Information Theory after 18 Years

After years of theoretical development, information theory may fulfill its engineering promise.

E. N. Gilbert

The identifying feature of information theory, the feature which distinguishes it from cybernetics or from other branches of statistical communication theory, is its use of a particular way of "measuring" information. Although earlier authors (*1*) had suggested various measures of information, C. E. Shannon is the real father of information theory. In a paper which appeared in 1948 (*2*), he not only defined a reasonable measure for information but also applied it to prove remarkable theorems which provide criteria for evaluating and comparing different communication systems. These theorems came at an advanced stage of the communications art; a radio engineer in 1948 could use amplitude modulation, frequency modulation, phase modulation, single side-band modulation, or pulse code modulation, or he could invent something new without much trouble. The need for good ways of evaluating communication systems was evident, and Shannon's paper received immediate attention.

Not all the attention came from communication engineers, however. Mathematicians found his paper a gold mine of statistical and combinatorial problems, partly because Shannon, writing for engineering readers, had not proved his theorems in the greatest possible generality or with the meticulous rigor that some mathematicians require (*3*), but mainly because he raised difficult mathematical questions which even today are unanswered. Physicists (*4*) were interested in a new interpretation of entropy as information. Psychologists found that the new information measure gave a convenient quantitative estimate of the difficulty of certain experimental tasks. Other applications have been to linguistics (*5*), music (*6*), cryptography (*7*), and gambling (*8*). The response to Shannon's paper was so great that by 1953 the Institute of Radio Engineers formed a Professional Group on Information Theory with a journal of its own. *Information and Control*, another journal devoted to information theory, began publication in 1958.

Information theory was a glamor science for many years. It was popularly supposed that information theory held the key to progress in remote fields to which in fact it did not apply. Interesting evidence of this remains in some of the editorials (*9*) which appeared in the *Transactions of the Professional Group on Information Theory* between 1955 and 1959: the editors seem appalled by a flood of worthless interdisciplinary papers submitted in the name of information theory.

As late as 1959 one respected mathematician (*10*) publicly questioned whether information theory was mathematically sound. On the other hand, many practical engineers had the private opinion that it was too much a mathematical idealization to be of help to them in real-life problems. It has been hard to answer the latter objection; few real communication systems designed since 1948 show decisive influence by information theory. Moreover, earlier inventions, such as the Morse code, the vocoder (*11*), the compandor (*12*), and pulse code modulation (*13*), demonstrate that engineers can achieve by "common sense" results close in spirit to information theory. One answer (*14*) to the objection is that information theory provides insights which guide the engineer without solving his problem in detail. It is also argued (*15*) that information theory will soon become an essential tool in the design of increasingly complex digital communication systems.

There are fashions in science. Who in 1900 foresaw the neglect which modern mathematicians accord quaternions and elliptic functions? By the year 2000 information theory may exist only in a few unread definitive treatises, preserved by college librarians as forlorn monuments to misspent lives. To avert this end, information theory must supply something more tangible than vague insights. I hope to show that the gap between theory and practice is closing.

Codes

As a concrete setting for the definition of information, let us consider the following basic communication problem. Messages must be sent over a wire as sequences of electrical pulses, as in telegraphy. Suppose that each pulse must be one of two allowed pulses—say, a positive pulse or a negative pulse—and that (unlike the dot and dash pulses of telegraphy) the two basic pulses require equal times for transmission. It is customary to use binary digits 0, 1 to represent the two pulses. The messages come from a source which may not produce binary digits directly. Instead, the messages are sequences composed of certain letters $L_1, L_2, \ldots, L_K$. For illustration, I often take the L_i to be the 27 letters of the English alphabet, counting the word space (the hyphen of Table 1, column 1) as a letter. However, for other sources the L_i might be decimal digits, speech phonemes, or other symbols. In order to transmit these messages one must invent a code which represents messages as sequences of binary digits.

Only "uniquely decipherable" codes will be allowed; that is, two different messages must not encode into the same binary sequence. The three codes of Table 1 encode English messages, letter by letter, into binary digits. For example, code 2 transmits the message BE- as 011110010000. Each code in Table 1 has the following property, which guarantees unique decipherability: the digits of one letter never appear as the leading digits of a different letter (consider, by contrast, the possibilities for confusion in the three-letter code A = 01, B = 010, C = 1). Not all uniquely decipherable codes have this property (consider the code A = 1, B = 10, C = 00), and in

Dr. Gilbert is a member of the Mathematics and Statistics Center of Bell Telephone Laboratories, Murray Hill, New Jersey.

general it takes some calculating to decide whether a given code is uniquely decipherable (*16*).

Morse code uses a letter space—really a third kind of digit—as a synchronizing sign to prepare the receiver for the start of the next letter. A possible danger, in a strictly binary system, is that the receiver may accidentally begin decoding in the middle of a letter and thereafter continue out of step with the transmitter. Information theory usually ignores such timing problems, and they need not be serious. Code 2 of Table 1 is actually Morse code with 1, 01, and 00 substituted for dot, dash, and letter space; then 100 identifies the end of a letter. Code 1 has some synchronizing ability because it is not "exhaustive"—that is, there exist sequences, such as 11011. . ., which are not encoded forms of any real messages. Ultimately the decoder may attempt to decode 11011 and realize that it is out of step. "Comma-free codes" (*17*) have been deliberately designed to be nonexhaustive to such an extent that the decoder cannot decode a single letter when it is out of step. Comma-free codes were invented originally for a theory of genetic structure. Unlike codes 1 and 2, code 3 is exhaustive. However, code 3, like most other codes in which $L_1, L_2, \ldots, L_K$ have unequal numbers of digits, has a self-synchronizing ability (*18*) which returns the decoder to synchronism after a few letters.

