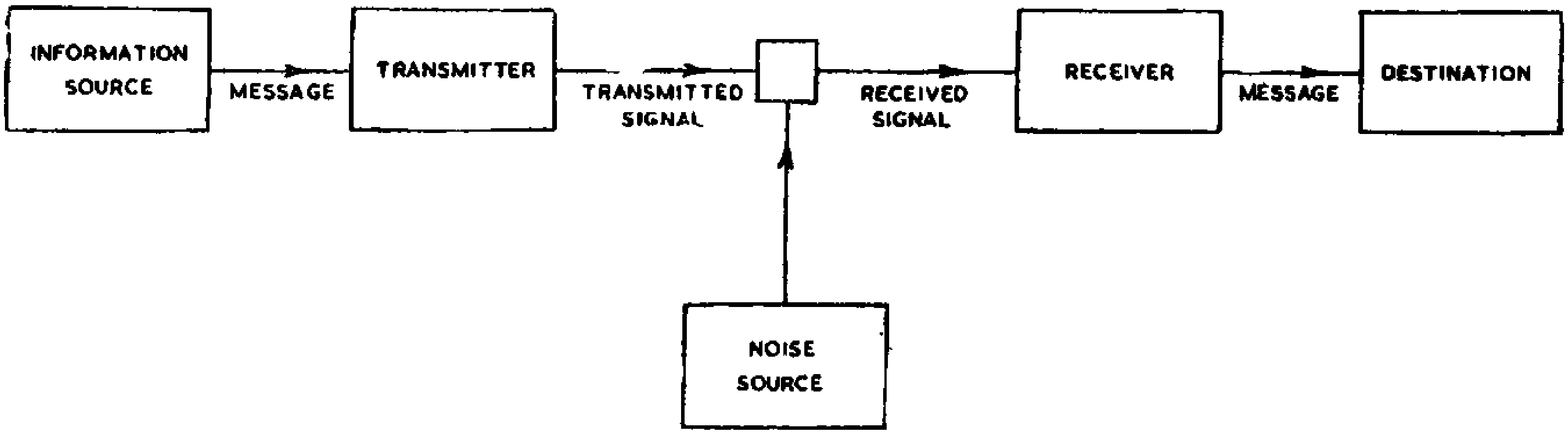

FIG. 1.

Operationally, we may characterise the different elements in this system as follows:

Ignoring all questions concerning the motivation of the messages, we may regard the function of the information source is to generate sequences of symbols or patterns which constitute the messages. We shall first deal with discrete sources which use only a discrete set of symbols like the letters of the alphabet and then take up continuous sources which produce continuously variable patterns like those of the speech-waves. The messages ordinarily used in practice convey a meaning because of the fact that the symbols used, *i.e.*, the written words, spoken sounds, etc., are associated with certain concepts, but we must here make a distinction between the *significant*, the symbol and the *signifie*, the entity which is symbolized. In the development of the telecommunication theory, the semantic aspects of the messages were set aside as the engineering of communication systems does not depend upon the meaning of the messages and much less upon their motivation. We should thus look upon messages which communication systems handle as an ensemble of sequences of symbols. It is the statistical rather than the semantical aspects of the messages that concern us. We are thus led to regard an information source as a stochastic or a Markoff process generating its messages symbol after symbol.

In electrical communication the original message symbols are frequently converted into a different set of symbols more suitable for transmission over the communication medium or channel. In telegraphy the letters of the message are converted into the sequence of dots and dashes for signalling according to the Morse code; in telephony the acoustic wave is converted into the corresponding electrical signal. In the above representation the transmitter which may involve a human operator thus performs the operation of producing the sequence of signal symbols at its output from the message symbols fed to its input. Functionally, the transmitter is an operator which maps the message space on the signal space.

The intervening medium between the transmitter and the receiver is called the channel. The channel may be susceptible to noise in which case the signal reaching the receiver differs from the transmitted signal. The noise, like the message, may be regarded as the output of a stochastic or a Markoff process. For instance, the noise in a pair of telephone wires may be due to cross-talk from an adjacent pair. It may be simply due to the random thermal motions of the electrons in the circuit elements. In any event, the effect of noise is to alter the signal in an unpredictable manner except in a statistical sense. It will be seen that the ability of a channel to transmit messages depends upon the band of frequencies it can transmit (the bandwidth of the channel) and the level of the signal relative to that of the noise.

The function of the receiver is to reconstruct the original message from the received signal and hence it may be thought of as an inverse operator to the transmitter. If the received signal is badly perturbed by noise, correct reconstruction of the message from the received signal may not be possible and there remains some uncertainty about the original message. The destination is obviously the terminating point for the message and may be a recording device like a photographic film or a magnetic tape or even a human sense organ like the ear or the eye. It comes into consideration here because the resolving power of the terminating device determines the degree of detail that we need transmit.

