XXII. SENSORY AIDS RESEARCH*

Prof. S. J. Mason
D. A. Cahlander
E. S. Davis
J. Dupress
W. G. Kellner
D. G. Kocher
W. B. Macurdy
M. A. Pilla
M. M. Schiffman
D. E. Troxel

A. AN INFORMATIONAL ANALYSIS OF DISCRETE HUMAN COMMUNICATION SYSTEMS

1. Introduction

An important problem confonting the designer of a sensory aid concerns the comparison of one sensory aid with another. One reasonable procedure is to separate the problem of extracting information from the real world from that of transmitting this information to the handicapped person. In this report we shall focus our attention on the second problem. In our analysis we shall consider a system that is capable of presenting L distinct stimuli, x_i, to which there correspond uniquely L distinct responses, y_j. Such a system, consisting of a stimulator and human receiver, will be termed a discrete human communication system. It is assumed that we know the first-order stimulus probabilities, $p(x_i)$, and that successive stimuli are statistically independent. To proceed we must know something about the relation of the response ensemble to the stimulus ensemble. One way of obtaining this information is to conduct an experiment in which successive stimuli are drawn from a known probability distribution and the responses that are elicited from the human receiver are tabulated in a confusion matrix. The rows of this matrix correspond to the stimuli presented, and the columns, to the responses that are evoked. A response of y_j to a stimulus of x_i is recorded by indexing the ij term of the confusion matrix up by one. Then at the conclusion of our experiment, the ij term of the confusion matrix divided by the total number of stimuli presented is an estimate of the joint probability $p(y_j, x_i)$. Thus we can construct an estimated conditional probability matrix $(p(y_j|x_i))$ from the confusion matrix. We shall take this conditional probability matrix to be a description of the channel consisting of the stimulator system and the human receiver. It is important to note that this channel matrix is not necessarily independent of the input stimulus probability distribution. However, for the present, let us fix the input distribution and consider the resultant channel matrix.

The channels with which we are concerned differ from those ordinarily treated by the methods of information theory in that they contain human receivers. Obvious complications spring to mind – humans have memory, have time-variant characteristics, and so forth. But all is not lost. Some significant simplifications are made possible

*This work was supported in part by the National Institutes of Health (Grant MH-04737-02); in part by the National Science Foundation (Grant G-16526); and in part by the U.S. Air Force (Electronic Systems Division) under Contract AF19(628)-258.

by the inclusion of humans in the system. Perhaps the most important simplification is that a workable system must be characterized by a fairly low over-all probability of error (say, less than 20 per cent). Humans characteristically tend to arrange things, in this case stimuli, in some natural grouping. Thus if there are L-1 possible errors that can be made when a particular stimulus is presented, a human is not very likely to make a large number of different erroneous choices. Instead, he is more likely to cluster his responses within a group that corresponds to stimuli that are somewhat similar to the presented stimulus.

The channel matrix and associated input distribution represents a description of the performance of the communication system. However, it is a rather unwieldy description if one is interested in evaluating performance under variations of stimulator parameters or for different human receivers. A much more manageable situation results if one reduces the channel description to a single number. There are many ways to accomplish this data reduction, but here we choose to calculate the average mutual information shared between the stimulus and response ensembles.[1] Unfortunately, it is very difficult to evaluate mutual information for large channel sizes (for example, L = 64) that are common in discrete human communication systems. A straightforward way of getting around the computational difficulties is to use modern computers. However, it is both inconvenient and expensive to run computer programs for day-to-day plotting of learning curves. Also, the large number of significant figures resulting from computer calculations tends to be somewhat misleading. One of the dangers of using a number is the propensity to put too much faith in its accuracy. It should be remembered at this point that the channel matrix, on which the information calculation is based, is but an estimate of the "true" channel matrix. Accordingly, it makes little sense to grind out the mutual information to 10 decimal places, or (in most cases) even to 3. With this in mind, it is then reasonable to try to develop a simplified computational procedure that will enable one to arrive at an approximate information transfer (mutual information) with an error in the neighborhood of that inherent in the test data. We have derived the following upper and lower bounds which use the simplifications afforded by the inclusion of humans in the system. Following this discussion of the bounds is a proposed model channel that is used to estimate the information transfer of a discrete human communication system.

