#2

296

BOUNDS TO THE ENTROPY OF TELEVISION SIGNALS

JACK CAPON

LOAN COPY only

TECHNICAL REPORT 296

MAY 25, 1955

RESEARCH LABORATORY OF ELECTRONICS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS

The Research Laboratory of Electronics is an interdepartmental laboratory of the Department of Electrical Engineering and the Department of Physics.

The research reported in this document was made possible in part by support extended the Massachusetts Institute of Technology, Research Laboratory of Electronics, jointly by the Army (Signal Corps), the Navy (Office of Naval Research), and the Air Force (Office of Scientific Research, Air Research and Development Command), under Signal Corps Contract DA36-039 sc-42607, Project 102B; Department of the Army Project 3-99-10-022.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

RESEARCH LABORATORY OF ELECTRONICS

Technical Report 296 May 25, 1955

BOUNDS TO THE ENTROPY OF TELEVISION SIGNALS

Jack Capon

This report is based on a thesis submitted to the Department of Electrical Engineering, M.I.T., 1955, in partial fulfillment of the requirements for the degree of Master of Science.

Abstract

Statistics of pictures were measured, including the second-order probabilities of successive cells and the autocorrelation function. These quantities, obtained by independent methods, are used to compute an upper bound to the entropy per sample. The results are compared and used to estimate the amount by which the channel capacity required for television transmission might be reduced through exploitation of the statistics measured.

INTRODUCTION

This research is an application of statistical communication theory to television transmission. An upper bound to the entropy per symbol is obtained by two independent methods. It is shown that this quantity not only provides an insight into the nature of pictures but furnishes a theoretical limit to the efficiency of picture-coding methods.

In Section I the basic concepts of information theory are introduced. The wastefulness of nonstatistical versus statistical coding is brought out in a general discussion of the coding problem.

Section II explains the scanning process and the standards used in commercial television transmission. The bandwidth requirement is shown to arise from the necessity of transmitting sharp changes in brightness such as those that appear at the edges of objects. It is shown that these changes occur in only a small portion of the picture. Large parts of the picture exhibit very low gradients in brightness. This phenomenon and the similarity of successive frames cause the video signal to have a low information content; since a picture with a high information content would be a random pattern, the low information content is assumed to be essential to the nature of pictures.

Section III contains details of the measurement and calibration procedures. The effect of the noise on the probability measurement is discussed. An exact analysis for deriving the signal probabilities from those of the signal plus noise is given, but it is pointed out that this method is vulnerable to experimental errors. Upper and lower bounds to the entropy per symbol are obtained and compared with the upper bound obtained by autocorrelation function methods. This, a true upper bound, will be the actual bound only if the process is gaussian.

Results are presented in Section IV, along with experimental procedures and computations. Various sources of error are discussed.

Section V contains an analysis of the results obtained. It is shown that the information content of pictures depends upon the amount of detail in the picture and the intensity range covered. To obtain a better estimate of the entropy per symbol from the autocorrelation function measurement, an entropy conditioned by three previous samples instead of one is computed. The entropy of the difference signal is also discussed.

The appendices contain derivations that could not properly be included in the main body of the paper.

Previous investigators have measured bounds to the entropy per symbol. W. F. Schreiber (4), of Harvard University, measured second-order probabilities of pictures and computed a bound to the entropy per symbol. E. R. Kretzmer (5), of Bell Telephone Laboratories, measured first-order probabilities and autocorrelation functions of pictures and computed a bound to the entropy per symbol; but their results were not obtained from the same set of pictures. This experiment is an extension of their work but the second-order probabilities, autocorrelation functions, and the bounds to the entropy per symbol were derived from one set of pictures.

I. CODING AND INFORMATION THEORY

1. BASIC CONCEPTS

The theory of communication has taught engineers to examine the content of the messages that they are transmitting from a statistical viewpoint. In the light of this new theory, engineers think of messages in terms of probability of occurrence. It is intuitively obvious that the reception of a signal of small probability conveys much more information than the reception of a signal of great probability. Examples occur frequently in everyday life. For example, a report of snow in Boston in August conveys much more information than a report of fair and mild weather.

One of the basic contributions of information theory is the definition of a measure of information that conforms to one's intuitive feelings about information. The basic concepts used in our discussions are cited for convenient reference.

In accordance with our previous statement about probability of occurrence, we can say that the information conveyed by a message x_i selected from a set X is (see ref. 1, Chap. II, p. 10)

$$I(x_i) = -\log p(x_i)^*$$

where $p(x_i)$ is the probability of occurrence of message x_i.

One sees that for $p(x_i) = 1$, that is, certainty about x_i, the information conveyed is zero. This obeys our intuitive feelings about information, for if an event is known to be certain to occur, its occurrence does not communicate any useful knowledge. On the other hand, as $p(x_i)$ is made smaller and smaller, increasing the uncertainty about x_i, the information conveyed becomes larger and larger. This is also in line with our intuitive feelings of what a measure of information should be.

In a similar manner, the information conveyed by a message y_j from a set Y, about a message x_i from a set X is (see ref. 1, Chap. II, p. 4)

$$I(x_i; y_j) = \log \frac{p(x_i|y_j)}{p(x_i)}$$

where $p(x_i|y_j)$ is the conditional probability of x_i if y_j is known.

This definition is also in accordance with our feelings about information. If the a posteriori is increased so that y_j specifies x_i more completely, one would expect the information conveyed by y_j about x_i to be increased. This is seen to be the case from the definition. We observe that, since the maximum value of $p(x_i|y_j)$ is unity, the maximum value for $I(x_i; y_j)$ is $-\log p(x_i)$. That is, the message y_j can convey no more information about x_i than that which x_i itself conveys. This is another obvious condition

*Logarithms are to the base two unless otherwise specified.

that one would expect the definition of information to obey.

Entropy is defined as the average amount of information that must be provided in order to specify any particular message x from a set X. It is represented by

$$H(X)^{*} = - \sum_{X} p(x) \log p(x)$$

The properties of the entropy that are of importance to the discussion are: If p(x) vanishes at point x_i in the set X, then the term corresponding to x_i also vanishes: $\lim p(x_i) \log p(x_i) = 0$, where $p(x_i) \to 0$. Thus H(X) is equal to zero when and only when p(x) vanishes at all points but one, in which case p(x) = 1 at the remaining point. This situation corresponds to the case in which only one particular message can ever be transmitted; then, no information can be provided about it because it is completely specified from the start. One might also say that the uncertainty regarding the message set is zero.

For any given set of symbols, the entropy is maximum when all symbols occur with equal probability. If the number of symbols is N, then from the formula for H(X) it is apparent that the maximum value of H(X) is given by log N. In this case the uncertainty about the message set is greatest. It is of interest to note that as the probabilities of the messages become more unequal, the entropy decreases. This is a very important property and will be referred to frequently in following sections.

For the sake of completeness a conditional entropy will be defined as

$$H(Y|X) = - \sum_{X} \sum_{Y} p(x; y) \log p(y|x)$$

where p(x; y) is the probability of joint occurrence of events x and y.

This quantity can be interpreted as the average amount of information required to specify the event y if event x should be known. An example of this is found in the transmission of printed English. In this case, the set X corresponds to a particular letter of interest; the set Y to the letter succeeding it. Suppose that we are interested in the letter q, so that the set X corresponds to just this one symbol. We are fairly certain that if we receive a "q" the succeeding letter will be a "u." Thus, of the twenty-seven symbols in the Y space, u is very probable, and all others improbable. This means that the entropy, or average information, is small – as mentioned before. Let us now examine the letter c, and let the X space now consist of just this symbol. If we know the occurrence of a c, we cannot be too sure of what the succeeding letter will be. It is likely to be a, e, i, o, u, or y and unlikely to be b, d, f, g, or t. In this example

*X in H(X) signifies that an average over the set X has been performed. Although the entropy is a function of the probabilities p(x), H(X) will be the notation used to denote the entropy of a set X with probabilities p(x).

the probabilities of the symbols are more alike than they were in the preceding one. Thus the entropy will be larger than before. This concept of how the conditional entropy changes for different conditional probability distributions is a most important one and, as will be seen later, forms the foundation for the measurements in this experiment.

The final quantity to be defined will be the channel capacity. It is the maximum average value of the mutual information provided by a message y about a message x. Thus

$$C = \left[\sum_{X} \sum_{Y} p(x;y) \log \frac{p(x|y)}{p(x)} \right]_{max}$$

The $p(x|y)$ represents the probability of recognizing x, the transmitted symbol, given the received signal y. It is other than one because of the noise in the channel. The maximizing is done with respect to the transmitter symbol probabilities, $p(x)$.

For a more complete discussion of channel capacity see reference 2. For the present let us say that the channel capacity of a noisy channel is the maximum possible rate of transmission of information when the source is properly matched to the channel.

2. CODING

Coding, in its most general form, is any process by which a message or message waveform is converted into a signal suitable for a given channel. Frequency modulation, single-sideband modulation and pulse-code modulation are examples of coding procedures; any modulator is an example of a coding device.

There are, in general, two classes of coding processes and devices: those that make no use of the statistical properties of the signal and those that do. Almost all of the processes and devices used in present-day communication belong to the 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 sequences, and 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. These two types of coding will now be discussed.

Messages can be either continuous waves like speech, music, and television; or they can consist of a succession of discrete characters, each with a finite set of possible values, such as English text. Continuous signals can be converted to discrete signals by the process of sampling and quantizing (3). This permits us to talk about them as equivalent from the communication engineering viewpoint. It is generally easier to think in terms of discrete messages. Thus, quantization of continuous signals, if they occur, will be assumed, and we shall think of our messages as always being available in discrete form.

Suppose we have the message set $x_1, x_2, \ldots, x_8$, with the probabilities given in Table I. If we did not make use of the statistics of this message and assumed that all

Table I.

Statistical Versus Nonstatistical Coding

Message	Probability of Occurrence	Nonstatistical Coding	Statistical Coding
x_1	1/2	000	1
x_2	1/4	001	01
x_3	1/8	010	001
x_4	1/16	011	0001
x_5	1/32	100	00001
x_6	1/64	101	000001
x_7	1/128	110	0000001
x_8	1/128	111	0000000

messages were equally likely, the code of column 3 of Table I would be the result. On the average, the number of symbols per message is

$$N = 3\left(\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \frac{1}{64} + \frac{1}{128} + \frac{1}{128}\right) = (3 \times 1)$$

$$N = 3 \text{ symbols/message}$$

If, now, a statistical coding scheme (Shannon-Fano code) is used (see ref. 1, Chap. III), one arrives at the code of column 4. Although some of the code words are longer, it will be found that the average number of symbols is

$$N = 1\left(\frac{1}{2}\right) + 2\left(\frac{1}{4}\right) + 3\left(\frac{1}{8}\right) + 4\left(\frac{1}{16}\right) + 5\left(\frac{1}{32}\right) + 6\left(\frac{1}{64}\right) + 7\left(\frac{1}{128}\right) + 7\left(\frac{1}{128}\right)$$

$$= 1.98 \text{ symbols/message}$$

which is considerably less than the 3 symbols per message obtained previously. This is a vivid example of how wasteful nonstatistical coding can be.

One of the teachings of information theory is that given a set of messages with certain probabilities, one can set a lower bound to the average number of symbols per message that can be used to code the messages. This lower bound is given by the entropy H(X) associated with the probabilities of the set when a binary alphabet is used (1).

Similarly, if one were to code the second of a pair of messages from a knowledge of the first message, the lower bound would be set by the entropy of their conditional probabilities:

$$H(Y|X) = -\sum_{X} \sum_{Y} p(x;y) \log p(y|x)$$

where y is the message succeeding x.

Suppose, now, that one is interested in coding the n^{th} symbol from a knowledge of the (n-1) preceding symbols. The lower bound to the number of symbols per message is given by

$$H(A_n|A_{n-1};A_{n-2};\ldots;A_1) = -\sum_{A_1} \cdots \sum_{A_n} p(a_1;\ldots;a_n)$$

$$\times \log p(a_n|a_{n-1};\ldots;a_1)$$

where a_i is the i^{th} symbol in the sequence, and A_i is the space containing the random variable a_i.

In the limit as $n \to \infty$, this expression approaches the entropy per symbol (see ref. 1, Chap. III, pp. 24-25) of the ensemble $H(A|A^{\infty})$ formed by all sequences of length n. From a practical point of view, what does this quantity mean? It tells how many bits of information must be added, on the average, to specify the n^{th} symbol of a sequence from a knowledge of the (n-1) preceding symbols. How large n is, in general, is determined by how far the statistical dependence extends.

The direct measurement of this quantity is extremely difficult, if not impossible, since it entails the measurement of high-order probabilities. What one does in practice is to find an upper bound for it. It can be shown (see ref. 1, Chap. III, pp. 21-25) that the n^{th} order conditional entropy for increasing n, constitutes a series of successive approximations to it. Hence, in this experiment the second-order probabilities and autocorrelation function were measured and from these an upper bound to the entropy per symbol was computed. As we shall point out later these measurements are made on signals derived from pictures.

II. THE PHILOSOPHY OF PRESENT-DAY TELEVISION TRANSMISSION

1. PRESENT STANDARDS

A picture can be approximated by an array of dots. The smaller the individual dots are, the greater the resolution of the picture, and the closer the approximation. For a given resolution, the intensity of a dot is a function of two space variables, and of time if the picture is changing. The basic problem in the transmission of pictures is the coding of this function of three variables for transmission through a one-dimensional (time) channel. This is done at present by means of a process known as scanning. The picture is broken up into 525 horizontal lines, and the intensity of every cell in the line is specified. All television pictures have an aspect ratio of four to three, which means that the ratio of frame width to frame height is four thirds. Thus, if the resolution is the same in the horizontal and vertical directions, the picture consists of (525) (525) (4/3) or approximately 366,000 cells. To create the illusion of continuous motion, the picture is scanned 30 times per second. Actually, an interlace scheme is used, whereby the odd lines are scanned in one sixtieth of a second, and the even lines in the next sixtieth of a second. This, essentially, has the effect of increasing the flicker frequency from 30 cps to 60 cps.

An estimate of the bandwidth required can be found by assuming that every cell is different from the succeeding cell in intensity, and that the resulting video waveform is a square wave. Accordingly, for a 525-line picture with an output ratio of 4:3, the period of this square wave will be

$$T = \frac{2}{\frac{(525)(4/3)(525)}{1/30}}$$

$$= 0.182\ \mu\text{sec (neglecting blanking time)}$$

To pass the fundamental of this square wave, our system must have a bandwidth of

$$W = \frac{1}{T} = 5.5\ \text{Mc/sec}$$

In practice, the Federal Communications Commission allots 6 Mc/sec to a station and the picture signal occupies 4 Mc/sec of this. Hence, the assumption made about equal vertical and horizontal resolutions is not correct. However, in this experiment the assumption will be made, since the bandwidth of all components used is large enough to give this resolution.

Thus the problem of coding the picture has been solved but at the expense of using a large bandwidth.

