Efficient Coding

By B. M. OLIVER

(Manuscript received May 14, 1952)

This paper reviews briefly a few of the simpler aspects of communication theory, especially those parts which relate to the information rate of and channel capacity required for sampled, quantized messages. Two methods are then discussed, whereby such messages can be converted to a "reduced" form in which the successive samples are more nearly independent and for which the simple amplitude distribution is more peaked than in the original message. This reduced signal can then be encoded into binary digits with good efficiency using a Shannon-Fano code on a symbol-by-symbol (or pair-by-pair) basis. The usual inefficiency which results from ignoring the correlation between message segments is lessened because this correlation is less in the reduced message.

INTRODUCTION

The term coding, as applied to electrical communication, has several meanings. It means the representation of letters as sequences of dots and dashes. It means the representation of signal sample amplitudes as groups of pulses having two or more possible amplitudes as in pulse code modulation. Lately, it has also come to be the generic term for any process by which a message or message wave is converted into a signal suitable for a given channel. In this usage single-sideband modulation, frequency modulation and pulse code modulation are examples of encoding procedures, while microphones, teletypewriters and television cameras are examples of encoding devices.

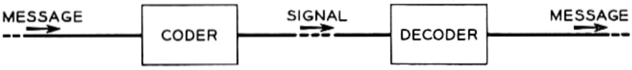
This is a nice concept, but it is useful to distinguish between two classes of encoding processes and devices: those which make no use of the statistical properties of the signal, and those which do. In the first class, the encoding operation consists simply of a one-to-one conversion of the message into a new physical variable, as a microphone converts sound pressure into a proportional voltage or current, or of the one-to-one remapping of the message into a new representation without regard to probabilities, as by ordinary amplitude, frequency or pulse code modulation. In ordinary PCM for example, the message samples are converted into groups of on-or-off pulses. The particular combination of pulses in any group depends only upon the amplitude of the particular sample, not upon any other property of the message, and the same time is allotted to each group, regardless of the probability of that group or of the amplitude it represents. Almost all the processes and devices used in present day communication belong to this first class. In the second class, the probabilities of the message are taken into account so that short representations are used for likely messages or likely subsequences, longer representations for less likely ones. Morse code, for example, uses short code groups for the common letters, longer code groups for the rare ones.

Processes of the first class we may call non-statistical coding processes, or simply modulation or remapping processes. The time of transmission is the same for all messages of the same length, and all messages are handled by the system with equal facility (or difficulty). These processes require no memory and have a small and constant delay. They are inefficient in their use of channel capacity.

Processes of the second class we may call statistical encoding processes. These processes in general require memory. The time of transmission of messages of the same length may be different so that if messages are to be accepted and delivered by the system at constant rates, variable delays may be necessary at the sending and receiving ends. They are more efficient in their use of channel capacity. It is with this second type of process that this paper is concerned, although processes of the first type may be used as component steps. Thus we consider systems of the type shown in Fig. 1, with the accent on the word "efficient".

TRANSMISSION CIRCUITS AND THEIR VOCABULARIES

Communication circuits or channels can, of course, differ in many respects. Either the peak signal power or the average signal power may be limited. The transmission may be uniform over the band or vary with frequency; it may be constant or subject to selective fading. The noise may be gaussian thermal or shot noise uniform across the band, or peaked at some frequency; or it may be largely impulse noise or erratic



SIGNAL = EFFICIENT DESCRIPTION OF MESSAGE

Fig. 1—Reversible statistical encoding.

static discharges. The best type of signal for one channel may be very poor for another.

In the following sections it is assumed that the channel transmission characteristic is flat in amplitude and delay over a definite band and zero outside. It is also assumed that the channel has a definite peak signal power limitation, and that the noise is white gaussian noise. Such a channel is no mere academic ideal. It is in fact quite closely approached in practice by many circuits. Moreover, the conclusions based on these assumptions can usually be modified or extended to other actual cases, such as that of noise with non-uniform spectral distribution (as for example the coaxial cable).

If the bandwidth of the channel is W, we can (using single sideband modulation, if necessary) transmit over it without distortion from frequency limitation signals containing frequencies from 0 to W (or $-W$ to W in the Fourier sense). Such a wave can assume no more than $2W$ *independent* amplitudes per second. Any set of samples of the wave taken at regular intervals $\frac{1}{2W}$ serves to specify the wave completely. The wave may be thought of as a series of $(\sin x)/x$ pulses centered on the samples and of proportional height, and indeed the wave may be reconstructed from the samples in this fashion. This is the well-known sampling theorem[1]. Thus a message source of bandwidth W can supply at most $2W$ independent symbols (samples) per second, and this same number can be transmitted as overlapping, but independently distinguishable pulses by a circuit of bandwidth W.

Since, as will appear later, channels which are to transmit signals resulting from efficient statistical encoding must be relatively invulnerable to noise, we shall assume that the pulses on the channel are quantized. This allows regenerative repeatering to be used to eliminate the accumulation of noise[1]. If there are b quantizing levels, and if the levels are sufficiently separated so that the probability of noise causing incorrect readings is negligibly small, then the capacity of the channel in bits/sec is[2]

$$C = 2W \log_2 b. \tag{1}$$

Such a circuit talks in an alphabet of b "letters" and uses a language in which all combinations of these letters are allowed. There are no forbidden or impossible "words". The circuit has a vocabulary of b one-letter words, b^2 two-letter words, b^n n-letter words. The basic inefficiency in present day electrical communication is that we build circuits with unrestricted vocabularies and then send signals over them which use only a tiny fraction of this vocabulary. If all the letters of the written alphabet were used with equal probability and if all combinations of letters were allowed, then many words which are now long could be made shorter, and written text would be less than one third as long as English. Similarly, if we could arrange to let our circuits use their entire vocabulary with equal probability, they could describe our messages with much less time (or bandwidth) on the average.

EXCHANGE OF BANDWIDTH AND SIGNAL TO NOISE RATIO

It was the advent of wide band FM, and other modulation methods which exchange bandwidth for signal-to-noise ratio, which revealed the inadequacy of earlier concepts of information transmission and ultimately led to the development of modern communication theory, or information theory[2].

One of the more familiar results of this theory is the expression for the maximum capacity of a channel disturbed by white noise:

$$C = W \log_2 \left(1 + \frac{P}{N}\right) \tag{2}$$

in which C is the capacity in bits/sec, W is the bandwidth and P/N the ratio of average signal power to average noise power. This capacity can only be approached, never exceeded, and is only reached when the signal itself has the statistics of a white noise. The expression sets a limit for practical endeavor, and also gives the theoretical rate of exchange between W and P/N.

A practical quantized channel, operated so that the loss of information due to incorrectly received levels is negligible requires about 20 db more peak signal power than the average signal power of the ideal channel to attain the same capacity[1]. However, bandwidth and signal-to-noise ratio are still exchanged on the same basis. For example, a satisfactory television picture could be sent over a channel with, say, 100 levels. This would require a (peak) signal to rms noise ratio of some 40 + 20 = 60 db. The bandwidth could be halved by a sort of reverse PCM: by using one pulse to represent two picture elements. But there are 10,000 combinations of two samples each of which can have any of 100 values. Hence the new combination pulse would need 10,000 distinguishable levels and this would require a signal to noise ratio of 80 + 20 = (2 × 40) + 20 = 100 db.