2. Derivation of a Lower Bound

We can express the information transfer as

$$I(X;Y) = H(X) - H(X|Y). \tag{1}$$

We wish to consider the probability of error in the bounds, thus it is necessary to define a decoding scheme. We shall only consider maximum mutual-information decoding that is equivalent to maximum likelihood decoding:

$$I(x;y) \equiv \log \frac{P(y|x)}{P(y)}. \tag{2}$$

We can now define a probability of error

$$P(e) = \sum_{k} P(x_k) \sum_{j \neq k} P(y_j|x_k) = \sum_{j} P(y_j)\, P(e|y_j), \tag{3}$$

where

$$P(e|y_j) \equiv 1 - P(x_j|y_j) = \sum_{k \neq j} P(x_k|y_j). \tag{4}$$

It is convenient to define the entropy

$$H(e) \equiv -P(e) \log P(e) - [1-P(e)] \log [1-P(e)] \tag{5}$$

for the binary-choice error and no error, and the corresponding conditional entropy

$$H(e|Y) \equiv \sum_{j} P(y_j)\, H(e|y_j), \tag{6}$$

with

$$H(e|y_j) = -P(e|y_j) \log P(e|y_j) - [1-P(e|y_j)] \log [1-P(e|y_j)]. \tag{7}$$

We shall also define C_j as the number of nonzero off-diagonal conditional probabilities, $P(y_j|x_j)$, in column j and C_{max} as the maximum of C_j for any column.

THEOREM 1: The mutual information satisfies the inequality

$$I(X;Y) \geqslant H(X) - H(e|Y) - P(e) \log C_{max} \geqslant H(X) - H(e) - P(e) \log C_{max}. \tag{8}$$

PROOF 1: H(X|Y) is, by definition,

$$H(X|Y) \equiv -\sum_{j} P(y_j)\, H(X|y_j), \tag{9}$$

with

$$H(X|y_j) \equiv \sum_{k} P(x_k|y_j) \log P(x_k|y_j) \tag{10}$$

the equivocation when the point of the space Y is given. This equivocation can be rewritten as

$$H(X|y_j) = -P(x_j|y_j) \log P(x_j|y_j) - \sum_{k\neq j} P(x_k|y_j) \log P(x_k|y_j)$$

$$- [1-P(x_j|y_j)] \log [1-P(x_j|y_j)] - [1-P(x_j|y_j)] \log \frac{1}{[1-P(x_j|y_j)]}. \tag{11}$$

Applying Eqs. 4 and 7, we have

$$H(X|y_j) = H(e|y_j) - \sum_{k\neq j} P(x_k|y_j) \log P(x_k|y_j) - \sum_{k\neq j} P(x_k|y_j) \log \frac{1}{[1-P(x_j|y_j)]}$$

$$= H(e|y_j) - [1-P(x_j|y_j)] \sum_{k\neq j} \frac{P(x_k|y_j)}{1 - P(x_j|y_j)} \log \frac{p(x_k|y_j)}{1 - P(x_j|y_j)}$$

$$= H(e|y_j) + P(e|y_j) \sum_{k\neq j} \frac{P(x_k|y_j)}{P(e|y_j)} \log \frac{p(x_k|y_j)}{P(e|y_j)}. \tag{12}$$

The summation on the right-hand side of Eq. 12 is recognized as the entropy of an ensemble consisting of C_j points, and thus, its value cannot exceed $\log C_j$. It follows that

$$H(X|y_j) \leq H(e|y_j) + P(e|y_j) \log C_j. \tag{13}$$

Since

$$\log C_{max} \geq \log C_j, \tag{14}$$

we have

$$H(X|y_j) \leq H(e|y_j) + P(e|y_j) \log C_{max}. \tag{15}$$

Averaging over the ensemble Y, we have

$$H(X|Y) \leq H(e|Y) + P(e) \log C_{max}. \tag{16}$$

Substitution of inequality (16) in Eq. 1 and use of the fact that

$$H(e|Y) \leq H(e) \tag{17}$$

yields the inequality (7), which was to be proved.

