where $W(\nu/(l/W))$ is the conditional density function of Doppler shifts for illuminated dipoles, given that they provide a delay l/W at time t. Since we have assumed ν independent of time, $W(\nu/(l/W))$ is also time independent. For purposes of discussion assume that the density function of Doppler shifts is the same at all illuminated prolate spheroidal shells. Then

$$W\left(\nu \,/\, \frac{l}{W}\right) = W(\nu)$$

$$R_{t,s}\left(f_0, \frac{l}{W}\right) = \int \frac{f_0}{\tau} e^{-j2\pi f_0 \Delta} W\left(\frac{f_0 \Delta}{\tau}\right) d\Delta \tag{74}$$

$$= \int e^{-j2\pi\tau f_0} W(\nu)\, d\nu = R(\tau),$$

where $R(\tau)$ is the characteristic function corresponding to the probability density function of Doppler shifts. Using (74) and (71) in (70), we obtain

$$C_{ll}(t, t+\tau) = \frac{1}{W} K\left(\frac{l}{W}\right) C(\tau) R(\tau). \tag{75}$$

The parameter τ_0 is seen here to be the value of τ beyond which the characteristic function of Doppler shifts is negligible. Examination of (72) then shows that

$$\Delta_0 = \frac{\nu_{\max} \tau_0}{f_0}, \tag{76}$$

where $\nu_{\max}$ is the maximum absolute Doppler shift possible. The inequality (59) then becomes

$$W \ll \frac{f_0}{\tau_0 \nu_{\max}}. \tag{77}$$

As a final note to this paper it should be pointed out that if the average number of illuminated dipoles in a prolate spheroidal shell is sufficiently large, then the corresponding tap gain function will be a nearly normally distributed process because of the Central Limit Theorem. More generally, the set of tap gain functions will have statistics approaching those of a set of dependent complex valued Gaussian processes. Then, satisfaction of the inequality $W\,\Delta_0 \ll 1$ allows one to model the orbital dipole channel as a continuum of independent scatterers having complex-valued Gaussian statistics.

On the Effectiveness of Receptors in Recognition Systems*

T. MARILL†, MEMBER, IRE, AND D. M. GREEN‡

Summary—**In the type of recognition system under discussion, the physical sample to be recognized is first subjected to a battery of *tests*; on the basis of the test results, the sample is then assigned to one of a number of prespecified categories.**

The theory of how test results should be combined to yield an optimal assignment has been discussed in an earlier paper. Here, attention is focused on the tests themselves. At present, we usually measure the effectiveness of a set of tests empirically, *i.e.*, by determining the percentage of correct recognitions made by some recognition device which uses these tests. In this paper, we discuss some of the theoretical problems encountered in trying to determine a more formal measure of the effectiveness of a set of tests; a measure which might be a practical substitute for the empirical evaluation. Specifically, the following question is considered: What constitutes an effective set of tests, and how is this effectiveness dependent on the correlations among, and the properties of, the individual tests in the set?

Specific suggestions are considered for the case in which the test results are normally distributed, but arbitrarily correlated. The discussion is supported by the results of experiments dealing with automatic recognition of hand-printed characters.

* Received December 17, 1961; revised manuscript received, April 10, 1962. This research has been sponsored by Electronics Research Directorate, AFCRL, under Contract AF 19(604)-6632 and AF 19(628)-227.

† Bolt Beranek and Newman Inc., Cambridge, Mass.

‡ Bolt Beranek and Newman Inc., Cambridge, Mass., and Massachusetts Institute of Technology, Cambridge, Mass.

INTRODUCTION

WE CANNOT HOPE, in this introduction, to do justice to the many papers, ideas and techniques that have been contributed to the young but already large field of automatic recognition. An extensive bibliography is given by Minsky [13]. Interesting discussions have been given by Bledsoe and Browning [2], Bomba [3], Doyle [5], Grimsdale, *et al.*, [6], Harmon [7], Hawkins [8], Highleyman [9], Minsky [14], Rosenblatt [15], Selfridge [17], Sherman [18], Uhr and Vossler [20] and Unger [21].

The present discussion applies to a specific type of recognition model, namely, the type that derives from, or is related to, *statistical classification theory* (Anderson [1]). Models of this type have been investigated with a certain degree of success by Chow [4], Marill and Green [12], Sebestyen [16], Smith and Klem [19] and Welch and Wimpress [22]. These models are of the following sort (see Fig. 1): In the first stage of the process, the input (a pattern, or object, or event) is subjected to a battery of *tests*, each of which generates a number. The term "receptor" is applied to the part of the system that performs this first stage of processing. The output of

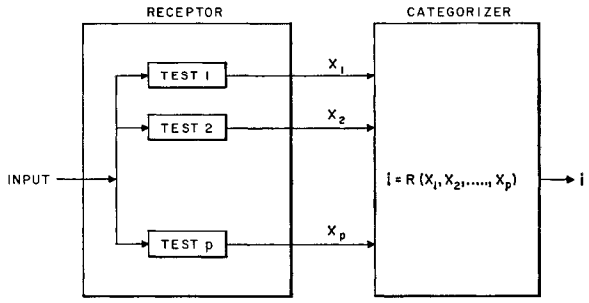

Fig. 1—Schematic representation of a recognition system.

the receptor may be represented as a p-dimensional column vector

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix},$$

where p is the number of tests, and the ith component X_i is the output of the ith test. As an aid to intuition, we may find it useful to think of each test as measuring some property of the input; accordingly we speak of the various X_i as measurements, and of $\mathbf{X}$ as a measurement vector.