Information

Table 1 gives estimates of the probability p_i with which letter L_i occurs in English text (*19*). If a code uses N_i binary digits to encode L_i, then D, the mean number of digits per letter, is $p_1N_1 + p_2N_2 + \ldots + p_KN_K$. The means, D, for codes 1, 2, and 3 of Table 1 are, respectively, 5, 4.868, and 4.120 digits per letter. Given $p_1, \ldots, p_K$ for a source, the uniquely decipherable code which encodes messages letter by letter and has the smallest D can be constructed by a procedure developed by D. Huffman (*20*). For the probabilities in Table 1, code 3 is the minimizing code. In general the minimizing code has $N_i \cong -\log_2 p_i$, as may be verified for Table 1 by comparing columns 5 and 6. Then the minimum value of D is near a number H defined as the following sum:

$$H = -\sum_i p_i \log_2 p_i. \qquad (1)$$

For Table 1, H equals 4.08. Actually the minimum D satisfies $H \leq D \leq H + 1$, in general. Note the formal resemblance between H and the entropy function of statistical mechanics.

Instead of encoding letter by letter one might encode other units—say, blocks B letters long. Huffman's algorithm provides an efficient code if the probabilities of occurrence of the blocks (27^B in number, in the case of English) are known. Equation 1 gives an estimate of the minimum mean number of digits per block if the p_i are reinterpreted as block probabilities. Generally the mean number of digits needed per letter decreases as the encoded unit lengthens. Word-frequency counts show that 2.14 digits per letter would suffice to encode English word by word. An experiment by Shannon (*21*) indicated that about one digit per letter might suffice to encode English in blocks of 100 letters.

In general, even complicated codes cannot represent the messages of a source by arbitrarily small numbers of digits per letter. The mean numbers of digits per letter used by all possible codes for a given source have some greatest lower bound. The information rate of the source is defined to be this greatest lower bound. The information rate is given in units of "bits" (a contraction of "binary digits") per letter. Shannon's experiment (*21*) suggests that the information rate of an English-text source is near one bit per letter. When a source produces letters at a given rate—say, n letters per second—its information rate, in bits per letter, may be multiplied by n to express the information rate in bits per second.

The information rate of certain sources can be given simply. If a source produces letter L_i, with probabilities of occurrence p_i, independently, then H, given by Eq. 1, is its ininformation rate. Such a source, using English letters with the probabilities of occurrence given in Table 1, produces messages like the following typical sample:

LT-EEED-TK-O-IT-EHILIGD-
SI-RESCO-ENT

A source which produces successive letters independently and with equal probabilities ($p_i = 1/K$) has the highest information rate of all sources using the K letters $L_1, \ldots, L_K$. For this "maximizing source," Eq. 1 simplifies to $H = \log_2 K$ bits per letter. Messages from the maximizing source for the English alphabet have incorrect letter probabilities and so look even less like English than the sample given above does. The high rate $\log_2 27 = 4.755$ bits per letter reflects the absence of all the usual constraints (U follows Q, E is more common than J, and so on) which make the future of real English text somewhat predictable from its past. Since the messages of a real English source contain only 20 percent as much information per letter as this maximizing source produces, English is some-

Table 1. Binary codes for English text.

L_i	p_i	Code 1	Code 2	Code 3	$-\log_2 p_i$
-	.1859	00000	00	000	2.4
A	.0642	00001	10100	0100	4.0
B	.0127	00010	0111100	011111	6.3
C	.0218	00011	01101100	11111	5.5
D	.0317	00100	011100	01011	5.0
E	.1031	00101	100	101	2.9
F	.0208	00110	1101100	001100	5.6
G	.0152	00111	0101100	011101	6.0
H	.0467	01000	111100	1110	4.4
I	.0575	01001	1100	1000	4.1
J	.0008	01010	101010100	0111001110	10.4
K	.0049	01011	0110100	01110010	7.7
L	.0321	01100	1011100	01010	5.0
M	.0198	01101	010100	001101	5.7
N	.0574	01110	01100	1001	4.1
O	.0632	01111	01010100	0110	4.0
P	.0152	10000	10101100	011110	6.0
Q	.0008	10001	010110100	0111001101	10.4
R	.0484	10010	101100	1101	4.4
S	.0514	10011	11100	1100	4.3
T	.0796	10100	0100	0010	3.7
U	.0228	10101	110100	11110	5.5
V	.0083	10110	1110100	0111000	6.9
W	.0175	10111	1010100	001110	5.8
X	.0013	11000	01110100	0111001100	9.6
Y	.0164	11001	011010100	001111	5.9
Z	.0005	11010	01011100	0111001111	11.0

times said to be 80 percent redundant.

Like the English source, most other real sources are very redundant. A television source is a good example. A rectangular array of 200,000 dots, each dot having one of 64 brightnesses, is a good approximation to a television picture. To encode television dot by dot (essentially what is done in practice) takes six binary digits per dot or 1,200,000 binary digits per picture. Actually there is a strong tendency for dots near each other to have the same brightness. Successive pictures also tend to be almost identical. These constraints make the information rate of television much lower than 1,200,000 bits per picture (*22*).

Really efficient codes for redundant sources like English text or television could be very valuable. Although information theory shows, in principle, how to find the codes, there are enormous practical difficulties, first in measuring the needed probabilities and afterward in using the more complicated codes. Coding for facsimile has been the most successful effort. A reasonably simple code was designed in 1957 to transmit drawings and diagrams by facsimile; it uses about 12 percent as many digits as a conventional system (*23*). This facsimile system has as yet found no commercial acceptance.