Q.E.D.

In particular, if the probability distribution of the L inputs is uniform, we have an easy bound to calculate

$$I(X;Y) \geq \log L - H(e) - P(e) \log C_{\max}. \tag{18}$$

The equality sign in (13) holds only when

$$P(x_k|y_j) = \frac{1 - P(x_j|y_j)}{C_j} \qquad \text{or zero for all } k \neq j. \tag{19}$$

For the equality sign to hold in (15) and (16) we must have $C_j = C_{\max}$ for all j, in addition to (19).

EXAMPLE 1: Take $P(x_k) = \frac{1}{L}$ for all k and $P(y_j|x_k) = p$ for $j = k$.

Let $C_j = C_{\max}$ for all j, and let $P(y_j|x_k) = \frac{1-p}{C_{\max}}$ when it is nonzero.

Now

$$H(X) = -\sum_{j=1}^{L} \frac{1}{L} \log \frac{1}{L} = \log L$$

and

$$P(y_j) = \frac{1}{L}\left(p + C_{\max}\left(\frac{1-p}{C_{\max}}\right)\right) = \frac{1}{L};$$

thus $H(Y) = \log L$. Also,

$$H(XY) = \sum_{j=1}^{L} -\frac{p}{L} \log \frac{p}{L} - C_{\max} \frac{1-p}{LC_{\max}} \log \frac{1-p}{LC_{\max}}$$

$$= -p \log p - (1-p) \log (1-p) + \log L + (1-p) \log C_{\max}.$$

Since $P(e) = 1 - p$, we have

$$H(XY) = H(e) + \log L + P(e) \log C_{\max}.$$

Thus

$$I(X;Y) = H(X) + H(Y) - H(XY) = \log L - H(e) - P(e) \log C_{\max},$$

which is the lower bound. A sample channel matrix of order L = 4 is shown below. Note that this is a special case of a doubly uniform channel.

Uniform input distribution

$$\begin{array}{c|cccc} & & & Y & \\ \hline & .9 & 0 & .05 & .05 \\ & .05 & .9 & .05 & 0 \\ X & 0 & .05 & .9 & .05 \\ & .05 & .05 & 0 & .9 \end{array}$$

$$\begin{aligned} I(X;Y) &= \log 4 - H(.1) - .1 \log 2 \\ &= 1.431 \text{ bits.} \end{aligned}$$

EXAMPLE 2: Consider the following channel.

$$\begin{array}{ccc} & & \text{Y} \\ \begin{array}{l}\text{Uniform}\\ \text{input}\\ \text{distribution}\end{array} & \text{X} & \left[\begin{array}{cccc} .9 & 0 & .05 & .05 \\ .05 & .85 & .05 & .05 \\ 0 & .1 & .9 & 0 \\ .05 & .05 & 0 & .9 \end{array}\right] \end{array}$$

Here, $C_j = C_{\max} = 2$. Also, $P(e) = 1 - \frac{1}{4}\left[3(.9)+.85\right] = 0.1125$, so that

$$I > \log 4 - H(.1125) - .1125 \log 2$$

or

$$I > 1.380 \text{ bits}.$$

Calculation of the actual mutual information yields I = 1.386 bits.