It is evident that while bandwidth compression by non-statistical or straight signal remapping means is not an impossibility, it is nevertheless impractical when the signal to noise ratios are already high. What we should really try to do is make our descriptions of our messages more efficient so that less channel capacity is required in the first place. The saving can then be taken either in bandwidth or in signal-to-noise ratio, whichever fits the requirements of our channels best.

MESSAGES

Messages can either be continuous waves like speech, music, or television; or they can consist of a succession of discrete characters each with a finite set of possible values, such as English text. Because a finite bandwidth and a small added noise are both permissible, continuous signals can be converted to discrete signals by the processes of sampling and quantizing[1]. This permits us to talk about them as equivalent from the communication engineering viewpoint. Since many of the principles which follow are easier to think of with discrete messages and since quantization of the channel is assumed for reasons already stated, we shall think of our messages as always being available in discrete form.

Let S = the symbol (or sample) rate of the message

$$W_0 = \frac{S}{2} = \text{the original bandwith of the continuous message}$$

ℓ = number of quantizing levels.

Then if all the message samples were independent and if all quantizing levels were equally likely, the information per sample would be

$$H_0 = \log_2 \ell \text{ bits} \tag{3}$$

the information rate would be

$$H_0' = S \log_2 \ell \text{ bits/sec} \tag{4}$$

and the message would use the full capacity of a channel with ℓ quantizing levels, and bandwidth $S/2$. Or by remapping k message samples (with the ℓ possible levels) into $\left(\frac{\log \ell}{\log b}\right) k$ samples, a channel with b levels and bandwidth $W = S/2 \left(\frac{\log \ell}{\log b}\right)$ could be loaded to full capacity.

However, it is not true that the successive samples of typical messages are independent, nor is it true that the various sample amplitudes are in general equiprobable. If these things *were* true, speech and music would sound like white noise, pictures would look like the snowstorm a TV set produces on an idle channel. Written text would look like WPEIPTNKUH WFIOZ—: a random sequence of letters. The statistics of the message, in particular the correlations between the various samples, greatly reduce the number of sequences of given length which are at all likely. As a result the information rate is less, and fewer bits per second are required to describe the average message.

A sequence of M binary digits can describe any of 2^M possible messages. Conversely any of N messages can be described by $\log_2 N$ binary digits. The information rate, H, of a message source is therefore given by

$$H = \lim_{n \to \infty} \frac{\log N}{n} \text{ bits/symbol}$$

where N = number of message sequences of length n. If the successive symbols of the message are *independent* but *not equiprobable*, then a long sequence will contain x_1 symbols of type 1, x_2 of type 2, etc. The number of possible combinations of these symbols will be

$$N = \frac{n!}{\prod_j x_j!},$$

$$\text{so that } \log N = \log n! - \sum_j \log x_j!$$

For large enough n, all the x_j will be large also and we may write, by Stirling's approximation

$$\log N \to \log \sqrt{2\pi n} + n \log n - n - \sum_j [\log \sqrt{2\pi x_j} + x_j \log x_j - x_j]$$

But since $\sum x_j = n$, and since for large n, $x_j \to p(j)n$ where $p(j)$ is the probability of the j^{th} symbol, we have

$$\log N \to \log \sqrt{2\pi n} + n \log n - n \sum_j \log \sqrt{2\pi x_j} - n \sum_j p(j) \log p(j) - n \log n + n$$

$$H_1 = \lim_{n \to \infty} \frac{\log N}{n} = -\sum_j p(j) \log p(j) \tag{5}$$

which is the expression Shannon derives more rigorously[2]. H_1 is a maximum when all the $p(j)$ are equal to $1/\ell$. Then $H_1 = \log_2 \ell = H_0$. The more unequal the $p(j)$, i.e., the more peaked the probability distribution, the smaller H_1 becomes.

If the successive samples are not independent, the message source will pass through a sequence of states which are determined by the past of the message*. In each state there will be a set of *conditional* probabilities describing the choice of the next symbol. If the state is i and the conditional probability (in this state) of the next symbol being the j^{th} is $p_i(j)$, then the information produced by this selection is

$$H_i = -\sum_j p_i(j) \log p_i(j). \tag{6}$$

The average rate of the source is then found by averaging (6) over all states with the proper weighting; thus

$$H = \sum_i p(i)H_i = -\sum_i p(i) \sum_j p_i(j) \log p_i(j). \tag{7}$$

The greater the correlation between successive symbols or samples of a message, the more peaked the distributions $p_i(j)$ become on the average, and this results in a lower value for H. As Shannon points out, the information rate of a source, as given by (7), is simply the average uncertainty as to the next symbol when all the past is known. But in a properly operating communication channel the past of the message is available at both ends, so that it should be possible to signal over the channel at the rate H bits/message symbol, rather than H_0 as we now do. In present day communication systems we ignore the past and pretend each sample is a complete surprise.

By completely efficient statistical coding it should be possible to reduce the required channel capacity by the factor H/H_0. Whether or not this improvement can be actually reached in practice depends upon the amount of past required to uniquely specify the state of the message source. If long range statistical influences exist, then long segments of the past must be remembered. If there are m symbols in the past which determine the present state and each symbol has ℓ possible values, there will be ℓ^m states *possible* (although only 2^{mH} of these are at all probable for large m). If m is large the number of possible states becomes fan-

* In a philosophical sense the state of a message source may be dependent on many other factors besides the past of the message. If the source is a human being, for example, the state will depend on a large number of intangibles. If these could really be taken into account the resulting H for the message might be quite low. If the universe is strictly deterministic one might say that H is "really" always zero. When we describe the drawing of balls from the urn in terms of probabilities, we admit our ignorance as to the exact detail of the mixing operation which has occurred in the urn. Likewise the information rate of a source is a measure of our ignorance of the exact state of the source. From a communication engineering standpoint, the knowledge of the state of the source is confined to that given by the past of the message.

tastically large and complete statistical encoding becomes an economic impossibility if not a technical one.

Let B_i^k be a particular combination (the i^{th}) of k symbols in the past of the message. Each of these combinations at least partially determines the state of the system. Hence we can write an approximation to (7):

$$F_k = -\sum_i p(B_i^k) \sum_j p_{B_i^k}(j) \log p_{B_i^k}(j) \tag{8}$$

$F_k \to H$, as $k \to \infty$. If only m symbols in the past influence the present state, then k need only be as great as m, in order that $F_m = H$. In any case the sequence F_1, F_2, $\cdots$ F_k is monotone decreasing. Naturally one should always pick the k symbols in the past which exert the greatest effect upon the present state, i.e. which cause $p_{B_i^k}(j)$ to be as highly peaked as possible, on the average. In English these would be the immediately previous letters; in television, the picture elements in the immediate space-time vicinity of the present element.

Suppose we break the message up into blocks of length k. Each of these blocks may be considered to be a character in a new (and huge) alphabet. If we ignore any influences from previous blocks, i.e. if we consider the blocks to be independent, then the information per block will be simply

$$-\sum_i p(B_i^k) \log p(B_i^k). \tag{9}$$

Since there are k symbols per block, the information per symbol, G_k is

$$G_k = -\frac{1}{k} \sum_i p(B_i^k) \log p(B_i^k). \tag{10}$$

As $k \to \infty$, $\mathrm{G}_k \to H$, since the amount of statistical influence ignored (between blocks) becomes negligible compared with that taken into account.

