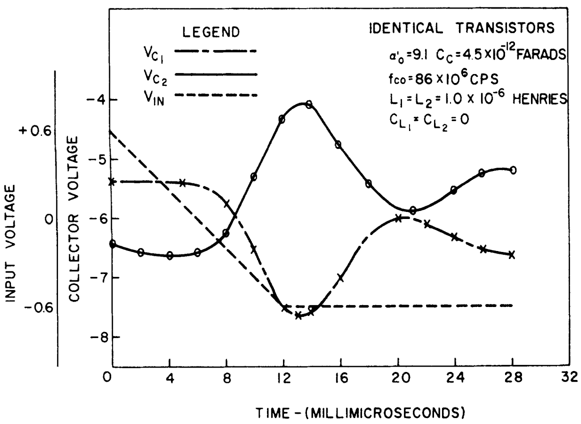

Fig. 6—Unloaded response of the circuit in Fig. 3.

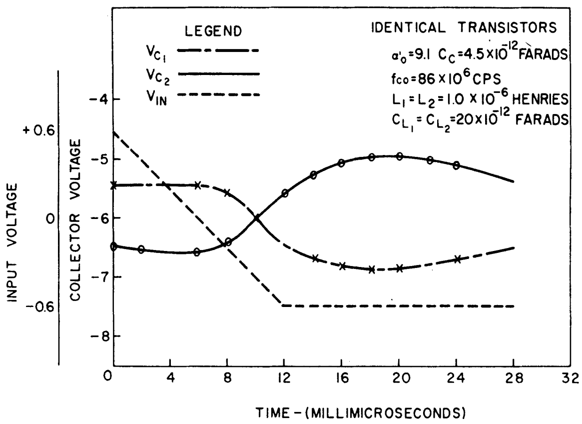

Fig. 7—Loaded response of the circuit in Fig. 3.

ibility of such an approach in determining the response of transistor switching circuits. It should not be inferred that the method is a substitute for the preliminary design of circuits, or a means of gaining insight into the operation of a circuit. It should be considered instead as a substitute for the final bench setup used to obtain data. The program needs further work before its use can be widely advocated. At present the program requires 15 to 20 minutes to obtain a complete response, and its capacity is a circuit containing up to ten transistors. The next phase of this program will concentrate on methods of reducing the computation time and making the routine more flexible.

Acknowledgment

The work reported here is being done at the IBM Poughkeepsie Product Development Laboratory. The author wishes to acknowledge the valuable assistance of the transistor computation group in completing this phase successfully.

An Optimum Character Recognition System Using Decision Functions*

C. K. CHOW†

Summary—**The character recognition problem, usually resulting from characters being corrupted by printing deterioration and/or inherent noise of the devices, is considered from the viewpoint of statistical decision theory. The optimization consists of minimizing the expected risk for a weight function which is preassigned to measure the consequences of system decisions. As an alternative, minimization of the error rate for a given rejection rate is used as the criterion. The optimum recognition is thus obtained.**

The optimum system consists of a conditional-probability densities computer; character channels, one for each character; a rejection channel; and a comparison network. Its precise structure and and ultimate performance depend essentially upon the signals and noise structure.

Explicit examples for an additive Gaussian noise and a "cosine" noise are presented. Finally, an error-free recognition system and a possible criterion to measure the character style and deterioration are presented.

* Manuscript received by the PGEC, June 3, 1957.
† Burroughs, Corp., Paoli, Pa.

Introduction

Character recognition has been receiving considerable attention as the result of the phenomenal growth of office automation and the need for translating human language into machine language.[1,2] Broadly speaking, the character printed in conventional form and size on the document (checks, etc.) is first converted to electrical signals, and sufficient information is then extracted from the latter. The purpose of the recognition system is based on the observed data and on *a priori* knowledge of the signal and noise structure

[1] K. R. Eldredge, F. J. Kamphoefner, and P. H. Wendt, "Automatic input for business data processing system," *Proc. Eastern Joint Computer Conf.*, pp. 69–73; December 11, 1956.

[2] E. C. Greanias and Y. M. Hill, "Considerations in the design of character recognition devices," 1957 IRE National Convention Record, pt 4, vol. 5 pp. 119–126.

to identify which of the possible characters is present, or to reject if the data are ambiguous.

The over-all performance of the recognition system depends not only upon itself, but also upon the number of characters to be recognized, the character style, and noise statistics. In this paper the character style and noise statistics are assumed given and adequate, and the purpose of the paper is to obtain an optimum recognition system. For convenience, the recognition problem is considered one of statistical inference, so that useful results in decision theory can be applied.[3–5] To accomplish this, the notion of risk is employed and proper weights are assigned to various types of error, rejection, and correct recognition to measure the consequences of decisions. This results in an optimum system which minimizes the expected (average) risk function and includes a possible alternate system with a minimum error rate. The results reveal the explicit structure of an optimum system which is determined by the *a priori* noise statistics, the signal structure, and the preassigned weights.