In the second stage of the process, the measurement vector is assigned by a decision procedure to one of a finite number of categories. The term "categorizer" is applied to the part of the system that performs this second stage of processing. The output of the categorizer may be represented as an integer $i = 1, 2, \cdots, m$, under the convention that output j is to mean that the categorizer has assigned the input to the jth of the m permissible categories.

The theory of optimal categorizers has been discussed by the authors in an earlier paper [12]. Here, our concern will be with the first stage of the process, the receptor. Insofar as the present theoretical formulation is concerned, there is no limit to the effectiveness of a receptor in a given situation. There is, therefore, no theory of the "optimal" receptor to match the theory of the optimal categorizer. What we shall attempt to do is to understand what is meant by the effectiveness of a set of tests and how this effectiveness depends on the properties of the individual tests and their interrelations. We shall not be concerned with what appears to be the much more difficult problem of how an effective set of tests actually may be discovered.

Geometric Interpretation: Criterion of Effectiveness

A receptor consisting of p tests may be thought of as mapping each input into a point of a p-dimensional space. Corresponding to each input category there exists, then, a set of points in the space. Since we wish to consider each category as having potentially infinitely many members, and since we place no restriction on the range of the output of any test, points from any category may fall anywhere in the space. It is fruitful, therefore, to think of each category as generating a certain *density* of points at each location in the space, or, in other words, to think of the (multivariate) probability density function corresponding to each category. Now, it is quite clear that if the mapping generated by the tests is such that, at each location in the space, the density due to category i is equal to the density due to category j, then *no* categorizer will be able to discriminate between these two categories. The tests that generate these separate mappings are then ineffective. On the other hand, if, by this mapping, category i generates a high density in one region of the space while category j generates a high density in a quite different region, then it should be possible to discriminate between these two categories, and the tests are said to be effective.

Roughly speaking, then, we may say that a set of tests generates a probability density corresponding to each input category, and that the ability of the tests to discriminate between two given categories depends upon the "distance" between the densities. We wish to consider, therefore, the formulation of such a "distance" measure in terms of the parameters of the densities involved. This distance will then reflect the effectiveness of the tests.

We must first ask what requirements we wish such a measure of effectiveness to fulfill. There is, it would seem, only one sensible criterion: *the measure must reflect the error probability of the recognition system incorporating the tests in question.* To put it more specifically: let ρ be a set of tests which maps input category i into probability density p_i $(i = 1, \cdots, m)$; let P be a recognition system having ρ as receptor, and let e_{ij} be the probability that P will make an error in discriminating between category i and j (assuming these two categories to be *a priori* equally likely). We then assert that the relevant measure must be so formulated that the "distance" between p_i and p_j is large or small accordingly as e_{ij} is small or large.

Even this does not pin down the requirements sufficiently. The effectiveness of a recognition system depends, after all, not only on the tests used, but also on the adequacy of the decision-logic, or categorizer. Hence, the error probability e_{ij} is undefined so long as the cate-

gorizer is undefined. This difficulty may be circumvented in a theoretically satisfactory manner by agreeing that the categorizer in question be *optimal* with respect to ρ (in the sense of [12]).

In this view, then, the theory of the receptor is intertwined with the theory of the optimal categorizer. This implies that the former theory will be tractable only in those areas in which there is available an explicit formulation for the latter. Such a formulation was given in [12], following Anderson [1], for the special case in which the probability densities are normal. Accordingly, the discussion given below is also restricted to that case.

Normal Variables and Equal Covariance Matrices

Assume that we are dealing with a receptor which, for any i, maps inputs of category i into the (multivariate) normal density with mean $\mathbf{\mu}_i$ and with covariance matrix[1] $\mathbf{V}$. (This matrix being the same for all categories.) Thus, when the input is from category i, the measurement vector $\mathbf{X}$ has density (Anderson [1]):

$$p_i(\mathbf{X}) = \frac{1}{(2\pi)^{p/2}|\mathbf{V}|^{1/2}} \exp\left[-\tfrac{1}{2}(\mathbf{X} - \mathbf{\mu}_i)'\mathbf{V}^{-1}(\mathbf{X} - \mathbf{\mu}_i)\right].$$

Consider the quantity α_{ij} defined for two densities p_i and p_j as follows (Mahalanobis [11]):

$$\alpha_{ij} = (\mathbf{\mu}_i - \mathbf{\mu}_j)'\mathbf{V}^{-1}(\mathbf{\mu}_i - \mathbf{\mu}_j). \tag{1}$$

If $\mathbf{V}$ is the identity matrix, then α_{ij} represents the squared distance between the means of p_i and p_j. If $\mathbf{V}$ is not the identity, α_{ij} still has some aspects of a distance:

1) $\alpha_{ij} > 0$, $i \neq j$, $\mathbf{V}$ being positive definite
2) $\alpha_{ii} = 0$
3) $\alpha_{ij} = \alpha_{ji}$.

