[13] S. S. Wolff, J. L. Gastwirth, and H. Rubin, "The effect of autoregressive dependence on a nonparametric test," *IEEE Trans. Information Theory* (Correspondence), vol. IT-13, pp. 311–313, April 1967.

[14] M. Rosenblatt, "A central limit theorem and a strong mixing condition," *Proc. Natl. Acad. Sci.*, U.S.A., vol. 42, pp. 43–47, 1956.

[15] A. Kolmogorov and I. Rozanov, "On a strong mixing condition for stationary Gaussian processes," *Teor. Veroyatnost. i Primenen*, vol. 5, 1960.

[16] Yu. A. Rozanov, "An application of the central limit theorem," *Proc. 4th Berkeley Symp. on Mathematical Statistics and Probability.* Berkeley: University of California Press, 1961, pp. 445–454.

[17] W. Hoeffding, "A class of statistics with asymptotically normal distributions," *Ann. Math. Stat.*, vol. 19, pp. 293–325, 1948.

[18] H. Chernoff and I. R. Savage, "Asymptotic normality and efficiency of certain nonparametric test statistics," *Ann. Math. Stat.*, vol. 29, pp. 972–994, 1958.

[19] E. L. Lehmann, *Testing Statistical Hypotheses.* New York: Wiley, 1959, p. 257.

[20] M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics.* New York: Hafner, 1961.

[21] R. F. Pawula, "A modified version of Price's theorem," *IEEE Trans. Information Theory*, vol. IT-13, pp. 285–288, April 1967.

On Optimum Recognition Error and Reject Tradeoff

C. K. CHOW, SENIOR MEMBER, IEEE

***Abstract*—The performance of a pattern recognition system is characterized by its error and reject tradeoff. This paper describes an optimum rejection rule and presents a general relation between the error and reject probabilities and some simple properties of the tradeoff in the optimum recognition system. The error rate can be directly evaluated from the reject function. Some practical implications of the results are discussed. Examples in normal distributions and uniform distributions are given.**

I. Introduction

The ERROR rate and the reject rate are commonly used to describe the performance level of pattern recognition systems. A complete description of the recognition performance is given by the error–reject tradeoff, i.e., the functional relation of the error rate and reject rate at all levels. An error or misrecognition occurs when a pattern from one class is identified as that of a different class. The error is sometimes referred to as a substitution error or undetected error. A reject occurs when the recognition system withholds its recognition decision, and the pattern is rejected for exceptional handling, such as rescan or manual inspection.

Because of uncertainties and noise inherent in any pattern recognition task, errors are generally unavoidable. The option to reject is introduced to safeguard against excessive misrecognition; it converts potential misrecognition into rejection. However, the tradeoff between the errors and rejects is seldom one for one. Whenever the reject option is exercised, some would-be correct recognitions are also converted into rejects. We are interested in the best error–reject tradeoff in the optimum rejection scheme.

An optimum rejection scheme was derived in [1]. The error–reject tradeoff curves have been used to describe and compare the empirical performances of recognition methods [2] and [3], and they have also been found useful in the actual system design of an optical page reader [4]. However, few theoretical results on the error–reject tradeoff are available.

This paper first describes an optimum rejection rule and then derives a general relation between the error and reject probabilities. The error rate can be directly evaluated from the reject function. This result provides a basis for calculating the error rates from the empirical rejection curve without actually identifying the errors. Some simple properties of the optimum tradeoff are presented. Examples in normal distributions and uniform distributions are given.

II. Optimum Recognition Rule

A recognition rule is optimum if for a given error rate (error probability) it minimizes the reject rate (reject probability). It is known [1] that the optimum rule is to reject the pattern if the maximum of the a posteriori probabilities is less than some threshold. More explicitly, the optimum recognition rule δ is given as

$$\text{(a)} \qquad \delta(d_k \mid v) = 1 \qquad (k \neq 0) \tag{1}$$

i.e., to accept the pattern v for recognition and to identify it as of the kth pattern class whenever

$$p_k F(v \mid k) \geq p_j F(v \mid j) \qquad \text{for all} \quad j = 1, 2, \cdots n,$$

Manuscript received March 24, 1969; revised June 2, 1969. Some results of the paper were described at the IEEE 1968 International Workshop on Pattern Recognition, Delft, the Netherlands; a summary of the paper was presented at the 1969 International Symposium on Information Theory, Ellenville, N. Y.