If d is the number of binary digits required to specify a message n symbols long, then as $n \to \infty$, $d/nH \to 1$. For large n there are thus 2^{nH} messages which are at all likely out of $2^{nH_0} = \ell^n$ possible sequences (in an ℓ letter alphabet). The probability that a purely random source will produce a message (i.e., a sequence with all the proper statistics) is therefore

$$p \cong 2^{-n(H_0-H)} \tag{11}$$

for large n. Even if $H_0 - H$ is small, $p \to 0$ rapidly for large n. This is why white noise never produces anything resembling a picture on a television screen, for instance. For in television signals, $H_0 - H > 1$

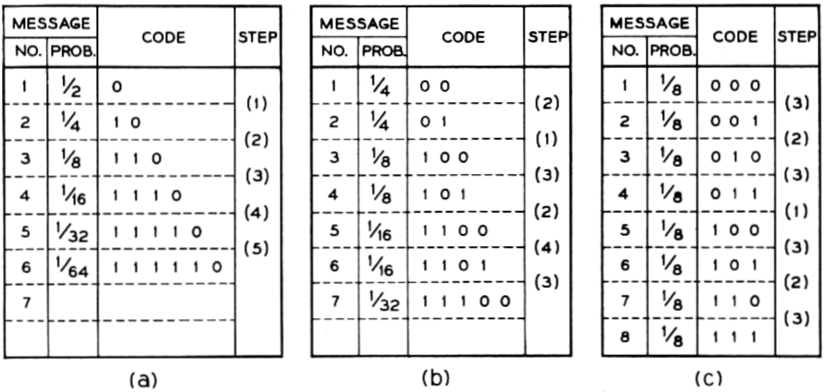
even for very complicated picture material, and $n \cong 250{,}000$ for a single frame.

As given by (11), p, also represents the fraction of the possible signals on a channel of ℓ levels which are likely ever to be used by messages of length n without statistical encoding.

STATISTICALLY MATCHED CODES

Since a sequence of binary digits can be remapped by a non-statistical process into a channel with b quantizing levels, or indeed into a wide variety of other signalling alphabets, it suffices to consider statistical coding processes and codes which reduce the message to a sequence of binary digits. An efficient code is then one for which the average number of binary digits, H_c, per message symbol lies between H_0 and H. As the efficiency increases $H/H_c \rightarrow 1$, so this ratio may be taken as an efficiency index. With highly efficient processes, the sequences of binary digits produced will have little residual correlation, i.e., they will be nearly random sequences. Since the encoding process must be reversible the receiver must be able to recognize the beginnings and ends of code groups. Since we have at our disposal only zeros and ones, the divisions between code groups must either be marked by a special code group reserved for this purpose, or else the code must have the property that no short code group is duplicated as the beginning of a longer group.

A code which satisfies this latter requirement and which is capable of unity efficiency is the so-called Shannon-Fano code, developed independently by C. E. Shannon of Bell Telephone Laboratories and R. M. Fano of the Massachusetts Institute of Technology. This code is constructed as follows: One writes down all the possible message sequences of length k in order of decreasing probability. This list is then divided into two groups of as nearly equal probability as possible. One then writes *zero* as the first digit of the code for all messages in the top half, *one* as the first digit for all messages in the bottom half. Each of these groups is again divided into two subsets of nearly equal probability and a zero is written as the second digit if the message is in the top subsets, a one if it is in the bottom. The process is continued until there is only one message in each subset. Fig. 2a shows the code which results when this process is applied to a particularly simple probability distribution $p(B_i^k) = (1/2)^i$. Here each code group is a series of ones followed by a zero. The receiver knows a code group is finished as soon as a zero appears. Although the longer groups contain mostly ones, their probability is less and on the average as many zeros are sent as ones.



(a)

MESSAGE NO.	MESSAGE PROB.	CODE	STEP
1	½	0	(1)
2	¼	1 0	(2)
3	⅛	1 1 0	(3)
4	1/16	1 1 1 0	(4)
5	1/32	1 1 1 1 0	(5)
6	1/64	1 1 1 1 1 0	
7			

(b)

MESSAGE NO.	MESSAGE PROB.	CODE	STEP
1	¼	0 0	(2)
2	¼	0 1	(1)
3	⅛	1 0 0	(3)
4	⅛	1 0 1	(2)
5	1/16	1 1 0 0	(4)
6	1/16	1 1 0 1	(3)
7	1/32	1 1 1 0 0	

(c)

MESSAGE NO.	MESSAGE PROB.	CODE	STEP
1	⅛	0 0 0	(3)
2	⅛	0 0 1	(2)
3	⅛	0 1 0	(3)
4	⅛	0 1 1	(1)
5	⅛	1 0 0	(3)
6	⅛	1 0 1	(2)
7	⅛	1 1 0	(3)
8	⅛	1 1 1	

Fig. 2—Shannon-Fano codes for three different distributions. The successive bisections are indicated by the dashed lines and the number gives the step at which that bisection took place.

If the successive message segments are independent, the code will generate a random sequence of zeros and ones. Fig. 2b shows the code which results with another distribution. Here the termination of each code group is more complicated but the non-duplicative property exists so the receiver can still identify the groups. Fig. 2c shows the code which results when all the $p(B_i^k)$ are equal. It is the ordinary binary code.

The length of each code group is equal to log $1/p(B_i^k)$, for the cases shown in the figures. This is true in general so long as it is possible to divide the list into subgroups which are of exactly equal probability.

When this is not possible, some code groups may be one digit longer as Shannon shows. The average number of digits per message symbol using this code is therefore given by

$$-1/k \sum_i p(B_i^k) \log p(B_i^k) \leq H_c \leq -1/k \sum_i p(B_i^k) \, [-1 + \log p(B_i^k)]$$

$$G_k \leq H_c \leq G_k + 1/k.$$

For large k, $H_c \rightarrow G_k \rightarrow H$ and the efficiency approaches unity. With small k, H_c increases both because the smaller list of messages cannot be so accurately divided repeatedly into equal probability subsets (so-called "granularity" trouble), and also because more statistics are ignored between the shorter blocks.

The ordinary binary code provides a statistical match between message source and channel only if the various message blocks B_i^k have equal probability $p(B_i^k) = 1/2^n$, and are mutually independent. With $k = 1$, $p(B_i^k) = p(j)$ and the "blocks" are merely the successive symbols.

Ordinary PCM is statistically matched only to a random message source with flat distribution.

If the messages from a source are characterized by frequent long runs of symbols of the same type (e.g., long runs of zeros) an obvious saving is possible by sending the value of the symbol only once, together with a code group which gives the length of the run. This is commonly known as run length coding. The remaining sections of the message (between runs) may then either be sent directly (i.e., merely remapped by a non-statistical process) or they may be encoded by some other statistical process, if this seems warranted. In the latter case we have a mixed coding procedure. The codes representing run lengths must either be set apart from the remainder of the signal by "punctuating" codes, or identifiable by some distinguishing characteristic.

Run length coding may be generalized to take care of other common sequences besides runs of a single symbol. Any commonly occurring sequence of symbols may be considered a "run" and treated in the same fashion. More complicated code groups will be required to specify the type of run, if a large variety is accommodated this way. Ultimately, the distinction between this type of coding and Shannon-Fano coding becomes rather nebulous, especially if a fixed maximum length of run is permitted, for then all possible messages of this length may be considered "runs" and simply encoded by the Shannon-Fano code.