We wish to determine the relation between the quantity α_{ij} and the probability of error.

Let the random vector $\mathbf{X}$ have one or the other of two probability densities, p_i or p_j, accordingly as the input is from category i or j. The random variable defined by

$$L(\mathbf{X}) = \frac{p_i(\mathbf{X})}{p_j(\mathbf{X})} \tag{2}$$

is the likelihood ratio of the vector $\mathbf{X}$. Note that $L(\mathbf{X})$ is a scalar quantity.

If we consider that $\mathbf{X}$ has one or the other of two normal distributions, the one having mean $\mathbf{\mu}_i$ and covariance matrix $\mathbf{V}$, the other having mean $\mathbf{\mu}_2$ and the same covariance matrix, then the logarithm of the likelihood ratio is

$$\begin{aligned} u_{ij} &= \log_e L(\mathbf{X}) \\ &= \mathbf{X}'\mathbf{V}^{-1}(\mathbf{\mu}_i - \mathbf{\mu}_j) - \tfrac{1}{2}(\mathbf{\mu}_i + \mathbf{\mu}_j)'\mathbf{V}^{-1}(\mathbf{\mu}_i - \mathbf{\mu}_j). \end{aligned} \tag{3}$$

As was seen in [12], the optimal categorizer for the case of normal variables and equal covariance matrices calculates the functions u_{ij}. If a decision is to be made between category i and category j, it decides for the former if u_{ij} is positive, and for the latter if u_{ij} is negative (assuming that categories i and j are equally probable). Hence, the probability of error is given by

$$\begin{aligned} e_{ij} &= \tfrac{1}{2}\Pr(u_{ij} > 0 \mid \text{category } j) \\ &\quad + \tfrac{1}{2}\Pr(u_{ij} < 0 \mid \text{category } i). \end{aligned}$$

It is readily shown[2] that in this case the error probability is given by the following quantity (note the quantity α_{ij} in the lower limit of integration):

$$e_{ij} = \int_{\frac{1}{2}\sqrt{\alpha_{ij}}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\tfrac{1}{2}y^2\right] dy.$$

Hence, the probability of error e_{ij} is a monotonically decreasing function of α_{ij}. We notice that the function which relates the "distance" α_{ij} to the error probability e_{ij} is the univariate normal distribution with zero mean and unit variance. The effectiveness with which a receptor discriminates between categories i and j may thus be identified with the number α_{ij}, in the case of normal variables and equal covariance matrices.

The following interesting question may be raised: let ρ, consisting of tests $t_1, \cdots, t_p$, have effectiveness α_{ij} for discriminating between category i and j; how is this effectiveness changed if a new test t_{p+1} is added to the set? The answer to this question can be obtained directly from the definition of α_{ij}. The result is as follows:

Let $t_1, \cdots, t_p$ have effectiveness α_{ij}. Let t_{p+1} be a new test whose output has mean m_i or m_j accordingly as the input is from category i or j, and variance σ^2 in either case. Moreover, let $\mathbf{\gamma}$ be the vector of covariances that obtain between the output of t_{p+1} and the output of each of the other tests (*i.e.*, let the kth component of $\mathbf{\gamma}$ be the covariance between t_k and t_{p+1}). Then if α_{ij}^* is the effectiveness of the augmented set $t_1, \cdots, t_p, t_{p+1}$, we have that

$$\alpha_{ij}^* - \alpha_{ij} = \frac{[(m_i - m_j) - (\mathbf{\mu}_i - \mathbf{\mu}_j)'\mathbf{V}^{-1}\mathbf{\gamma}]^2}{\sigma^2 - \mathbf{\gamma}'\mathbf{V}^{-1}\mathbf{\gamma}}.$$

This quantity is the incremental effectiveness due to the addition of t_{p+1}. Note that when the output of t_{p+1} is uncorrelated with the output of the other tests, *i.e.*, when $\mathbf{\gamma}$ is the zero vector, then the incremental effectiveness is $(m_i - m_j)^2/\sigma^2$, which is precisely the effectiveness of test t_{p+1} alone. In general, however, the increment is seen to depend on the interrelations of all the tests in a fairly complex manner.

The Divergence

We should like to gain some intuitive apprcciation of the quantity α_{ij} and to generalize this quantity to other cases, particularly to the case of normal variables with unequal covariance matrices.

[1] $\mathbf{V}$ is taken to be a positive definite (symmetric) matrix of order $p \times p$, $\mathbf{x}$ and $\mathbf{\mu}_i$ are taken to be column matrices of order $p \times 1$. The inverse of a matrix $\mathbf{A}$ is indicated by $\mathbf{A}^{-1}$, the transpose by $\mathbf{A}'$, the determinant by $|\mathbf{A}|$, and the trace (*i.e.*, the sum of the diagonal elements) by tr $\mathbf{A}$.

[2] Anderson [1], p. 135.