Channels

Although the information rate of a source was defined originally in terms of numbers of binary digits per letter, this rate also will determine how fast messages can be sent when symbols other than binary digits are transmitted. A channel might, for example, transmit messages as decimal digits 0, 1, . . . , 9. It is now assumed that each kind of transmitted symbol requires a known time for transmission, not necessarily the same for all symbols (dots and dashes, for example, have different transmission times). Again, the rate, in letters per second, at which messages from the source are transmitted depends on the code.

Consider, first, noiseless channels—channels for which the received symbol always agrees with the transmitted symbol. A good code for a noiseless channel should handle messages at a high rate. A noiseless channel has a capacity, C, measured in bits per second, for which Shannon proved the following. No source can be encoded to signal over the channel at an information rate exceeding C bits per second. However, every source with a nonzero information rate H bits per letter can be encoded to signal at a letter rate n letters per second, so that the information rate nH, in bits per second, is as close to C as is desired.

A noiseless channel which transmits K different kinds of symbols, each requiring a time T seconds for transmission, has channel capacity $C = (\log_2 K)/T$.

A channel is called "noisy" if the symbols which emerge at its output cannot be predicted with certainty from the symbols which were transmitted at the input. The binary symmetric channel and the Gaussian channel, which are described in the next two paragraphs, are the noisy channels which have received the most theoretical attention.

The binary symmetric channel accepts binary digits at its input and emits binary digits at its output. Both 0 and 1 require T seconds for transmission. For each digit the channel makes an independent random choice to decide whether to transmit the digit correctly or incorrectly. Both 0 and 1 have the same probability p of being in error. This simple model approximates some digital channels very well. In practice, other kinds of binary noise are often encountered. In particular, errors may come in "bursts"—isolated clusters of several erroneous digits close together. The independent errors of the binary symmetric channel do not tend to group into bursts.

The Gaussian channel transmits real numbers. A real number x at the input is received as $x + y$, where y is a second real number representing noise; y is chosen independently of x and from a Gaussian (normal) distribution of mean zero and a given variance N. Not all long sequences $x_1x_2, \ldots$ of real numbers are accepted by this channel, only those for which the average of the x_i^2 is less than a given number S. This channel is a useful model of a radio link in which the signal-to-noise power ratio at the receiver is S/N. To anyone accustomed to picturing a radio signal as a function $x(t)$, representation of signals as discrete sequences of real numbers may seem odd. Nevertheless, if the radio channel occupies a frequency band W cycles per second wide, a "sampling" theorem (*24*) applies, which reconstructs a signal $x(t)$ exactly by interpolating between the discrete sequence of sample values $x(k/2W)$ ($k = \ldots,-1,0,1, \ldots$). Thus, sample values are sent once every $T = 1/2W$ seconds. The Gaussian noise of this channel is a good representation both of the thermal noise within a radio receiver and of some kinds of static. It is sometimes used, for want of a better simple model, to represent interference from other radio stations. However, it cannot account for multipath distortion and fading.

Noise

In designing a code for a noisy channel one must compromise between speed and reliability. For example, suppose one must transmit a long sequence of decimal digits over the Gaussian channel. A message—say, 301992 . . .—could be encoded as a single real number, $x = .301992 \ldots$, which the Gaussian channel transmits in just T seconds. However, the noise would make all but the first few received digits unreliable. One could send the digits in groups of three, as .301,.992, . . .; in twos, as .30,.19, .92, . . .; singly, as .3,.0,.1,.9,.9,.2, . . .; or perhaps two times each for still greater reliability, as .3,.3,.0,.0,.1, .1,.9, ... It seems that one must use slower and slower codes to achieve greater reliability. Suppose one is satisfied if, on the average, a prescribed small nonzero fraction ϵ of the received letters are decoded incorrectly. The fastest allowable signaling rate then decreases as ϵ decreases. As ϵ approaches zero, must not the signaling rate also approach zero? Shannon's 1948 paper gave a surprising "No" in answer to that question, by means of the following coding theorem.

Even a noisy channel has a capacity C which is not, in general, zero. The properties of this capacity are similar to those of the capacity of a noiseless channel. If one must signal over the channel at an information rate R exceeding C bits per second, then the error probability must exceed some positive number $\epsilon(R)$; there is no hope of obtaining $\epsilon < \epsilon(R)$ at that high rate. However, if, for any source, an error probability ϵ and a rate R less than C are given, then there exists a way of coding the source for the channel so that the information rate is more than R bits per second and, moreover, so that the probability of error is less than ϵ.

The capacity of the binary symmetric channel is

$$C = [1 + p \log_2 p + (1 - p) \log_2 (1 - p)]/T \quad (2)$$

bits per second. The Gaussian channel has capacity

$$C = W \log_2 (1 + S/N) \quad (3)$$

bits per second.

Complexity

The channel capacity represents an ideal rate against which to compare rates really achieved. Such comparisons usually reveal that much improvement is possible. To obtain an improvement one expects to use a more complex code. The practical question is, How much improvement can be bought for a moderate increase in complexity?

A "block code" breaks a message from the source into blocks of some fixed length, say B seconds, and then encodes block by block. In the original coding theorem rates near C were obtained as B approached infinity. Large B requires that the encoder and the decoder process message data in large amounts. Thus, B is one index of the complexity of a block code. Recent proofs of the coding theorem show the influence of the parameter B more clearly. For a given source rate R and block length B, these proofs provide bounds on the probability, P, that a block will be decoded incorrectly.