A Measure of Information

Having described the basic elements of a generalized communication system we proceed to develop the fundamental concept of information. When we ignore the meaning and look at the message as the output of a stochastic process (as a cryptanalyst does when deciphering a cryptogram) we begin to notice that the various symbols of the message (letters, etc.) are not entirely random but exhibit certain statistical regularities such as the constant frequencies of the letters, etc. The statistical properties as defined, for instance, by the probabilities of the occurrence of the different symbols, the transition probabilities between successive symbols, etc., enable us to define a quantitative property of the message called the information content or entropy of the message. We are certainly not using the word information here in an unconventional sense although we intend to attach a numerical measure to the information contained in the message. To be sure, the information content of a message is not the same thing as its meaning; meaning or rather its comprehension has a subjective side while the information is a measurable quantity without reference to the meaning of the message.

To develop a quantitative measure of information consistent with its commonsense usage, notice that we seek information only when we are in doubt, which arises when there are a number of alternatives or choices and we are uncertain of the outcome of the event. We go to an enquiry office (rightly called information office in America) to remove our doubts; we consult weather forecasts for information whether it will be rain or sunshine; we are, in fact, seeking information (in the form of data) when we perform experiments whether they are launching of sputniks or the testing of nuclear weapons. On the other hand, if an event can happen in only one way, there is no choice or uncertainty about it and no information is called for either. Obtaining information is equivalent to making a choice thereby removing the *a priori* doubt. Choice, uncertainty or doubt and information thus all come to possess the same measure.

Any written message in English can thus be regarded as the result of a sequence of different choices from the 26 letters of the English alphabet. It will be convenient to regard the space between words as a symbol by itself and consider the alphabet to have 27 letters instead. We can thus form no more than 27^N different messages of N symbols length. Only a very tiny fraction of these correspond to the conventional use of English and are thus used in preference to the others. What is more, we can associate with each message a certain *a priori* probability if we have sufficient knowledge of the statistics of the message symbols, *i.e.*, their probabilities, etc.

How much information does a given message contain? First, we need a unit to measure information and then the probability measure of the message. The most elementary type of choice we have is the choice between two equally probable alternatives (*e.g.*, the choice between the heads and tails in the tossing of a coin). For reasons which will become more convincing as we proceed, we shall choose the logarithm to the base 2 of the number of alternatives as the amount of information H associated with the choice so that, in the binary choice referred to above, we obtain one unit of information ($H = \log_2 2 = 1$) which is designated as a bit. If there are N equally probable alternatives, we obtain $H = \log_2 N$ bits of information with the specification of any one of them. In this case since the probability of any of the alternatives is $p = 1/N$, the definition is equivalent to choosing the negative logarithm of the probability of the event as the measure of the information associated with the selection.

The various symbols, however, do not occur with equal probability (the letter *e* has the highest frequency of occurrence, about 13% and *z* has the least, about 0·09%). If different symbols have different probabilities p_i ($i = 1, 2, \ldots n$), and occur independently of each other, the average amount of information per symbol (of the event *x*, say), may be shown to be given by

$$H(x) = - \Sigma p_i \log_2 p_i \text{ bits.} \quad (1)$$

The only requirements imposed by Shannon in obtaining this measure are that (*i*) H should be a continuous function of the p_is, (*ii*) when all the p_is are equal, *i.e.*, each $p_i = 1/n$, it should be a monotonic increasing function of *n*, and (*iii*) that if the choice be made in successive stages, the weighted sum of the individual values of H associated with each stage must be equal to the value of H obtained by direct selection. Notice, however, that the expression (1) can be interpreted as the weighted average of the information obtained with the selection of each symbol, the weight factor being the probability p_i of the symbol. The formal resemblance of this expression for the amount of information to the entropy of a thermodynamical system which can have *n* different complexions with probabilities p_i cannot escape notice here. For this reason the term entropy is frequently used to refer to the average amount of information associated with a set of alternatives.

We digress for a short while here to examine the stochastic character of actual messages. Shannon and Miller have given striking demonstrations of how we approach actual languages by merely choosing successive letters or words incorporating longer and longer probabiliy constraints in their choice (see references 1 and 8).