System Approach to the Problem

One practical application of a character recognition system for business documents is to read arabic numerals and selected symbols printed in magnetic ink. A method[1] for achieving this is shown in Fig. 1. The character is first passed through the field of a permanent magnet where it is magnetized in a given direction before being scanned by the read head. From the read head, the printed character is converted into an electrical signal corresponding to the differentiation of the plane area of the character. The function of the recognition system is to examine the amplitude-time signal obtained by the read head and to decide which of the possible characters is being recognized.

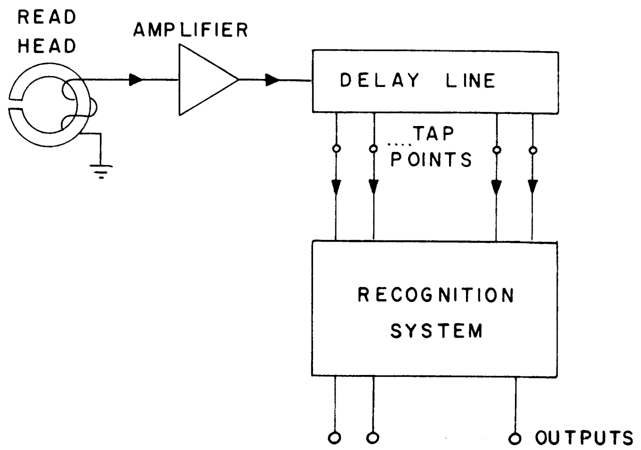

Fig. 1—A recognition system.

It is convenient, at times, to deal with the sampled data rather than the continuous time waveforms. By the sampling theory, if the number of samples is sufficiently large, little information carried by the continuous signal is lost. As shown in Fig. 1, the signal from the read head is first amplified and then fed into a tapped delay line. This serves as a means for sampling and acts as a temporary storage device to convert the series information into parallel information. Although not essential, sampled data are used in the following discussion.

Let the vector $v=(v_1, v_2, \cdots v_s)$ (subscript s being the number of samples) denote voltages on the taps of the delay line at the instant of sampling. (See Appendix I for the meaning of the Symbols.) The vector $a_i=(a_{i1}, a_{i2}, \cdots, a_{is})$ denotes the true sampled signal associated with the ith character where $i=1, 2, \cdots, c$, c being the number of possible characters to be recognized. The vector v constitutes the input to the recognition system. It is assumed that the characters are distinct, *i.e.*, all a_i's are different.

In a simple form, the recognition system may consist of c separate channels, one for each character. Each channel obtains a weighted sum of v_i's, with properly chosen weights, b_{ij}. The output of the ith channel is

$$X_i(v) = \sum_{j=1}^{s} b_{ij}v_j. \tag{1}$$

This operation may be realized by a summing amplifier and possibly with some inverters to provide negative weights, if required. One possible set of weights is:

$$b_{ij} = \frac{a_{ij}}{\left[\sum_{j=1}^{s} a_{ij}^2\right]^{1/2}}. \tag{2}$$

The recognition system is known as a correlation network when the weights are defined by (2).

If the printing is perfect, and the reading devices are noiseless, the observed data v will be identical to one of the a_i's and therefore, it can easily be shown that the right channel of the correlation network has the largest (algebraic) output. Consequently, the recognition system identifies the character with absolute accuracy by the channel having the highest output. However, in practice, there are always, to some degree, deteriorations in printing and inherent noise in the devices. Therefore, the observed data v generally will not be identical to any of the a_i's. In view of this, ambiguities arise which may result in possible misrecognition. To safeguard against the occurrence of error, the recognition system should have provisions for examining the degree of ambiguity and making rejects when required. This function can be achieved in various ways; *e.g.*, whenever the next highest output of the correlation network exceeds some preassigned fraction of the highest output, the system will reject, otherwise the system

[3] A. Wald, "Statistical Decision Functions," John Wiley & Sons, Inc., New York, N. Y., 1950.
[4] D. Van Meter and D. Middleton, "Modern statistical approaches to reception in communication theory," IRE Trans., vol. PGIT-4, pp. 119–145; September, 1954.
[5] D. Middleton and D. Van Meter, "On optimum multiple-alternative detection of signals in noise," IRE Trans., vol. IT-1, pp. 1–9; September, 1955.

identifies the character by the channel having the highest output.

The system described above merely represents one of many possible recognition systems and is not necessarily optimum. A basic problem in the design of recognition systems is to evaluate the system performance in the presence of printing deterioration and inherent noise and to obtain an optimum system. Optimum performance depends primarily upon the character style and permissible deterioration. Greanias and Hill in a recent paper[2] describe the effects of character style and printing deterioration on the character recognition problem from the viewpoint of matching the character with an ideal character and further propose definitions for character quality and style factors. In this paper, the discussion is confined to the problem of obtaining an optimum recognition system for a given set of adequately styled characters and known statistics of character deterioration. The recognition problem is considered to be that of testing multiple hypotheses in the statistical inference. Consequently, the design and evaluation of a recognition system is comparable to a statistical test. Results of decision theory can be applied.[3–5].