Consider again the logarithm of the likelihood ratio for the case of normal variables with equal covariance matrices, as given by (3). On the assumption that the input is from category i, the expected value of u_{ij} is

$$E[u_{ij}|i] = \mathfrak{y}_i'\mathbf{V}^{-1}(\mathfrak{y}_i - \mathfrak{y}_j) - \tfrac{1}{2}(\mathfrak{y}_i + \mathfrak{y}_j)'\mathbf{V}^{-1}(\mathfrak{y}_i - \mathfrak{y}_j)$$
$$= \tfrac{1}{2}(\mathfrak{y}_i - \mathfrak{y}_j)'\mathbf{V}^{-1}(\mathfrak{y}_i - \mathfrak{y}_j). \quad (4)$$

Whereas, if the input is from category j, the corresponding value is

$$E[u_{ij}|j] = -\tfrac{1}{2}(\mathfrak{y}_i - \mathfrak{y}_j)'\mathbf{V}^{-1}(\mathfrak{y}_i - \mathfrak{y}_j). \quad (5)$$

Subtracting (5) from (4), we have

$$E[u_{ij}|i] - E[u_{ij}|j] = (\mathfrak{y}_i - \mathfrak{y}_j)\mathbf{V}^{-1}(\mathfrak{y}_i - \mathfrak{y}_j) = \alpha_{ij}.$$

Hence, we see that α_{ij} can be interpreted as the difference between the two expected values of the log likelihood ratio of $\mathbf{X}$, for the particular case of normally distributed variables with equal covariance matrices. A generalization of α_{ij} to the case of arbitrary distributions may be expressed by

$$E[\log L(\mathbf{X}) \,|i] - E[\log L(\mathbf{X})|\, j] \quad (6)$$

where $L(\mathbf{X})$ is defined by (2) for arbitrary densities p_i and p_j.

It has recently been argued by Kullback [10] that the quantity defined by (6) is indeed an appropriate "distance" between *arbitrary* distributions. This quantity (which is given very general definition by Kullback) is asserted to be "a measure of the difficulty of discriminating" between two distributions,[3] and given the name "divergence."

Divergence is a quantity defined for pairs of probability distributions or densities. In the context of recognition systems, such a pair of distributions comes into being as a result of applying a set of tests to two categories of inputs. We will write

$$\text{Div} \quad (i, j \mid t_1, \cdots, t_p)$$

to represent the divergence between the two distributions that are generated when tests $t_1, \cdots, t_p$ are applied to the members of categories i and j.

Certain properties of divergence may be noted (see [10] for a more rigorous setting, and for proofs).

1) Div $(i, j \mid t_1, \cdots, t_p) > 0, i \neq j$
2) Div $(i, i \mid t_1, \cdots, t_p) = 0$
3) Div $(i, j \mid t_1, \cdots, t_p) =$ Div $(j, i \mid t_1, \cdots, t_p)$
4) Divergence is additive for independent tests. That is, if, regardless of whether the input is from category i or j, $t_1, \cdots, t_p$ have statistically independent outputs, then

$$\text{Div}\,(i, j \mid t_1, \ldots, t_p) = \sum_{k=1}^{p} \text{Div}\,(i, j \mid t_k).$$

5) Adding new tests to a set of tests never decreases the divergence; *i.e.*,

$$\text{Div}\,(i, j \mid t_1, \ldots, t_p) \leq \text{Div}\,(i, j \mid t_1, \ldots, t_p, t_{p+1}).$$

[3] Kullback [10], p. 6.

Normal Variables and Unequal Covariance Matrices

Consider now the case in which the receptor maps the input categories into normal densities, no further restrictions being imposed. Thus, as in the case of normal variables with equal covariance matrices, we allow that the outputs of the tests be arbitrarily intercorrelated; here we allow further that the system of intercorrelations be different for inputs belonging to different categories.

Let p_i be the normal density with mean $\mathfrak{y}_i$ and covariance matrix $\mathbf{V}_i$ and let p_j be the normal density with mean $\mathfrak{y}_j$ and covariance matrix $\mathbf{V}_j$. The divergence for this case[4] is given by

$$\text{Div}\,(i, j \mid t_1, \cdots, t_p) = \tfrac{1}{2}\,\text{tr}(\mathbf{V}_i - \mathbf{V}_j)(\mathbf{V}_j^{-1} - \mathbf{V}_i^{-1})$$
$$+ \tfrac{1}{2}\,\text{tr}(\mathbf{V}_i^{-1} + \mathbf{V}_j^{-1})(\mathfrak{y}_i - \mathfrak{y}_j)$$
$$(\mathfrak{y}_i - \mathfrak{y}_j)'. \quad (7)$$

Notice that if we set $\mathbf{V}_i = \mathbf{V}_j$, we obtain

$$\text{Div}\,(i, j \mid t_1, \cdots, t_p) = \alpha_{ij}.$$

While there is in the present instance no simple function that relates divergence to the probability of error, certain bounds may be set. Thus, an examination[5] of the univariate case yields the bounds indicated in Fig. 2. For a given value of divergence, the probability of correct recognition (*i.e.*, one minus the probability of error) is constrained to lie between the two indicated curves. The upper curve, it may be noted, also gives the exact relation between probability of correct recognition and divergence for the multivariate normal case with equal covariance matrices.