2. REDUNDANCY IN TELEVISION TRANSMISSION

According to the dictionary, a picture is a description so vivid as to suggest a mental image. To convey its meaning to someone's mind, the picture must be well ordered and must contain few boundaries and vast regions of slowly varying intensity. If this were not so, one would be looking at a picture of white noise (similar to looking at a snow storm) or else at a picture that did not make much sense. However, according to our definition this would not be a picture, since it conveys no meaning. With these ideas in mind, we can say that knowing part of a picture makes it possible to draw certain inferences about the remainder. As a matter of fact, the brightness value of a picture point in a frame is very likely to be nearly equal to the brightness value of the preceding point; the only exceptions occur at boundaries, which, as we know, occupy a small part of the picture. Besides cell-to-cell correlation, there is also correlation between successive frames, which tends to lower the information content of the video signal. If one were to examine these successive frames, he would find that they are almost exactly alike. That this is so is evident from the fact that subject material cannot move very fast. How far can one move in one-thirtieth of a second! In general, then, these picture points are related in a statistical manner, imparting a type of redundancy to the point-by-point description that makes up the video signal.

Commercial television uses a nonstatistical coding scheme and thus does not take advantage of the statistical relationships in the signal. As we observed before (see Table I), this can be costly from the point of view of efficiency of transmission. By means of statistical coding a reduction in the number of symbols per message could be achieved. This means a reduction in the channel capacity, since fewer symbols have to be transmitted. The channel capacity of a channel of band W disturbed by white noise of power N when the average transmitter power is limited to P is given (see ref. 2, pp. 66-68) by

$$C = W \log \left(1 + \frac{P}{N}\right) \text{bits/sec}$$

If by means of a proper coding scheme the channel capacity could be reduced, a potential saving in bandwidth or power, or both, exists. It was pointed out earlier that the greatest reduction in channel capacity is foretold from knowledge of the n-order conditional entropy, where n is large and extends over as many previous symbols as have statistical influence. As we also pointed out, it is extremely difficult to measure probability distributions of an order higher than two. However, an upper bound to the entropy per symbol can be obtained by measuring the second-order probabilities, that is, the probability that a pair of adjacent cells will be at given intensity levels.

An upper bound to the entropy per symbol can also be obtained from the autocorrelation function of the picture, as will be shown later. This bound, realized for a gaussian process, will be greater than the bound obtained by the other method and will equal it only if the probability distribution is truly gaussian. It will be interesting to compare

these two values obtained by independent methods.

Previously, Schreiber (4) measured second-order probabilities and from the entropy determined that a potential saving of a factor of two or three in channel capacity could be achieved. Kretzmer (5) measured first-order probabilities and autocorrelation functions and found that a saving of a factor of two could be effected in the channel capacity.

The following calculation gives an idea of how large the channel capacity is. Goodall (6) found that for a 32-level system, contours would be masked with an input peak-to-peak signal to rms noise ratio of 40 db. For a 4 Mc/sec bandwidth this implies a channel capacity of

$$C = W \log_2 \left(1 + \frac{P}{N}\right) = (4 \times 10^6)\,(40 \log_2 10) = 530 \text{ million bits/sec}$$

The magnitude of this figure becomes evident when one realizes that the maximum rate of information reception of the human brain is about 45 bits per second (7). This is another form of redundancy but the exploitation of this form of redundancy is impractical, since it would require knowing which part of the picture each observer was looking at, that is, which part of the television screen should display the 45 bits. Clearly, this is an impossible feat.

3. A SUMMARY OF AIMS

A list of the quantities to be measured and computed is given for reference purposes.

1. The second-order probability of successive cells in a typical picture is measured. An upper bound to the entropy per symbol is computed, and from this an estimation of the saving in channel capacity is made.

2. The autocorrelation function of the same picture is measured, and an upper bound to the entropy per symbol is computed. This value is then compared to the previous one, which was determined by an independent method.

III. MEASUREMENT PROCEDURES

The first part of this section deals with the second-order probability measurement; the last part with the autocorrelation function. It is assumed that a picture is composed of (525) (525) (4/3) = 366, 000 cells, corresponding, respectively, to the 525 scanning lines used in television transmission and the (525) (4/3) = 700 resolvable elements in a horizontal line. The resolution in horizontal and vertical directions is assumed to be the same. It is also assumed that each cell can have any one of 32 levels of intensity from white to black. This is a reasonable assumption, since it has been found (6) that a 32-level picture is a good approximation for the original.

1. METHOD FOR MEASURING THE JOINT PROBABILITY

The measurement of a probability is basically a counting process. Thus, the most direct method of measuring the joint probability is to cut the picture into 366, 000 parts and count the occurrence of each joint event. This procedure is, of course, laborious.

In this experiment television methods are used. The technique for the measurement consists of displaying on the face of a cathode-ray tube a brightness pattern that is proportional to the second-order probability. This method was used by Schreiber (4) in his experiment. Kretzmer (5) also used it in his measurement of first-order probability distributions. He employed a clever photographic process to obtain the probability distribution directly from the cathode-ray tube without having to use a photometer to measure the light from the face of the cathode-ray tube. A detailed analysis of this technique with reference to the block diagram in Fig. 1 follows. (For a detailed description of the equipment see ref. 8.)

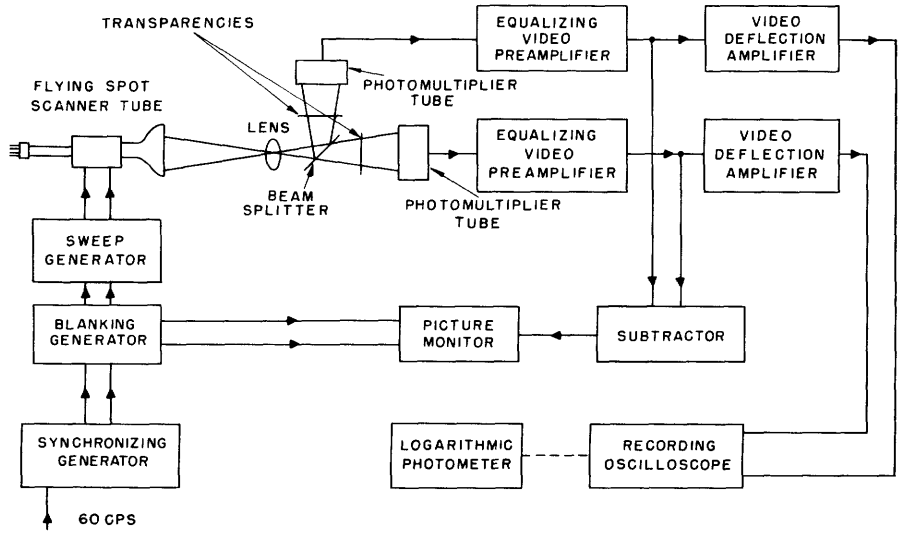

Fig. 1. Block diagram for the measurement of second-order probabilities.

The outputs of the synchronizing generator are 60-cps and 15,750-cps pulses. They key the blanking generator that produces the blanking and synchronizing pulses delayed by the proper amount. The synchronizing pulses initiate the sweeps of the sweep generator, which drives the deflection coils of the flying-spot scanner tube. This forms a raster of 525 lines on the face of the tube. The light emitted by this raster is focused by the lens on the two transparencies, simultaneously, by means of the beam splitter. The beam splitter consists of a membrane coated with a thin layer of aluminum, so that part of the impinging light is reflected and the rest transmitted. The transparencies are mounted in holders so that one can be positioned with respect to the other. In this manner, the transparencies can be scanned exactly, in register or out of register. That is, when cell i, j is being scanned in one picture, then either cell i, j or cell $(i + \Delta i, j + \Delta j)$ in the other can be scanned at the same time.

The light signals generated as a consequence of the scanning are converted to electric signals by the photomultipliers. The video signals are amplified in the video preamplifiers and deflection amplifiers, and applied to the horizontal and vertical deflection plates of a recording scope. If the two signals are identical, the resultant pattern will be a 45° line, assuming equal gain and phase-shift through the amplifiers. To get an accurate indication for this condition, the outputs of the preamplifiers are diverted to a subtractor and the difference signal is applied to a picture monitor. If the two transparencies are in register, a complete null will be observed on the monitor's cathode-ray tube. If the pictures are not in register, a few outlines will be seen, indicating the parts of the pictures that are not being scanned simultaneously.

One of the transparencies is then shifted out of register with the other by a one-cell displacement. The 45° line on the recording scope then spreads out because the two video signals are no longer identical. A typical brightness pattern is shown in Fig. 2. If one were to examine the brightness of any cell i, j, he would find it proportional to the joint probability that one cell in the picture is at intensity i and the succeeding cell at

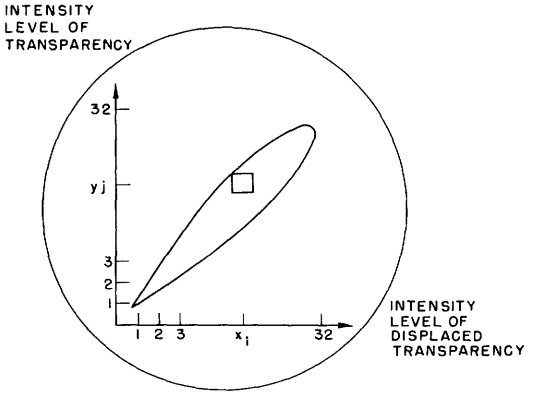

Fig. 2. Typical brightness pattern as seen on recording oscilloscope.

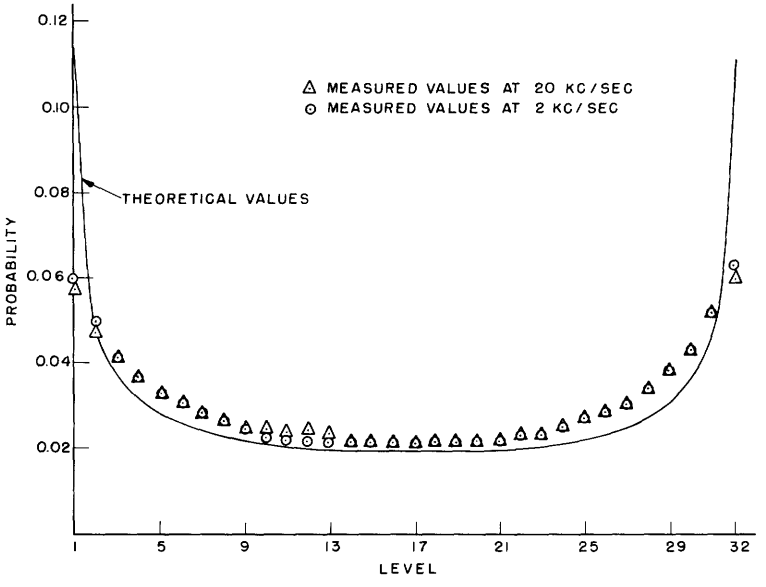

Fig. 3. Measured probabilities of sinusoidal waves for 32 levels from -1 to +1.

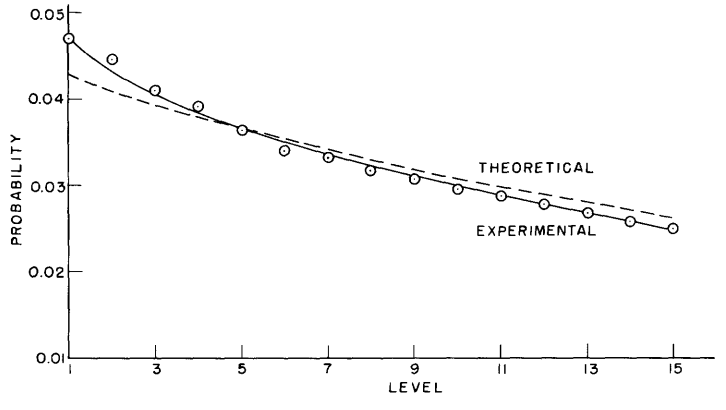

Fig. 4. Measured probabilities of exponential waves for 15 levels: frequency, 300 cps; R, 0.47 megohms; C, 7000 μμf.

intensity level j. This can be seen as follows. The dwell time of the beam on square i, j is an indication of how long the picture is at level i at the same time that its displaced version is at level j. Since the displaced version is only one cell away from the original picture, the dwell time is proportional to the joint probability p(i, j). The dwell time, though, determines the brightness at any particular point on the scope, so that the brightness of any cell i, j is in turn proportional to p(i, j). The brightness is measured by a photometer, as shown. Since 32 levels of intensity are assumed, 1024 probabilities will be measured.

The equipment is quite versatile in that probabilities of cells displaced by more than one unit shift can be measured easily. This can be done by merely shifting one transparency by the required amount. The measurement need not be limited to pictures. The equipment will measure the joint probability of any two signals whose highest frequency component is less than 10 Mc/sec, and whose lowest is greater than 30 cps (corresponding to the bandpass of the video amplifiers).

To check the calibration of the photometer, as well as the video amplifiers and recording oscilloscope, known probabilities of electrical waves were measured. These included both sinusoidal and exponential waves. The derivation of the theoretical probabilities is given in Appendix I. Fairly close agreement between the theoretical and measured values is obtained, as Figs. 3 and 4 show. Four different repetition rates were used in the exponential-wave test: 300 cps, 1 kc/sec, 10 kc/sec, and 60 kc/sec. Only the results for 300 cps are shown, since they were typical of the curves obtained at the other frequencies.

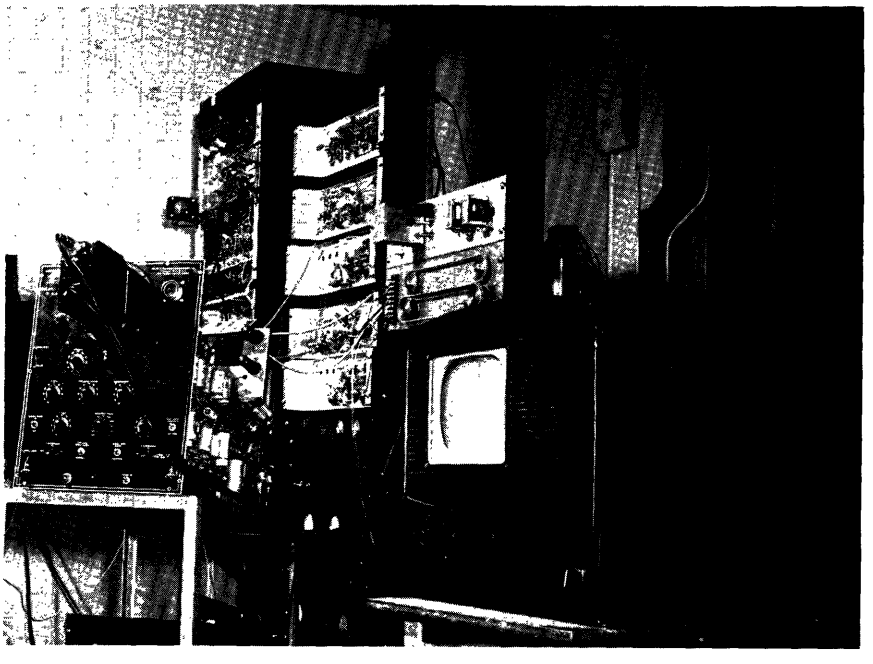

Fig. 5a. Electronic components.