No optimal general solution of the coding problem is known. That is, one cannot say in all cases exactly what coding procedure one should use with a given message source to produce the most efficient encoding for a given complexity of apparatus. Several procedures have been devised which seem suitable for certain types of messages and these are discussed in the following sections.

n-GRAMMING

The application of the Shannon-Fano code to a block of k symbols of a message in an ℓ letter alphabet requires that ℓ^k different codes be used. The receiver must be able to recognize each of these and to regenerate the proper message block when a particular code is received. If ℓ is on the order of 10 to 100 as is typically the case, we very quickly run out of room to house the receiver and money to build it with. On the other hand, if k is small, say on the order of 1 to 3, considerable statistical information between blocks is ignored. These considerations led to the development of a class of encoders known as n-grammers. The name stems from the fact that they operate on the n-gram statistics of the message, to produce a reduced signal having more nearly independent symbols, but (in return) a highly peaked simple probability distribution which allows savings with Shannon-Fano coding on a symbol-by-symbol ($k = 1$) basis.

The simplest member of this class is the monogrammer. It is basically merely a re-ordering device. The operation may be best understood by the following example. Suppose someone supplied us with English text encoded into a quantized pulse signal as follows:

Symbol	Pulse height
Space	0
A	1
B	2
C	3
D	4
etc.	etc.

Now the letter frequencies in English are shown in Fig. 3. Merely to save average power in our channel we might wish to convert this signal into one in which the pulse height is not alphabetical, but in which the most common symbol is sent as a pulse of zero height, the next most common as a pulse of unit height, etc. In other words, we would like the following representation:

Symbol	Pulse height
Space	0
E	1
T	2
A	3
etc.	etc.

The device shown in Fig. 4 will accomplish this translation. The original signal is applied to the vertical deflecting plates of a cathode ray tube. The rest position of the spot corresponds to "space", i.e. no pulse. A pulse one unit high deflects the spot to *A*, a pulse two units high deflects the spot to *B*, etc.

Now in front of these spot positions we place a number of light attenuating filters. In front of the "space" position we place an opaque mask. Hence when the spot is deflected to "space" the photocell receives no light and no pulse is sent. In front of the "*E*" position we place a mask having one unit of transmission. So although *E* is received as a pulse 5 units high, it is sent as a pulse of unit height. In front of the

"*T*" position we place a mask with two units transmission, and so on. The signal amplitudes as received are thus re-ordered in the desired fashion.

The resulting signal has lower average power and this can sometimes be an advantage, particularly if several such signals are to be sent over a common channel by frequency division. In this case the extreme rarity of occurrence of high peak powers on all channels simultaneously means

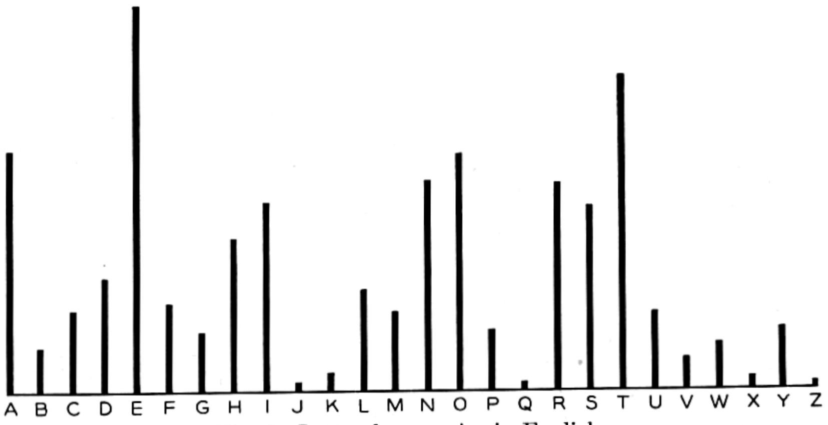


Fig. 3—Letter frequencies in English.

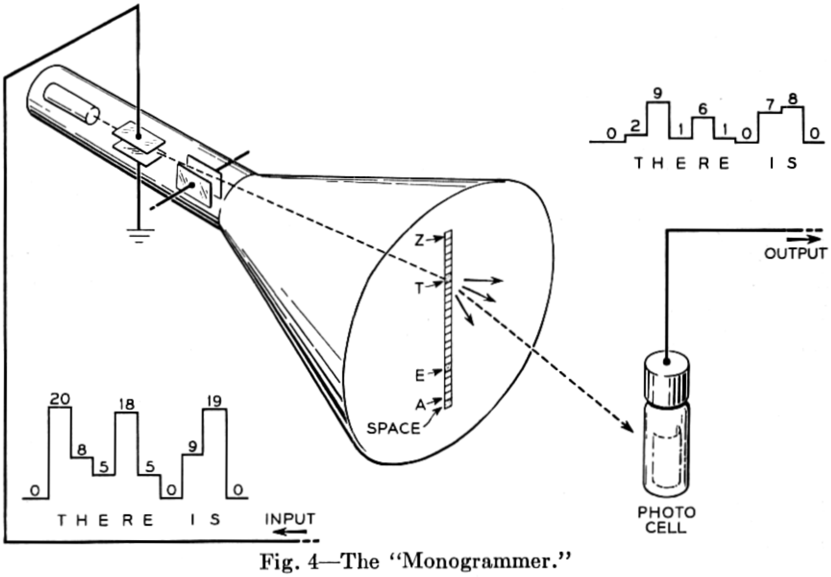


Fig. 4—The "Monogrammer."

that the system can be designed to have a lower *peak* power capacity. The signal out of the monogrammer can be remapped into binary digits using a Shannon-Fano code, pulse by pulse. However, this could have been done equally well with the original signal merely by rearranging the code groups in the coder tube. It is when we extend the principle to digrams and trigrams that the potentialities of the system become evident.

We can easily take account of the influence of the preceding message symbol. To do this we apply the signal to the vertical plates as before, and to the horizontal plates we apply the signal delayed by an amount equal to the time between successive pulses as shown in Fig. 5. Thus the beam is deflected *vertically* by the *present* message symbol, and *horizontally* by the *previous* message symbol. Whereas before we used a single column of optical filters chosen in accordance with the simple probabilities of the letters, we now have 27 columns, one for each letter and one for the space. The filters in each column are chosen in accordance with the *conditional* probabilities which apply when the corresponding letter was the previous symbol. For example, in the "Q" column (last letter Q), and the "U" row (present letter U) the mask would be opaque, since U is most common after Q. In general, the transmission of cell ij, in the i^{th} column and j^{th} row, is proportional to the rank of the entry for $p_i(j)$ when the entire distribution (conditioned on i) is ordered in a monotone decreasing sequence. The amplitude distribution of the output pulses

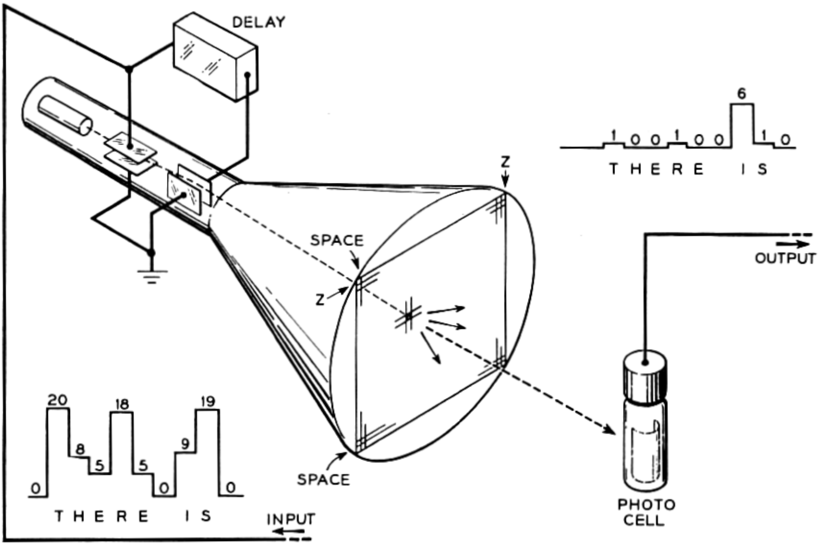


Fig. 5—The "Digrammer."

from the digrammer will be more peaked toward zero amplitude than that of the monogrammer. This is illustrated by the signals in the figures. At the receiver the same type of device, but with an inverse mask can be used to convert the signal back to its original form.