Figure 1 illustrates the kind of result now obtainable (*25*). The figure relates to Gaussian channels. For a given signal-to-noise ratio S/N, it shows bounds on the largest information rate R attainable with error probability $P = 10^{-4}$. The dashed curves are lower bounds on R, the solid curves are upper bounds. A pair of curves appears for each of the three values $2WB = 5$, 25, and 101. Even when $2WB = 101$ (that is, when messages are encoded into blocks of 101 real numbers), about 1.6 times as much power (2 decibels) is required as the capacity formula (Eq. 3) indicates.

Given a channel with capacity C and a source of rate $R < C$, the best achievable error probability P approaches zero as the block length B approaches infinity. The manner in which P approaches zero is now known fairly precisely, especially when R is near C. For a certain range of rates,

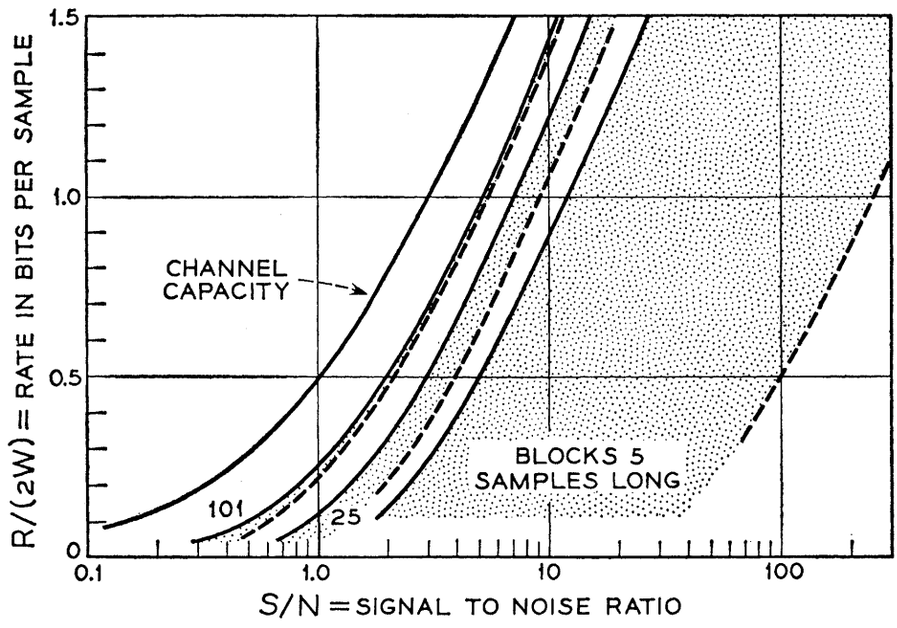

Fig. 1. Bounds on the best rate achievable on the Gaussian channel when, at most, one block in 10^4 may be in error.

$R_{crit} \leq R < C$, the error probability satisfies the inequalities

$$\exp\{-B[E(R) + \delta]\} \leq P \leq \exp[-BE(R)] \quad (4)$$

where $E(R)$ is a known function of R and δ is a term which approaches zero as B approaches infinity. At rates R below R_{crit}, P also has exponential upper and lower bounds but the difference between the best-known exponents does not approach zero. These bounds on probabilities of error are the result of a series of difficult studies (*26*). Recently R. G. Gallager (*27*) has given a simple and elegant treatment.

When $R > C$, frequent errors are unavoidable. The situation is curious because P actually approaches 1 as B increases (*28*).

Random and Systematic Codes

In order to achieve the high signaling rates promised by the coding theorem and its variants one must use a well-designed code. Unfortunately the coding theorem does not really tell how to construct codes. It says, in effect, "The codes exist; now you go find them." Turning to the proofs of the coding theorem for a clue, one finds that the codes used in these proofs are based on random processes.

Figure 2 shows a random construction simpler than most random codes. In Fig. 2 the problem was to pack circles, without overlap, into a square. The packing procedure was a sequence of random trials. In each trial a point within the square was randomly chosen as a possible center of a circle. If the circle with that center overlapped none of the circles already packed, then the new circle was packed. Packing of the 16 circles shown required a few hundred trials.

Figure 2 can be a code for a Gaussian channel. The coordinates (x_1,x_2) of each center serve as a pair of real numbers to be transmitted. The list of all 16 pairs is then a block code of block length $B = 2/(2W)$ seconds. To use the code, messages from the source may first be encoded into binary digits —say by Huffman's procedure. Afterward the 16 possible blocks of four binary digits are associated with the 16 pairs of real numbers. A received pair (z_1,z_2) is interpreted as the nearest center. By packing spheres one keeps the transmitted points far enough apart so that the most likely noises cause no trouble.

The encoder for Fig. 2 is basically a code book, such as is used by cryptographers, with 16 entries. Better random codes require much larger code books. A random code corresponding to the curves $2WB = 101$ of Fig. 1 would be represented by a set of random points in 101-dimensional space. If the code signaled at a rate of one bit per sample, the code book would contain $2^{101} = 2.5 \times 10^{30}$ entries.

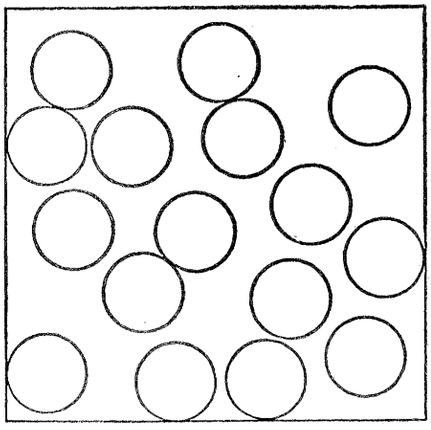

Fig. 2. Random packing.