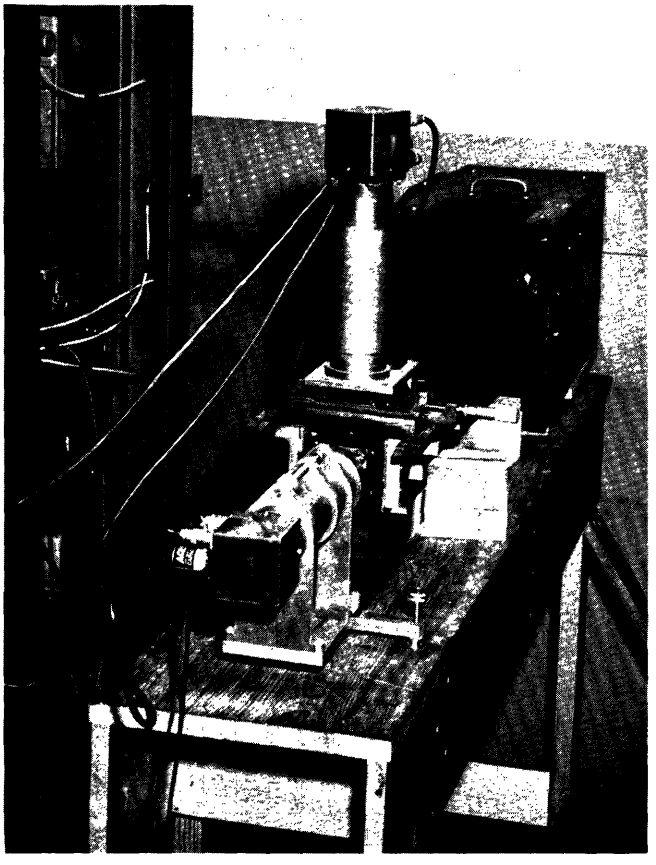

Fig. 5b. Optical system.

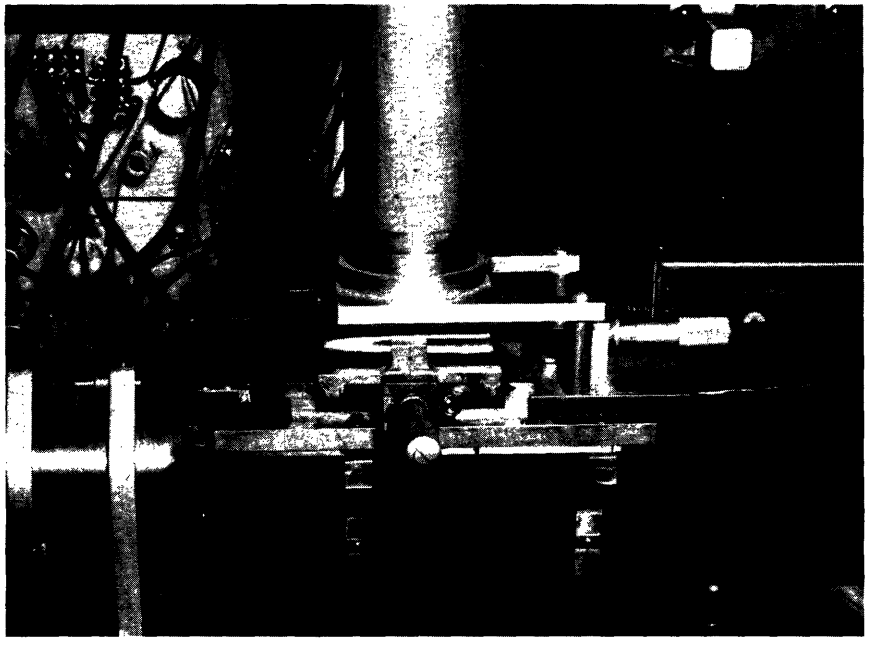

Fig. 5c. Close-up view of positioning controls.

The equipment is shown in Fig. 5. Figures 5(a) and 5(b) show the electronic components and optical components. Figure 5(c) is a close-up of the transparency positioning controls.

2. BOUNDING THE ENTROPY

The probabilities measured are not the true probabilities of the video signal but the probability of the signal plus noise. The main source of the noise is the original light signal itself. This is because light travels in the form of discrete photons and generates a signal that is quite noisy. The interesting point about this noise is that it is not statistically independent of the signal. The noise and signal are linearly independent (uncorrelated), but the variance of the noise is a function of the signal. That is, the noise variance at all instants of time is different, depending on the value of the light intensity at that instant. Another source of noise is the result of thermal emission from the photocathode of the photomultiplier, known as "dark" current. This is, however, a minor source of noise. Noise is also introduced in the photomultiplier tube because the electron stream is not continuous but flows in discrete steps. This is the familiar shot-noise effect found in the temperature-limited diode. It is also only a minor source of noise.

At first one might be inclined to say that the noises in the two channels were not statistically independent. But this is not true. Any given photon in the original light does not know whether the beam splitter will reflect it or transmit it. Hence, the noise in one channel in no way depends on the noise in the other, and they are thus statistically independent for any given light intensity.

Since the occurrence of noise poses a problem in that it masks the probability of the signal, some method must be used to extract the true probability. Several avenues of approach are available. Among these is the possibility of obtaining the signal probabilities from the measurements. The method for doing this is given in Appendix II. It is shown that given $p(S_1+N_1; S_2+N_2)$ and $p(N_1; N_2)$, it is possible to get $p(S_1; S_2)$, with the assumption that the signal is independent of the noise and that the noises are independent. The last assumption we know to be true, but the first is not true. However, the simplifications obtained by making this assumption are justifiable in view of the small error involved because of the slight correlation of noise and signal. This method, however, is very vulnerable to experimental error as we will show later. If we let $F_{(S+N)}(u)$, $F_S(u)$, and $F_N(u)$ be the characteristic functions of the signal plus noise, the signal, and the noise respectively, then (see ref. 9, Chap. III, p. 19)

$$F_{(S+N)}(u) = F_S(u)\, F_N(u)$$

The noise is essentially "white" with a gaussian probability density. Thus $F_N(u)$ is of the form $\exp(-u^2)$. In terms of circuit theory, this means that the signal probability has been passed through a filter whose characteristic is gaussian. To get the signal back it must be passed through a filter whose characteristic is of the form

$\exp(u^2)$, where u corresponds to ω. If there are any errors in $F_{(S+N)}(u)$, they will be enlarged by $\exp(u^2)$. The most serious errors will occur at the tails of the function where u is largest. The small probabilities will then have very little meaning. For this reason, this method is only stated; not used.

It was decided to obtain two bounds for the upper bound to the entropy per symbol. One will be an upper and the other a lower bound. It is shown in Appendix III that the upper bound to the entropy per symbol $H(S_2|S_1)$ can be bounded by

$$H(S_2 + N_2|S_1 + N_1) \geq H(S_2|S_1) \geq H(S_2 + N_2|S_1 + N_1) - H(N_2) - H(N_1)$$

if the signal and noise are statistically independent. It is also shown that if the signal and noise are not statistically independent, the lower bound will not be affected, in that the inequality sign is just made stronger. For the upper bound, some dependence of noise on signal can be allowed. However, the inequality sign is weaker.

Schreiber (see ref. 4, pp. 36-37) used a different method for evaluating the signal entropy from the signal-plus-noise entropy. In his analysis he assumes signal and noise to be statistically independent, and also assumes the signal to have a gaussian probability distribution. The first assumption, as we have seen, is reasonable, but the second may or may not be, depending upon what the distribution function for any particular picture looks like. However, his last assumption is not used in this analysis.

3. THE OPTICAL CORRELATOR

The device shown schematically in Fig. 6 is used to obtain the autocorrelation function (5) of pictures. The light source is placed at the focal point of the first condensing lens, and thus all the light that is transmitted by the lens is composed of parallel rays. These rays pass through the glass slides and are collimated again by the second condensing lens onto the photoelectric cell.

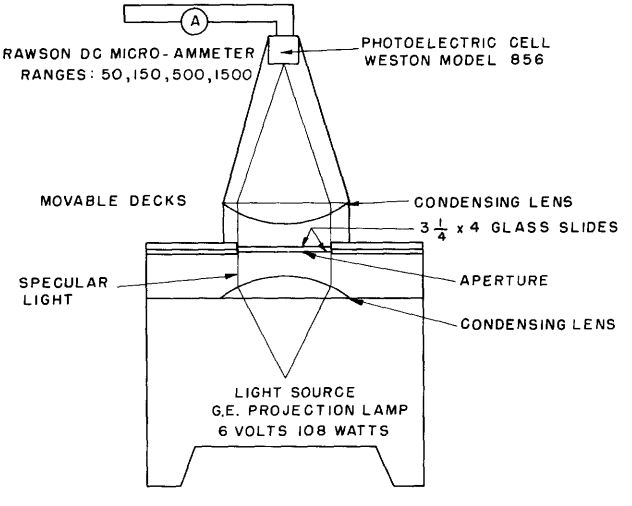

Fig. 6. Schematic representation of optical correlator.

To see how the autocorrelation function is obtained, one proceeds as follows. The autocorrelation function of a function of time f(t) is defined as

$$\phi_{11}(\tau) = \overline{f(t)\, f(t+\tau)}$$

where the bar indicates an average taken over all time for various values of the time shift τ. For a picture transparency the optical transmission is a function of the vector $\vec{r}$, where $\vec{r} = r \angle\theta$. Hence, when one shines light through two pictures that have a relative shift $\Delta\vec{r}$, one is measuring

$$\phi_{11}(\Delta\vec{r}) = \overline{T(\vec{r})\, T(\vec{r} + \Delta\vec{r})}$$

where $\Delta\vec{r} = \Delta r \angle\phi$; and $T(\vec{r})$ is the optical transmission function of picture transparency.

The averaging is the result of this inherent property of the photoelectric cell. The expression given above is seen to be the autocorrelation function of the picture evaluated for a vector shift $\Delta\vec{r}$.

The apparatus is not limited to the measurement of autocorrelation functions but may be used to measure crosscorrelation functions by merely replacing one of the two identical slides by a different one. This is quite useful in the determination of the

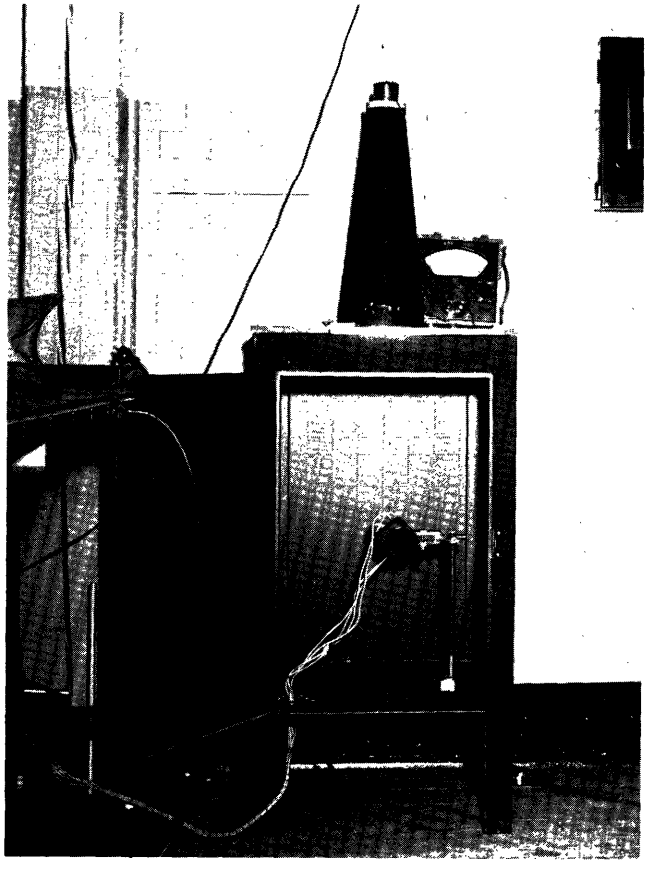

Fig. 7. Optical correlator.

correlation between successive frames in either television or motion pictures.

A picture of the correlator is shown in Fig. 7.

4. THE AUTOCORRELATION FUNCTION BOUND TO THE ENTROPY PER SYMBOL

It is shown in Appendix IV that the normal random process with the second moments a_{ij} has the maximum entropy of all processes with the same set of second moments. Thus, by assuming a process is gaussian, one can get an upper bound to the entropy per symbol.

The gaussian process is completely specified if its autocorrelation function is known. The entropy of the process can thus be expressed in terms of the autocorrelation function. The general expression for the entropy per symbol of the multivariate normal distribution in terms of its correlation coefficients is derived in Appendix V. The bound obtained from a knowledge of the preceding sample is shown to be

$$H(X_2|X_1) = \frac{1}{2} \log 2\pi e\sigma^2 + \frac{1}{2} \log \left[1 - \left(\frac{\rho_{12}}{\sigma^2}\right)^2\right]$$

This bound may or may not lie between the bounds obtained by the other method, depending on the extent to which the noise affects the results. If it lies between the aforementioned bounds, it will give a tight bound to the entropy per symbol. If it does not, it will be a measure of how different the process is from a gaussian process.

An upper bound can be obtained by the use of an indefinite number of previous samples, as well as from one previous sample. The formula in this case is (see Appendix V):

$$H(X_N|X_{N-1}, \ldots, X_1) = \frac{1}{2} \log 2\pi e \frac{\left|\rho_{ij}^N\right|}{\left|\rho_{ij}^{N-1}\right|}$$

The bound for $N > 3$ may conceivably be smaller than both of the bounds obtained by the probability measurement. It will certainly be less than the correlation bound obtained for $N = 2$.

5. STANDARDIZATION OF TRANSPARENCIES

It is of the utmost importance that all the transparencies used in the measurements be alike. At first this may seem like a trivial problem. However, anyone who is familiar with photographic processes will acknowledge that it is difficult to make pictures that are alike. They may look the same to the naked eye, but compared on a densitometer they are very different. The differences are caused by several factors: differences in photographic emulsion, differences in exposure time, and differences in development time. Any differences in the development process, such as the temperature of the solution and the amount of solution used, will give rise to different transparencies.

The following procedure was used to determine whether or not the transparencies were similar. A diffuse density step tablet was printed on the 35-mm and $4 \times 3\ 1/4$ glass slides. This tablet consisted of a negative whose optical transmittance varied in steps along its length. We measured the transmittance of each step on each transparency and determined whether they were the same or differed by a constant multiplier. Only transparencies which satisfied these conditions were used in the measurements. This amounted to making certain that the gamma (see ref. 10, pp. 195-201) of the pictures was the same.

The quantity γ is defined as

$$\gamma = \frac{\partial D}{\partial(\log_{10} E)}$$

where D is the density, the ratio of incident to reflected light, and E is the exposure of any given point on film.

If γ is unity, a picture which is directly proportional to the original has been made. However, if γ is not unity, then one has a picture that corresponds to the original signal raised to the power γ. Thus it is seen that unity gamma, or a gamma which is the same for all transparencies, is essential.