The digrammer can, with a little assistance, supply all the data required to prepare the encoding mask. If typical signals from the message source are applied to the cathode ray tube (without mask) for a long time, and a time exposure is made of the face of the tube, a lattice of spots will be obtained on the film. These spots will be dense where the high probability combinations occur and less dense elsewhere. The order of decreasing density in each column is noted, and the filter transmissions are arranged in the same order.

It is, of course, not necessary to use a phosphor, optical filters, and a photocell. An array of targets each of which connects to the appropriate tap on a load resistor might be simpler and more efficient. The cathode ray tube itself can be replaced with an appropriate diode switching network. Relay networks could be used for low-speed operation.

At the digrammer level we run out of new dimensions to use in the cathode ray tube. The principle can, however, be extended to trigramming and general n-gramming. For example, tetragramming could be accomplished by using a bank of ℓ^2 digrammers all in parallel, and all deflected by the present and previous samples. Only one of these tubes would be turned on at a time however. Which one this was would depend on the other two previous symbols of the tetragram. These (by additional delays) would be applied to the deflecting plates of a master switching tube having an array of target plates in place of a mask. Depending on the particular combination of signal samples applied to this tube, the beam would strike a particular target. The target current would then be used to turn on the beam of a particular digrammer tube, namely the one with the proper mask for that particular combination of two past symbols.

The complete array of equipment is admittedly rather staggering, but then, rather efficient coding should result. In practice it would probably be found that the masks of many of the tubes would be so similar that little gain resulted from differentiating between them. That is, the state of the message source might be nearly equivalent for several past combinations. In these cases, the group of tubes could be replaced with one having the best average mask, and the corresponding targets on the switching tube then tied together. This compromise would be particularly warranted for those tubes which were rarely used anyway. By these tricks it should be possible to keep the growth of equipment down to something approaching 2^{nH} rather than ℓ^n.

The output signal from the n-grammer will be, as we have seen, a series of pulses with an amplitude distribution very peaked toward zero and small pulses. If ℓ ,the alphabet, is large, these pulses can be efficiently encoded into a Shannon-Fano code. For small alphabets, granularity trouble can be reduced by remapping the output pulses two-by-two into pulses of base ℓ^2, and then encoding these into the Shannon-Fano code.

The output signal from the n-grammer with English text as the input message is a pulse amplitude representation of the type of "reduced text" one gets by using running n-gram prediction on English, as described by Shannon[3].

More efficient encoding would result if the properly matched Shannon-Fano code for *each particular conditional distribution* were applied to the output pulses, rather than using the same code for all of them. The efficiency of the coding operation would then be close to F_n as given by (8) (take $k = n$). This would add a great deal to the complexity and with most signals it is felt the gain would be small. If all the conditional distributions were alike after ordering, the improvement would be nil.

English text was used as the message in describing the n-gramming technique to emphasize the fact that it is a powerful general method which works even when the conditional probability distributions of a message are disorderly, multimodal affairs. It is obviously suited to other types of messages as well. Its main drawback is the complexity of apparatus required.

PREDICTIVE-SUBTRACTIVE CODING

When the conditional probability distributions of a message are unimodal (or merely strongly peaked as a rule in the vicinity of a particular sample amplitude) it is not necessary to re-order the distributions in order to obtain a reduced message for coding. The distributions may then merely be shifted along the amplitude scale until their modes are near zero (or their second moments about zero are nearly minimum). This shifting can be accomplished by computing from the preceding $(n - 1)$ gram the amplitude at which this mode or mean is located, and then subtracting this computed amplitude (or the nearest quantizing level) from the actual amplitude of the present sample. The difference in each case is a symbol whose amplitude distribution is peaked in the vicinity of zero amplitude. Fig. 6 shows a block schematic of a system using predictive-subtractive coding. In an actual system the reduced signal would ordinarily be encoded into Shannon-Fano code groups before transmission over the channel.

If s_0 is the present sample amplitude, and $s_1, s_2, s_3 \cdots s_n$ are previous sample amplitudes we compute a predicted value, s_p, for the present sample which is given by

$$s_p = f(s_1, s_2, \cdots s_n) \pm \delta$$

where $\delta < \frac{1}{2}$ quantizing level. If the conditional probability distribution for the present sample is $p_{s_1 \cdots s_n}(s_0)$, then the difference, or output, or "error" signal, ϵ, will have the conditional distribution $p_{s_1 \cdots s_n}(\epsilon + s_p)$ for this particular case. The simple distribution is then the weighted average over all cases, i.e.

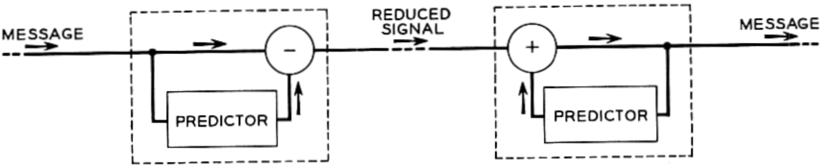


Fig. 6—Predictive-subtractive coding.

$$p(\epsilon) = \sum p(s_1, s_2, \cdots s_n) p_{s_1 \cdots s_n}(\epsilon + s_p)$$

where the sum is over all combinations of $s_1, s_2 \cdots s_n$.

Predictive-subtractive coding has especial merit when a simple function can be used for computing s_p. This is often the case. When the function is simply a weighted sum of the past sample amplitudes, i.e. when

$$s_p = (as_1 + bs_2 + cs_3 + \cdots) \pm \delta$$

we have what is known as *linear prediction*. Of course, linear prediction can always be used, but it may not be good enough with some types of messages.

As Wiener has shown the coefficients $a, b, c \cdots$ which minimize $\overline{\epsilon^2}$ are readily computed. For simplicity, assume only two message samples, s_1 and s_2, from the past are to be used. We then have

$$\epsilon = s_0 - s_p$$

$$\epsilon = s_0 - as_1 - bs_2$$

$$\epsilon^2 = s_0^2 + a^2 s_1^2 + b^2 s_2^2 - 2as_0s_1 + 2ab\, s_1s_2 - 2b\, s_0s_2$$

Now

$$\overline{s_0^2} = \overline{s_1^2} = \overline{s_2^2} = A_0$$

$$\overline{s_0 s_1} = \overline{s_1 s_2} = A_1$$

$$\overline{s_0 s_2} = A_2$$

where A_0, A_1, and A_2 are the values of the auto-covariance of the message wave at displacements of 0, 1, and 2 sampling periods. Thus

$$\overline{\epsilon^2} = (1 + a^2 + b^2)A_0 + 2(ab - a)A_1 - 2bA_2 .$$

The autocorrelation (normalized auto-covariance) is given by $\phi_i = \dfrac{A_i}{A_0}$. A_0 is proportional to the average power in the message wave, so the ratio $\rho = \dfrac{\overline{\epsilon^2}}{A_0}$ is the ratio of the power in the error signal to the power in the original message wave. Thus:

$$\rho = (1 + a^2 + b^2) + 2(ab - a)\phi_1 - 2b\phi_2$$

$$\frac{\partial\rho}{\partial a} = 2a + 2(b - 1)\phi_1 = 0$$

$$\frac{\partial\rho}{\partial b} = 2b + 2a\phi_1 - 2\phi_2 = 0$$

from which

$$a = \frac{\phi_1(1 - \phi_2)}{1 - \phi_1^2}, \qquad b = \frac{\phi_2 - \phi_1^2}{1 - \phi_1^2}.$$