In order to judge the relative merit of recognition systems, some criterion of evaluating system performance must be established. The error rate of the system for a given rejection rate is used as the performance criterion for cases where no distinction is made among misrecognitions. Cases may arise where different misrecognitions have different consequences; *e.g.*, the registering of a four as an asterisk may not be as serious an error as registering it as a nine. The criterion of minimum error rate is then no longer appropriate. Instead, the criterion of minimum risk[3] is employed. Proper weights are assigned to measure the consequences of errors, rejections, and correct recognitions. These weights indicate the loss incurred by the system for every possible decision. The loss, which should be regarded as negative utility, may actually represent loss in dollars or unit of utility in measuring the consequence of the system decision. The over-all performance of the system is judged by its expected (or average) risk.

In the following discussion, an optimum system which minimizes the expected risk is derived, and a system having minimum error rate is obtained. Examples are presented for illustration purposes. An error-free system and a possible criterion for judging character style and deterioration are also presented.

The Expected Risk

The vector $a_i = (a_{i1}, a_{i2}, \cdots, a_{is})$ in the s-dimensional space denotes the true sampled signal associated with the ith character $(i = 1, 2, \cdots, c)$, where c and s are respectively, the number of possible characters to be recognized and the number of samples. Let $p = (p_1, p_2, \cdots, p_c)$ be the *a priori* distribution of characters (p_i is the *a priori* probability that the ith character occurs). Then, evidently,

$$\sum_{i=1}^{c} p_i = 1, \qquad p_i > 0. \tag{3}$$

The received data are denoted by a s-components vector $v = (v_1, v_2, \cdots, v_s)$. It is the signal corrupted by factors such as the deterioration of printing and inherent noise of the devices. *A priori* noise statistics and the manner in which various signals and noise are combined determine precisely the conditional probability density $F(v \mid a_i)$ of the observed data v when a_i is the incoming signal.

The space of decisions available to the recognition system consists of $c+1$ possible decisions $d_0, d_1, d_2, \cdots, d_c$. The quantity $d_i (i \neq 0)$ is the decision that the ith character is present while d_0 is the decision for reject. A basic problem in statistical decision theory is the selection of a proper decision rule δ. The rule is expressed as a vector function of the data v, namely, $\delta(v) = (\delta(d_0 \mid v), \delta(d_1 \mid v), \delta(d_2 \mid v) \cdots \delta(d_c \mid v))$ with $c+1$ components, and satisfies the restriction that:

$$\sum_{i=0}^{c} \delta(d_i \mid v) = 1 \qquad \text{for all } v, \tag{4}$$

and

$$\delta(d_i \mid v) \geq 0 \qquad \text{for all } i \text{ and } v. \tag{5}$$

The quantity $\delta(d_i \mid v)$ is the probability that, for a given observed data v, the decision d_i will be made.

In order to judge the relative merits of the decision rules it is necessary to assign the weight function $W(a_i, d_j)$. This is a function of a_i and d_j, which is the loss incurred by the system if the decision d_j is made when a_i is the true signal. This measure of consequence for various d_j under various a_i is a datum of the problem and is given in advance. Let the weight function be:

$$W(a_i, d_j) = w_{ij} \qquad \begin{array}{l} i = 1, 2 \cdots c \\ j = 0, 1, 2 \cdots c, \end{array} \tag{6}$$

where $w_{ii} (i \neq 0)$ is the weight of correct recognition of the ith character; $w_{ij} (i \neq j \neq 0)$ is the weight of misreading the ith character as the jth one, and $w_{i0} (i \neq 0)$ is the weight of rejecting the ith characters. Therefore, it is required that

$$w_{ij} > w_{i0} > w_{ii} \qquad (i \neq j \neq 0). \tag{7}$$

Usually, w_{ij} is much larger than w_{i0} since the most serious consideration in design of a character recognition system is the occurrence of undetected errors.

In general, w_{ij}'s may all differ, so that various misrecognitions, rejections, and correct recognitions can be properly weighted. The expected risk for any decision rule δ is

$$R(p, \delta) = \sum_{i=1}^{c} \sum_{j=0}^{c} \int_V \delta(d_j \mid v) p_i w_{ij} F(v \mid a_i) dv, \tag{8}$$

with integration over the entire observation space V.