What is really needed is a code with a systematic mathematical structure, so that the code book can be replaced by relatively concise rules. In the packing problem a concise rule might be "all sphere centers have integer coordinates." This rule adequately describes cubic packing without individual listing of the centers. However, in high-dimensional space the cubic point lattice provides neither dense sphere packings (*29*) nor good codes.

Parity Checks

Systematic codes have been designed most successfully for digital channels. In the usual technique for designing codes for binary channels, parity checks are used. A parity check is a constraint requiring that the sum of the digits in certain positions be an even number. A simple example is a code based on the use of blocks $(x_1,x_2, \ldots ,x_7)$ of seven binary digits for which the three check sums $S_1 = x_1 + x_3 + x_5 + x_7$, $S_2 = x_2 + x_3 + x_6 + x_7$, and $S_4 = x_4 + x_5 + x_6 + x_7$ must all be even numbers. When this code is used, x_3, x_5, x_6, and x_7 can be four digits of some binary message; x_1, x_2, and x_4 are then determined by the parity conditions. Thus this code can signal at a rate of 4/7 bit per digit. This is the earliest example of an "error-correcting code" (*30*). The receiver can reconstruct the block even when one of the seven received digits is incorrect. To do so the receiver forms the check sums S_1, S_2, and S_4 from the received digits. The position of the incorrect digit will be the sum of the subscripts of the sums S_i which are not even numbers (for example, if S_2 and S_4 are odd numbers, then x_6 is incorrect).

A similar code having m checks corrects any single error in a block 2^m-1 digits long. It is more difficult to design codes to correct larger numbers of errors within a block, but now several families of these codes are available (*31*). One of the most versatile codes, the Bose-Chaudhuri-Hocquenghem code (*32*), corrects any pattern of t or fewer errors in a block of 2^m-1 digits and requires at most mt parity-check conditions. For any of these codes the encoder is a relatively simple device that uses the check constraints to compute digits. The decoder performs a more complicated logical task; often it requires considerable ingenuity to design a simple machine for decoding an error-correcting code (*33*).

Parity checks are also useful against other kinds of noise. Several kinds of burst-correcting codes are available (*34*; see also *35*). Other codes correct erasures (*36*) (an "erasure" replaces a binary digit by a blank). For a very simple example of a code which corrects bursts of erasures, let blocks be 12 digits long and let there be four parity check sums, $x_1 + x_5 + x_9$, $x_2 + x_6 + x_{10}$, $x_3 + x_7 + x_{11}$, and $x_4 + x_8 + x_{12}$. If a single burst of erasures 4 digits or less in length occurs, each check sum contains at most one erased digit; the receiver can recompute the missing digits from the check-sum relations.

Parity checks have also been used in "recurrent" codes (*35*), which do not have a block structure. A typical example of a recurrent code, by D. W. Hagelbarger, has a check sum $S_k = x_{2k-12} + x_{2k-6} + x_{2k+1}$ for each integer k. The digits x_2, x_4, x_6, . . . , with even subscripts, can be the digits of some binary message; the remaining digits, x_1, x_3, x_5, . . . , are then computed from the check-sum relations. This code corrects bursts which are separated by at least 19 correct digits; a burst may be any pattern of errors contained in a cluster of up to six consecutive digits. The decoding is extremely simple. To decode message digit x_{2k} the decoder first computes the check sums S_{k+3} and S_{k+6} from the received digits. If both check sums are odd numbers, the decoder concludes that x_{2k} is in error; if at least one check sum is an even number, the decoder concludes that x_{2k} is correct.

Onward to Capacity

It is still not possible to signal with a low probability of error at rates arbitrarily close to channel capacity with the known systematic codes. For example, with a code of block-length b digits on the binary symmetric channel there would be about pb errors in each block. One might try to use an error-correcting code which corrects slightly more than pb errors. However, it is now known (*37*) that if $\frac{1}{4} \leq p \leq \frac{3}{4}$, the rate of even the fastest error-correcting block code must approach zero as b increases. Thus, to approach channel capacity one must deliberately allow occasional patterns of pb errors to go uncorrected. With one exception (*38*), the known families of systematic codes for the binary symmetric channel either have zero rate or high probability of error in the limit of large block length.

The approach to channel capacity presents the clear challenge of a Mount Everest or a 4-minute mile. A solution would excite great theoretical interest. Its practical importance might be less that one at first imagines, both because it undoubtedly would require complicated codes and also because real channels are seldom as simple as theoretical models. To use a real channel at rates near its capacity, one must have an accurate statistical description of the channel noise. Such statistics are hard to obtain. When telephone circuits were measured for transmission of binary digits (*39*), the errors tended to occur in bursts but were otherwise so scarce that a large number of data had to be collected. Simple models (*40*) succeeded in simulating some of the gross features of the data, but the true noise is much more complicated. When the channel noise is only partially understood one may prefer to correct some of the most common errors and to forget about achieving channel capacity.

Feedback

Coding is not always the only approach to a problem of communication-system design. The cost of coding and decoding equipment must be balanced against the cost of improving the channel. For example, instead of buying sophisticated coding equipment for a Gaussian channel, one might bet-

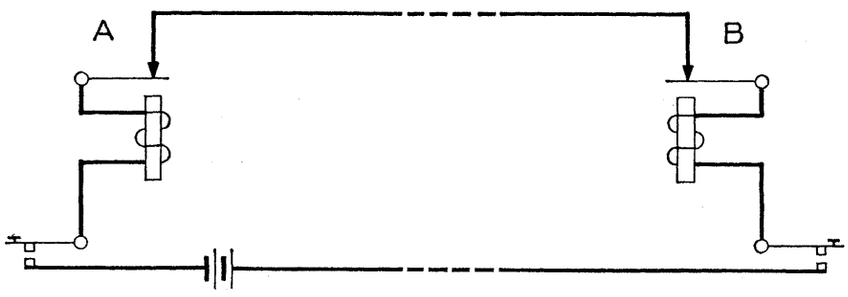

Fig. 3. A two-way channel.

ter spend the money for increased transmitter power or antenna gain.