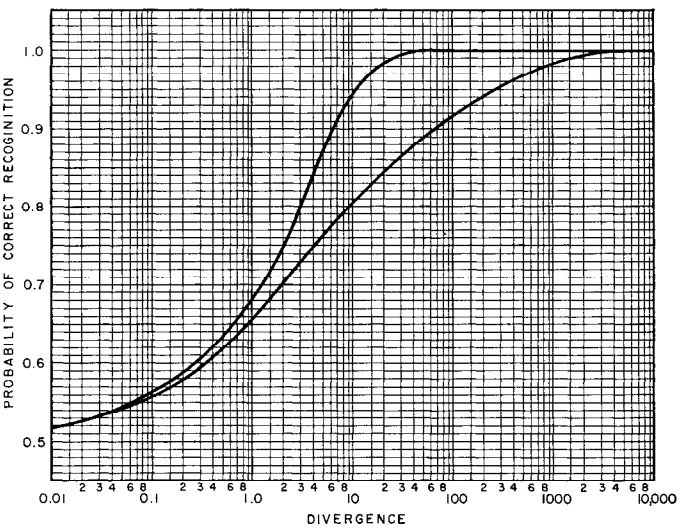

Fig. 2—Upper and lower bounds on probability of correct recognition as a function of divergence, for normal variables with unequal covariance matrices. Upper curve also represents the exact relation for normal variables with equal covariance matrices. Lower curve determined from univariate case.

[4] Kullback [10], p. 190.
[5] Performed with the aid of a Monte-Carlo-type computer program.

Some Experimental Results

Motivated by the preceding considerations, several experiments were performed. The raw data for these experiments were obtained by asking subjects to print the letters A, B, C and D by hand. Each of 40 subjects produced five examples of each of the four letters. Our entire sample, then, consisted of 800 individual characters. Each character was printed within a two-inch square.

Our tests consisted of determining the distance, along predetermined paths, from the edge of the square field to the edge of the character. These tests were performed by hand. Specifically, a grid of intersecting lines was overlaid on each character. The ith test ($i = 1, \cdots, 8$) consisted of determining the distance, along the ith line segment, from the edge of the field to the edge of the character.[6] Thus, each character yielded a vector of eight measurements. The entire sample, therefore, produced 800 vectors of eight measurements each. The four letters A, B, C and D represent the four input categories used in these experiments.

Two programs for the LGP-30 computer were prepared in connection with these experiments. The first program was used for computing the divergence, based on a specific set of tests, between specified input categories. Thus, for example, this program could be used to calculate Div (A, $C \mid t_1, t_3, t_5$), *i.e.*, to calculate the divergence between categories A and C as based on the first, third and fifth tests.

The divergence is calculated by this program in accordance with (7) (normal variables, unequal covariance matrices). The program demands the following quantities as inputs: the two "full" (8×1) mean vectors for the two categories in question, the two "full" (8×8) covariance matrices for the same categories and an indication of which particular tests the divergence is to be based on.

The second program was designed to be an optimal categorizer for the case of normal variables and unequal covariance matrices. The program could be set up to assign any vector $\mathbf{X}$ of p components ($p \leq 8$) to one of m categories ($m \leq 4$), given the appropriate m covariance matrices $\mathbf{V}_i$ (each of order $p \times p$) and the appropriate m mean vectors $\mathfrak{y}_i$ (each of order $p \times 1$). The rule of operation of this program was as follows:[7] assign vector $\mathbf{X}$ to category k if and only if $l_k(\mathbf{X})$ is the largest of the m quantities $l_i(\mathbf{X})$ given by

$$l_i(\mathbf{X}) = -(\mathbf{X} - \mathfrak{y}_i)' \mathbf{V}_i^{-1} (\mathbf{X} - \mathfrak{y}_i) - \log_e \mid \mathbf{V}_i \mid$$
$$i = 1, \cdots, m.$$

The covariances among the outputs of all eight tests were computed separately for each of the four categories, the resulting matrices being given in Table I(A–D). The corresponding four mean vectors are given in Table II.

[6] This method of analysis of characters is identical to that used in [12], which may be consulted for further details and illustrations.

[7] This rule is equivalent to, but computationally much simpler than, the one given in [12] for the case of normal variables and unequal covariance matrices.

TABLE I-A

Covariance Matrix for Letter A Based on Eight Tests

1.034	1.281	0.351	−0.293	0.098	0.301	0.141	1.336
	1.967	0.664	−0.219	0.259	0.556	0.276	2.094
		7.138	1.192	2.726	1.116	0.678	2.097
			2.269	1.367	0.146	0.201	−0.308
				5.727	1.280	0.933	2.107
					2.941	1.949	2.197
						1.577	1.229
							6.606

TABLE I-B

Covariance Matrix for Letter B Based on Eight Tests

4.792	4.417	4.244	2.406	1.798	0.790	0.785	2.993
	5.074	4.636	2.798	1.824	0.639	0.644	2.799
		5.428	3.224	2.111	0.903	1.131	2.943
			5.287	3.006	1.326	1.897	2.648
				3.574	2.229	2.471	1.915
					4.008	2.405	1.106
						4.507	1.727
							3.972