A typical 'sentence' of zero-order approximation is obtained by choosing all the letters with equal probability and independently:

XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGXYD QPAAMKBZAACI-BZLHJQD

The first order approximation is obtained by choosing the letters independently but with frequencies as in English:

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

The second order approximation incorporates the transition probabilities between successive letters and hence has the same digram (*i.e.*, letter-pair) frequencies as in English:

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

The third order approximation ensures correct trigram structure:

IN NO IST LAT WHEY CRATCIT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOA CTIONA OF CRE

Higher order approximations with letters cannot be constructed due to lack of statistics regarding tetragrams, etc., but Miller, following Shannon, used words instead of letters and constructed approximations up to the seventh order.

Zero Order: BETWIXT TRUMPETER PEBBLY COMPLICATION VIGOROUS TIPPLE CAREEN OBSCURE ATTRACTIVE CONSEQUENCE EXPEDITION PANE UNPUNISHED

First Order: IS TO WENT BIPED THE OF BEFORE LOVE TURTLEDOVES THE SPINS AND I OF YARD THAN ASK WENT GREEK YESTERDAY

Second Order: SUN WAS NICE DORMITORY IS I LIKE CHOCOLATE CAKE BUT I THINK THAT BOOK IS HE WANTS TO SCHOOL THERE

Third Order: FAMILY WAS LARGE DARK ANIMAL CAME ROARING DOWN THE MIDDLE OF MY FRIENDS LOVE BOOKS PASSIONATELY EVERY KISS IS FINE

Fourth Order: WENT TO THE MOVIES WITH A MAN I USED TO GO TOWARDS THE HARVARD SQUARE IN CAMBRIDGE IS MAD FUN FOR

Seventh Order: SAID THAT HE WAS AFRAID OF DOGS MARKED WITH WHITE SPOTS AND WITH BLACK SPOTS COVERING IT THE LEOPARD DID

These were obtained as follows: To obtain the third order approximation, for instance, Miller chose the sequence of the first two words from a text and asked a person to supply the next word to form a sentence so that the transition from the first pair of words to the third takes place as in common English. This word is noted along, the first word is concealed and the last two now given to another person and the next word obtained. The process is repeated using different persons every time a word is obtained. The other approximations are constructed in a similar manner. The resemblance to natural English increases at each stage although the messages are not purposive and motivated and thus strengthens the conviction that natural languages can be represented by sufficiently complex stochastic processes.

The examples above show that the first order approximation which takes only the letter frequencies into account is a rather poor one for real languages. The expression for the entropy can, however, be generalised readily to include constraints between symbols. If the choice of each letter depends upon the preceding one and only on that, we can define the entropy per symbol from digrams of the type ij as

$$H = \tfrac{1}{2} H(xy) = -\tfrac{1}{2}\sum_{i,j} p(i,j)\log_2 p(i,j) \quad (2)$$

where $p(i, j)$ is the probability of the digram ij or as

$$H = H_x(y) = -\sum p(i) \sum p_i(j)\log_2 p_i(j) \quad (3)$$

where $p_i(j)$ is the transition probability from the ith symbol to the jth and $p(i)$ is the probability of the occurrence of the ith symbol. The extension to the case where the choice of a symbol depends upon the preceding $n-1$ symbols is now obvious. We obtain the average entropy per symbol by considering the probability of blocks of n symbols or the transition probabilities from blocks of $n - 1$ symbols length to the next one. We may also interpret the relation (3) as the average entropy of the second symbol weighted in accordance with the probability of occurrence of the first. As the information obtained is the same as the uncertainty removed, it also measures how uncertain we are, on the average, of the next symbol knowing the previous one. More generally it is called the conditional entropy of the second event y relative to the first event x and is denoted by $H_x(y)$. The expression

$$H(x, y) = -\sum_{i,j} p(i,j)\log_2 p(i,j)$$

may likewise be interpreted as the entropy or uncertainty of the joint event xy.

Some Properties of the Information Measure

We shall now exhibit some properties of the measure of information which support the claim that the entropy of a set of probabilities (as defined above) measures the choice or the uncertainty associated with them in accordance with our intuitive requirements:

(*i*) If there are n possible ways an event can happen with probabilities $p_1, p_2, \ldots p_n$, then the entropy H is a maximum when all the p_is are equal as may be seen by maximizing H. This is obviously the most uncertain situation.

(*ii*) H vanishes when all the p_is are zero except one which is unity. There is no choice, no uncertainty and no information either.

(*iii*) It may be shown from our definitions of $H(x)$, $H(x, y)$, $H_x(y)$, etc., that the relations

$$H(x,y) \leq H(x) + H(y) \quad (4)$$

$$H(x,y) = H(x) + H_x(y) = H(y) + H_y(x) \quad (5)$$

and hence

$$H_x(y) \leq H(y), \quad H_y(x) \leq H(x) \quad (6)$$

hold good.