The author is with the Thomas J. Watson Research Center, IBM Corporation, Yorktown Heights, N. Y. 10598, and is currently visiting the Department of Electrical Engineering, Massachusetts Institute of Technology, Cambridge, Mass. 02139.

and

$$p_kF(v \mid k) \geq (1 - t) \sum_{i=1}^{n} p_iF(v \mid i) \tag{2}$$

(b)

$$\delta(d_0 \mid v) = 1 \tag{3}$$

i.e., to reject the pattern whenever

$$\max_i \, [p_iF(v \mid i)] < (1 - t) \sum_{i=1}^{n} p_iF(v \mid i) \tag{4}$$

where v is the pattern vector, n is the number of classes, $(p_1, p_2, \cdots, p_n)$ is the a priori probability distribution of the classes, $F(v \mid i)$ is the conditional probability density for v given the ith class, d_i $(i \neq 0)$ is the decision that v is identified as of the ith class while d_0 is the decision to reject, and t is a constant between 0 and 1 $(0 \leq t \leq 1)$. The probability of error, or error rate, is

$$E(t) = \int_V \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \delta(d_j \mid v)p_iF(v \mid i) \, dv \tag{5}$$

and the probability of reject or reject rate is

$$R(t) = \int_V \delta(d_0 \mid v) \sum_{i=1}^{n} p_iF(v \mid i) \, dv \tag{6}$$

where V is the pattern space. Both the error and reject rates are implicit functions of the parameter t.

The probability of correct recognition is

$$\begin{aligned} C(t) &= \int_V \sum_{i=1}^{n} \delta(d_i \mid v)p_iF(v \mid a_i) \, dv \\ &= 1 - E(t) - R(t) \end{aligned} \tag{7}$$

and the probability of acceptance (or acceptance rate) is defined as

$$A(t) = C(t) + E(t). \tag{8}$$

Now let $m(v)$ denote the random variable

$$m(v) = \frac{\max_i p_iF(v \mid i)}{F(v)} \tag{9}$$

where $F(v)$ is the absolute probability density

$$F(v) = \sum_{i=1}^{n} p_iF(v \mid i). \tag{10}$$

The variable $m(v)$ is the maximum of the a posteriori probabilities of the classes given the pattern v.

The optimum rule δ can then be restated as:

1) accept the pattern v whenever

$$m(v) \geq 1 - t, \tag{2'}$$

2) reject the pattern v whenever

$$m(v) < 1 - t. \tag{4'}$$

The optimum rule is to reject the pattern v whenever its maximum of the a posteriori probabilities $m(v)$ is small (less than $1 - t$). The optimality can be seen by observing that $m(v)$ is the conditional probability of correctly recognizing a given pattern v. A detailed proof is given in [1].

III. Rejection Threshold

The parameter t in the decision rule will be called the "rejection threshold." For any fixed value of t $(0 \leq t \leq 1)$, the decision rule δ partitions the pattern space V into two disjoint sets (or regions) ($V_A(t)$ and $V_R(t)$) where (2) and (4), respectively, hold, namely:

$$V_A(t) = \{v \mid m(v) \geq (1 - t)\} \tag{11}$$

$$V_R(t) = \{v \mid m(v) < (1 - t)\}. \tag{12}$$

Without loss of generality, it will be assumed that $F(v)$ is nonzero over the entire space V, otherwise the set over which $F(v)$ is zero is first deleted. V_A and V_R are called, respectively, the "acceptance region" and the "reject region" of the decision rule. An example is depicted in Fig. 1(a) where the shaded region is V_R and the unshaded region is V_A.

In terms of these regions, the various probabilities can be written as

$$R(t) = \int_{V_R(t)} F(v) \, dv \tag{6'}$$

$$A(t) = \int_{V_A(t)} F(v) \, dv \tag{8'}$$

and

$$\begin{aligned} C(t) &= \int_{V_A(t)} \max_i \, [p_iF(v \mid i)] \, dv \\ &= \int_{V_A(t)} m(v)F(v) \, dv \end{aligned} \tag{7'}$$

$$\begin{aligned} E(t) &= \int_{V_A(t)} \left\{ \sum_{i=1}^{n} p_iF(v \mid i) - \max_i \, [p_iF(v \mid i)] \right\} dv \\ &= \int_{V_A(t)} [1 - m(v)]F(v) \, dv. \end{aligned} \tag{5'}$$

We shall now present some simple properties of the rejection threshold t.