With these values of a and b:

$$\rho = 1 - \phi_1^2 - \frac{(\phi_1^2 - \phi_2)^2}{1 - \phi_1^2}.$$

If $\phi_2 = \phi_1^2$, then the expressions simplify to

$$a = \phi_1, \qquad b = 0, \qquad \rho = 1 - \phi_1^2 .$$

As can easily be shown, if $\phi(x) = e^{-\alpha|x|}$, then all the coefficients except a are zero, and a has the value $e^{-\alpha}$. In other words, if the autocorrelation function is of exponential shape, the previous sample *alone* is needed for linear prediction. Samples before this add no further information as to the location of the mean of the conditional distributions.*

It happens that in typical television signals the autocorrelation for small displacements shows a very nearly exponential behavior. Thus linear prediction on the basis of the previous picture element alone is a natural method for television, particularly in view of the simplicity of apparatus required.

Linear prediction is easily instrumented. Fig. 7 shows in block schematic form the essentials of a linear predictor. Samples of the message are applied to a delay line. Taps along this line separated by the intersymbol time of the message, or multiples thereof, make the desired past symbols available. The signals from these taps are merely attenuated by amounts corresponding to the coefficients a, b, c $\cdots$ and added. A differential summing amplifier is shown to allow for negative coefficients, and also to accomplish the subtraction of the predicted sample amplitude from the present sample amplitude.

A complete linear predictor-subtractor is nothing but a transversal (time domain) filter whose impulse response is

$$f(t) = \delta(t) - a\delta(t - \tau) - b\delta(t - 2\tau) \cdots$$

and whose equivalent frequency response is therefore

$$F(\omega) = 1 - ae^{-i\omega\tau} - be^{-2i\omega\tau} \cdots$$

where τ is the delay between taps. If, for example, simple previous value prediction is used ($a = 1; b, c \cdots = 0$)

$$F(\omega) = 1 - e^{-i\omega\tau} = 2i \sin \frac{\omega\tau}{2} e^{-\frac{i\omega\tau}{2}} .$$

* From the preceding expression for ρ, we see that $\rho = 0$ (i.e., perfect prediction is possible) if:

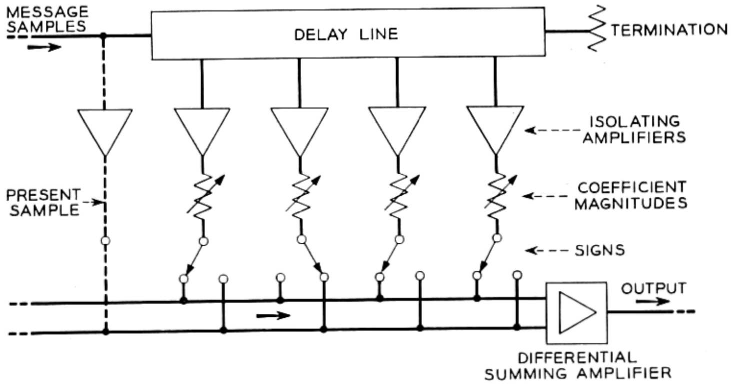
$$(\phi_1^2 - \phi_2)^2 = (1 - \phi_1^2)^2$$
$$\phi_1^2 - \phi_2 = \pm(1 - \phi_1^2)$$
$$\begin{cases} \phi_2 = 1 \\ \phi_2 = 2\phi_1^2 - 1. \end{cases}$$

If $\phi_2 = 1$, the message samples alternate between two independent but constant values. For this case $a = 0$, $b = 1$. If $\phi_2 = 2\phi_1^2 - 1$ the autocorrelation is a cosine wave so the message consists of samples of a sinusoid. In this case $a = 2\phi_1$, $b = -1$. If ϕ_1 is nearly unity, the sinusoid is of low frequency, and the prediction approaches "slope" prediction (i.e. extrapolation of a straight line through the last two samples).

In any case where perfect prediction is possible the wave is periodic and therefore $H = 0$.

It is often argued that linear prediction is therefore nothing more than pre-distortion (frequency-wise). If the message is unquantized and unsampled, and if the signal from the predictor is applied to the channel as straight amplitude or single side-band modulation, the allegation is certainly true. Pre-distortion is a perfectly valid way of improving the statistical match between message, and channel, and destination as the optimum filter theory of Weiner and Lee shows. On the other hand, when the message is sampled and quantized, and when the output of the linear predictor is further encoded into a sequence of binary digits, and these are possibly remapped onto a higher base for the channel, then the information is being handled digitally throughout, and the usual reasons for a certain type of predistortion no longer apply. The best linear predictor will usually be quite different for the two cases. Even though analogue operations (such as subtraction of amplitudes) are used for convenience, the quantization makes the operation discrete and hence equivalent to a digital process.

At the beginning of this section, we were a little vague as to whether the prediction should shift the modes or the means of the conditional distributions to zero amplitude. If the object of the prediction-subtraction operation is to minimize the *power* in the error signal, then certainly the means should be shifted to zero. The coefficients as determined from the autocorrelation function do this aside from quantizing granularity. They specify an optimum least-square predictor, i.e., one which tends to minimize $\overline{\epsilon_j^2} = \sum_j j^2 p(j)$.



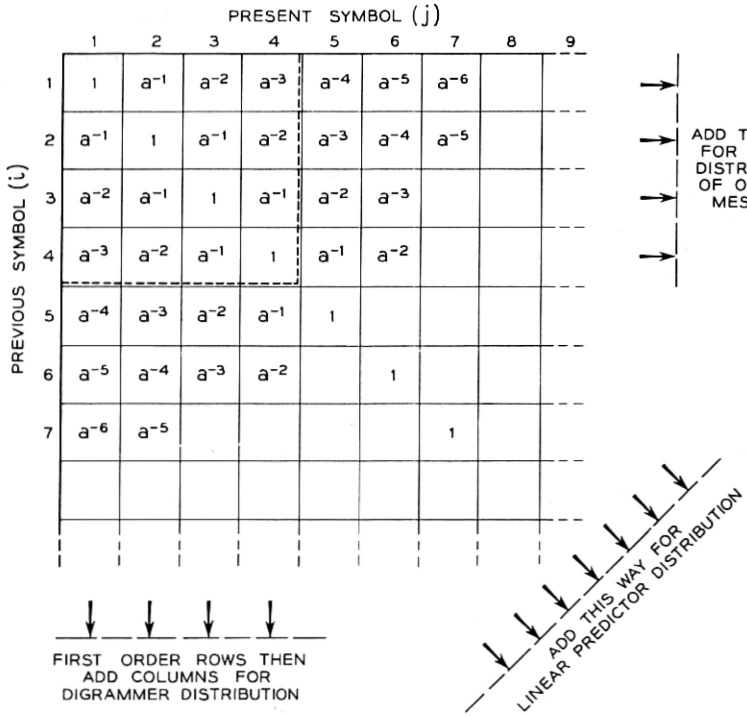
IF PRESENT SAMPLE IS SUBTRACTED, OUTPUT WILL BE "ERROR" SIGNAL

IF PRESENT SAMPLE IS OMITTED, OUTPUT WILL BE "PREDICTION"

Fig. 7—A linear predictor.