The Minimum Risk System

The problem is then to choose a decision rule to minimize the average risk. By using (4), and since $\int_V F(v|a_i)dv = 1$ for all i, (8) may be written as:]

$$R(p, \delta) = R_0(p) + R_1(p, \delta), \tag{9}$$

where

$$R_0 = \sum_{i=1}^{c} p_i w_{i0}, \tag{10}$$

$$R_1 = \int_V \sum_{j=0}^{c} \delta(d_j | v) Z_j(v) dv, \tag{11}$$

and

$$Z_j(v) = \begin{cases} \sum_{i=1}^{c} (w_{ij} - w_{i0}) p_i F(v | a_i); \; j = 1, 2 \cdots c \\ 0 \qquad \text{for } j = 0. \end{cases} \tag{12}$$

The symbol R_0 will express the expected risk when rejection is made for all recognition and R_1 is that part of R which may be adjusted through the choice of δ. Evidently

$$R_1(p, \delta) \geq \int_V \min_j [Z_j(v)] dv, \tag{13}$$

and the equality sign holds if, and only if, the decision rule is chosen as:

$$\delta^*(d_k | v) = 1,$$
$$\delta^*(d_j | v) = 0 \qquad \text{for all } j \neq k \tag{14}$$

whenever

$$\min_j [Z_j(v)] = Z_k(v). \tag{15}$$

This is the optimum decision rule δ^* (the Bayes strategy) which minimizes the expected risk and is nonrandomized since its components are either zero or one. Therefore, R_1 for this decision rule is always nonpositive, and its expected risk (the Bayes risk) is no greater than R_0. The expected risk for the optimum decision rule, δ^*, is

$$R(p, \delta^*) = \sum_{i=1}^{c} p_i w_{i0} + \int_V \min_j [Z_j(v)] dv. \tag{16}$$

Eqs. (14) and (15) reveal that the optimum system for character recognition consists of a computer which evaluates $F(v|a_i)$'s; $(i = 1, 2, \cdots, c)$ for an observed data v; computes the various $Z_j(v)$ $(j = 1, 2, \cdots, c)$; examines and compares these $Z_j(v)$ $(j = 0, 1, 2, \cdots, c)$; selects the smallest (algebraically) one, say $Z_k(v)$; and finally makes the decision d_k [having the same subscript as $Z_k(v)$]. Of course, this method of setting up the computing procedures is not unique; *e.g.*, any ordering-preserving transformation may be used. In any event, the system must be equivalent to the above.

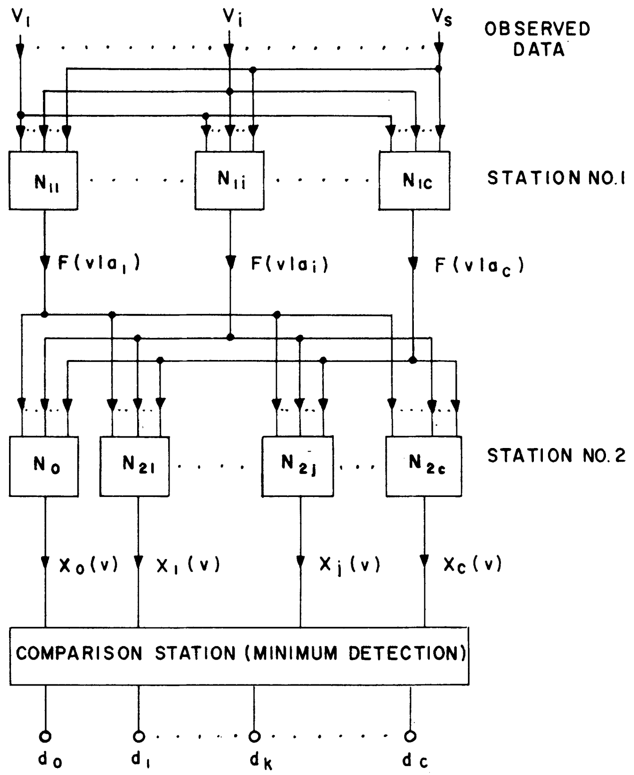

Fig. 2—Functional diagram of an optimum system.

The functional diagram of the optimum system is shown in Fig. 2. Station No. 1 consists of c similar component networks. Each network receives the observed data $v = (v_1, v_2, \cdots, v_s)$ and computes the corresponding conditional probability density $F(v|a_i)$ as its output. This operation depends only upon the *a priori* knowledge of signal and noise structure and on the observed data v; it does not depend upon the weight function $W(a_i, d_j)$ or on *a priori* probability distribution of signals, p.

The outputs of station No. 1 are fed to station No. 2, which consists of c character channels $N_{2j}(j \neq 0)$, and one rejection channel, N_0. They perform the linear operation of weighting each input and then the summing of all weighted inputs. The weights are $p_i w_{ij}$'s and $p_i w_{i0}$'s. The output of the rejection channel is

$$X_0(v) = \sum_{i=1}^{c} w_{i0} p_i F(v | a_i), \tag{17}$$

while the outputs of character channels are

$$X_j(v) = \sum_{i=1}^{c} w_{ij} p_i F(v | a_i) \quad (j = 1, 2 \cdots c). \tag{18}$$

The comparison station receives X's from station No. 2, examines all its inputs, and makes decision by selecting the algebraically smallest of the $c+1$ X's. If the rejection channel has the smallest one, the system rejects. If one of the character channels has the smallest output (say $X_k(v)$, $(k \neq 0)$) then the system recognizes the signal as the kth character.