The first of these states that the amount of information (or the uncertainty) of the joint event xy is equal to or less than the sum of the informations (or the uncertainties) associated with the individual events x and y. The second relation means that the uncertainty of the joint event xy is equal to or less than the uncertainty of the event x *plus* the uncertainty of the event y knowing x and *vice versa*. The last statement asserts that the uncertainty $H(y)$ of the event y knowing the event x is equal to or less than the uncertainty of y

without a knowledge of x and *vice versa*. The equality obtains when x and y are independent events.

Redundancy

These considerations regarding the entropy of a message lead us to the important concept of redundancy, a knowledge of which enables us to design suitable codes for transmission of messages. We have seen that the entropy is a maximum when all the symbols are independent and equiprobable. In English language the maximum entropy per smybol would be $\log_2 27 = 4 \cdot 76$ bits (corresponding to the zero-order approximation), but the entropies calculated on the basis of letter, digram and trigram frequencies turn out to be 4·03, 3·31 and 3·10 bits per symbol. There is evidence that if we consider constraints extending over longer sequences, the entropy reduces down to about 1·5 bits per symbol. The ratio of the actual entropy obtained in a given message to the maximum possible entropy is called the relative entropy and one *minus* the relative entropy the redundancy in the message. The farther the possibilities p_is are removed from the equiprobable case, the greater is the redundancy, the extreme case being one in which all the p_is except one are zero. The letter u after q is an example of this extreme case of redundancy. In English one finds a redundancy of 15% on the basis of letter frequencies alone and about 30% on the basis of digram frequencies. Consideration of longer sequences leads to redundancies as much as 70% or more. Just as the entropy measures how uncertain we are on the average about the outcome of an event, the redundancy is an average measure of our confidence in the outcome. A highly redundant source produces less information per symbol than a less redundant source and conversely to convey the same information we need the least number of symbols when the alphabet is used without any redundancy. Redundancy, however, insures the message against misrepresentation. Any letter after q can always be corrected as u. The price we pay for securing the correct transmission is an increase in the length of the message and the resulting slower rate of transmission.

The immediate significance of redundancy to the telecommunication problem lies in the fact that some of the redundancy may be removed in the process of encoding the message into the signal for transmission. A code is a unique correspondence between the message symbols and the signal symbols or between groups of them in a one-to-one fashion. There are, however, severe restrictions set by noise and the inevitable delays in encoding on the extent to which the redundancy in a message can be removed.

Rate of Generation of Information and Channel Capacity

When we realise that the generation of each of the message symbols requires a finite time, we can also define the time rate at which information is produced by the stochastic process. If the durations of the different symbols are t_i then the average rate of generation of information is

$$H' = H/\Sigma\, p_i t_i \quad \text{bits/sec.} \tag{7}$$

i.e., the average entropy per symbol divided by the average duration of the symbols. The rate of transmission of information over the channel is likewise determined by the duration t'_i of the channel symbols constituting the signal. (We are assuming here that the physical properties of the channel permit the transmission of the signal symbols at least as fast as they are produced.) This rate is not necessarily the maximum possible rate at which information can be transmitted over the channel with the given set of channel symbols as these may not be occurring with the optimal frequencies. The channel capacity is defined as the maximum rate at which information can be transmitted over the channel, given the durations of the channel symbols and the constraints that must be obeyed. This capacity rate of transmission is to be achieved by suitably assigning the probabilities of the different symbols and their transitions with due regard to their duration. One of the basic theorems of communication theory tells that there exists a code by which the output of a source producing entropy at the rate of H bits per symbol can be transmitted over a channel of capacity C bits per second at the maximum possible rate of $[(C/H) - \epsilon]$ symbols per second and ϵ can be made arbitrarily small. The emphasis here is on the possibility of a code by which the information produced by the source can be transmitted at the full capacity of the channel. To approach this limiting value, in general, increasingly long delays are needed in coding and decoding as longer and longer sequences have to be examined. One simple example of a perfectly matching code may be given: Let a source produce the four symbols A, B, C, D with probabilities 1/2, 1/4, 1/8 and 1/8 respectively so that the entropy works out at 7/4 bits per symbol according to (1). A binary (*i.e.*, a two-symbol) code can be construed as follows:

Write the original symbols in order of decreasing probability and divide them first into two groups (here A and B, C, D) of equal (or as nearly equal as possible) probability and assign 0 to the first and 1 to the second group. Proceed to subdivide each group and assign additional binary digits 0 and 1 in the same way to the subdivisions until each symbol is given a unique representation as illustrated below.