IV. COMPUTATIONS AND RESULTS

1. THE JOINT PROBABILITY MEASUREMENT

Since it is of the utmost importance that the gamma be unity, or the same for both transparencies, the first step in the experimental procedure is to measure the gamma and make certain that either one of these conditions holds. This is done by masking the transparency except for the step tablet. The pattern is then scanned vertically. That is to say, the step tablet is scanned from its densest step to its lightest step as the flying-spot scanner beam travels downward. The resultant video waveform is then viewed on an oscilloscope, and the steps in intensity are observed. When these steps are the same or differ by a constant multiplier for a pair of transparencies, one can be sure that the transparencies are suitable for the measurement.

The masking is then removed from the picture portion of the glass slide and the step wedge is so masked that it does not interfere with the measurements. The transparencies are put into their holders and positioned until they are in register. A sensitive indication of this condition is a complete void on the picture monitor. The transparencies are then shifted from register by a one-cell displacement. Since for equal vertical and horizontal resolutions there are 700 cells in a horizontal line, and since the picture is 1 1/3-inches wide, one slide is shifted 0.00190 inch past the other. The brightness pattern on the face of the recording oscilloscope is divided into 1024 cells, and the brightness of each cell is measured. This measurement is repeated three times and the results are averaged. The data obtained with the photometer are in arbitrary units and are linearly related to the logarithm of the corresponding probability. To convert this reading to actual brightness, we use the equation

$$D = 10^{3(1-R)}$$

where D are the data proportional to brightness and thus to probability, and R is the reading taken with the photometer.

This conversion equation is a consequence of the manner in which the photometer was calibrated.

The number recorded in the i^{th} row and j^{th} column is called D(i; j) and is proportional to the corresponding probability, p(i; j). The calculation of the conditional entropy from these values is performed as follows: since

$$\sum_{i,j} p(i;j) = 1$$

$$p(i;j) = \frac{D(i;j)}{\sum_{i,j} D(i;j)} = \frac{D(i;j)}{k}$$

The conditional entropy is

$$H(Y|X) = -\sum_{i,j} p(i;j) \log \frac{p(i;j)}{p(i)}$$

where

$$p(i) = \sum_{j} p(i;j)$$

$$= \sum_{j} \frac{D(i;j)}{k}$$

$$H(Y|X) = -\sum_{i,j} p(i;j) \log p(i;j) + \sum_{i} p(i) \log p(i)$$

The complete procedure can be summarized as follows:

1. Add each column to obtain a column sum, and then add the column sums; this total sum yields the constant k.
2. Divide each D(i; j) by k to obtain p(i; j).
3. Divide each column sum by k to obtain p(i).
4. From an entropy table obtain (-p log p) for each p(i; j) and p(i) (see ref. 11).
5. Sum all terms of the form -p(i; j) log p(i; j).
6. Sum all terms of the form -p(i) log p(i).
7. Subtract 6 from 5 to obtain the conditional entropy.

As we mentioned previously, this conditional entropy is not the true entropy $H(S_2|S_1)$, but is $H(S_2 + N_2|S_1 + N_1)$. It is proved in Appendix III that the latter entropy forms an upper bound to the former entropy. To obtain the lower bound, the entropies of the noises are required. Since the actual noise entropy is difficult to obtain, one must calculate an entropy that is always greater than the noise entropy. The procedure is as follows: The variance of the noise is

$$\sigma^2 = \overline{(N-\bar{N})^2} = \int_N \int_S (N-\bar{N})^2 \, p(N;S) \, dN dS$$

$$= \int_S p(S) dS \int_N (N-\bar{N})^2 \, p(N|S) \, dN$$

The second integral given above represents the noise variance for a given value of signal and is designated by σ_S^2.

Since the instantaneous noise power is directly proportional to the instantaneous

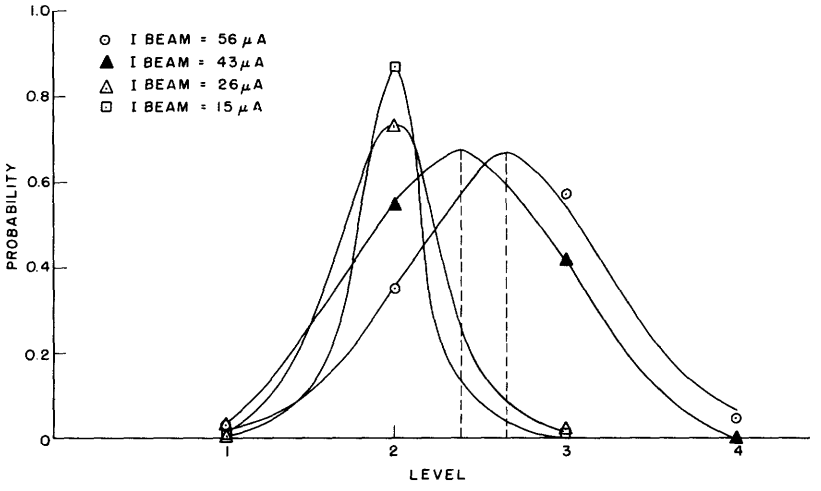

Fig. 8. Conditional noise probabilities (dashed lines designate value of mean).

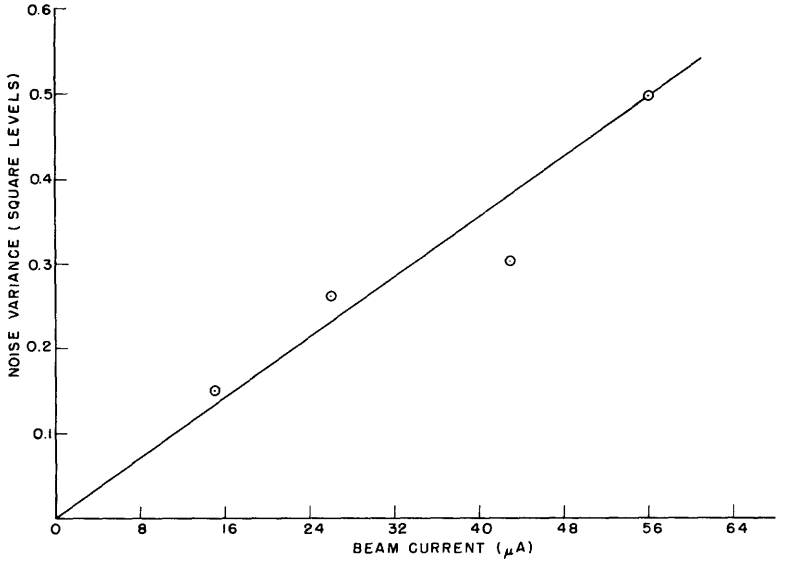

Fig. 9. Noise variance versus beam current.

signal, σ_S^2 will be proportional to the instantaneous signal. Thus

$$\sigma^2 = \int_S \sigma_S^2 \, p(S)dS = k \int_S Sp(S)dS$$

$$= k\bar{S}$$

Hence, the variance of the noise is the variance that will occur at the mean of the signal. Now that the variance is known, an upper limit to the noise entropy can be set by assuming that its probability density is gaussian (see Appendix IV). The entropy of a first-order gaussian process is (see Appendix V)

$$H = \frac{1}{2} \log 2\pi e \sigma^2$$

Thus if σ^2 is known, the entropy can be calculated.

To check the fact that noise variance goes down with signal, and thus with beam current, probabilities of noise for various beam currents were measured, and the variances computed were plotted against the beam current. The results are shown in Figs. 8 and 9. Figure 9 shows that a linear trend appears.

The smooth curves through the discrete probabilities of Fig. 8 have no real meaning. They were drawn to look gaussian with the same variance as the discrete probabilities and centered on the same mean. The entropies of the discrete and continuous distributions were calculated, and it was found that they differed at most by 5 per cent. Thus, very little error is incurred by assuming the noise is continuous. This is undoubtedly due to the fact that entropy is a slow function of the shape of probability distributions and depends more on variance. As a matter of fact, it is likely that very little error in entropy is incurred by assuming the noise has a gaussian distribution for the same reason.

In the lower bound

$$H(S_2|S_1) \geqslant H(S_2 + N_2|S_1 + N_1) - H(N_1) - H(N_2)$$

$$H(N_1) = H(N_2)$$

since the noises have the same statistics. The lower bound becomes

$$H(S_2|S_1) \geqslant H(S_2 + N_2|S_1 + N_1) - \log 2\pi e \sigma^2$$

The complete procedure follows.

1. From the previously computed values for p(i), determine the mean value of i, denoted by $\bar{i}$.

2. With this mean value of signal, and thus beam current, pick off the variance of the noise from the graph of Fig. 9.

3. Insert this quantity in the inequality given above to derive the lower bound to the entropy.

Fig. 10a. Subject No. 1.

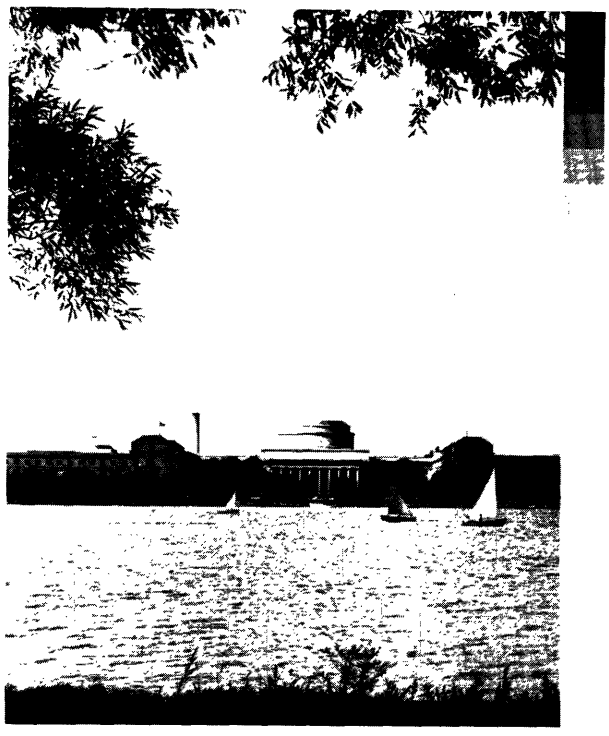

Fig. 10b. Subject No. 2.

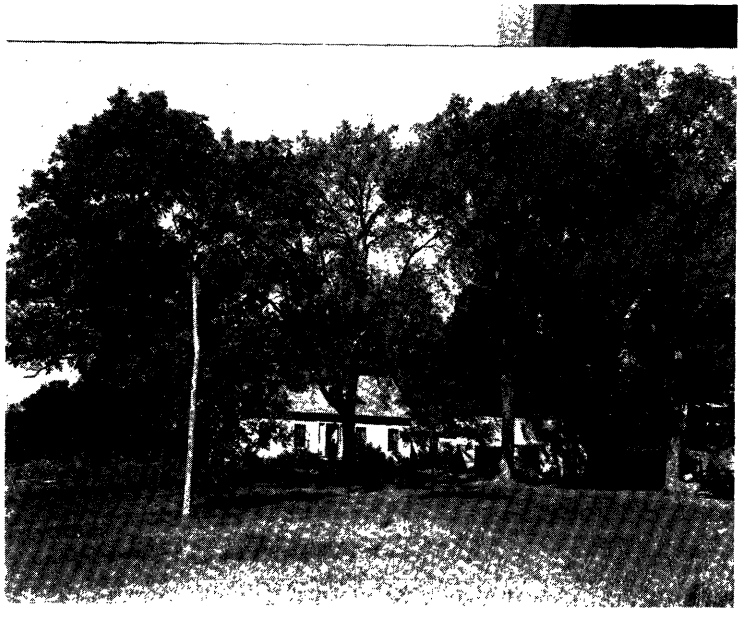

Fig. 10c. Subject No. 3.

Fig. 10d. Subject No. 4.

Fig. 10e. Subject No. 5.

The experimental and computational procedure is now complete. The results are given in Table II. The five subjects mentioned in this table are shown in Figs. 10(a)-10(e); their data sheets are given in Figs. 11(a)-11(e).

Table II.

Computed Upper and Lower Bounds to Entropy per Symbol

Subject	Upper Bound (bits)	Lower Bound (bits)
1	1.365	0.6725
2	2.075	1.450
3	0.982	0.357
4	1.696	0.976
5	0.758	0.713

2. THE AUTOCORRELATION FUNCTION MEASUREMENT

In this measurement, not only is it essential that the glass slides have the same gamma, but that they have the same gamma as the 35-mm transparencies of the previous measurement. The gamma of the 3 1/4 × 4 transparencies is measured by masking everything but the step tablet and measuring the light transmitted by each step. When two slides have the same set of readings as obtained previously for the 35-mm transparencies, they are suitable for measurement. It is significant to note that the gamma is measured in a manner that is different from the previous measurement. This is as it should be, since what is essential is to reproduce the actual conditions under which the experiment is performed.

It is important that the area of the aperture remain constant as the slides are shifted relative to each other. Since a total shift of 0.5 inch is to be used, a 0.25-inch strip along each edge of the aperture is masked off, so that as one transparency is moved the aperture area will stay the same. The two slides are then put in their holders and positioned in register. This condition is indicated by a maximum of transmitted light and thus a maximum in the meter reading. One slide is shifted horizontally by an amount ΔS. The reading corresponding to this is the autocorrelation function for a shift of ΔS. This is repeated several times until a relative shift of 0.5 inch is obtained. The procedure is repeated for vertical shifts. A contour of constant autocorrelation is then obtained at a convenient value between zero and a 0.5-inch shift.