Since $X_j(v)$ $(j = 0, 1, \cdots, c)$ is equal to $Z_j(v) + X_0(v)$, this system makes decisions in accordance with δ^* as defined by (14) and (15), and thus minimizing the expected risk.

Probabilities of Error and Rejection

The expected risk provides a means for evaluating the performance of a recognition system. At times, it may be desirable to compute the probabilities of error, rejection, and correct recognition as an auxiliary set of merit figures. They are obtained for any decision rule δ as follows:

The probability of correct recognition is:

$$P_c(\delta) = \int_V \sum_{i=1}^{c} \delta(d_i \mid v) p_i F(v \mid a_i) dv; \tag{19}$$

the probability of rejection, or rejection rate, is:

$$P_r(\delta) = \int_V \delta(d_0 \mid v) \sum_{i=1}^{c} p_i F(v \mid a_i) dv; \tag{20}$$

and the probability of misrecognition, or error rate, is:

$$P_e(\delta) = 1 - P_c - P_r. \tag{21}$$

Eqs. (19)–(21) result directly from the fact that $\int_V \delta(d_0|v)F(v|a_i)dv$ and $\int_V \delta(d_i|v)F(v|a_i)dv$ are, respectively, the conditional probabilities of rejection of the ith character and correct recognition of the ith character.

Criterion of Minimum Error Rate

Cases may arise in which the criterion of judging the system performance is the magnitude of its error rate for a given rejection rate. In using this criterion, the optimum recognition system is the one which, for a given rejection rate, α, has a minimum error rate. The optimum decision rule is obtained as: (See Appendix II for proof.)

$$\delta^{**}(d_k \mid v) = 1 \qquad (k \neq 0) \tag{22}$$

whenever

$$p_k F(v \mid a_k) \geq p_j F(v \mid a_j) \qquad \text{for all } j \neq k, \text{ and}$$

$$p_k F(v \mid a_k) \geq \beta \sum_{i=1}^{c} p_i F(v \mid a_i), \tag{23}$$

and

$$\delta^{**}(d_0 \mid v) = 1, \tag{24}$$

whenever

$$\beta \sum_{i=1}^{c} p_i F(v \mid a_i) > p_j F(v \mid a_j) \quad \text{for all } j (j = 1, 2 \cdots c), \tag{25}$$

where $\beta (0 \leq \beta \leq 1)$ is a nonnegative constant determined by the condition that $P_r(\delta^{**}) = \alpha$; namely,

$$\int_V \delta^{**}(d_0 \mid v) \sum_{i=1}^{c} p_i F(v \mid a_i) dv = \alpha. \tag{26}$$

The constant β increases with increasing α, and $P_e(\delta^{**})$ and $P_c(\delta^{**})$ are monotonic decreasing functions of α.

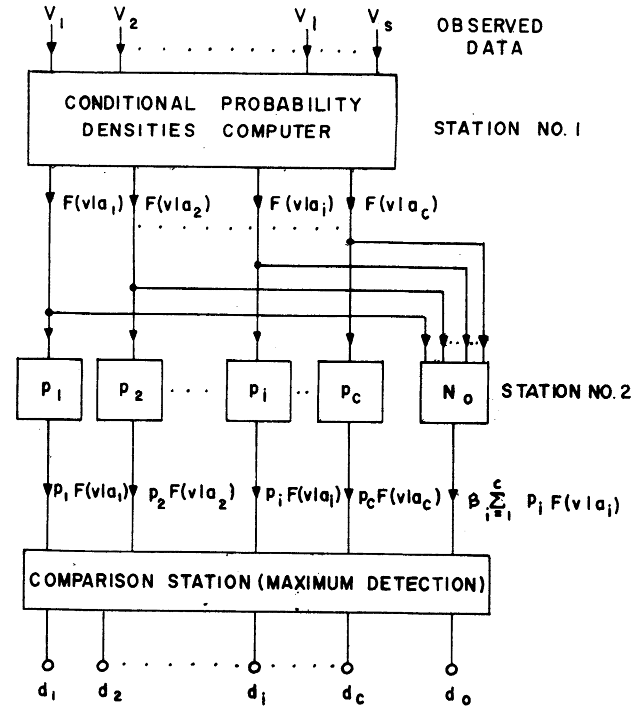

Fig. 3—Functional diagram of the system having the minimum error rate.

The change in the constant β provides a control over the error-reject ratio.

The optimum rule δ^{**} provides the basis for the functional diagram of the system of minimum error rate as shown in Fig. 3. The first station is identical to that of the minimum risk system (see Fig. 2) which computes the conditional probability densities $F(v|a_i)$'s.

The second station for this system is somewhat similar to that shown in the functional diagram for the minimum risk system (see Fig. 2). The c-character channels perform the weighting operation and have $p_i F(v|a_i)$ as outputs. The rejection channel, N_0, performs the operation of weighting and summing and has $\beta \sum_{i=1}^{c} p_i F(v|a_i)$ as its output. All of these $c+1$ outputs are nonnegative. The comparison station then examines these outputs and selects the largest. If the output of the rejection channel is the largest, the system rejects; otherwise the system will identify the character by the channel having the largest output.