Message symbols		Signal Symbols
A		0
	1st div.	
B		1 0
	2nd div.	
C		1 1 0
	3rd div.	
D		1 1 1

It will be seen that a typical long message in A, B, C, D will result in producing the channel symbols 0 and 1 with equal probabilities so that they carry maximum information. We must remember that with such ideal coding which removes all redundancy any error in transmission cannot be corrected.

Noise in Communication Systems

To make our discussion realistic we must include the effect of noise in transmission, which perturbs the transmitted signal in an unpredictable way. Thus if the transmitter produces the symbol i at the input to the channel, there is no certainty that it will be received as i at the receiver; all that we have are conditional probabilities $p_i(j)$ that if the symbol i is transmitted, it will be perturbed into the symbol j at the receiving end. If the noise is not bad, the $p_i(i)$s will be nearly unity and all other $p_i(j)$s will be small, the noiseless case being the special one for which $p_i(j) = \delta_{ij}$. Therefore, when a message is received over a noisy channel there will be some uncertainty of what the transmitted message was.

How much information is conveyed by each symbol under these conditions? Assuming for simplicity that the input symbols are independent and further that noise affects each symbol independently, we can compute the entropy input per symbol H(x) on the basis of the probabilities of the input symbols and likewise, the entropy H(y) of the output symbols. Knowing the $p_i(j)$s, which characterise the noise source, from previous statistics, we can also compute the conditional entropy which measures

$$\mathrm{H}_x(y) = -\sum_i p(i) \sum_j p_i(j) \log p_i(j)$$

the average uncertainty of transmission. It is therefore proper to define the actual information transmitted over the channel as the received information H(y) less the uncertainty $\mathrm{H}_x(y)$ in its transmission. Thus the average information transmitted per symbol is H(y) − $\mathrm{H}_x(y)$ or in view of the relation (5), H(x) − $\mathrm{H}_y(x)$. The latter expression may be interpreted as the amount of information sent less the uncertainty that remains of the transmitted message after the message is in hand. To use such a noisy channel to its capacity we must maximize H(x) − $\mathrm{H}_y(x)$ by assigning the probabilities $p(i)$ of the input symbols in the optimal way. Roughly speaking, if all the symbols are of the same duration, those symbols which are least disturbed by noise are to be used most frequently.

Continuous Systems

The messages and signals discussed so far use a discrete alphabet. There are, however, messages like those due to speech-waves, etc., which are conventionally regarded as being continuous and also continuously variable. The problem of developing a measure of the information produced by such continuous messages may be approached in two different ways: The messages constitute an ensemble with a probability density measure $p(x)$, where x is some statistic such as the pressure amplitude of the different possible speech-wave forms at some fixed instant. For such continuous distributions Shannon has formulated the entropy as

$$\mathrm{H}(x) = -\int_{-\infty}^{\infty} p(x) \log p(x)\, dx. \qquad (8)$$

In the continuous case also the entropy possesses the properties analogous to (*i*) to (*iii*) of the discrete distributions. A particularly interesting property of the continuous distribution is that if the standard deviation is fixed at some value σ, the entropy in (8) has the maximum value $\log \sqrt{2\pi e}\,\sigma$, when $p(x)$ is gaussian, *i.e.*,

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-x^2/2\sigma^2}. \qquad (9)$$

A more practical approach to the entropy (and the channel capacity) in the continuous case is based on the fact that in practice there is a least upper bound W to the frequencies produced by any source of information. There is an important theorem, called the sampling theorem, which states that if a function of time $f(t)$ contains only frequencies between 0 and W cycles per second, then it is completely determined by specifying its values (*i.e.*, the ordinates) at discrete points 1/2 W seconds apart. Thus we need only state a finite number 2 W of ordinates to specify a function over one second. Moreover, the accuracy of specification need not be greater than the resolving power of the ultimate destination or the level of the ambient noise fluctuation to which the channel may be subject. Thus both the abscissæ and the ordinates of the function can only assume discrete values. This double quantization reduces the continuous case in practice to the discrete case.

Capacity of a Noisy Channel

One of the most important contributions of modern information theory is the determination of the maximum possible rate at which information can be transmitted over a channel perturbed by white thermal noise of power N. Thermal noise of power N is characterized by the fact that its amplitudes are independent and have the gaussian probability distribution (9) with standard deviation $\sigma = \sqrt{N}$. Therefore, the entropy produced by the noise source in one second is $2\,W \log \sqrt{2\,\pi e\,N}$. Given a signal of power P, to carry maximum information it must also assume a gaussian distribution with the standard deviation $\sigma = \sqrt{P}$. As the transmitted signal power and the noise power add directly during the course of transmission, the received signal will have the power $P + N$, and will also have a gaussian distribution with $\sigma = \sqrt{P + N}$. The entropy of the received signal [corresponding to $H(y)$ in the discrete case] per second will be $2\,W \log \sqrt{2\,\pi e (P + N)}$. The channel capacity C is obtained by subtracting from this the uncertainty due to noise. Thus

$$C = W \log \frac{P + N}{N} \text{ bits/sec.} \qquad (10)$$

This is known as the Shannon-Hartley Law. Regarded as an exchange relation between the channel capacity, bandwidth and signal-to-noise ratio, it shows what is possible under ideal conditions.