3. Derivation of an Upper Bound

THEOREM 2: The mutual information satisfies the inequality

$$I(X;Y) \leqslant \log L - \sum_i P(x_i)\, H(p_{ii}), \tag{20}$$

where L is the order of the channel and

$$H(p_{ii}) \equiv -P(y_i|x_i) \log P(y_i|x_i) - [1-P(y_i|x_i)] \log [1-P(y_i|x_i)]. \tag{21}$$

PROOF 2: We can express the information transfer as

$$I(X;Y) = H(Y) - H(Y|X).$$

We always have, of course, the fact that

$$H(Y) \leqslant \log L. \tag{22}$$

Now,

$$H(Y|X) = \sum_i P(x_i)\, H(Y|x_i)$$

and

$$\begin{aligned} H(Y|x_i) &= -\sum_j P(y_j|x_i) \log P(y_j|x_i) \\ &\geqslant -P(y_i|x_i) \log P(y_i|x_i) - [1-P(y_i|x_i)] \log [1-P(y_i|x_i)] = H(p_{ii}). \end{aligned} \tag{23}$$

Thus we have

$$H(Y|X) \geqslant \sum_{i} P(x_i)\, H(p_{ii}). \tag{24}$$

Substitution of (22) and (24) in (1) yields the inequality (20), which was to be proved.

Q.E.D.

The equality signs in (23) and (24) hold only when there is only 1 off-diagonal nonzero conditional probability in each row. Obviously, the equality sign in (4) holds only when the output distribution $P(y_j)$ is uniform.

EXAMPLE 3: Consider the channel treated in Example 1. The upper bound is

$$I \leqslant \log L - \sum_{i=1}^{L} \frac{1}{L} H(p)$$

or

$$I \leqslant \log L - H(p) = \log L - H(e).$$

The difference between this upper bound and the actual mutual information is simply the term

$$P(e) \log C_{max} = .1 \text{ bit}$$

for the sample 4 × 4 channel so that

$$I < \log 4 - H(.1)$$

$$I < 1.531 \text{ bits}.$$

EXAMPLE 4: Consider the channel treated in Example 2. The upper bound is

$$I < \log 4 - \frac{3}{4} H(.1) - \frac{1}{4} H(.15)$$

or

$$I < 1.496 \text{ bits}.$$

4. Proposed Model Channel

A first-order approximation to a channel describing a discrete human communication system is shown in Fig. XXII-1. Each of the major diagonal terms of this conditional probability matrix is taken to be equal to p. The off-diagonal terms in each row may be either zero or a constant. If r_i is defined as the number of nonzero off-diagonal terms in the i^{th} row, then this constant is $(1-p)/r_i$. The order of the matrix is L, so that we have $1 \leqslant r_i \leqslant L - 1$. No assumptions have been made about the probability of any

$$\begin{array}{cc} & \text{Y Response} \\ \begin{array}{c}\text{X}\\ \text{Stimulus}\end{array} & \left[\begin{array}{ccccccccc} p & \cdot & 0 & \frac{1-p}{r_1} & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \frac{1-p}{r_i} & \cdot & p & 0 & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & \cdot & & & & & \\ & & & & \cdot & & & & \\ & & & & & \cdot & & & \\ & & & & & & \cdot & & \\ & & & & & & & \cdot & \\ & & & & & & & & p \end{array}\right] \end{array}$$

Fig. XXII-1. Model channel.

particular off-diagonal term being zero. The zeros inserted in Fig. XXII-1 serve solely as an example. The probability of error for this model channel is 1 - p when maximum likelihood decoding is used.

A simple estimate for the information transfer with uniform input probability distribution will be based on the following upper bound.

THEOREM 3: The information transfer for the model channel shown in Fig. XXII-1, with uniform input probability distribution, satisfies the following inequality.

$$I(X;Y) \leq \log L - H(e) - \frac{P(e)}{L} \sum_i \log r_i \tag{25}$$

PROOF 3: The mutual information can be expressed as

$$I(X;Y) = H(X) + H(Y) - H(XY). \tag{26}$$

The input distribution is uniform. We have

$$H(X) = \log L \tag{27}$$

and, of course, we always have

$$H(Y) \leq \log L.$$

H(XY) will now be evaluated as the negative sum over the whole matrix of

$$\frac{p_{ij}}{L} \log \frac{p_{ij}}{L}.$$

Consider, first, the sum on row i.