Since it is the correlation coefficient that is of interest, it must be calculated from the measurements. It is defined as

$$\rho(\Delta\vec{r}) = \phi_{11}(\Delta\vec{r}) - \overline{x_1} \cdot \overline{x_2}$$

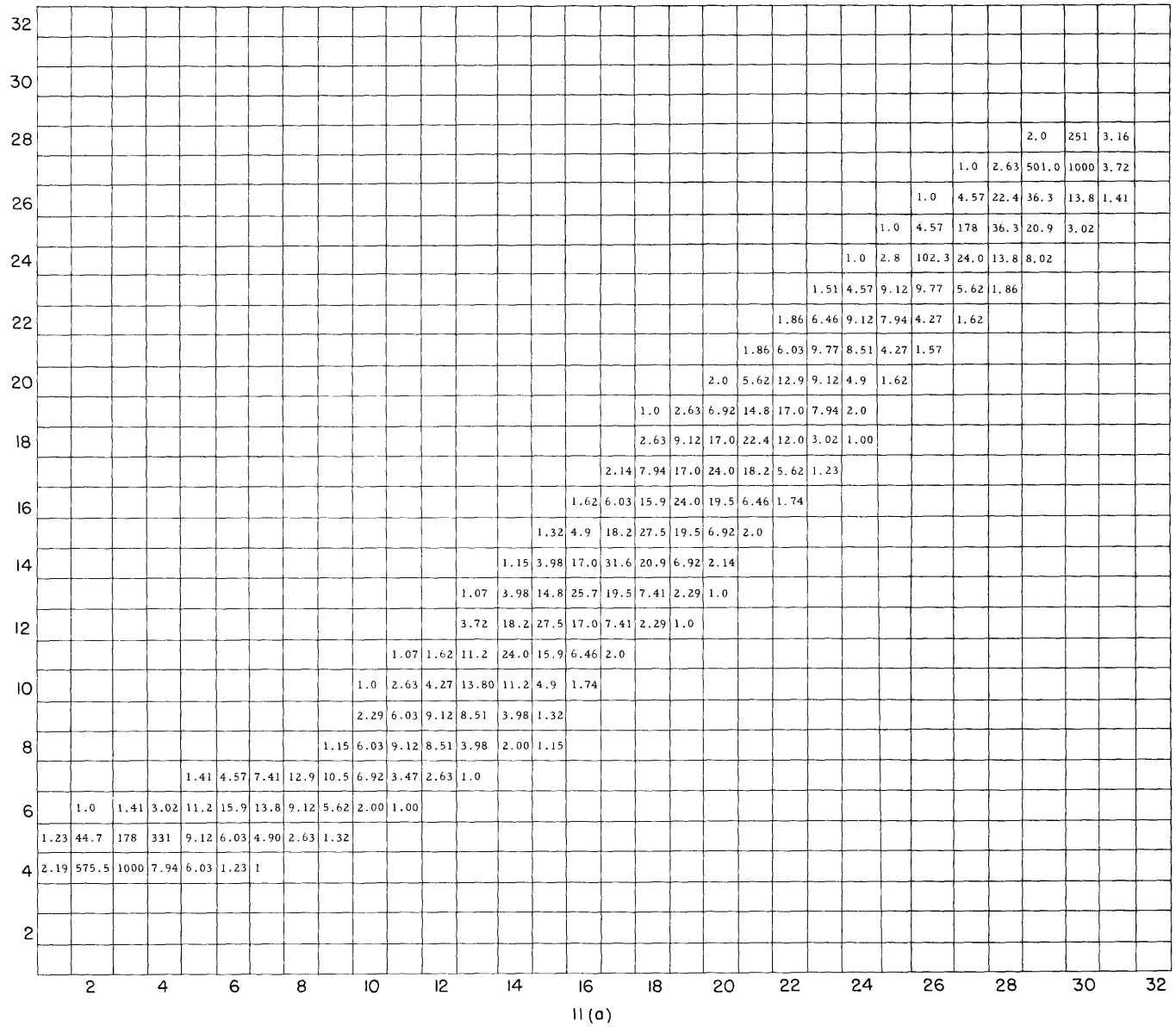

Fig. 11a. Data sheet for subject No. 1 (vertical shift, 0.0019 inch).

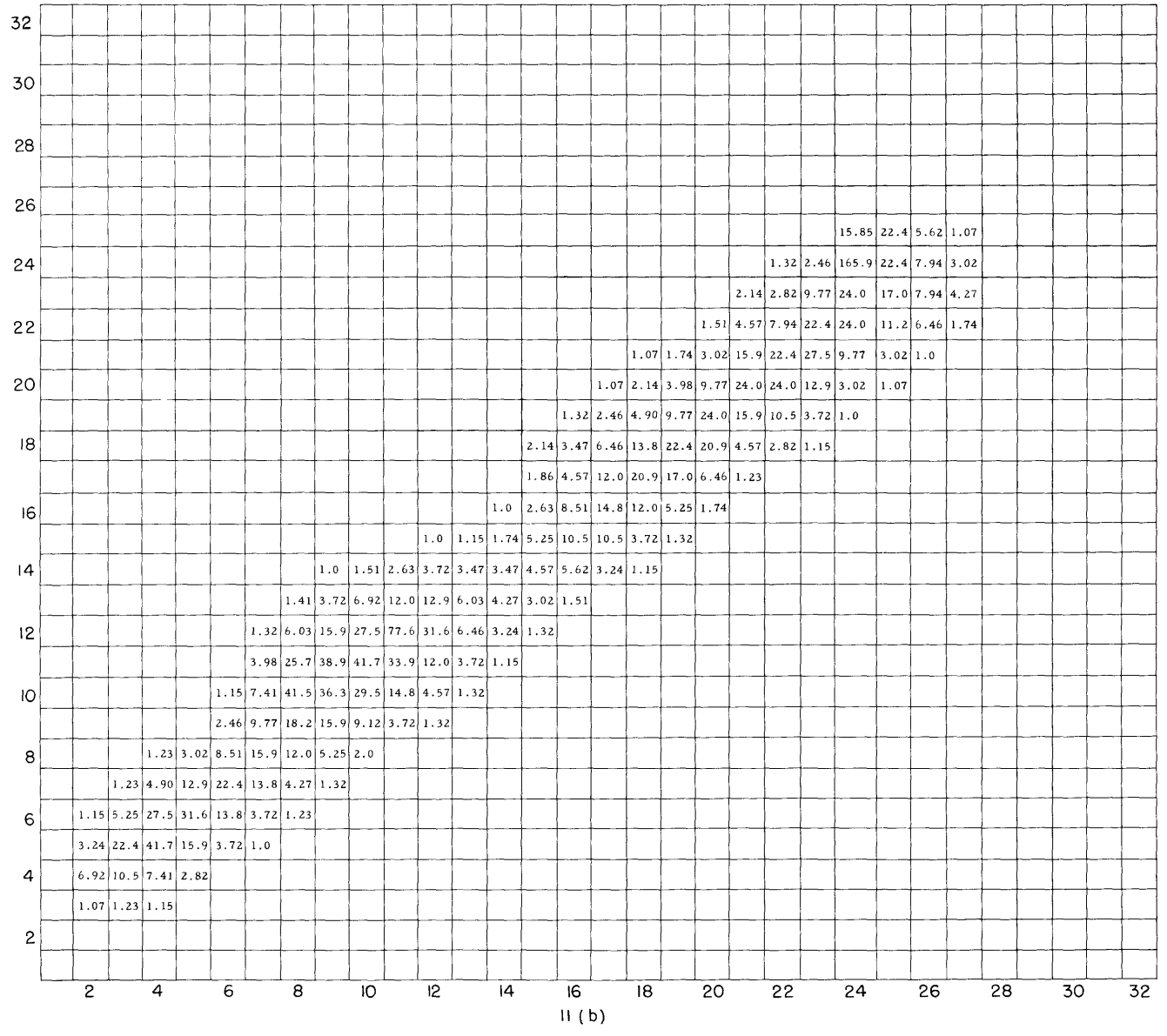

Fig. 11b. Data sheet for subject 2 (vertical shift, 0.0019 inch).

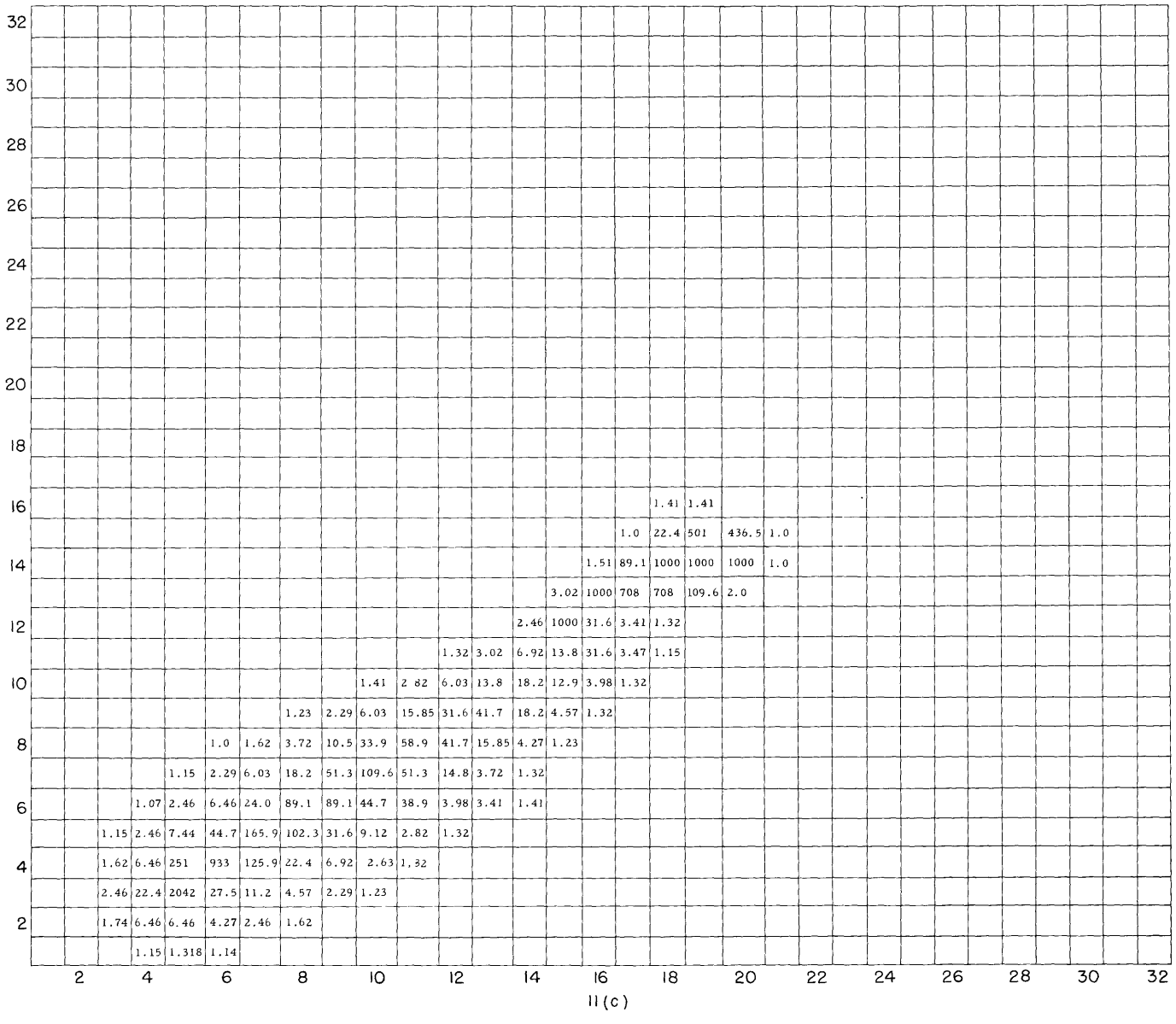

Fig. 11c. Data sheet for subject No. 3 (horizontal shift, 0.0019 inch).

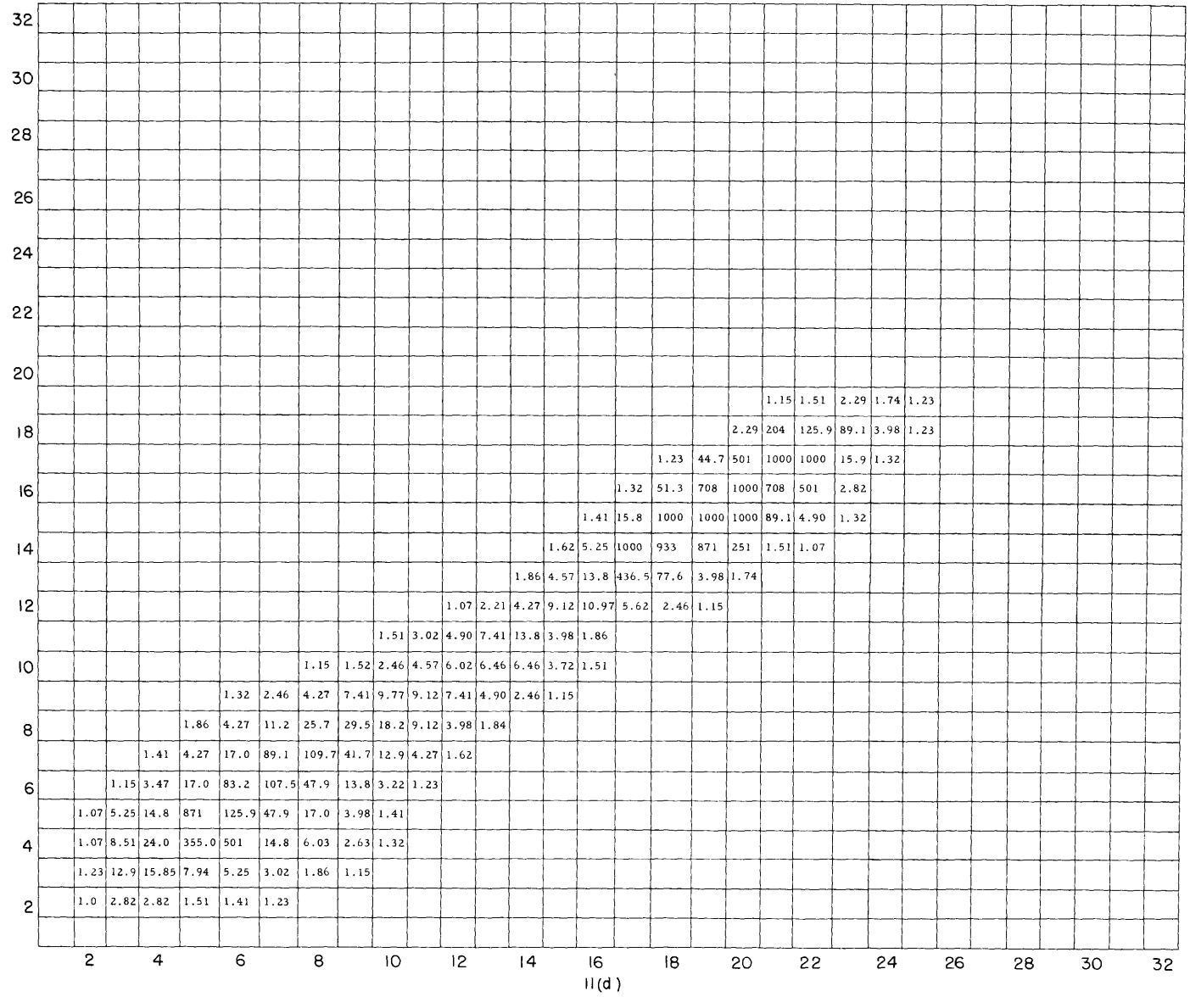

Fig. 11d. Data sheet for subject No. 4 (horizontal shift, 0.0019 inch).