Information Theory Outside Telecommunication Engineering

The science of telecommunication abuts on many scientific disciplines, in particular, linguistics, psychology, neurophysiology and others not to speak of its obvious connections with certain branches of physics like acoustics and statistical mechanics. It stems out of the fact that in the telecommunication chain, the ultimate source of information and the ultimate destination happen to be nearly always a human being. We are thus confronted with the problem of matching the telecommunication channel to the human channel. The concepts of telecommunication theory provide valuable analogies in the field of human communication and we can ask questions like "How is information transmitted in the human being and what are his capacities as a channel ?" We have no satisfactory way of answering these questions yet but these problems are receiving attention in several fields. Models of communication in the human being, beginning with the physical stimulus and ending with the behavioural response, have been put forward following the general framework of the telecommunication systems. Some recent experimental work by Miller, Licklider, Pierce and others indicates that human beings cannot probably take in information at a rate much greater than about 50 bits per second through the sense organs like the ear and the eye. The importance of the problem of matching the telecommunication and the human channels becomes obvious when we observe that our present telecommunication systems utilize channel capacities several orders of magnitudes greater than that of the human channel. In the outline that follows (which necessarily belongs to the controversial) we can only hope to direct attention to some of the problems in this sphere. We must caution, however, that speculative hypotheses and theorizations which appear frequently in literature should not be mistaken to have any accepted standing.

Language and Human Communication

Communication by speech and writing have been going through a process of evolution long before we could extend their scope by telecommunication and form, even today, the bulk of our communicative activity. Speech or (written) language usually appears as the input and output of most telecommunication systems which is but one reason for our interest in these. Even before the development of information theory, a very considerable body of statistical data regarding the characteristics of speech and language has been gathered over a number of years. The growth of telephone systems following the development of the vacuum tube initiated systematic investigations into the physical characteristics of speech. The needs of the military for secret communication by codes and for deciphering enemy cryptograms provided an early incentive for the study of the relative frequencies of different letter symbols, digrams, trigrams, etc. With information theory came the realization that these statistical data define a measurable property of speech and language. The point of view has also emerged that speech and language may themselves be regarded as codes for certain conceptual entities. The question of what are the principles which underlie this coding process has received some attention recently.

The first concern of information theorists with speech and language has almost always been to determine the extent of redundancy in a given message. The proper exploitation of redundancy offers one of the most hopeful means of more effective utilization of the channel capacities. Hence the search for the information bearing elements (or shortly *ibes*) of speech. The fact that some saving in channel capacity required to transmit intelligible speech may be possible was appreciated for some time. One of the most successful practical attempts in this direction was the Bell Telephone Laboratories' *Vocoder* which indicates that a bandwidth of some 300 cps. may be adequate for intelligible communication as compared with the nearly 3,000 cps. used in telephony. That speech is highly redundant has been amply established by a number of experiments with speech-waves during the last decade. The intelligibility of speech was tested on speech-sounds, syllables, and words using transmission systems which distort it in a variety of ways, *e.g.*, by clipping off the positive and negative peaks till the wave becomes nearly rectangular, by interrupting the wave at a rapid rate, etc. Even when the wave form is severely distorted, speech retains its intelligibility to a remarkable degree, indicating the presence of a great deal of redundancy. This is not surprising as we already saw that the output of a continuous source has maximum entropy when it corresponds to white noise. It is correct, though uncomplimentary, to say that human speech fails to be informative to the extent that it falls short of noise. How much of this redundancy can be removed and how best the remaining can be utilized in a given situation are questions which will continue to engage our attention.

Semantic and Pragmatic Questions

Following Shannon's example, information theory has been developed (as we have done above) without attention to the meaning of the messages. This point of view was adequate as long as one's problems were strictly confined to telecommunication systems, but when once the human terminal is considered, we can no longer ignore the semantic and pragmatic problems involved. Obviously, the use of the concepts of information theory in the semantic and pragmatic fields needs both clarification and caution. As one illustration, consider a message like 'The sun will rise in the east tomorrow'. One feels intuitively that the message is highly redundant because one has no *a priori* doubt about the truth of the statement and yet the statement is not redundant in the same sense that the letter *u* is redundant after *q*. The redundancy here is at the semantic level and not at the syntactic level as in the case of *u* after *q*. There is evidently a need for a more general theory of information which includes the meaning of the messages. Bar Hillel and Carnap have advanced, however, a theory of semantic information in the limited sense that their theory takes into account the concepts or the entities to which the symbols refer. It does not, however, take into account the meaning of the messages.