$$S_i = \sum_j \frac{p_{ij}}{L} \log \frac{p_{ij}}{L} = \frac{p}{L} \log \frac{p}{L} + \frac{1-p}{Lr_i} \log \frac{1-p}{Lp_i}$$

$$= \frac{p}{L} \log \frac{p}{L} + \frac{1-p}{L} \log \frac{1-p}{L} - \frac{1-p}{L} \log r_i .$$

Now

$$-H(XY) = \sum_i S_i = p \log \frac{P}{L} + (1-p) \log \frac{1-p}{L} - \frac{1-p}{L} \sum_i \log r_i .$$

As $P(e) = 1 - p$, we have

$$H(XY) = \log L + H(e) - \frac{P(e)}{L} \sum_i \log r_i . \qquad (28)$$

Substitution of (27), (22), and (28) in (26) yields the desired inequality (25), which was to be proved.

Q.E.D.

The inequality sign is necessary solely because of Eq. 22. If the total probability of error is small (say less than 10-15 per cent), then one can reasonably expect that $p(y)$ will be fairly independent of y. $H(Y)$ does not vary rapidly as $p(y)$ departs from a uniform distribution; thus one can reasonably expect that the bound (Eq. 25) is rather tight.

On the basis of Theorem 3, the following formula is proposed as an estimate of the information transfer for a channel describing a discrete human communication system when the input stimulus distribution is uniform.

$$\hat{I}(X;Y) = \log L - H(e) - \frac{P(e)}{L} \sum_i \log r_i . \qquad (29)$$

EXAMPLE 5: Consider the channel treated in Example 1 with $r_i = C_{max}$. The estimate is

$$I(X;Y) = \log L - H(e) - P(e) \log C_{max} .$$

This is identical with the lower bound that is the actual information transferred. Thus for the sample channel of order four, we have $\hat{I} = 1.43$ bits per stimulus.

EXAMPLE 6: Consider the channel treated in Example 2. The estimate is

$$\hat{I}(X;Y) = \log 4 - H(.1125) - \frac{.1125}{4} (\log 2 + \log 3 + \log 1 + \log 2) = 1.39 \text{ bits per stimulus.}$$

5. Experimental Results

As we have shown, the primary purpose of the bounds that have been derived and the proposed estimate is to circumvent the laborious computations that are inherent in the calculation of mutual information. The price to be paid for a simplified estimate (Eq. 29) is uncertainty about its accuracy. One indication of accuracy is obtained by calculating the upper and lower bounds. The worst possible percentage of error is given by

$$E = 100\frac{\text{estimate} - \text{lower bound}}{\text{lower bound}} \tag{30}$$

or

$$E = 100\frac{P(e)\left[\log C_{\max} - \frac{1}{L}\sum_i \log r_i\right]}{\log L - H(e) - P(e)\log C_{\max}}. \tag{31}$$

It follows that

$$E \leqslant \frac{P(e)\log C_{\max}}{\log L - H(e) - P(e)\log C_{\max}}. \tag{32}$$

This represents an easily calculable upper bound to the error. When a nonuniform input distribution is used, a reasonable estimate of the mutual information is obtained by choosing the mid-point between the bounds.

Some feeling for the "tightness" of the bounds and the accuracy of the estimate can be achieved by inspection of Figs. XXII-2 and XXII-3, in which the bounds, estimates, and actual mutual information are compared for various error probabilities. The stimuli for the smaller channel (L=5) are pulsed sine waves of different frequencies in a background of white noise, while the stimuli for the larger channel (l=16) consist of patterns of poke probes presented to the right index finger.

Results for a channel of order L = 64, for which the stimuli consist of poke probes to the right index finger, are:

measured probability of error = 0.095;
upper bound = 5.58 bits;
actual information = 5.36 bits;
estimate = 5.36 bits; and
lower bound = 5.23 bits.