TABLE I-C

Covariance Matrix for Letter C Based on Eight Tests

1.638	2.153	1.482	1.695	−0.557	−2.443	−0.710	1.983
	3.596	2.461	2.436	−0.591	−3.711	−0.493	2.434
		2.500	2.834	−0.665	−2.621	0.248	1.738
			4.704	−0.629	−2.913	0.576	2.471
				19.000	0.896	8.622	−0.254
					5.856	1.357	−2.915
						20.800	−0.622
							3.214

TABLE I-D

Covariance Matrix for Letter D Based on Eight Tests

5.116	4.736	4.058	1.821	1.109	1.289	1.029	2.232
	5.684	4.523	2.311	1.273	1.328	1.151	2.425
		6.117	2.525	1.321	1.501	1.274	2.191
			4.432	2.481	2.179	1.080	1.784
				2.134	2.325	1.017	1.030
					4.099	2.019	1.803
						1.872	2.081
							3.806

TABLE II

Mean Measurement-Vectors Based on Eight Tests

μ'_A	= [7.825	6.750	5.835	8.525	6.615	7.065	7.865	4.435]
μ'_B	= [5.760	5.715	5.705	4.150	6.225	6.960	6.750	3.910]
μ'_C	= [6.610	5.060	5.980	3.975	9.020	14.685	10.640	4.175]
μ'_D	= [6.120	6.285	5.850	4.365	6.340	4.675	6.260	4.440]

Experiment 1

This experiment dealt with the discrimination between the letters A and B. First, various "effective" subsets of the set of eight tests were generated, by the following method: Calculate the divergence for the 8 possible subsets consisting of 7 tests each. Pick the subset yielding the largest divergence. Now, starting with the set of 7 tests already picked, calculate the divergence for the 7 possible

TABLE III

DISCRIMINATION OF A VS B : DIVERGENCE AND OBSERVED PERCENT OF CORRECT RECOGNITION FOR VARIOUS SETS OF TESTS

	Divergence	Percent Correct Recognition
All Tests	Div $(A, B\mid t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8) = 36.1$	Pr $(C) = 97.25\%$
Effective Subsets	Div $(A, B\mid t_1, t_2, t_3, t_4, t_5, t_6, t_7\ ) = 33.1$ Div $(A, B\mid t_1, t_2, t_3, t_4, \quad t_6, t_7\ ) = 29.8$ Div $(A, B\mid t_1, t_2, \quad t_4, \quad t_6, t_7\ ) = 25.2$ Div $(A, B\mid t_1, t_2, \quad t_4, \quad\quad t_7\ ) = 19.9$ Div $(A, B\mid t_1, t_2, \quad t_4, \quad\quad\quad\ ) = 19.0$ Div $(A, B\mid t_1, \quad\quad t_4, \quad\quad\quad\ ) = 11.9$ Div $(A, B\mid \quad\quad t_4, \quad\quad\quad\ ) = 6.53$	Pr $(C) = 95.75\%$ Pr $(C) = 93.5\ \%$ Pr $(C) = 91.5\ \%$ Pr $(C) = 89.75\%$
Ineffective Subsets	Div $(A, B\mid t_1, t_2, t_3, \quad t_5, t_6, t_7, t_8) = 21.2$ Div $(A, B\mid \quad t_2, t_3, \quad t_5, t_6, t_7, t_8) = 12.7$ Div $(A, B\mid \quad t_2, t_3, \quad t_5, t_6, \quad t_8) = 5.01$ Div $(A, B\mid \quad\quad t_3, \quad t_5, t_6, \quad t_8) = 1.40$ Div $(A, B\mid \quad\quad t_3, \quad t_5, t_6, \quad\quad\ ) = 0.582$ Div $(A, B\mid \quad\quad t_3, \quad\quad t_6, \quad\quad\ ) = 0.099$ Div $(A, B\mid \quad\quad t_3, \quad\quad\quad\quad\ ) = 0.041$	Pr $(C) = 62\ \%$ Pr $(C) = 58.75\%$ Pr $(C) = 54.75\%$
Individual Tests	Div $(A, B\mid t_1, \quad\quad\quad\quad\quad\ ) = 3.93$ Div $(A, B\mid \quad t_2, \quad\quad\quad\quad\ ) = 0.861$ Div $(A, B\mid \quad\quad t_3, \quad\quad\quad\quad\ ) = 0.041$ Div $(A, B\mid \quad\quad\quad t_4, \quad\quad\quad\ ) = 6.53$ Div $(A, B\mid \quad\quad\quad\quad t_5, \quad\quad\ ) = 0.148$ Div $(A, B\mid \quad\quad\quad\quad\quad t_6, \quad\ ) = 0.051$ Div $(A, B\mid \quad\quad\quad\quad\quad\quad t_7, \ ) = 1.14$ Div $(A, B\mid \quad\quad\quad\quad\quad\quad\quad t_8) = 0.188$	Pr $(C) = 54.75\%$ Pr $(C) = 89.75\%$

subsets consisting of 6 tests each. Pick the subset yielding the highest divergence; etc. In this manner, effective subsets of 7, 6, $\cdots$, 1 tests were obtained (see Table III).