Another possibility is to provide a "feedback channel" from the receiver to the transmitter. In its simplest usage a feedback channel permits the receiver to ask for retransmission of questionable parts of the message. For example, when a parity check block code is used on a binary channel, blocks containing parity check failures could be retransmitted. In this way an error-correcting code can overcome many additional error patterns. This system is especially advantageous when the noise occurs in bursts of perhaps a dozen errors separated by many thousands of correct digits [as observed in sending digital data over telephone circuits (*41*)]. Then a feedback channel of very low capacity suffices.

An extended coding theorem covers feedback channels and, even more generally, two-way channels, in which messages travel in both directions, possibly interfering with one another (*42*). An example illustrates some of the complications possible in such a theory. Suppose the two-way binary channel is just a noiseless telegraph wire and battery connected in series with a key and buzzer at each of two stations, A and B, as shown in Fig. 3. The two buzzers sound (signifying binary digit 1) only when both keys are closed. Operator B may close his key permanently to receive messages from A, but then the signaling rate from B to A is zero. If operators B and A both have messages to send they might transmit in turns, alternately sending one digit and closing the key to receive one. However, A and B can increase their average signaling rates, and still not make errors, if they adopt a suitable system in which both signal at the same time (*43*).

When a forward channel of capacity C and a feedback channel are both available one might hope to send information in the forward direction faster than C bits per second without a high probability of error. This is shown to be impossible if the forward channel is "memoryless" (both the binary symmetric channel and the Gaussian channel are "memoryless") (*44*). Instead, the feedback channel simplifies the coding and replaces the exponent $E(R)$ in Eq. 4 by a larger function (*45*). A simple code is now known which uses feedback to signal over the Gaussian channel with a low probability of error and at a rate as close as desired to the channel capacity (*46*).

Psychology

Information theorists and psychologists have some interests in common just because humans are important sources and recipients of messages. Shannon's estimate of the information rate for English came not from measurements of the impossibly complicated statistical properties of English but, instead, from an experiment in which a human tried to guess letters of a text (*21*). In addition, psychologists have tried to use the information measure in rating experimental tasks.

It was hoped that one could measure a subject's short-term memory in terms of a fixed number of bits, regardless of the kind of item to be remembered. However, it was found that a typical subject, who might remember 9 binary digits after reading them once, remembered 8 decimal digits (25 bits), 7 English letters (33 bits), or 5 monosyllabic words (50 bits) (*47*).

In other experiments, rates at which humans process information were measured. In such studies the subject is a kind of channel; his instructions are the messages to be transmitted, and his performance is what is received. Various tasks were studied—sight reading at the piano, typing, tapping targets, marking squares, reading word lists—in an attempt to achieve as high a signaling rate as possible (hence, to realize the full "capacity of the human channel") (*48*). The best rates obtained were about 40 to 50 bits per second. These results are hard to interpret, but they do make one wonder what possible use a television viewer makes of the millions of bits of information he receives each second.

Psychologists' interest in information theory reached a peak about 10 years ago (*49*). Nowadays its limitations for psychology are better understood, and interest has waned somewhat. The information measure is still used as a convenient statistic, like a mean or a variance, but it is interpreted with more caution (*50*).

Conclusion

In this article I portrayed that aspect of modern information theory which relates to explicit coding systems intended to signal at high rates. I omit other, more theoretical parts of the subject. Information theory is a very active area of investigation in the U.S.S.R., but there the emphasis is on mathematical results (*51*), and these I had to omit.

As an engineering subject, information theory has flourished for 18 years because of the promise it gave of improved communication systems. The results are still almost exclusively on paper. Nevertheless, the paper work has come closer to practicalities. Experimental systems which use some of the new codes have been tested, and some coders and decoders are now commercially available. They may be in widespread use in a few years. Meanwhile, a page count in the journals devoted to information theory shows that the field is still growing.

References and Notes

1. L. Szilard, *Z. Physik* **53**, 840 (1929); H. Nyquist, *Bell System Tech. J.* **3**, 324 (1924); R. V. L. Hartley, *ibid.* **9**, 535 (1928); R. A. Fisher, *Proc. Cambridge Phil. Soc.* **22**, 700 (1925).
2. C. E. Shannon, *Bell System Tech. J.* **27**, 379, 623 (1948).
3. J. L. Doob, *Math. Rev.* **10**, 133 (1949).
4. L. Brillouin, *J. Appl. Phys.* **3**, 334, 338 (1951).
5. Y. Bar-Hillel and R. Carnap, in *Communication Theory*, W. Jackson, Ed. (Academic Press, New York, 1952); C. Cherry, *On Human Communication* (Wiley, New York, 1957); C. E. Cherry, M. Halle, R. Jakobson, *Language* **29**, 34 (1953); B. Mandelbrot, *Word*