The percentage of error of the estimate with respect to the actual information is

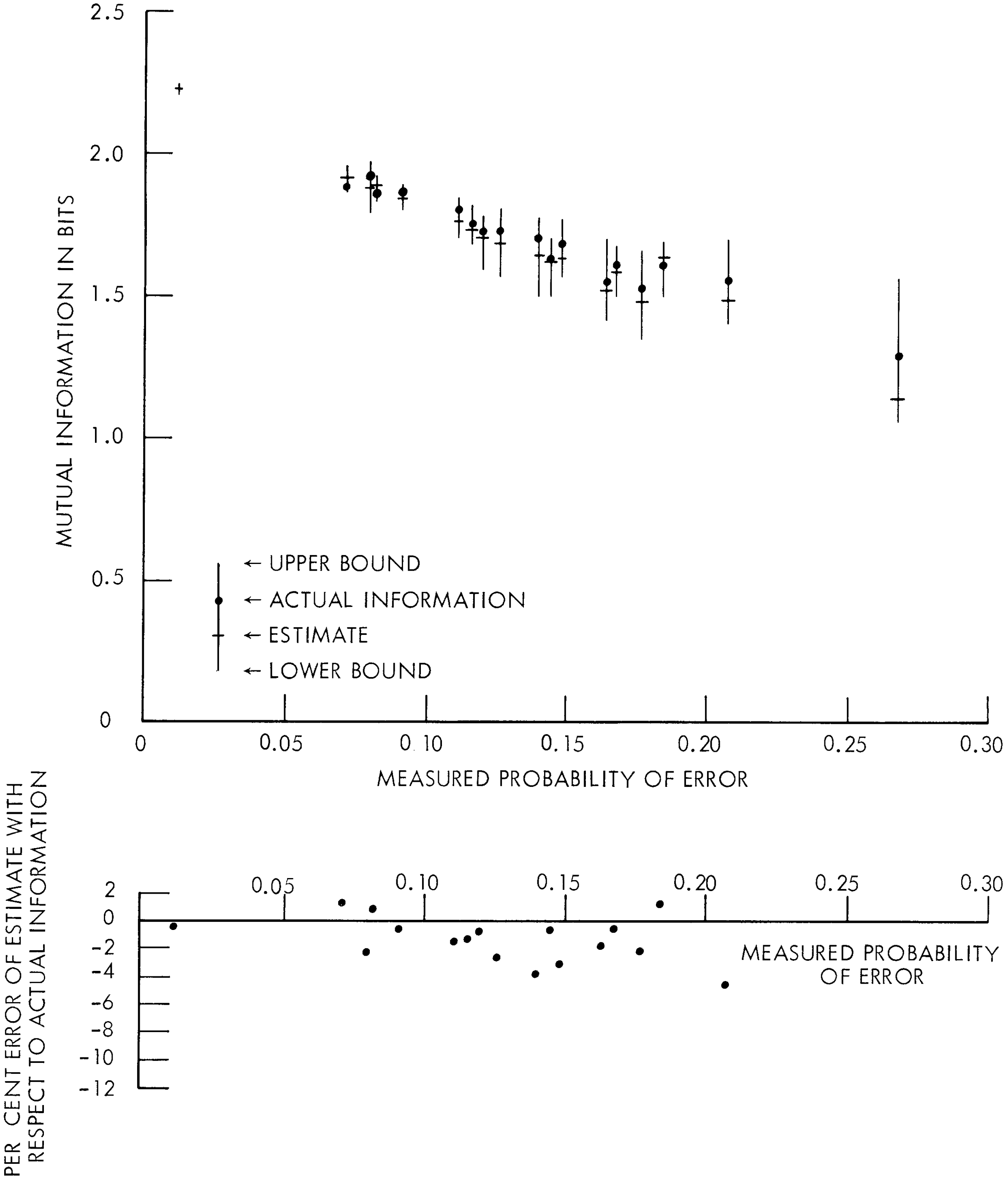

Fig. XXII-2. Uniform input distribution (L=5).