It can be shown that the system depicted in Fig. 2 reduces to the system shown in Fig. 3 when β is replaced by $(w_m - w_r)/(w_m - w_c)$, and the following weight function is used.

$$W(a_i, d_j) = \begin{cases} w_c & \text{for } i = j \neq 0 \\ w_r & \text{for } i \neq 0, j = 0. \\ w_m & \text{for } i \neq j \neq 0. \end{cases} \tag{27}$$

Examples

1) As an illustration, consider a condition where the signals and noise are additive, and the noise has independent normal distribution. To be explicit, the probability density function of the noise of the jth component of the ith character is taken as:

$$\frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp\left\{-\frac{(v_j - a_{ij})^2}{2\sigma_{ij}^2}\right\}, \tag{28}$$

where $v_j - a_{ij}$ is the noise and σ_{ij}^2 is the given variance.

The conditional probability density $F(v|a_i)$, under the assumption that noise is statistically independent, is

$$F(v \mid a_i) = \frac{\exp\left\{-\sum_{j=1}^{s} \frac{(v_j - a_{ij})^2}{2\sigma_{ij}^2}\right\}}{(2\pi)^{s/2} \prod_{j=1}^{s} \sigma_{ij}}. \tag{29}$$

Therefore, the last expression dictates the precise structure of station No. 1 for the optimum system. Each component circuit performs the operations of taking differences, squaring, weighting, summing, and taking exponential. This station is common for a minimum risk or minimum error rate system. The structures of the second station and comparison station are indicated in Figs. 2 and 3.

2) In this example, the signal and noise structure are such that the conditional probability density for a given length, $|v|$, of v is directly proportional to the cosine of the angle θ between vectors v and a_i for $|\theta| < \pi/2$ and is zero elsewhere, and that the distributions of $|v|$ for given a_i are identical for all i.[6] [It is denoted as $f(|v|)$.] In other words, $F(v|a_i)$ can be written as

$$\begin{aligned} F(v \mid a_i) &= \rho f(|v|) \frac{a_i \cdot v}{|a_i| \cdot |v|} \quad \text{for } a_i \cdot v > 0 \\ &= 0 \qquad\qquad\qquad\qquad \text{elsewhere}, \end{aligned} \tag{30}$$

where ρ is a constant independent of i and is determined by the fact that $\int_V F(v|a_i)dv = 1$, and $a_i \cdot v$ denotes the scalar product of vectors a_i and v.

An inspection of the optimum decision rule (δ^* or δ^{**}) reveals that the system remains optimum if the first station is to compute $T[F(v|a_i)]$ instead of $F(v|a_i)$, where T is defined as

$$\begin{aligned} T(v \mid a_i) &= \frac{|v|}{\rho f(|v|)} F(v \mid a_i) \\ &= \begin{cases} \dfrac{a_i}{|a_i|} \cdot v = \sum_{j=1}^{s} b_{ij} v_j & \text{for } a_i \cdot v > 0 \\ 0 & \text{for } a_i \cdot v \le 0, \end{cases} \end{aligned} \tag{31}$$

where b_{ij}'s are constants [see (2)]. This operation can be easily realized. Each component of the first station is simply a correlation network followed by a half-wave rectifying circuit. The circuit passes the positive output of the correlation network unaltered and converts its negative output into zero.

The above results also indicate that the recognition system described in the second section of this paper is not optimum for the particular signal and noise structure as given in examples 1 or 2.

[6] This particular signal and noise structure was suggested by the author's colleague, I. M. Sheaffer, Jr., Burroughs Corp., Paoli, Pa.

Error-Free System

For convenience, let V_i denote the set (or region) of all possible observation v when the ith character is present, and let $\tilde{V}_i$ be the largest subset of V_i so that $\tilde{V}_i$'s are nonoverlapping. If noise distributions are so truncated and the signal vectors a_i's are so placed that all $\tilde{V}_i$'s are nonempty, then an error-free system for character recognition does exist. Evidently for any observed data v belonging to $\tilde{V}_k$, only $F(v|a_k)$ is nonzero while all others are zero; the character can then be identified with certainty. On the other hand, if the data v do not belong to any one of the $\tilde{V}_i$'s, then more than one of the $F(v|a_i)$'s will be nonzero. This results in the data being ambiguous for recognition purpose, and an error-free system will reject. Symbolically, the error-free decision rule is:

$$\delta(d_k \mid v) = 1 \quad \text{if } F(v \mid a_k) > 0$$

and

$$F(v \mid a_i) = 0 \quad \text{for all } i \neq k, \tag{32}$$

and

$$\delta(d_0 \mid v) = 1, \quad \text{otherwise},$$

the rejection rate is determined by the probability measures of $\tilde{V}_i$'s, namely $\int_{\tilde{V}_i} F(v|a_i)dv$. The latter is determined by the character style and allowable deterioration. The character style may be considered ideal and the control over the printing perfect, if the resultant $\int_{\tilde{V}_i} F(v|a_i)dv$ is unity for all i, and all characters with allowable deterioration can then be recognized with neither an error nor reject. In this sense, the probability measures of $\tilde{V}_i$'s may be used to evaluate the combined quality of the character style and printing.