Next, "ineffective" subsets of the set of 8 tests were generated by the following complementary method: Calculate the divergence for the 8 subsets of 7 tests each. Pick the subset yielding the *smallest* divergence. Now, starting with the set of 7 tests already picked, calculate the divergence for the 7 subsets of 6 tests each. Pick the subset yeilding the smallest divergence; etc. In this manner, ineffective subsets consisting of 7, 6, $\cdots$, 1 tests were obtained (see Table III).

For purposes of comparison, the divergence for each of the eight individual tests were also calculated (see Table III).

In the second phase of the experiment, eight of the selected sets of tests were each coupled, so to speak, to a categorizer, and each of the eight systems so formed was given the task of recognizing 200 samples of the letter A and 200 of the letter B.[8] The percentage of samples correctly recognized is given in the right side of Table III. The results, plotted in graphic form, are given in Fig. 3. The solid curves of Fig. 3 are those of Fig. 2.

[8] The particular choice of these eight sets was dictated by a desire to economize on computer time while still covering the total available range of divergences. The original plan called for obtaining the per cent correct recognition for the set consisting of all eight tests and for both the "effective" and "ineffective" sets consisting of 6, 4, 2 and 1 tests. This plan was carried out with the exception that the "ineffective" set of six tests was omitted, since its divergence nearly duplicated that of another set which had already been run (the "effective" set of two tests).

Experiment 2

In the preceding experiment the nature of the discrimination, A vs B, was held constant, and the receptor of the system was varied. In the present experiment, the receptor consisting of tests 4 and 6 was used throughout, and the discrimination was varied. All six possible two-fold discriminations among the four letters were investigated: A vs B, A vs C, A vs D, B vs C, B vs D and C vs D. The results are plotted in Fig. 4, each point being based on 400 samples.

Experiment 3

As a matter of general interest, although not directly related to the preceding discussion, the categorization program was set up for the four-fold discrimination: A vs B vs C vs D, based (as in Experiment 2) on tests 4 and 6. Two hundred samples of each letter were used. The results are given in Table IV.

Discussion

Several aspects of these results may be noted. Experiment 1 shows that, for any number n, the set of n tests having high divergence yields higher percentage of correct recognition than the set of n tests having low divergence. In fact, the rank ordering of various sets of tests by divergence is equal to their rank ordering by percentage of correct recognition.

From the graphic representations of Figs. 3–4, we see that the points, except at the low end of the graph where sampling error may be expected to be greater, fall within the curves of Fig. 2. For the type of situation under

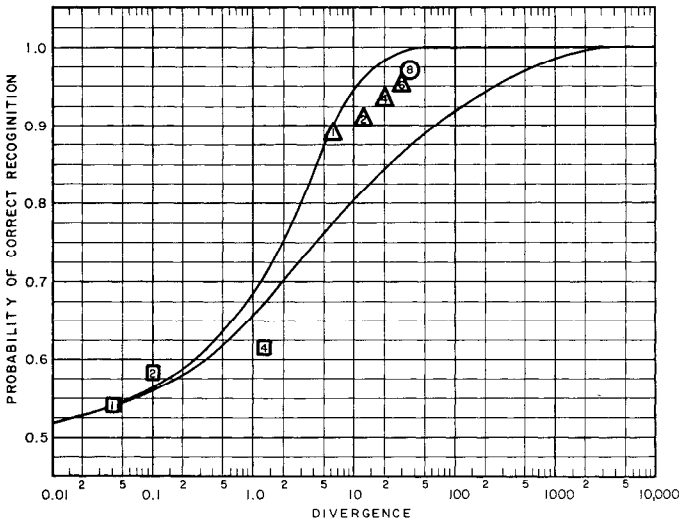

Fig. 3—Discrimination of A vs B. Observed probability of correct recognition as a function of divergence, for eight receptors investigated in Experiment 1. The solid curves are those of Fig. 2. The triangular and square points represent, respectively, so-called "effective" and "ineffective" subsets. The number within the points indicate the number of tests used in the discrimination.

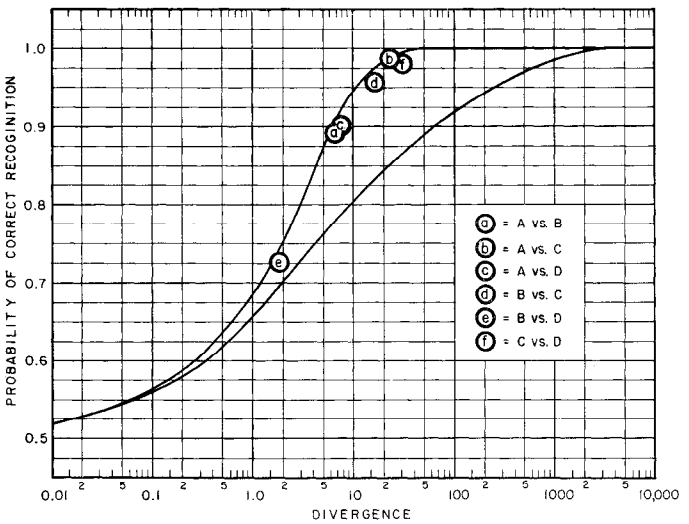

Fig. 4—Discrimination using a receptor consisting of tests 4 and 6. Observed probability of correct recognition as a function of divergence, for the six two-fold discriminations investigated in Experiment 2. Solid curves are those in Fig. 2.