1) Both the error and reject rates are monotonic in t.
2) t is an upper bound of the error rate.
3) i is a differential error-reject tradeoff ratio [see (20)].

A. Monotonicity

It follows immediately from the definitions of (11) and (12) that for any t_1 and t_2 in [0, 1] if $t_1 < t_2$, then

$$V_A(t_1) \subset V_A(t_2)$$

and

$$V_R(t_1) \supset V_R(t_2).$$

All the integrands in the integrals of (5′) to (8′) are

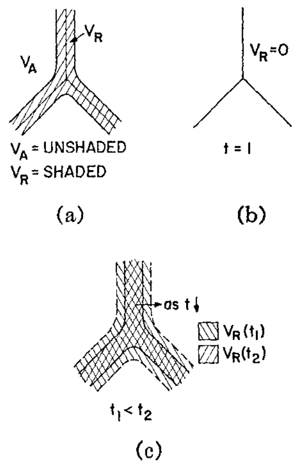


Fig. 1. Reject regions in the pattern space.

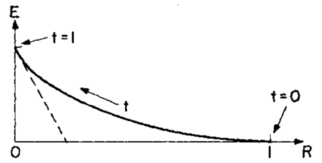


Fig. 2. Error–reject tradeoff curve.

nonnegative, hence if the domain of integration expands, the value of the integral increases. More specifically, if $t_1 < t_2$, then $V_A(t_1) \subset V_A(t_2)$ and $V_R(t_1) \supset V_R(t_2)$; therefore, $E(t_1) \leq E(t_2)$ and $R(t_1) \geq R(t_2)$. In other words, E increases and R decreases with increasing t. In particular, when $t = 0$, $E = 0$, and when $t = 1$, $A = 1$ and $R = 0$. Whenever $t \geq 1 - 1/n$, $R = 0$.

B. *An Upper Bound of Error Rate*

We shall now show that

$$E(t) \leq t.$$

For any v in $V_A(t)$, we have

$$m(v) \geq (1 - t).$$

Therefore (5′) gives

$$E(t) \leq \int_{V_A(t)} F(v)\, dv.$$

Hence

$$E(t) \leq tA(t) \leq t.$$

IV. Error-Reject Tradeoff

A complete description of the performance of recognition systems is given by the error–reject tradeoff, i.e., the functional relation of E and R at all levels. A typical tradeoff curve is given in Fig. 2. Since both E and R of the optimum recognition systems are monotonic functions of the rejection threshold t, one can compute the tradeoff E versus R from $E(t)$ and $R(t)$.

We shall now show that the rejection function $R(t)$ alone suffices to completely characterize the optimum recognition performance. In other words, E can be derived from $R(t)$. The central result is a simple functional relation between E and R; $E(t)$ is a Stieltjes integral of t with respect to $R(t)$, namely,

$$E(t) = -\int_{t=0}^{t} t\, dR(t). \tag{13}$$

This relation is valid for all optimum decision rules as defined in (1)–(4). No explicit forms for the density functions $F(v/i)$ are required in deriving the integral. If $R(t)$ is differentiable with respect to t, then the above Stieltjes integral reduces to the ordinary Riemann integral

$$E = \int_R^1 t(R)\, dR. \tag{14}$$

It is noted that the integral of (14) is not always meaningful; for example, when $R(t)$ is discontinuous (as in the case of Section VI-B). On the other hand, the Stieltjes integral of (13) always exists, as $R(t)$ is always bounded, monotonic, and thus of bounded variation.