Conclusion

The decision theory has been successfully applied to the problem of character recognition. By employing the concept of risk, differences in consequences for various decisions have been taken into consideration. A rejection channel has been introduced to examine the degree of ambiguity of input signal and make rejections when necessary.

As developed, the structure and performance of an optimum system depend upon the signal and noise statistics; therefore, *a priori* knowledge of these statistics is required. Usually, a realistic estimate of noise statistics is not easy to obtain. However, it is sincerely felt that the requirement for high grade performance in character recognition warrants the expenditures in this direction.

Quite often an optimum system may prove to be too expensive for mechanization. Nevertheless, the results presented in this paper are considered useful in that they provide insight into the recognition problem and furnish an ideal system, which actual recognition circuitry may be patterned after.

Although it is recognized as being beyond the scope of this paper, it is worth mentioning that one practical approach to the over-all problem would be to design adequately the character style and to control properly the printing process so that a reliable system would not be too far fetched or difficult to ultimately realize.

Appendix I.

List of Symbols

$a_i = (a_{i1}, a_{i2}, \cdots, a_{is})$, s-dimensional vector associated with the ith character, $(i=1, 2, \cdots, c)$.
$a_{ij} = j$th sample of the signal of the ith character.
$|a_i| = (\sum_{j=1}^{s} a_{ij}^2)^{1/2}$, the length of vector a_i.
c = number of characters.
d_0 = decision that rejection be made.
d_j = decision that the signal is the jth character $(j=1, 2, \cdots, c)$.
$f(|v|)$ = probability density of the length of v.
$F(v|a_i)$ = conditional probability density for the observed data v when a_i is the incoming signal.
i, j, k = indexes.
N_{1i} = a network of station No. 1, $i=1, 2, \cdots, c$.
N_{2i} = a network of station No. 2, $i=1, 2, \cdots, c$.
N_0 = the network of rejection channel.
P_c = probability of correct recognition.
P_r = probability of rejection (rejection rate).
P_e = probability of misrecognition (error rate).
p_i = *a priori* probability that the ith character occurs, $(i=1, 2, \cdots, c)$.
$p = (p_1, p_2, \cdots, p_c)$.
$R(p, \delta)$ = expected risk of the system; $R = R_0 + R_1$.
$R_0(p)$ = expected risk of the system when rejection is made for all recognition.
$R_1(p, \delta)$ = part of R which is dependent upon δ.
s = number of samples.
T = functional transformation.
V = s-dimension observation space.
V_i = set of all v when the ith character is present.
$\tilde{V}_i$ = largest subset of V_i such that $\tilde{V}_i \cap \tilde{V}_j = 0$ for all $j \neq i$.
$v = (v_1, v_2, \cdots, v_s)$, a vector in V.
$|v| = (\sum_{i=1}^{s} v_i^2)^{1/2}$, the length of vector v.
$vi = i$th component of the observed data v.
$W(a_i, d_j)$ = weight function.
$w_{ij} = W(a_i, d_j)$.
w_c, w_r, w_m = weights.
$x_i(v)$ = output of the ith channel.
α = permissible rejection rate.
β = constant.
$\delta(v)$ = decision rule, $\delta = [\delta(d_0|v), \delta(d_1|v) \cdots \delta(d_c|v)]$.
$\delta^*(v)$ = optimum decision rule which minimizes the expected risk.
$\delta^{**}(v)$ = optimum decision rule which minimizes the error rate.
θ = angle.
ρ = a normalizing constant.
σ_{ij}^2 = statistical variance of noise.

Appendix II.

To Prove That δ^{**} Has a Minimum Error Rate for a Given Rejection Rate

Without loss of generality, it is assumed that the absolute probability density of the occurrence of v, namely, $\sum_{i=1}^{c} p_i F(v|a_i)$ is nonzero over the entire observation space V. Otherwise, the set over which $\sum_{i=1}^{c} p_i F(v|a_i)$ is zero is first deleted.

Let $m(v)$ be the subscript such that

$$\max_i [p_i F(v \mid a_i)] = p_m F(v \mid a_m) \tag{33}$$

and let $\delta^1(v)$ be any arbitrary decision rule having the same rejection rate as δ^{**}. It is to be proved that $P_e(\delta^1) \geq P_e(\delta^{**})$.