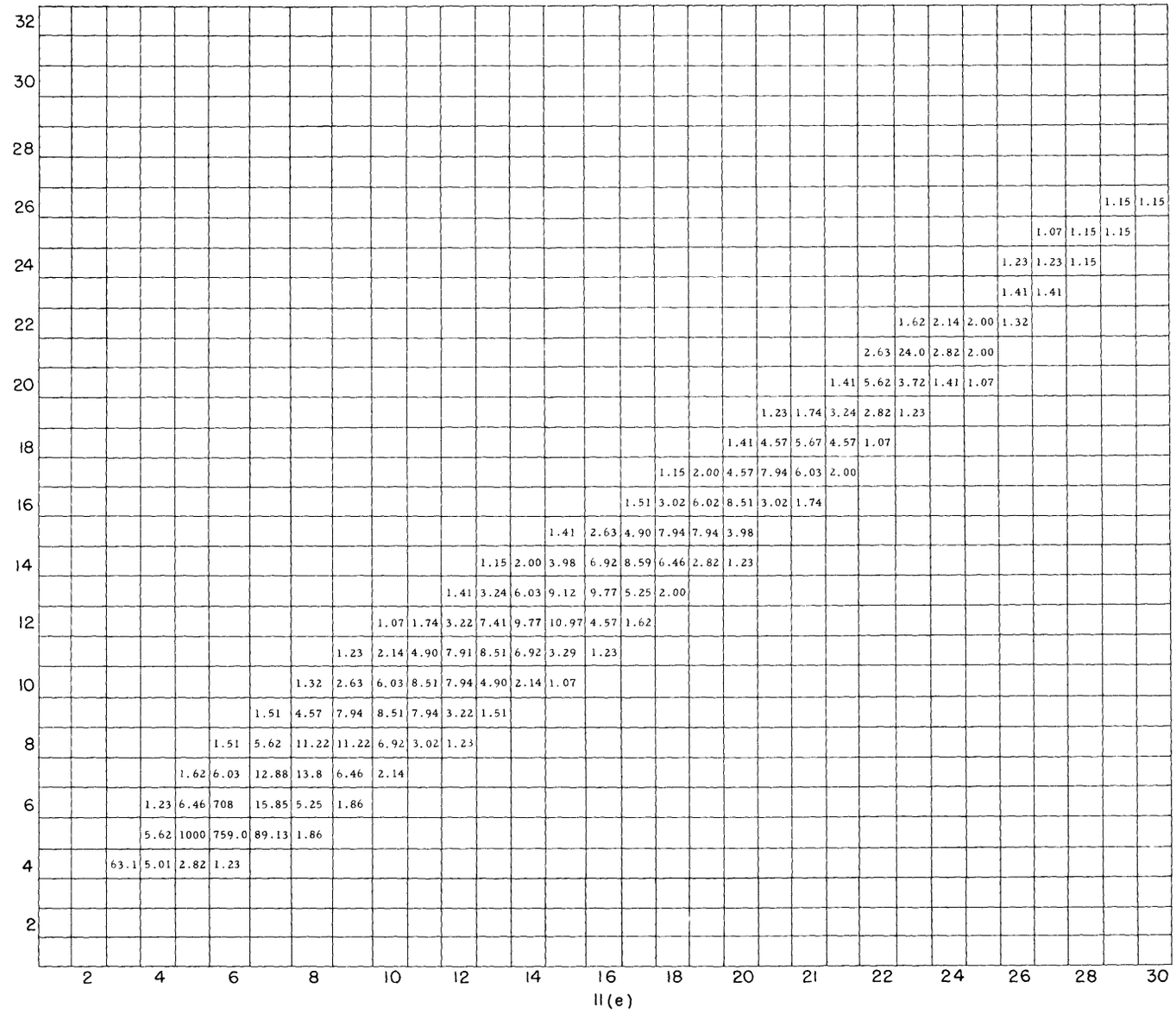

Fig. 11e. Data sheet for subject No. 5 (horizontal shift, 0.0019 inch).

where $\overline{x_1}$ is the mean transmittance of the first transparency and $\overline{x_2}$ is the mean transmittance of the second transparency.

As the slides are shifted, substantial amounts of new picture material are introduced into the aperture. Thus the mean transmittance of each transparency will not be constant. Additional measurements have to be made to determine this "floating" mean. This is done as follows:

$$\rho(\Delta\vec{r}) = kT_2(\Delta\vec{r}) - \frac{T_1(\Delta\vec{r})T_1(0)}{T}$$

where T is the transmission with both slides removed, $T_2(\Delta\vec{r})$ is the transmission through the two cascaded slides shifted by $\Delta\vec{r}$, and $T_1(\Delta\vec{r})$ is the transmission of a single slide with the same displacement $\Delta\vec{r}$.

To obtain the normalized correlation coefficient, division by the variance is necessary. It is given by

$$\sigma^2 = k\left(T_2(0) - \frac{T_1^2(0)}{T}\right)$$

The normalized correlation coefficient is given by

$$\frac{\rho(\Delta\vec{r})}{\sigma^2} = \frac{TT_2(\Delta\vec{r}) - T_1(\Delta\vec{r})\,T_1(0)}{TT_2(0) - T_1^2(0)}$$

Figures 12 and 13 are plots of the normalized correlation coefficient versus $\Delta\vec{r}$ for horizontal and vertical shifts, respectively. Figure 14 shows contours of constant normalized correlation coefficient for the five subjects.

It is shown in Appendix V that the conditional entropy is

$$H(Y|X) = \frac{1}{2}\log 2\pi e\sigma^2 + \frac{1}{2}\log\left[1 - \left(\frac{\rho_{12}}{\sigma^2}\right)^2\right]$$

The quantity (ρ_{12}/σ^2) is the value of the normalized correlation coefficient for a unit shift in the horizontal direction. The only quantity which is not known is σ^2. It is found as follows:

$$\sigma^2 = \left(\frac{T_2(0)}{T} - \frac{T_1^2(0)}{T^2}\right)(32)^2$$

Thirty-two levels from black to white are assumed. This presupposes that the intensity range of all subjects is the same. This, however, is not a true assumption, since it was found that not all the transparencies had the same intensity range. It was assumed that Subject No. 1 covered the entire range from black to white, and the variance of the other transparencies was adjusted to take into account the difference in intensity ranges. In the continuous case the entropy can be considered a measure of randomness relative

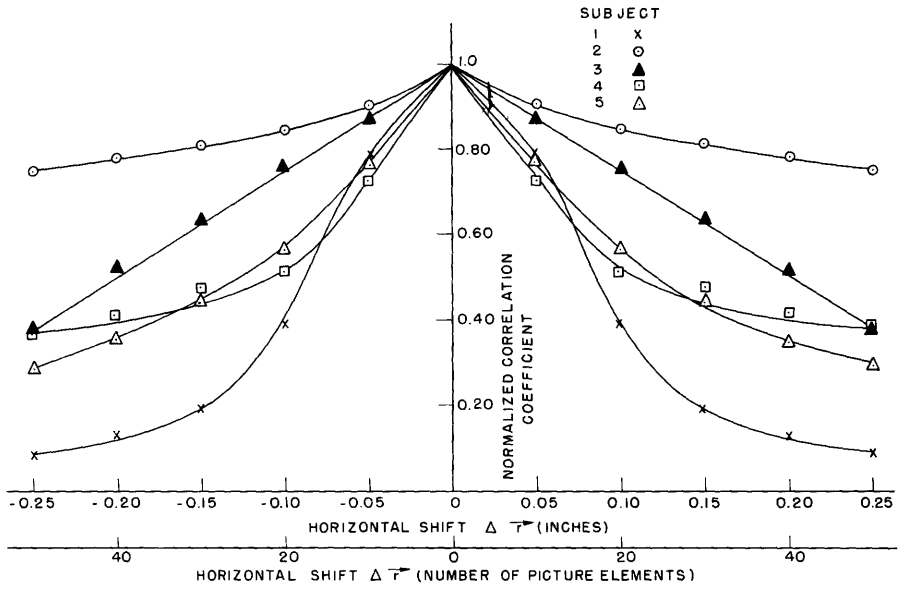

Fig. 12. Normalized correlation coefficient versus Δr for horizontal shifts.

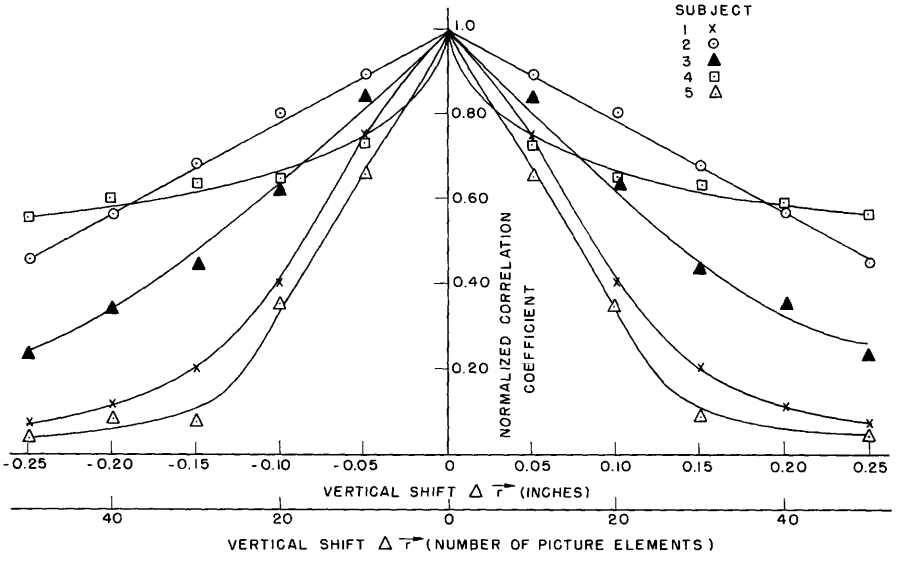

Fig. 13. Normalized correlation coefficient versus Δr for vertical shifts.

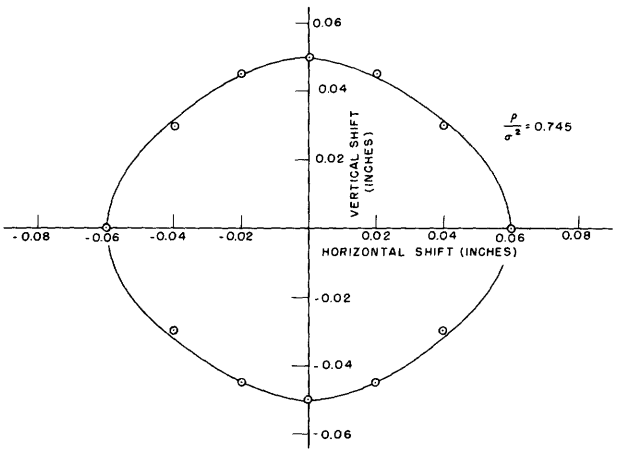

Fig. 14a. Isonormalized correlation coefficient contour for subject No. 1.

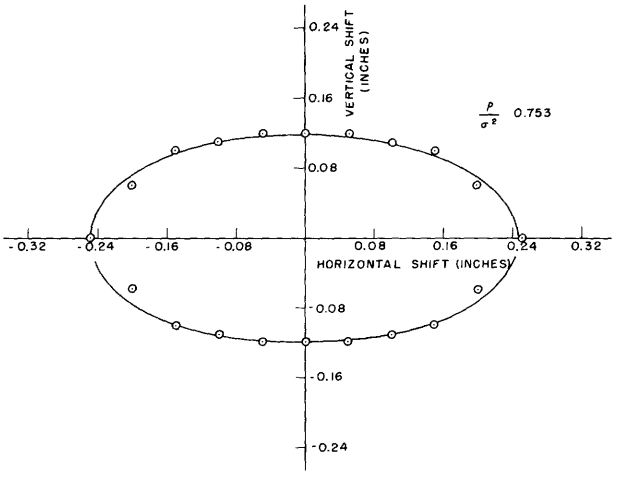

Fig. 14b. Isonormalized correlation coefficient contour for subject No. 2.

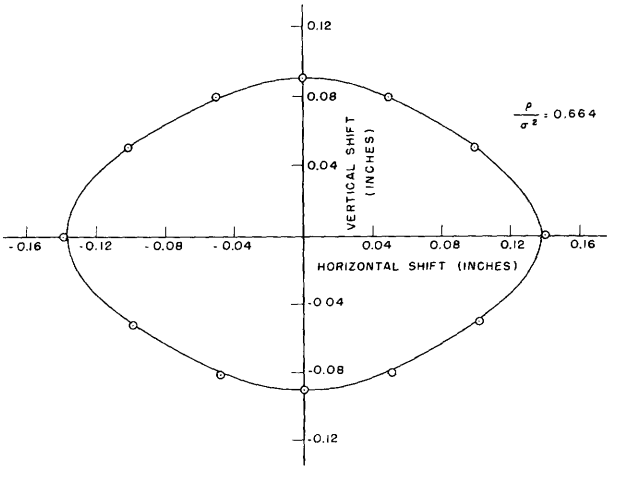

Fig. 14c. Isonormalized correlation coefficient contour for subject No. 3.

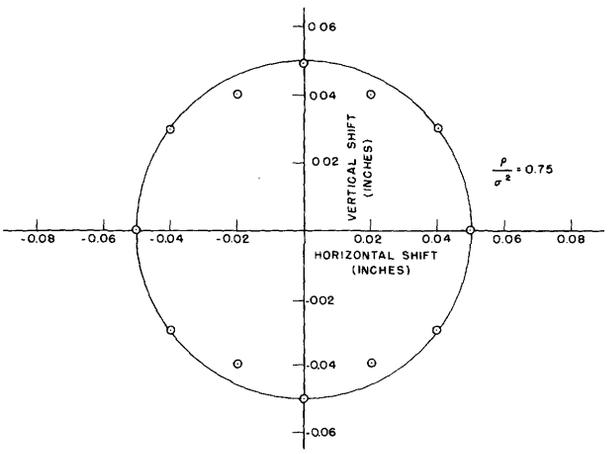

Fig. 14d. Isonormalized correlation coefficient contour for subject No. 4.

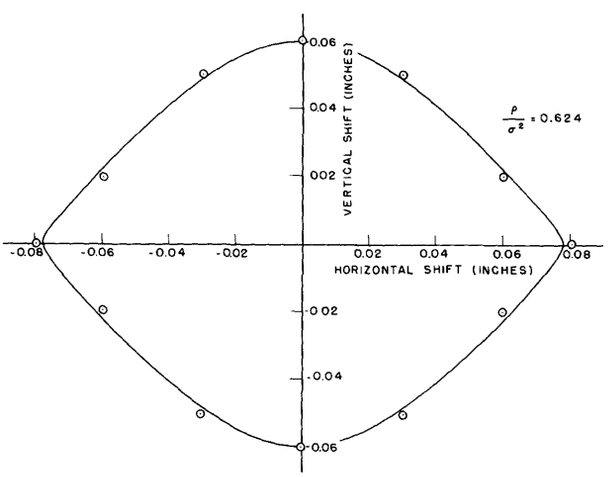

Fig. 14e. Isonormalized correlation coefficient contour for subject No. 5.

to an assumed standard only. Thus, to enable us to compare these results with those of the previous experiment, Subject No. 1 was taken as the standard.

This determines all quantities in the expression for $H(Y|X)$, and it may be computed. The results are given in Table III.

Table III.

Computed Autocorrelation Bound to Entropy per Symbol

Subject	Conditional Entropy (bits)
1	1.340
2	1.605
3	1.930
4	2.350
5	2.810

3. ACCURACY OF MEASUREMENTS

It is appropriate at this point to consider the various sources of error that are introduced in the experiment. The joint probability measurement will be considered first.

A primary source of error in this measurement is the phosphor of the flying-spot scanner tube. It has been assumed throughout that the luminosity is the same for all points on the phosphor. This, however, is not true, as we can see if a blank pattern is being scanned. The recording oscilloscope pattern in this case is oblong and elliptical in shape instead of being just a small circle. It is observed that the variation in luminosity of the phosphor is approximately 10 per cent. That is, the ratio of the difference of the maximum and minimum light to maximum light is 10 per cent. This is equivalent to a spread of three cells. This nonuniformity has the effect of changing the picture. The effect is not too bad, however, since on the monitor the picture looks very much like the original.

Lack of precision in reading brightness from the face of the recording oscilloscope is another source of error. Since the maximum error incurred by the photometer itself is only 5 per cent, it may be disregarded as a serious source of error. Random movement of the pattern on the recording oscilloscope also leads to errors. If the brightness of the recording oscilloscope were to change erroneously, an error would result.