10, 1 (1954); G. Herdan, *Type-token Mathematics* (Mouton, The Hague, Netherlands, 1960).
6. R. C. Pinkerton, *Sci. Amer.* **194**, 77 (1956); W. Fucks, *IRE (Inst. Radio Engrs.) Trans. Inform. Theory* **8**, 225 (1962).
7. C. E. Shannon, *Bell System Tech. J.* **28**, 656 (1949).
8. J. L. Kelly, Jr., *ibid.* **35**, 917 (1956).
9. P. Elias, *IRE (Inst. Radio Engrs.) Trans. Inform. Theory* **4**, 99 (1958); C. E. Shannon, *ibid.* **2**, 2 (1956).
10. J. L. Doob, *ibid.* **5**, 3 (1959).
11. H. Dudley, *J. Acoust. Soc. Amer.* **11**, 169 (1939).
12. R. C. Mathes and S. B. Wright, *Bell System Tech. J.* **13**, 315 (1934).
13. A. H. Reeves, *IEEE (Inst. Elec. Electron. Engrs.) Spectrum* **2**, 58 (1965); W. M. Goodale, *Bell System Tech. J.* **26**, 395 (1947); Reeves patented pulse code modulation in 1937.
14. J. R. Pierce, *IRE (Inst. Radio Engrs.) Nat. Conv. Record* **5**, pt. 2, 51 (1957).
15. R. M. Fano, *IRE (Inst. Radio Engrs.) Trans. Inform. Theory* **4**, 63 (1958).
16. A. A. Sardinas and G. W. Patterson, *IRE (Inst. Radio Engrs.) Nat. Conv. Record* **1**, pt. 8, 104 (1953); G. Bandyopadhyay, *Inform. and Control* **6**, 331 (1963).
17. S. W. Golomb, L. R. Welch, M. Delbrück, *Biol. Med. Danske Videnskabernes Selskab* **23**, 1 (1958); S. W. Golomb, B. Gordon, L. R. Welch, *Can. J. Math.* **10**, 202 (1958); E. N. Gilbert, *IRE (Inst. Radio Engrs.) Trans. Inform. Theory* **6**, 470 (1960); J. J. Stiffler, *IEEE (Inst. Elec. Electron. Engrs.) Trans. Inform. Theory* **11**, 107 (1965).
18. M. P. Schützenberger, *IRE (Inst. Radio Engrs.) Trans. Inform. Theory* **2**, 47 (1956); E. N. Gilbert and E. F. Moore, *Bell System Tech. J.* **38**, 933 (1959).
19. G. Dewey, *Relativ Frequency of English Speech Sounds* (Cambridge Univ. Press, Cambridge, 1923); F. Pratt, *Secret and Urgent* (Blue Ribbon, New York, 1942).
20. D. A. Huffman, *Proc. IRE (Inst. Radio Engrs.)* **40**, 1098 (1952).
21. C. E. Shannon, *Bell System Tech. J.* **30**, 50 (1951).
22. E. R. Kretzmer, *ibid.* **31**, 751 (1952).
23. W. S. Michel, W. O. Fleckenstein, E. R. Kretzmer, *IRE (Inst. Radio Engrs.) WESCON Conv. Record* **1**, pt. 2, 84 (1957).
24. J. M. Whittaker, *Interpolatory Function Theory* (Cambridge Univ. Press, Cambridge, 1935); C. E. Shannon, *Proc. IRE (Inst. Radio Engrs.)* **37**, 10 (1949).
25. This figure was adapted from curves in D. Slepian, *Bell System Tech. J.* **42**, 681 (1963), derived from formulas of C. E. Shannon, *ibid.* **38**, 611 (1959).
26. S. O. Rice, *ibid.* **29**, 60 (1950); C. E. Shannon, *ibid.* **38**, 611 (1959); A. Feinstein, *IRE (Inst. Radio Engrs.) Trans. Inform. Theory* **4**, 2 (1954); R. M. Fano, *Transmission of Information* (M.I.T. Press, Cambridge, 1961).
27. R. G. Gallager, *IEEE (Inst. Elec. Electron. Engrs.) Trans. Inform. Theory* **11**, 3 (1965).
28. J. Wolfowitz, *Inform. and Control* **3**, 89 (1960).
29. C. A. Rogers, *Packing and Covering* (Cambridge Univ. Press, Cambridge, 1964).
30. R. W. Hamming, *Bell System Tech. J.* **29**, 147 (1950).
31. M. J. E. Golay, *Proc. IRE (Inst. Radio Engrs.)* **37**, 657 (1949); D. E. Muller, *IRE (Inst. Radio Engrs.) Trans. Electron. Computers* **3**, 6 (1954); D. Slepian, *Bell System Tech. J.* **35**, 203 (1956); J. E. MacDonald, *IBM J. Res. Develop.* **4**, 43 (1960); I. S. Reed and G. Solomon, *J. Soc. Ind. Appl. Math.* **8**, 300 (1960); C. M. Melas, *IBM J. Res. and Develop.* **4**, 364 (1960); W. W. Peterson's *Error-Correcting Codes* (M.I.T. Press, Cambridge, 1961) is the standard general reference for error-correcting coding theory.
32. R. C. Bose and D. K. Ray-Chaudhuri, *Inform. and Control* **3**, 68 (1960); A. Hocquenghem, *Chiffres* **2**, 147 (1959).
33. I. S. Reed, *IRE (Inst. Radio Engrs.) Trans. Inform. Theory* **4**, 38 (1954); W. W. Peterson, *ibid.* **6**, 459 (1960).
34. E. N. Gilbert, *Proc. Symp. Appl. Math. No. 10* (1960), p. 291; N. Abramson, *IRE (Inst. Radio Engrs.) Trans. Inform. Theory* **5**, 150 (1959); J. J. Stone, *J. Soc. Ind. Appl. Math.* **11**, 74 (1963); A. D. Wyner, *IEEE (Inst. Elec. Electron. Engrs.) Trans. Inform. Theory* **9**, 124 (1963); E. Gorog, *IBM J. Res. Develop.* **7**, 102 (1963).
35. D. W. Hagelbarger, *Bell System Tech. J.* **38**, 1741 (1958); J. M. Wozencraft and B. Reiffen, *Sequential Decoding* (M.I.T. Press, Cambridge, 1961); A. D. Wyner, *IEEE (Inst. Elec. Electron. Engrs.) Intern. Conv. Record* **7**, pt. 4, 139 (1963); E. R. Berlekamp, *Inform. and Control* **6**, 1 (1963).
36. M. A. Epstein, *IRE (Inst. Radio Engrs.) Nat. Conv. Record* **6**, pt. 4, 56 (1958).
37. M. Plotkin, *IRE (Inst. Radio Engrs.) Trans. Inform. Theory* **6**, 445 (1960).
38. P. Elias, *ibid.* **4**, 29 (1954).
39. A. A. Alexander, R. M. Gryb, D. W. Nast, *Bell System Tech. J.* **39**, 431 (1960); R. Morris, *ibid.* **41**, 1399 (1962).
40. E. N. Gilbert, *ibid.* **39**, 1253 (1960); J. M. Berger and B. Mandelbrot, *IBM J. Res. Develop.* **7**, 224 (1963).
41. W. A. Malthaner, *Bell Lab. Record* **35**, 121 (1957); A. B. Brown and S. T. Meyers, *IRE (Inst. Radio Engrs.) Nat. Conv. Record* **6**, pt. 4, 37 (1958); W. R. Cowell and H. O. Burton, *Trans. AIEE* **80**, 577 (1961); A. B. Fontaine and R. G. Gallager, *Proc. IRE (Inst. Radio Engrs.)* **49**, 1059 (1961); R. J. Benice and A. H. Frey, Jr., *IEEE (Inst. Elec. Electron. Engrs.) Trans. Commun. Technol.* **12**, 146 (1964); B. Reiffen, W. H. Schmidt, H. L. Yudkin, *Trans. AIEE* **80**, 224 (1961).
42. C. E. Shannon, *Proc. Berkeley Symp. Math. Statistics Probability, 4th* (1961), vol. 1, p. 611.
43. If A and B use a two-letter code 01, 10, the only possible received pairs are 00, 01, and 10. Knowing what he transmitted and what he received, A or B can infer what the other transmitted. However, whenever the initial 1 of a 10 pair is received, A and B realize that they are both trying to send 10, and they therefore need not waste time sending the second digit 0. Both A and B can thus send at rates faster than ½ bit per digit. D. W. Hagelbarger invented this scheme.
44. C. E. Shannon, *IRE (Inst. Radio Engrs.) Trans. Inform. Theory* **2**, 8 (1956).
45. M. Horstein, *IEEE (Inst. Elec. Electron. Engrs.) Trans. Inform. Theory* **9**, 136 (1963); A. J. Viterbi, *Inform. and Control* **8**, 80 (1965).
46. This code appears in an unpublished manuscript by J. P. M. Schalkwijk based on his doctoral thesis at Stanford University in 1965.
47. G. A. Miller, *IRE (Inst. Radio Engrs.) Trans. Inform. Theory* **2**, 129 (1956); R. N. Shepard (*J. Verbal Learning Verbal Behavior*, in press) reports up to 600 bits of short-term memory for words and pictures.
48. P. M. Fitts, *J. Exp. Psychol.* **47**, 381 (1954); H. Quastler, Ed., *Information Theory in Psychology* (Free Press, Glencoe, Ill., 1955); G. C. Sziklai, *IRE (Inst. Radio Engrs.) Trans. Inform. Theory* **2**, 125 (1956); J. R. Pierce and J. E. Karlin, *Bell System Tech. J.* **36**, 497 (1957); see H. Jacobson [*Science* **112**, 143 (1950); *ibid.* **113**, 292 (1951)] for estimates of the capacities of the ear and eye alone.
49. R. D. Luce [*Development in Mathematical Psychology* (Free Press, Glencoe, Ill., 1960)] gives a review of the literature up to 1957.
50. W. R. Garner, *Uncertainty and Structure as Psychological Concepts* (Wiley, New York, 1962).
51. R. L. Dobrushin, *Proc. Berkeley Symp. Math. Statistics Probability, 4th* (1961), vol. 1, p. 211.