The problem of semantic information is not entirely separable from the information of the telecommunication problem, sometimes called the selective information to distinguish it from the semantic information. For, if we have the choice between two languages in which to transmit the same semantic information, we would naturally prefer to transmit it in the language which requires the smaller number of bits of selective information. Following this line of reasoning, in the course of their studies in Indian languages from the information theory point of view, the author and his co-workers have recently advanced the view that it is possible to compare the relative efficiencies of different languages for communication of semantic content without reference to its absolute value. To use a metaphor, translation from one language to another is a transformation of the code which leaves the semantic content, but not the selective entropy of the code symbols, invariant. The total number of bits of selective information contained in semantically equivalent materials in different languages are thus, in a sense, the appropriate measures of the efficiencies with which different languages encode semantic content into linguistic symbols. A preliminary comparative study of English and German languages, considered as alternate codes for communication of semantic content, revealed some interesting aspects of the process of translation. On counting the number of bits of information in samples of texts in English and their translations into German, it was observed that a bit in German is semantically equivalent to about 0·82 bits in English. On the other hand, when translations from German to English were examined, one bit in German was found to be equivalent to 0·94 bits in English. A unique ratio may, however, be obtained on the assumption that translation from one language into another involves a certain amount of ambiguity which is in the nature of noise or loss of information and therefore one uses additional bits of information in the language of translation to overcome this noise. On the basis of this assumption one finds that a bit in English has about the same semantic value as 1·15 bits in German (which makes English slightly more efficient) and that each bit requires an additional 0·065 bits to overcome the noise inherent in the process of translation. Comparative studies in statistical aspects of Indian languages from the information theory point of view also showed the possibility of a common telegraph code for the Indian languages. One may go further and regard the different scripts as different codes for the same phonetic pattern and thus compare their efficiencies for transcribing from the verbal to the orthographic form.

Informational and Thermodynamic Entropies

A large part of our scientific activity may be described, in a sense, as seeking, processing and using information. A theory of information, therefore, cannot but be of some significance to science at least on the philosophical plane. Brillouin, who has some provoking statements to make, has explored the relationship between the thermodynamical entropy of a physical system and the amount of information obtained when the state of the system changes. When an isolated system is left to itself, according to the second law of thermodynamics, the system can undergo only those changes which lead to an increase in its physical entropy or a degradation of its energy. At best the entropy remains constant if the change is a reversible one. Brillouin argues that information can only be obtained by letting the entropy of the system increase and the increase in entropy is always greater than the amount of information obtained.

Consider a system whose initial state could result, for instance, from any of the W_0 equally probable complexions and hence *a priori* probability $p_0 = 1/W_0$ and initial entropy $S_0 = k \log W_0$, k being Boltzmann's constant. As the system degrades in the course of a natural change, its final state can result from any of the $W_1 (> W_0)$, say, equally probable complexions with probability $p_1 = 1/W_1$, and its entropy becomes $S_1 = k \log W_1$. The change in entropy is

$$S_0 - S_1 = k \log_e (W_0/W_1) = k \log_e (p_1/p_0).$$

Information theory tells that an observation which involves a change in the probability from p_0 to p_1 brings $\log_2 (p_1/p_0)$ bits of information. We can, however, identify $k \log_e (p_1/p_0)$ itself with the amount of information I (by choosing a new unit of information) that could possibly be obtained from the system. Thus

$$S_1 = S_0 - I.$$

Brillouin now restates the second law of thermodynamics in a somewhat more generalized form by saying that "in any transformation of a closed system, the quantity entropy *minus* information must always increase or may at best remain constant" and hence argues that information can only be obtained from a system by letting its entropy increase. Notice that every observation is an irreversible one and hence involves an increase in the entropy in accordance with the above reasoning. Under the best of circumstances, the amount of information obtained is equal to the increase in the entropy so that in general

$$\Delta I \leq \Delta S.$$

Because of the opposite sign of information to entropy we might make use of information to decrease the entropy, but the overall balance still remains in favour of an increase in the entropy. Brillouin has seized upon this relationship to discuss the efficiency of an experimental observation which he defines as the ratio $\Delta I/\Delta S$ (< 1).

The concept of entropy and the second law of thermodynamics raised in the past many issues with philosophical implications which are still being debated. Information theory with its wide connections gives us · a new opportunity to re-examine these questions afresh. Information theory is thus more than a theoretical tool for the evaluation of communication systems; it has already come to stay as a way of thinking about a very wide class of problems.