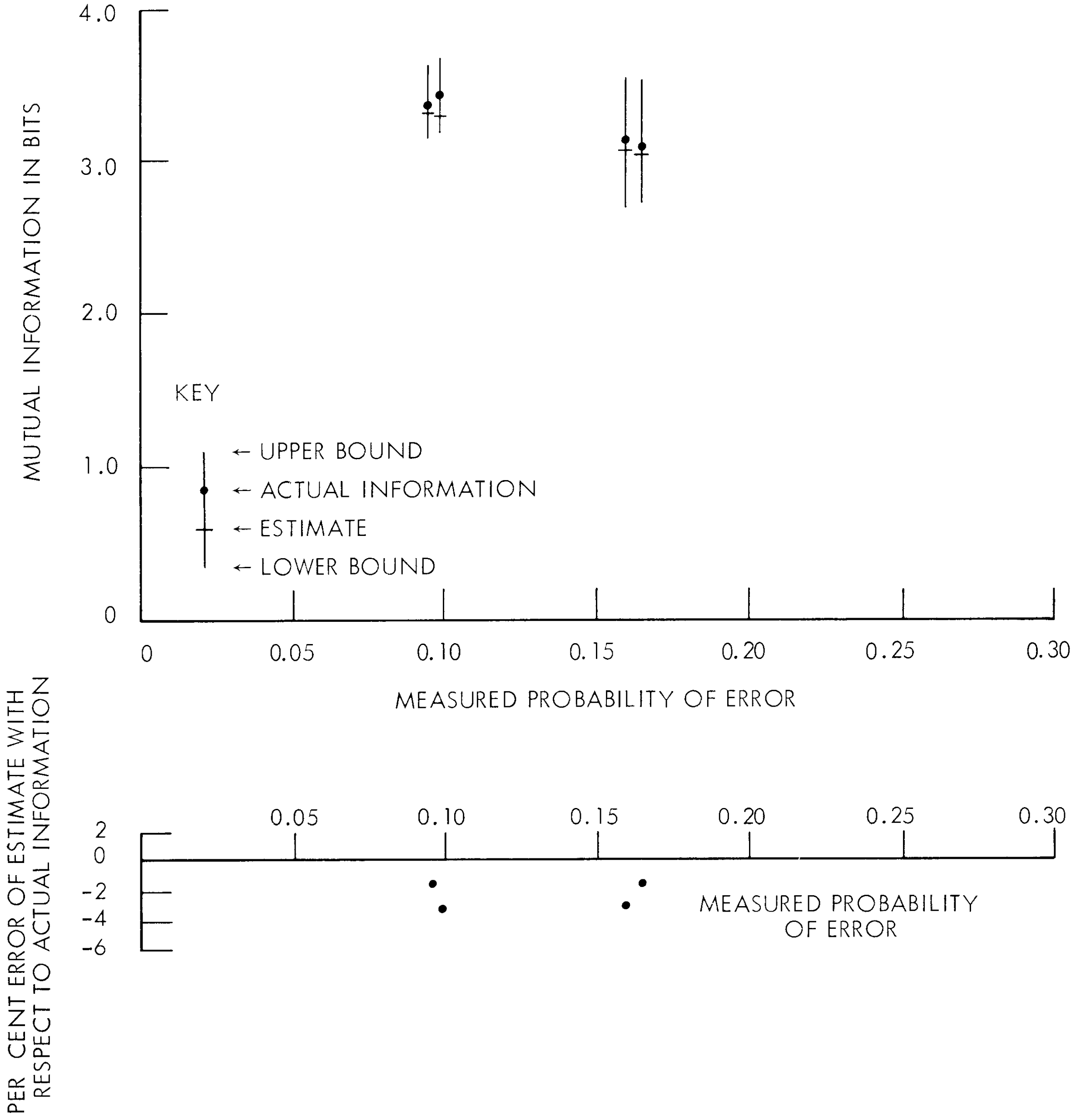

Fig. XXII-3. Uniform input distribution (L=16).

-0.15 per cent. The confusion matrix for this channel is a composite one formed by adding confusion matrices for 4 different human receivers.

D. E. Troxel

References

1. For a discussion of the applicability of the mutual-information measure to various psychophysical test situations, see F. Attneave, Applications of Information Theory to Psychology (Henry Holt and Company, New York, 1959), Chapter 4.