NEWS AND COMMENT

200-Bev: The Academy Committee Knew Where It Was Going

The burghers of the posh Chicago suburb of South Barrington last week forced the removal of their manicured acres from the small list of finalists in the great accelerator competition, charging, among other things, that an influx of scientists would "disturb the moral fiber of the community."

The residents, whose neighborhood was volunteered by the Governor of Illinois, did not specify the present condition of the fiber or the direction in which it might be disturbed. But outside of this admirably perverse refusal to trade rustic charm for a $375-million laboratory, there was little strong reaction to the announcement that, from 85 proposals, covering some 150 tracts in 43 states, an evaluation committee of the National Academy of Sciences had selected, for final consideration, six sites (or, depending on how you count, seven, since South Barrington was listed along with an alternate site in the Chicago area.) These were, Ann Arbor, Michigan; the Brookhaven National Laboratory, Upton, Long Island, N.Y.; Denver, Colorado; Madison, Wisconsin; the Sierra foothills, near Sacramento, California; and South Barrington, or Weston, near Chicago.

The proximity of the finalists to major northern universities or research centers inspired senators Sparkman, of Alabama, and Russell, of Georgia, to scriptural sarcasm, with Sparkman declaring, "For unto every one that hath shall be given, and he shall have abundance . . . ," and Russell concluding, "But from him that hath not shall be taken away even that which he hath." But though they were joined by Symington, of Missouri, and Mansfield, of Montana, in lamenting the geographic distribution of the finalists, the senators