The amount by which the pattern is positioned is determined by precision resistors that are accurate to 1 per cent. However, if for some reason the pattern were not positioned correctly, an error would result. Because of the small size of the aperture used in the photometer, any small shift in the pattern will lead to an appreciable error.

Finally, if the gammas of the transparencies were not the same, the results would be in error. In this case, one would be measuring the statistics of two pictures, one of which is the equivalent of the other raised to a power. The value of the exponent

would be proportional to the difference between the gamma of each transparency.

In the autocorrelation function measurement, the main source of error is the bulb. The light output of the bulb is not constant for long periods of time and has a tendency to drift. This is no doubt caused by variations in filament temperature. The photocell and microammeter do not introduce appreciable error when used properly. The photocell must be loaded by a constant resistance for the microammeter readings to have relative value. Thus, as the scale on the microammeter is changed, an external resistance is added to keep the load on the photocell constant.

V. ANALYSIS OF RESULTS

The joint probabilities measured are conditional in that they are measured for a particular picture. To get the actual joint probabilities, measurements must be made on an infinite number of pictures and the results averaged over this ensemble. The impracticability of this course is evident. What we actually do is to choose subjects having little to great detail. An average entropy is then obtained by assuming that each picture is representative of its set and thus occurs with probability one-fifth. This value is different from the entropy calculated from the averaged joint probabilities. The difference between the two will depend on how different the probabilities of the specific pictures are from the probabilities of the average picture. From the viewpoint of coding, the average entropy corresponds to coding all types of pictures, whereas the entropy of the average indicates a system in which the coding scheme is designed for average pictures. The choice between the two is determined only by the flexibility desired in the coding system.

From the probability measurement the average upper bound is 1.375 bits while the average lower bound is 0.834 bits. For present-day television, where equiprobable events are assumed, the entropy for a 32-level system would be log 32 or 5 bits. Thus, approximately 4 bits of redundancy are exposed. This means a potential reduction, by a factor of three to five, in the channel capacity required for television transmission. This reduction can be brought about only by means of an ideal coding scheme. As we mentioned previously, this reduction can be effected as a saving in either signal power or bandwidth, or both.

It is of interest to note that some of the more detailed pictures did not have a greater entropy than the less detailed pictures. In particular, Subjects No. 3 and No. 5 had less entropy than Subject No. 1. This is contrary to one's expectations, judging from a subjective estimate of the order of complexity of the pictures. However, one must also keep in mind the intensity range covered by each transparency. If there is great detail in a picture at low contrast, the picture contains little information. If a picture of a black dog in a coalbin at midnight is taken, there will be great detail but practically no information in the picture, as it will be nearly all black. However, if the same picture is taken with a flash attachment, then it will contain much information. We see how the information in a scene depends not only on the detail but the intensity range covered by the scene. This fact would have to be taken into account in the coding scheme used. Thus, one can either define an intensity range from white to black for all pictures and keep the gain constant, or vary the gain so that all pictures will cover the same equivalent intensity range. This second scheme is used in present-day television transmission. The black level is set at 75 per cent of the maximum amplitude of the carrier envelope. The gain is then adjusted so that the white portion of the picture has an amplitude which is approximately 15 per cent of the maximum amplitude of the carrier envelope. In the final analysis, the appearance of the picture is determined by how the

viewer adjusts his contrast (gain) control in his home receiver. In the experiment, however, the gain was kept constant, since the change in gain actually changes the picture. It is not surprising, then, that the entropy of some of the more detailed pictures was less than the entropy of less detailed pictures. This entropy would correspond to a coding scheme in which the intensity range for all pictures is not the same, but varies from picture to picture.

Table IV.

Corrected Joint Probability Estimates Compared with Autocorrelation Function Estimates

Subject	Upper Bound (bits)	Lower Bound (bits)	Autocorrelation Bound (bits)
1	1.415	0.696	1.340
2	2.205	1.540	1.605
3	1.170	0.426	1.930
4	1.910	1.101	2.350
5	0.855	0.735	2.810

In order to compare these results with those obtained in the autocorrelation function measurement, they must be corrected for not having the same intensity range. This is done by multiplying each entropy by log 32/log N, where N is the actual number of levels in the picture. The results are listed in Table IV and compared with the autocorrelation bound.

The average upper bound is now 1.511 bits, the average lower bound is 0.899 bits, and the average autocorrelation bound is 2.007 bits. The autocorrelation function measurement yields an average upper bound which is greater than the one obtained by the other method. This is what was expected.

The results obtained are in close agreement with those obtained by Schreiber. His calculations show an average entropy of 0.87 bits, which is in agreement with the results reported here.

The plots of normalized correlation coefficient versus shift are symmetrical only because they were drawn that way. In actuality they are not. The results for positive and negative shifts are averaged, and these average results plotted versus shift. The same procedure is followed in the plots of the isonormalized correlation-coefficient contours, in order to make them symmetrical. It is essential that the autocorrelation function by symmetrical since its mathematical definition requires this condition. The results obtained are similar to those obtained by Kretzmer in his previous investigations of autocorrelation functions of pictures.

In view of the large-scale redundancies in television signals, pointed out in Section II, one would expect greater savings in channel capacity than have been found.

Table V.

Computed Autocorrelation Bound to Entropy of Sample with Three Previous Known Samples

Subject	$H(X_4 \mid X_3; X_2; X_1)$ (bits)
1	1.010
2	1.340
3	1.895
4	1.990
5	2.195

As a matter of fact, the equipment can be used to predict greater savings. For example, in the autocorrelation function bound, if the entropy of a sample, given three previous samples, is computed, a lower value for the bound will be obtained than when only one sample is known. The three previous samples chosen are the one on the same horizontal line preceding the cell in question, the one directly above it on the preceding line, and the one succeeding the latter cell. The formula for the conditional entropy is, in this case (see Appendix V),

$$H(X_4 \mid X_3; X_2; X_1) = \frac{1}{2} \log (2\pi e) \frac{\left|\rho_{ij}^{4}\right|}{\left|\rho_{ij}^{3}\right|}$$

and entails the solution of a fourth- and third-order determinant.

The results are shown in Table V.

Another method of predicting a reduction in channel capacity is to examine the probability of the difference signal; that is, $p\left[(S_2 - S_1) \mid S_1\right]$. Once a symbol has been transmitted, the difference signal completely specifies the picture, and thus it makes sense to investigate it. If measurements are made normal to the main diagonal of the brightness diagram, an equivalent change of variables from $S_2 + N_2$, $S_1 + N_1$ to $(1/2^{1/2})(S_2 + S_1 + N_2 + N_1)$, $(1/2^{1/2})(S_2 - S_1 + N_2 - N_1)$ has been made.

The probability measured by proceeding normal to the main diagonal is $p(S_2 - S_1 + N_2 - N_1 \mid S_2 + S_1 + N_2 + N_1)$. For any given measurement normal to the main diagonal, the noise variance stays practically constant so that one may say that the difference signal is statistically independent of the noise. The variance of the difference signal plus noise will then be equal to the sum of the variances of the difference signal and noise. The variance of signal plus noise is measured at unit shift, and the noise variance is measured at zero shift. The difference between the two is the variance of the difference signal for a particular value of $S_2 + S_1 + N_2 + N_1$. The results are then averaged over $S_2 + S_1 + N_2 + N_1$. An upper bound to the entropy per symbol is obtained

by assuming the difference distribution is gaussian (see Appendix IV). This corresponds to an entropy $H\left[(S_2 - S_1)|(S_2 + S_1 + N_2 + N_1)\right]$ which is an upper bound to $H\left[(S_2 - S_1)|(S_2 + S_1)\right]$ (see Appendix III. Since S_2 and S_1 are almost the same, the latter entropy is equivalent to $H\left[(S_2 - S_1)|S_1\right]$. As mentioned previously, the difference given the preceding cell completely specifies the picture. The value of the upper bound to $H\left[(S_2 - S_1)|S_1\right]$ for the fifth subject is 0.11 bits. This is far below the maximum entropy of 5 bits.

There are many other ways in which the equipment might be used to predict large savings in channel capacity. The most fruitful approach, however, seems to be to deal with the difference signal. By including a delay line, the probabilities of differences could be measured directly.

In conclusion, what is of importance is not the actual prediction of channel capacity reduction, but the philosophy behind the measurement. A communication system was examined and found to be inefficient. By applying the ideas and concepts of information theory, an estimate of how efficient the system can be made is obtained. This approach is a powerful one, since for the first time the engineer can evaluate in a quantitative manner the system he uses to transmit information. However, in its present state, information theory can predict great savings in efficiency but does not show how to effect them. In this way it is similar to thermodynamics which will tell one how efficient a heat engine can be, but never how to build it. How these savings can be effected is left entirely to the imagination of the engineer. To use his channels more efficiently, a large amount of terminal equipment may be necessary. In certain cases, the savings may not warrant the additional complexity of equipment. However, in television where the bandwidth is 4 Mc/sec the additional complications may be worthwhile.

APPENDIX I

PROBABILITIES OF SINUSOIDS (see reference 8)

With reference to Fig. 15, the probability that the random variable x lies between levels x_i and x_{i+1} is

$$p(x_i \leq x \leq x_{i+1}) = \frac{\text{favorable chances}}{\text{total chances}}$$

$$= \frac{(\sin^{-1} x_{i+1} - \sin^{-1} x_i) + (\pi - \sin^{-1} x_i) - (\pi - \sin^{-1} x_{i+1})}{2\pi}$$

$$= \frac{\sin^{-1} x_{i+1} - \sin^{-1} x_i}{\pi}$$

Thirty-two levels are assumed for x; therefore

$$x_i \rightarrow 0, \pm\frac{1}{16}, \pm\frac{2}{16}, \ldots, \pm\frac{16}{16}$$

The probabilities obtained for the positive values for x_i will be equal to those obtained for the negative values of x_i. The probabilities are plotted in Fig. 3.

PROBABILITIES OF EXPONENTIALS (see reference 8)

With reference to Fig. 16, the probability that the random variable x lies between k and k+1, where $k > 0$, is

$$p\left[\exp(-a/2) + \frac{k}{N}\left[1 - \exp(-a/2)\right] \leq x \leq \exp(-a/2) + \frac{k+1}{N}\left[1 - \exp(-a/2)\right]\right] = \frac{\text{favorable chances}}{\text{total chances}}$$

$$= \frac{RC}{1/f} \ln\left[\frac{\exp(-a/2) + (k+1)\left[1 - \exp(-a/2)\right]/N}{\exp(-a/2) + k\left[1 - \exp(-a/2)\right]/N}\right]$$

$$= \left(\frac{1}{a}\right) \ln\left[1 + \frac{\exp(a/2) - 1}{N + k\left[\exp(a/2) - 1\right]}\right]$$

For $k = 0$, we have

$$p(x = 0) = \frac{1}{2}$$

where k is the index and varies in discrete steps from 0 to N-1, for an N-step measurement. The probabilities are shown in Fig. 4.

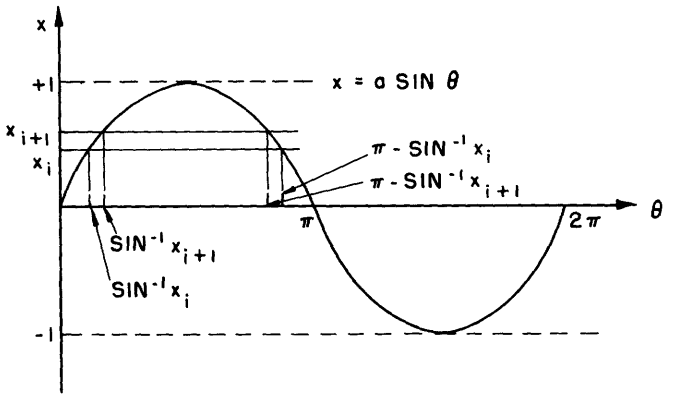

Fig. 15. Probabilities of sinusoids.

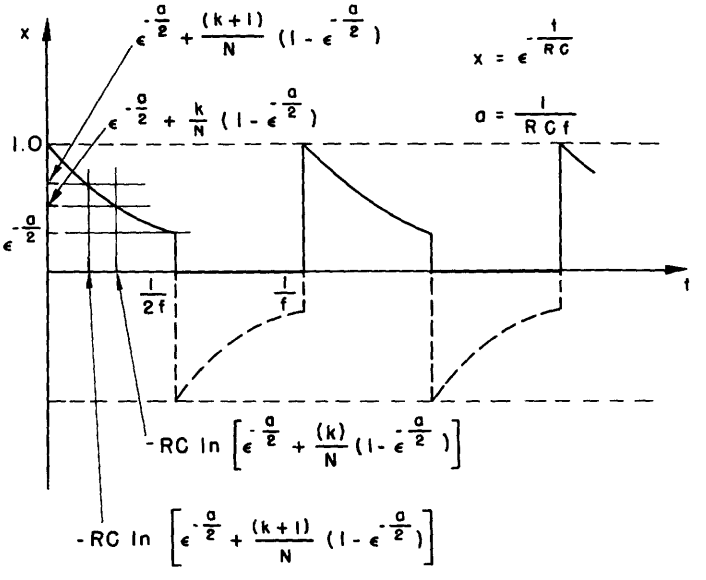

Fig. 16. Probabilities of exponentials.

APPENDIX II

ANALYSIS FOR OBTAINING $p(S^2|S^1)$ FROM $p\left[(S^2 + N^2);(S^1 + N^1)\right]$, $p(N^2)$, AND $p(N^1)$†

Let S_i^1 be the first signal at level i, S_i^2 be the second signal at level i, N_i^1 be the first noise at level i, and N_i^2 be the second noise at level i. Given the sets of probabilities $p\left[(S^2 + N^2);(S^1 + N^1)\right]$, $p(N^2)$, $p(N^1)$, find the set of probabilities $p(S^2|S^1)$, if:

1. signal and noise are statistically independent,
2. individual noises are statistically independent,
3. signals can assume any one of M levels,
4. noises can assume any one of N levels.

Proceed as follows:

$$p\left[(S^2 + N^2)_0;(S^1 + N^1)_0\right] = \text{probability that } S^2 \text{ is at level zero when } N^2 \text{ is at level zero, and } S^1 \text{ is at level zero when } N^1 \text{ is at level zero}$$

$$= p(S_o^2; N_o^2; S_o^1; N_o^1)$$

$$= p(S_o^2; S_o^1)\, p(N_o^2)\, p(N_o^1)$$

Since all quantities in the expression are known except $p(S_o^2; S_o^1)$, this probability can be obtained. Similarly

$$p\left[(S^2 + N^2)_1;(S^1 + N^1)_0\right] = p(S_1^2; N_o^2; S_o^1; N_o^1) + p(S_o^2; N_1^2; S_o^1; N_o^1)$$

$$= p(S_1^2; S_o^1)\, p(N_o^2)\, p(N_o^1) + p(S_o^2; S_o^1)\, p(N_1^2)\, p(N_o^1)$$

Since all quantities but $p(S_1^2; S_o^1)$ are known, this probability can be obtained. Proceeding in this manner, we can solve for the values of $p(S_2^2; S_o^1), \ldots, p(S_M^2; S_o^1)$. Then with the recurrence formula

$$p\left[(S^2 + N^2)_\ell;(S^1 + N^1)_k\right] = \sum_{i=0}^{\ell} \sum_{j=0}^{k} p(S_i^2; S_j^2)\, p(N_{\ell-i}^2)\, p(N_{k-j}^1)$$

we can obtain the other probabilities.