TABLE IV

RESULTS OF EXPERIMENT 3 : FOUR-FOLD DISCRIMINATION, USING TESTS 4 AND 6

		INPUT				
		A	B	C	D	
OUTPUT	A	184	27	1	30	242
	B	9	117	7	41	174
	C	0	7	189	0	196
	D	7	49	3	129	188
		200	200	200	200	800

discussion, then, these curves may be used to make approximate predictions of the ability of a set of tests to discriminate between given categories. Thus, for example, we may predict on the basis of divergence measurements that tests 4 and 6 are adequate for the discrimination of C vs D, but quite unsatisfactory for the discrimination of B vs D. Fig. 4 shows that this prediction is borne out.[9]

REFERENCES

[1] T. W. Anderson, "Introduction to Multivariate Statistical Analysis," John Wiley and Sons, Inc., New York, N. Y., 1958.
[2] W. W. Bledsoe and I. Browning, "Pattern recognition and reading by machine," *Proc. Eastern Joint Conf.*, pp. 225–232; December, 1959.
[3] J. S. Bomba, "Aplha-numeric character recognition using local operations," *Proc. Eastern Joint Conf.*, pp. 218–224; December, 1959.
[4] C. K. Chow, "Optimum character recognition system using decision functions," 1957 IRE WESCON CONVENTION RECORD, pt. 4, pp. 121–129. Also in IRE TRANS. ON ELECTRONIC COMPUTERS, vol. EC-6, pp. 247–254; December, 1957.
[5] W. Doyle, "Recognition of sloppy hand-printed characters," *Proc. Western Joint Computer Conf.*, pp. 133–142; May, 1960.
[6] R. L. Grimsdale, F. H. Summer, C. J. Tunis and T. Kilburn, "A system for the automatic recognition of patterns," PROC. IEE, vol. 106, pt. B, March, 1959.
[7] L. D. Harmon, "A line-drawing recognizer," *Proc. Western Joint Computer Conf.*, pp. 351–364; May, 1960.
[8] J. K. Hawkins, "Self-organizing systems—a review and commentary," PROC. IRE, pp. 31–48; January, 1961.
[9] W. H. Highleyman, "An analog method for character recognition," IRE TRANS. ON ELECTRONIC COMPUTERS, vol. EC-10, p. 502–512; September, 1961.
[10] S. Kullback, "Information Theory and Statistics," John Wiley and Sons, Inc., New York, N. Y., 1959.
[11] P. C. Mahalanobis, "On the generalized distance in statistics," *Proc. Nat'l. Inst. Sci. India*, vol. 12, pp. 49–55; 1936.
[12] T. Marill and D. M. Green, "Statistical recognition functions and the design of pattern recognizers," IRE TRANS. ON ELECTRONIC COMPUTERS, vol. EC-9, pp. 472–477; December, 1960.
[13] M. Minsky, "A selected descriptor-indexed bibliography to the literature on artificial intelligence," IRE TRANS. ON HUMAN FACTORS IN ELECTRONICS, vol. HFE-2, pp. 39–56; March, 1961.
[14] M. Minsky, "Steps toward artificial intelligence," PROC. IRE, vol. 49, pp. 8–30; January, 1961.
[15] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain," *Psychol. Rev.*, vol. 65, pp. 386–407; November, 1958.
[16] G. S. Sebestyen, "Recognition of membership in classes," IRE TRANS. ON INFORMATION THEORY, vol. IT-7, pp. 44–50; January, 1961.
[17] O. G. Selfridge, "Pandemonium: A paradigm for learning," *Proc. Symp. on Mechanisation of Thought Processes*, H.M.S.O., London, England, pp. 511–529; 1959.
[18] H. Sherman, "A quasi-topological method for machine recognition of line patterns," *Proc. Internatl. Conf. on Information Processing (ICIP)*, UNESCO House, Paris, France, pp. 232–238; 1959.
[19] Smith, J. E. Keith, and L. Klem, "Vowel recognition using a multiple discriminant function," *J. Acoust. Soc. Am.*, vol. 33, p. 358; March, 1961.
[20] L. Uhr and C. Vossler, "A pattern recognition program that generates, evaluates, and adjusts its own operators, *Proc. Western Joint Computer Conf.*, Los Angeles, Calif., pp. 555–561; May 9–11, 1961.
[21] S. H. Unger, "Pattern detection and recognition," PROC. IRE, vol. 47, pp. 1737–1752; October, 1959.
[22] P. D. Welch, and R. S. Wimpress, "Two multivariate statistical computer programs and their application to the vowel recognition problem," *J. Acoust. Soc. Am.*, vol. 33, pp. 426–434; April, 1961.

[9] It may also be of interest to observe in Table III that, of the eight possible subsets consisting of a single test, the method for picking "effective" subsets did indeed pick the one having the highest divergence, while the method for picking "ineffective" subsets did indeed pick the one with the lowest divergence. Unfortunately, situations can be devised in which this will not be the case. It may be conjectured that such situations are of a pathological sort; the matter is far from clear, however.