Power reduction is an index of merit when many reduced signals are to be sent by frequency division over one channel, as we have said. When the object is to reduce the channel capacity required for a single message source, then it is the upper bound entropy of the reduced signal which should be minimized, not the power. That is we want $-\sum_j p(j) \log p(j)$ to be minimized. For certain types of signals this requires the modes to be shifted to zero, although this is by no means a general rule. Shifting the modes to zero may actually increase the entropy of the "reduced" signal over that of the original message, by adding too many new symbol levels, as the example in the last section shows.

If the original message has ℓ quantizing levels, the reduced message after predictive-subtractive coding will in general contain more than ℓ levels since an error of more than $\frac{\ell}{2}$ can be made in either direction. An n-gramming operation, on the other hand, never increases the alphabet.


PRESENT SYMBOL (j)
PREVIOUS SYMBOL (i)
ADD THIS WAY FOR SIMPLE DISTRIBUTION OF ORIGINAL MESSAGE
ADD THIS WAY FOR LINEAR PREDICTOR DISTRIBUTION
FIRST ORDER ROWS THEN ADD COLUMNS FOR DIGRAMMER DISTRIBUTION

	1	2	3	4	5	6	7	8	9
1	1	a^{-1}	a^{-2}	a^{-3}	a^{-4}	a^{-5}	a^{-6}		
2	a^{-1}	1	a^{-1}	a^{-2}	a^{-3}	a^{-4}	a^{-5}		
3	a^{-2}	a^{-1}	1	a^{-1}	a^{-2}	a^{-3}			
4	a^{-3}	a^{-2}	a^{-1}	1	a^{-1}	a^{-2}			
5	a^{-4}	a^{-3}	a^{-2}	a^{-1}	1				
6	a^{-5}	a^{-4}	a^{-3}	a^{-2}		1			
7	a^{-6}	a^{-5}					1		

Fig. 8—Joint probability distribution (divide all coefficients by the sum over each array).

Other operations besides simple subtraction of the predicted symbol from the present symbol are of course possible. However, in most cases it would seem that if a more complicated operation were indicated, n-gramming would have provided a better start.

ILLUSTRATIVE EXAMPLE

Let us compare the operation of n-gramming and prediction-subtraction techniques on a hypothetical message. We will assume the message has digram statistics, but that longer range statistical influences either do not exist or are ignored. The statistics are then specified entirely by the joint probability distribution $p(i, j)$ of a pair of symbols. Let us assume that there are ℓ quantizing levels, and that

$$p(i, j) = Ka^{-|i-j|}$$

where a is a constant > 1, and K is given by

$$K = \left[\sum_{i,j} a^{-|i-j|}\right]^{-1}$$

K is the factor which assures that $\sum_{i,j} p(i, j) = 1$.

Thus the most likely level is that of the previous sample. A sample differing by one level is $1/a$ times as likely, one differing by m levels is a^{-m} times as likely. Figure 8 shows a plot of the *relative* values of $p(i, j)$ (neglecting the factor K). For $\ell = 4$, the total array would be the 4 x 4 portion enclosed by the dashed line. This sort of distribution is rather similar to those of typical television signals, as shown by preliminary measurements, although typical values of a have yet to be determined. With no statistical coding, the required channel capacity is

$$H_0 = \log_2 \ell \text{ bits/sample.}$$

If the simple distribution of individual samples is taken into account, the required channel capacity is reduced to

$$H_1 = -\sum p(i) \log p(i)$$

where

$$p(i) = \sum_j p(i, j) = \sum_i p(i, j) = p(j)$$

H_1 may be computed from the array of *relative* coefficients by adding the rows to form the sums

$$S_i = \frac{1}{K} \sum_j p(i, j) = \frac{p(i)}{K}\,.$$

In terms of these sums, we have

$$H_1 = \log \frac{1}{K} - K \sum_i S_i \log S_i .$$

Since, with the assumed distribution, the S_i are all nearly equal very little reduction in channel capacity is achieved by this step.

With linear prediction, the modes of the distributions ($i = j$) could be centered at zero merely by sending the difference between the present and previous sample (previous value prediction). This would give a reduced signal whose distribution may be found by adding the array along the diagonals. The required channel capacity is then given by:

$$H_L = -K\ell \log K\ell - 2\sum_{k=1}^{\ell-1} \frac{k(\ell - k)}{a^k} \log \frac{k(\ell - k)}{a^k}$$

The distribution of the signal from a digrammer is found by rearranging each row of the table in order of decreasing probability and then adding the resulting columns. Call these sums S_d . The digrammer output will thus require a channel capacity:

$$H_0 = -\sum_{d=1}^{\ell} KS_d \log KS_d$$

$$= \log \frac{1}{K} - K \sum_{d=1}^{\ell} S_d \log S_d$$

Lastly, the true rate of the source is given by

$$H = -\sum_i p(i) \sum_j p_i(j) \log p_i(j)$$

$$= \log \frac{1}{K} - H_1 + 2K \sum_{k=1}^{\ell-1} \frac{\ell - k}{a^k} k \log a$$

Values for the above quantities were computed for $a = z$ and $n = 2, 3, 4, 6, 8, 16, 32, \infty$. For the case of $a = 2$, we find that

$$K = [3\ell - 4(1 - 2^{-\ell})]^{-1}$$

and that as $\ell \rightarrow \infty$,

$$H_L , H_D , H \rightarrow \tfrac{4}{3} + \log_2 3 = 2.918 \text{ bits.}$$

The results are shown in the Table I and also are plotted in Fig. 9.

While H_0 and H_1 increase without limit as ℓ is increased, H_L , H_D , and H quickly approach a definite limit. This limit exists because we assumed that the decrease in joint probability as a function of *number*

Table I

Number of levels	H_0	H_1	H_L	H_D	H
1	0	0	0	0	0
2	1	1	1.252	0.918	0.918
3	1.585	1.583	1.777	1.437	1.422
4	2	1.995	2.074	1.764	1.750
6	2.583	2.575	2.381	2.157	2.131
8	3	2.988	2.552	2.370	2.343
16	4	3.99	2.768	2.678	2.641
32	5	$5 - \epsilon$	2.850	2.818	2.782
∞	$\log \infty$	$\log \infty$	2.918	2.918	2.918

(These figures were computed by slide rule so the fourth figure is not very significant.)

of levels off the diagonal was the same regardless of ℓ. In typical signals this is not true. The decrease is more apt to depend on *amplitude difference* and the finer the quantum step, the more levels a given difference represents. As a result, the probability will fall off less per level off the diagonal, and doubling ℓ will in general add one bit to H.