Once all $p(S_i^2; S_j^1)$ have been found

$$p(S_i^2|S_j^1) = \frac{p(S_i^2; S_j^1)}{\sum_i p(S_i^2; S_j^1)}$$

Thus all $p(S_i^2|S_j^1)$ can be found.

---

†Note that a shift from subscripts to superscripts has been made. Thus S^1 designates the first signal.

APPENDIX III

We wish to prove

$$H(S_2 + N_2 | S_1 + N_1) \geq H(S_2 | S_1)$$

where S_1 is the first signal, S_2 is the second signal, N_1 is the first noise, and N_2 is the second noise. We assume that the signal and noise are statistically independent, and that individual noises are statistically independent.

Proof: By factoring the joint probability distribution one obtains

$$H(S_2 | S_1 ; N_1 ; N_2) + H(N_1 ; N_2) = H(N_1 ; N_2 | S_1 ; S_2) + H(S_2 | S_1) \tag{1}$$

Let

$$Z_1 = S_1 + N_1$$

$$Z_2 = S_2 + N_2$$

Since the specification of signal plus noise for a given noise is the same as specifying just the signal for the given noise, we have

$$H(Z_2 | Z_1 ; N_1 ; N_2) = H(S_2 | S_1 ; N_1 ; N_2) \tag{2}$$

Substituting from Eq. 2 into Eq. 1

$$H(Z_2 | Z_1 ; N_1 ; N_2) = H(S_2 | S_1) + H(N_1 ; N_2 | S_1 ; S_2) - H(N_1 ; N_2) \tag{3}$$

Since the specification of additional symbols decreases the entropy

$$H(Z_2 | Z_1) \geq H(Z_2 | Z_1 ; N_1 ; N_2) \tag{4}$$

$$H(N_1 ; N_2 | S_1 ; S_2) = H(N_1 ; N_2) \tag{5}$$

if $N_1 ; N_2$ are statistically independent of S_1, S_2.

Substituting Eqs. 4 and 5 into Eq. 3 yields

$$H(Z_2 | Z_1) \geq H(S_2 | S_1) \tag{6}$$

It should be noted that some dependence of noise on signal is permissible with the inequality still holding. However, inequality certainly holds and is strongest when signal and noise are independent.

We wish to prove

$$H(S_2 | S_1) \geq H\left[(S_2 + N_2) | (S_1 + N_1)\right] - H(N_2) - H(N_1)$$

with the assumptions and notations of the preceding equations still valid. Again, by factoring the joint probability distributions,

$$H(Z_2 ; Z_1 ; S_2 ; S_1) = H(S_2 ; S_1) + H(Z_2 ; Z_1 | S_2 ; S_1) \tag{7}$$

Since the higher order entropy is always greater than any lower order entropy,

$$H(Z_2; Z_1; S_1; S_2) \geqslant H(Z_2; Z_1) \tag{8}$$

$$H(Z_2; Z_1 | S_2; S_1) = H(N_2; N_1 | S_2; S_1) = H(N_2; N_1) = H(N_2) + H(N_1) \tag{9}$$

from our previous assumptions.

Substitution of Eqs. 9 and 8 into Eq. 7 yields

$$H(S_2; S_1) \geqslant H(Z_2; Z_1) - H(N_2) - H(N_1) \tag{10}$$

From Eq. 6

$$H(S_1) \leqslant H(Z_1) \tag{11}$$

Subtracting Eq. 10 from Eq. 11 gives

$$H(S_2 | S_1) \geqslant H(Z_2 | Z_1) - H(N_2) - H(N_1) \tag{12}$$

Thus

$$H(Z_2 | Z_1) \geqslant H(S_2 | S_1) \geqslant H(Z_2 | Z_1) - H(N_1) - H(N_2) \tag{13}$$

APPENDIX IV

We wish to prove that the normal random process having the second moments a_{ij} has the maximum entropy of all processes having the same set of second moments (see ref. 2, pp. 55-56).

Proof: The mathematical formulation of the problem requires maximizing

$$H = -\int \cdots \int p(x_1, \ldots, x_n) \log p(x_i, \ldots, x_n)\, dx_1, \ldots, dx_n$$

Subject to the integral constraints

$$\int \cdots \int p(x_1, \ldots, x_n)\, dx_1, \ldots, dx_n = 1$$

$$\int \cdots \int x_i x_j\, p(x_1, \ldots, x_n)\, dx_1, \ldots, dx_n = a_{ij}$$

This is a standard problem of the calculus of variations (see ref. 12, pp. 139-144), and is solved by the method of Lagrangian multipliers. The problem requires satisfying the equation

$$\begin{aligned}
&\frac{\partial F}{\partial y} + \lambda_{11} \frac{\partial G_{11}}{\partial y} + \cdots + \lambda_{1n} \frac{\partial G_{1n}}{\partial y} \\
&\quad + \lambda_{21} \frac{\partial G_{21}}{\partial y} + \cdots + \lambda_{2n} \frac{\partial G_{2n}}{\partial y} \\
&\quad + \lambda_{n1} \frac{\partial G_{n1}}{\partial y} + \cdots + \lambda_{nn} \frac{\partial G_{nn}}{\partial y} \\
&\quad + \mu \frac{\partial G_{\mu}}{\partial y} = 0
\end{aligned}$$

where

$$F = p(x_1, \ldots, x_n) \log p(x_1, \ldots, x_n)$$

$$G_{ij} = x_i x_j\, p(x_1, \ldots, x_n)$$

$$y = p(x_1, \ldots, x_n)$$

$$G_{\mu} = p(x_1, \ldots, x_n)$$

and λ_{ij} and μ are Lagrangian multipliers which must be so chosen as to satisfy the constraints

$$\frac{1}{\ell n 2} + \frac{\ell n p}{\ell n 2} + \lambda_{11} x_1^2 + \ldots + \lambda_{1n} x_1 x_n$$

$$+ \lambda_{21} x_2 x_1 + \ldots + \lambda_{2n} x_2 x_n + \ldots$$

$$+ \lambda_{n1} x_n x_1 + \ldots + \lambda_{nn} x_n^2 + \mu = 0$$

$$p(x_1, \ldots, x_n) = e^{-\mu \ell n 2} e^{-\ell n 2 \lambda \underline{x} x]}$$

where

$$\lambda = \begin{bmatrix} \lambda_{11}, & \ldots, & \lambda_{1n} \\ \lambda_{21}, & \ldots, & \lambda_{2n} \\ \cdot & & \\ \cdot & & \\ \cdot & & \\ \lambda_{n1}, & \ldots, & \lambda_{nn} \end{bmatrix}$$

$$\underline{x} = \begin{bmatrix} x_1, & \ldots, & x_n \end{bmatrix}$$

$$x] = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix}$$

This has the form of an n-dimensional gaussian density function which can be written as

$$p(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2} |\rho|^{1/2}} e^{-1/2 \rho \underline{x} x]}$$

Where $|\rho|$ is the determinant whose elements are ρ_{ij}, and ρ is the matrix whose elements are

$$\frac{|\rho_{ij}|}{|\rho|}$$

$|\rho_{ij}|$ being a cofactor of $|\rho|$. Thus

$$e^{-\mu \ell n 2} = \frac{1}{(2\pi)^{n/2} |\rho|^{1/2}}$$

$$-\ell n\, 2\lambda = -\frac{1}{2} \rho$$

and

$$-\ell n\, 2\lambda_{ij} = -\frac{1}{2}\frac{|\rho_{ij}|}{|\rho|}$$

whence

$$p(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}\,|\rho|^{1/2}}\, e^{-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{|\rho_{ij}|}{|\rho|}x_i x_j}$$

$$\times \int \cdots \int x_i x_j\, p(x_1, \ldots, x_n)\, dx_1, \ldots, dx_n = \rho_{ij} = a_{ij}$$

Thus the elements ρ_{ij} of the determinant $|\rho|$ are the various second moments, and the theorem is proved.

APPENDIX V

GENERAL EXPRESSION FOR THE ENTROPY PER SYMBOL OF THE NORMAL RANDOM PROCESS IN TERMS OF ITS CORRELATION COEFFICIENTS (see reference 13)

Given

$$p(x_1, \ldots, x_n) = \frac{\exp\left[-\frac{1}{2|\rho|}\sum_{i=1}^{n}\sum_{j=1}^{n}|\rho_{ij}|\, X_i X_j\right]}{(2\pi)^{n/2}\,|\rho|^{1/2}}$$

with

$X_i = x_i - \overline{x}$

$\rho_{ij} = \overline{X_i X_j}$ = correlation coefficient

$|\rho|$ = determinant whose elements are ρ_{ij}

$|\rho_{ij}|$ = cofactor of determinant $|\rho|$

we have

$$H = -\int \cdots \int p(x_1, \ldots, x_n) \log p(x_1, \ldots, x_n)\, dx_1, \ldots, dx_n$$

and

$$-\log p(x_1, \ldots, x_n) = \log (2\pi)^{n/2}\,|\rho|^{1/2} + \frac{1}{2|\rho|}\sum_{i=1}^{n}\sum_{j=1}^{n}|\rho_{ij}|\, X_i X_j \log e$$

Substitution gives

$$H = \int \cdots \int \log (2\pi)^{n/2}\,|\rho|^{1/2}\, p(x_1, \ldots, x_n)\, dx_1, \ldots, dx_n$$

$$+\frac{\log e}{2|\rho|}\sum_{i=1}^{n}\sum_{j=1}^{n}|\rho_{ij}| \int \cdots \int X_i X_j\, p(x_1, \ldots, x_n)\, dx_1, \ldots, dx_n$$

Now,

$$\int \cdots \int p(x_1, \ldots, x_n)\, dx_1, \ldots, dx_n = 1$$

and

$$\int \cdots \int X_i X_j \, p(x_1, \ldots, x_n) \, dx_1, \ldots, dx_n = \overline{X_i X_j} = \rho_{ij}$$

By the Laplace development for $|\rho|$,

$$|\rho| = \rho_{1i} |\rho_{1i}| + \rho_{2i} |\rho_{2i}| + \ldots + \rho_{ni} |\rho_{ni}|$$

gives

$$H = \log (2\pi)^{n/2} |\rho|^{1/2} + \frac{\rho_{11} |\rho_{11}| + \ldots + \rho_{1n} |\rho_{1n}|}{2|\rho|} \log e$$

$$+ \frac{\rho_{21} |\rho_{21}| + \ldots + \rho_{2n} |\rho_{2n}|}{2|\rho|} \log e + \ldots$$

$$+ \frac{\rho_{n1} |\rho_{n1}| + \ldots + \rho_{nn} |\rho_{nn}|}{2|\rho|} \log e$$

The numerators of the last expressions are Laplace developments for the determinant $|\rho|$, and since we have n terms:

$$H = \log (2\pi)^{n/2} |\rho|^{1/2} \frac{n|\rho|}{2|\rho|} \log e$$

$$= \log (2\pi e)^{n/2} |\rho|^{1/2}$$

and

$$H(X_n | X_{n-1}, \ldots, X_1) = -\int_{x_1} \cdots \int_{x_n} p(x_1, \ldots, x_n) \log p(x_n | x_{n-1}, \ldots, x_1) \, dx_1, \ldots, dx_n$$

$$= \text{entropy/symbol}$$

Here

$$p(x_n | x_{n-1}, \ldots, x_1) = \frac{p(x_n, x_{n-1}, \ldots, x_1)}{p(x_{n-1}, \ldots, x_1)}$$

So

$$H(X_n | X_{n-1}, \ldots, X_1) = H(X_n, \ldots, X_1) - H(X_{n-1}, \ldots, X_1)$$

$$= \log (2\pi e)^{n/2} |\rho^n|^{1/2} - \log (2\pi e)^{(n-1)/2} |\rho^{n-1}|^{1/2}$$

$$= \frac{1}{2} \log 2\pi e \frac{|\rho^n|}{|\rho^{n-1}|} \qquad n = 2, 3, \ldots, \infty$$

This is the general expression for the entropy per symbol where $|\rho^n|$ is the determinant whose elements are ρ_{ij} and is of order n. For n = 2

$$H(X_2|X_1) = \frac{1}{2}\log 2\pi e \frac{\begin{vmatrix}\rho_{11} & \rho_{12}\\ \rho_{21} & \rho_{22}\end{vmatrix}}{\rho_{11}}$$

$$= \frac{1}{2}\log 2\pi e + \frac{1}{2}\log \frac{\rho_{11}\rho_{22} - \rho_{12}^2}{\rho_{11}}$$

where

$$\rho_{11} = \overline{X_1 X_1} = \sigma^2 = \rho_{22}$$

and

$$\rho_{12} = \overline{X_1 X_2}$$

$$H(X_2|X_1) = \frac{1}{2}\log 2\pi e + \frac{1}{2}\log\left(\sigma^2 - \frac{\rho_{12}^2}{\sigma^2}\right)$$

$$= \frac{1}{2}\log 2\pi e\sigma^2 + \frac{1}{2}\log\left[1 - \left(\frac{\rho_{12}}{\sigma^2}\right)^2\right]$$

Acknowledgment

The author is lastingly indebted to Professor Peter Elias. Throughout the course of this work, he was a constant source of inspiration and a counselor of enduring patience.

To Professor Harold R. Mimno, of Harvard University, the author expresses sincere appreciation for permission to use part of the equipment employed by Dr. William F. Schreiber in his investigations.

The author also wishes to thank Professor R. M. Fano for his continued interest in this project.

References

1. R. M. Fano, Class Notes for Course 6.574, M.I.T., 1953.
2. C. E. Shannon and W. Weaver, The Mathematical Theory of Communication (University of Illinois Press, Urbana, Illinois, 1949) pp. 38-48.
3. B. M. Oliver, J. R. Pierce, and C. E. Shannon, Proc. I.R.E. 36, 1324-1331 (1948).
4. W. F. Schreiber, Probability of television signals, Ph.D. Thesis, Harvard University, 1952.
5. E. R. Kretzmer, Bell System Tech. J. 31, 751-763 (1952).
6. W. M. Goodall, Bell System Tech. J. 30, 33-49 (1951).
7. Francis Bello, Information Theory, Fortune Magazine, p. 154, Dec. 1953.
8. Jack Capon, Bounds to the entropy of television signals, M.S. Thesis, Department of Electrical Engineering, M.I.T. (1955).
9. W. B. Davenport, Jr., Class Notes for Subject 6.573, M.I.T., 1952.
10. C. B. Noblette, Photography Principles and Practice (D. Van Nostrand Company, Inc., New York, 1938).
11. L. Dolansky and M. P. Dolansky, Technical Report 227, Research Laboratory of Electronics, M.I.T. (1952).
12. F. B. Hildebrand, Methods of Applied Mathematics (Prentice-Hall, Inc., New York, 1952).
13. Peter Elias, Proc. I.R.E., 39, 839 (1951).