For every $\delta^1(v)$, a decision rule $\delta^2(v)$ can be constructed as follows:

For every v,

$$\begin{aligned} \delta^2(d_0 \mid v) &= \delta^1(d_0 \mid v) \\ \delta^2(d_m \mid v) &= 1 - \delta^1(d_0 \mid v) = \sum_{i=1}^{c} \delta^1(d_i \mid v) \\ \delta^2(d_i \mid v) &= 0 \quad \text{for all } i \neq 0 \neq m. \end{aligned} \tag{34}$$

Evidently,

$$P_r(\delta^2) = P_r(\delta^1) = \alpha, \tag{35}$$

and

$$\begin{aligned} P_c(\delta^1) &= \int_V \sum_{i=1}^{c} \delta^1(d_i \mid v) p_i F(v \mid a_i) dv \\ &\leq \int_V \sum_{i=1}^{c} \delta^1(d_i \mid v) p_m F(v \mid a_m) dv \\ &= \int_V \delta^2(d_m \mid v) p_m F(v \mid a_m) dv \\ &= P_c(\delta^2). \end{aligned} \tag{36}$$

It follows from (35) and (36) that

$$P_e(\delta^2) \leq P_e(\delta^1). \tag{37}$$

That is, δ^2 is better than δ^1 (or at least as good) in the sense that for the same rejection rate δ^2 has an error rate smaller than, or equal to, that of δ^1.

The next step is to show that $P_e(\delta^{**}) \leq P_e(\delta^2)$. As shown in (22) ad (23), the decision rule δ^{**} partitions the observation space V into two nonintersecting regions, V_0^{**} and $V - V_0^{**}$, so that for every $v \epsilon V_0^{**}$

$$p_m F(v \mid a_m) < \beta \sum_{i=1}^{c} p_i F(v \mid a_i) \tag{38a}$$

$$\delta^{**}(d_0 | v) = 1, \tag{38b}$$

and for every $v \epsilon V - V_0^{**}$

$$p_m F(v \mid a_m) \geq \beta \sum_{i=1}^{c} p_i F(v \mid a_i) \tag{39a}$$

$$\delta^{**}(d_m \mid v) = 1. \tag{39b}$$

Let V_0^2 be the largest subspace of V such that $\delta^2(d_0 \mid v)$ is nonzero for all v belonging to V_0^2. V_0^2 is not properly contained in V_0^{**}. This follows readily from the condition that $P_r(\delta^{**}) = P_r(\delta^2)$. The latter may be written as:

$$\int_{V_0^{**}-V_0^{**}\cap V_0^2} \sum_{i=1}^{c} p_i F(v \mid a_i) dv + \int_{V_0^{**}\cap V_0^2} [1 - \delta^2(d_0 \mid v)] \sum_{i=1}^{c} p_i F(v \mid a_i) dv = \int_{V_0^2-V_0^{**}\cap V_0^2} \delta^2(d_0 \mid v) \sum_{i=1}^{c} p_i F(v \mid a_i) dv. \tag{40}$$

Substitution of (38) and (39) in (40) gives:

$$\int_{V_0^{**}-V_0^{**}\cap V_0^2} p_m F(v \mid a_m) dv + \int_{V_0^{**}\cap V_0^2} [1 - \delta^2(d_0 \mid v)] p_m F(v \mid a_m) dv \leq \int_{V_0^2-V_0^{**}\cap V_0^2} \delta^2(d_0 \mid v) p_m F(v \mid a_m) dv. \tag{41}$$

The equality sign prevails if, and only if, $p_m F(v \mid a_m)$ is equal to $\beta \sum_{i=1}^{c} p_i F(v \mid a_i)$ throughout the region $V_0^{**} \cup V_0^2$.

The probabilities of correct recognition of δ^{**} and δ^2 may be written respectively as:

$$P_c(\delta^{**}) = \int_{V-V_0^{**}} p_m F(v \mid a_m) dv = \int_{V-(V_0^{**}\cup V_0^2)} p_m F(v \mid a_m) dv + \int_{V_0^2-V_0^{**}\cap V_0^2} p_m F(v \mid a_m) dv, \tag{42a}$$

$$P_c(\delta^2) = \int_V \sum_{i=1}^{c} \delta^2(d_i \mid v) p_i F(v \mid a_i) dv = \int_{V-(V_0^{**}\cup V_0^2)} p_m F(v \mid a_m) dv + \int_{V_0^{**}-V_0^{**}\cap V_0^2} p_m F(v \mid a_m) dv + \int_{V_0^{**}\cap V_0^2} \delta^2(d_m \mid v) p_m F(v \mid a_m) dv + \int_{V_0^2-V_0^{**}\cap V_0^2} \delta^2(d_m \mid v) p_m F(v \mid a_m) dv. \tag{42b}$$

In accordance with (42), (41) is equivalent to $P_e(\delta^{**}) \leq P_e(\delta^2)$. Proof that $P_e(\delta^{**}) \leq \delta P_e(\delta^1)$ is thus completed.

Acknowledgment

This paper is based upon research performed in the Electromechanisms Department of the Burroughs Corporation Research Center at Paoli, Pa. The author wishes to thank his colleagues and Department Manager, G. A. Baird for their encouragement and assistance relative to this paper. The permission of the Burroughs Corporation for release of the material presented here is also acknowledged.