On the other hand, doubling the sampling rate will not in general double the required channel capacity, for the closer spaced samples will

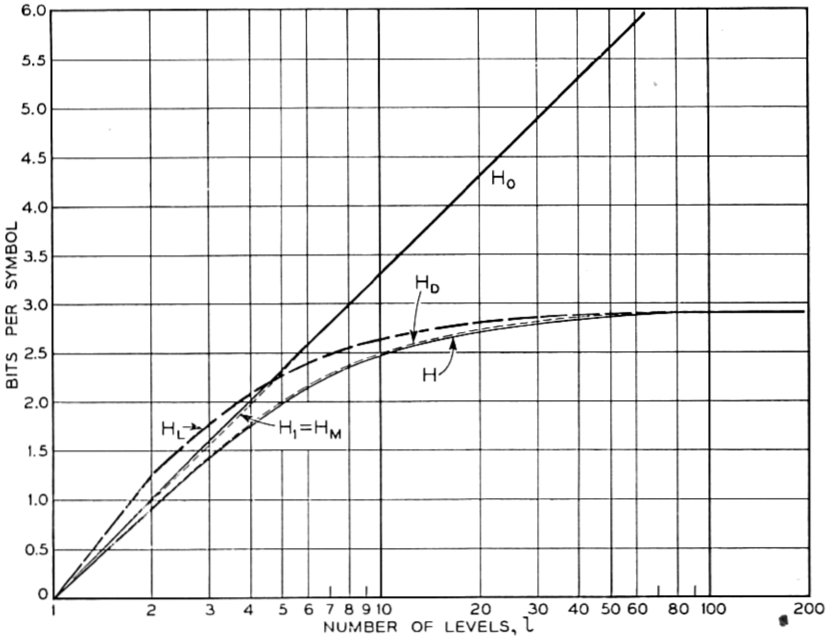


Fig. 9.

be more highly correlated. Thus in TV, doubling the horizontal resolution would not double the bandwidth for the same picture material if use were made of the statistics. (Of course, increased resolution in TV might encourage the use of more detailed scenes and this *would* increase the required bandwidth.)

It should also be noticed that for small ℓ, linear prediction actually makes matters worse. The increase in the number of levels in the error signal more than offsets the peaking of the distribution.

Since all the conditional distributions in this message are of similar shape (after ordering), H_D and H are almost the same, for all ℓ. The difference between H_L and H_D is slight except for small ℓ because the distribution we assumed is unimodal throughout.

Fig. 10 shows the simple probability distributions for (a) the original message, (b) the reduced signal from linear prediction, and (c) the reduced signal from the digrammer.

VARIABLE DELAY AND OTHER PROBLEMS

We have seen in the last two sections how it is possible to convert a message for which $H \ll H_0$ as a result primarily of intersymbol correlation, into a reduced signal for which $H \ll H_0$ as a result primarily of a highly peaked probability distribution in the individual symbols (i.e. one for which $H_1 \rightarrow H$). Since the operations are reversible, the true information rate, H, is preserved. In the original signal it was the *conditional* distributions which were peaked, while the simple distribution was relatively flat. In the reduced signal the *simple* distribution is peaked.

The result is that whereas a Shannon-Fano code would only have been effective on the original message if applied to blocks two or more symbols in length (and then it would ignore correlation between blocks), in the reduced signal the code will be effective on a symbol to symbol basis.

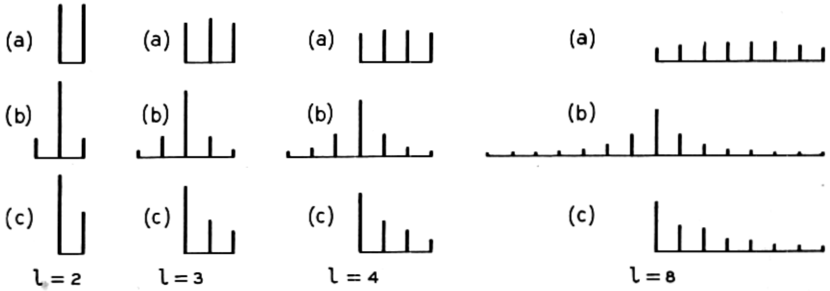


Fig. 10—Probability distributions: (a) Original, (b) After linear modal prediction, (c) After digramming.

The encoding of the reduced signal into binary digits presents no theoretical difficulties. A PCM type coder tube[4] with the appropriate Shannon-Fano groups built into it is all that is needed. The biggest practical complication arises out of the fact that the code groups are of different length. Some messages, such as written text, can be fed into the system as fast as it can handle them. The transmission time will then vary with the message complexity. Others, such as television are generated and must be accepted and delivered at a constant rate. One solution is then to take the binary digits in big and little batches as they come from the coder and store the surplus in a sort of pulse "surge tank" before they are sent over the channel at a regular rate. At the receiver, a similar sort of storage register is necessary as the pulses arrive over the channel at a regular rate and are used by the decoder at a varying rate. Devices which will perform this variable delay function satisfactorily for signals with relatively slow sampling frequency are available, and as the art progresses there is every reason to believe that high speed sampled signals like television can be handled also.

It will be noticed that the digramming or prediction operation, while it involves memory, does not introduce appreciable transmission delay. Each symbol of the reduced signal appears the moment the corresponding message sample is applied. The total transmission delay required for statistical coding thus depends upon how much variation is required in the variable delay units. This in turn depends upon the degree of stationarity in the "local information rate" of the message. For example, in television, if each line could be described (by the n-grammer and subsequent coder) in the same total number of binary digits, then the total delay variation and total delay would be less than one line time. Since this is not true, we either must have enough channel capacity to send in one line time the number of digits corresponding to the "worst" line, or enough variable delay to average the existing rate over many lines.

Probably the most practical solution is to provide sufficient channel capacity and variable delay to take care of all but a small fraction of the possible message sequences. Then when an unusual stretch of message continues long enough for the variable delay to be nearly all used up, the system should fail in some relatively harmless way. In television, the sampling rate could be momentarily reduced, for example. This would degrade the resolution in rare situations, but a small amount of this could be tolerated in return for transmission savings.

If long blocks of the message are efficiently encoded as a group, then an error in transmission may cause the whole block to be reproduced incorrectly. If n-gramming or prediction is used, then an error in transmission will cause the receiver to function improperly not only for that symbol, but its further n-gram decoding or prediction will also be disturbed. Thus errors of transmission are either spread over definite blocks, or propagate for a considerable time rather than being confined to the particular symbols sent in error. In fact, if the encoding were completely efficient, all received sequences would be possible messages, and a single error could convert the received message from the proper one into a completely different but possible one. With *no* redundancy there is no way to recognize an error. It is for these reasons that we have assumed a rugged (quantized) channel. In view of the eight to ten db more average power required in a quantized channel to achieve the same channel capacity as an ideal channel of the same bandwidth, considerable statistical saving must be possible before statistical coding may be warranted. This initial handicap of course does not apply to channels already designed to work on a digital basis for other reasons. Lastly, the use of error correcting codes[5] is a possibility. In these codes a small amount of redundancy is introduced in a particularly efficient fashion. As a result, a certain frequency of transmission errors can be tolerated without causing errors in the reproduced message.

REFERENCES

1. Oliver, Pierce, and Shannon, "The Philosophy of PCM," *Proc. Inst. Radio Engrs.*, Nov. 1948.
2. C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Tech. J.*, July and Oct. 1948; Shannon and Weaver, "The Mathematical Theory of Communication," University of Illinois Press, 1949.
3. C. E. Shannon, "Prediction and the Entropy of Printed English," *Bell System Tech. J.*, Jan. 1951.
4. R. W. Sears, "Electron Beam Deflection Tube for Pulse Code Modulation," *Bell System Tech. J.*, Jan. 1948; W. M. Goodall, "Television by Pulse Code Modulation," *Bell System Tech. J.*, Jan. 1951.
5. R. W. Hamming, "Error Detecting and Error Correcting Codes," *Bell System Tech. J.*, Apr. 1950.