Literature on information theory is very extensive and widely scattered. We cite some references which will enable the reader to pursue his way through this subject.

1. Shannon, C. E., "A Mathematical Theory of Communication, I & II," *Bell Syst. Tech. Jour.*, 1948, **27**, 379–423, 623–56.
2. Wiener, N., *Extrapolation, Interpolation and Smoothing of Stationary Time-Series*, Wiley & Sons Inc., New York, 1949.
3. *Symposium on Information Theory*, Ministry of Supply, London, 1950.

4. Jackson, W. (edited by), *Symposium on Communication Theory*, Butterworth Scientific Publications, London, 1953.

5. Cherry, C. (edited by), *Information Theory*, Butterworth Scientific Publications, London, 1955.

6. "Proceedings of Speech Communication Conference at M.I.T.," *Jour. Acous. Soc. Amer.*, 1950, 689–806.

7. Cherry, C., *On Human Communication*, Wiley & Sons Inc., New York, 1957.

8. Miller, G. A., *Language and Communication*, McGraw Hill, New York, 1951.

9. Goldman, S., *Information Theory*, Constable, London, 1953.

10. Brillouin, L., *Science and Information Theory*, Academic Press, New York, 1956.

11. C. C. I. R., *Bibliography on Communication Theory*, Union Internationale des Telecommunications, Geneve.

^{12}C AS REFERENCE NUCLIDE

THERE exist at present three scales of atomic masses of weights: (*i*) the absolute scale based on the gram, (*ii*) that defined by taking the mass of one atom of the nuclide ^{16}O equal to 16 units (the "Physical scale" of "atomic masses" or "nuclidic masses"), and (*iii*) that taking the average atomic masses of the isotope mixture of "natural" oxygen as 16 units (the "Chemical scale" of "atomic weights"). Of these, only the last two are in common and extensive use. The chemical scale is indefinite to the extent of the variation in the average atomic mass of oxygen from various natural sources (some 15 parts per million) resulting from variations in the relative proportions of ^{16}O, ^{17}O and ^{18}O.

Recently, proposals for improving this situation have been made and discussed. The necessity of matching the proper value of the Avogadro number with the mass values employed arises especially often in the domain of nuclear chemistry.

Proposals to unite the scales by adopting the physical scale for chemical atomic weights have been regarded with disfavour by many chemists because of the relatively large change, about 275 parts per million, which would have to be made in all of the quantities whose values depend on the size of the mole. There are many physicochemical data whose precision is greater than that and whose value would therefore have to be changed. On the other hand, the serious consideration which has been given by chemists to the proposal of a new unified scale based on $^{19}F = 19$, which would result in a change of 41 parts per million, indicates that many chemists would be willing to accept a unified scale if the atomic weights would not be changed by more than about this amount. There are relatively few chemical data bearing such high precision.

Fortunately, there is a possible scale definition which, as the basis of a unified scale, would suit chemists and by which, moreover, physicists would benefit greatly.

Evidently, that definition is to be preferred which allows most nuclidic masses to be expressed with the smallest errors, not only now but also in the foreseeable future. As is shown below, this purpose is fulfilled by taking ^{12}C as the reference nuclide. The best definition of the atomic mass unit is, accordingly,

Mass of ^{12}C equals exactly 12 atomic mass units. The unit defined in this way is 318 parts per million larger than the present physical mass unit and 43 parts per million larger than the present chemical one.

In the mass-spectroscopic determination of nuclidic masses, the most important substandard is ^{12}C. Not only do the doubly, triply, and quadruply charged atomic ions of ^{12}C occur at integral mass numbers so that they can be paired in doublets with nuclides having mass numbers 6, 4 and 3 respectively, but—much more important—no other element besides carbon can be found which forms molecular ions containing as many atoms of but one kind (up to 10 and more). Therefore, the scale $^{12}C = 12$ would allow many more direct doublet comparisons of masses, especially of heavy nuclides, with the reference nuclide than any other scale. 12-Carbon has the additional advantage that carbon forms many more hydrides than any other element, so that an easy reference line for doublets can be produced at almost every mass number up to $A \sim 120$. Many stable nuclides in the mass region $120 < A < 240$ can also be measured in reference to ^{12}C by pairing in doublets their doubly charged ions with singly charged ions of $^{12}C_n$ or of $^{12}C_nH_m$ fragments. Use can then be made of nuclear disintegration data to obtain accurate masses of many other, especially unstable nuclides.—*Science*, 20th June 1958, 127 (3312), 1431.