Consider a decremental change in the rejection threshold from t to $t - \Delta t$; the reject region expands from $V_R(t)$ to $V_R(t - \Delta t)$. Let $\Delta V_R(t)$ denote the incremental region $V_R(t - \Delta t) - V_R(t)$. For any v in $\Delta V_R(t)$, it was accepted at the threshold t and is now rejected at the lower threshold $t - \Delta t$. Equations (2) and (4) now give

$$(1 - t)F(v) \leq \max_i p_i F(v \mid i) < (1 - t + \Delta t)F(v) \qquad \text{for} \quad v \in \Delta V_R(t). \tag{15}$$

By integrating the last expression over the incremental region ΔV_R, one obtains

$$(1 - t)\, \Delta R \leq -\Delta C < (1 - t + \Delta t)\, \Delta R \tag{16}$$

where ΔR and ΔC are, respectively, the increments in the rejection rate and correct recognition rate, namely,

$$\Delta R = \int_{\Delta V_R} F(v)\, dv$$

$$\Delta C = \int_{\Delta V_R} \max_i\, [p_i F(v \mid i)]\, dv.$$

Of course, the increment in the error rate is simply

$$\Delta E = -\Delta R - \Delta C. \tag{17}$$

By substituting (17) into (16), one has

$$-t\Delta R \leq \Delta E < -(t - \Delta t)\, \Delta R. \tag{18}$$

One then sums (18) with t steadily decreasing throughout the range of interest from t to 0 to obtain

$$-\sum t\Delta R \leq E(t) \leq -\sum t\Delta R - \sum \Delta t\Delta R$$

and then lets Δt tend to zero. As Δt tends to zero, the last sum of the above expression vanishes and (13) is thus obtained.

When $R(t)$ is differentiable, (13) becomes

$$\begin{aligned} E(t) &= -\int_0^t t\frac{dR}{dt}\,dt \\ &= -\int_{R(0)}^{R(t)} t(R)\,dR \\ &= \int_R^1 t\,dR \end{aligned}$$

since $R(0) = 1$. This relation is depicted in Fig. 3. The area under the entire rejection curve is the (forced decision) error rate when no rejection is allowed. Through an integration by parts and as indicated in Fig. 3, (13) can also be written as

$$E = \int_0^t R(t)\,dt - tR(t). \tag{19}$$

Equation (14) gives

$$\frac{dE}{dR} = -t \leq 0. \tag{20}$$

The rejection threshold is the differential error-reject tradeoff. In particular, the initial slope of the error-reject curve is $-1 + 1/n$ or greater, while the final slope is 0. Equation 20 also gives

$$\frac{d^2E}{dR^2} = -\frac{dt}{dR} \geq 0. \tag{21}$$

The optimum error-reject curve is always concave upward and the slope increases from -1 to 0 as R increases from 0 to 1 (Fig. 2).

V. Rejection Threshold of a Minimum-Risk Rule

It is known [1] that the optimum decision rule given in (1) to (4) is also a minimum-risk rule if the cost function is uniform within each class of decisions, i.e., if no distinction is made among the errors, among the rejects, and among the correct recognition. The rejection threshold is then related to the costs as follows:

$$t = \frac{W_r - W_c}{W_e - W_c} \tag{22}$$

where W_e, W_r, and W_c are the costs for making an error, reject, and correct recognition, respectively. Usually $W_e > W_r > W_c$. The rejection threshold is simply the normalized cost for the rejection. We can take $W_c = 0$ and $W_e = 1$, and the minimum risk is

$$\text{risk}\ (t) = E(t) + tR(t) = \int_0^t R(t)\,dt \tag{23}$$

which is also depicted in Fig. 3.

VI. Examples

For numerical illustration, two examples are given here. In these examples, the pattern vector v is 1-dimensional and there are two pattern classes with equal a priori probability of occurrence, i.e., $p_1 = p_2 = \frac{1}{2}$. The examples are concerned with the normal distributions and uniform distributions, respectively.

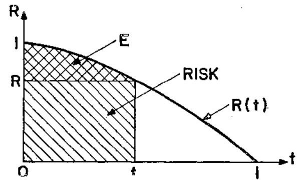


Fig. 3. Reject curve.

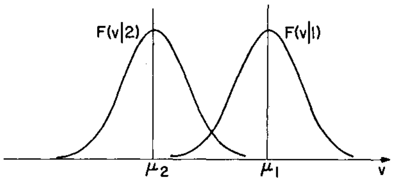


Fig. 4. Example in normal distribution.

For two classes, the condition for rejection, namely, (4), can never be satisfied when $t > \frac{1}{2}$; hence the reject rate is always zero if the reject threshold t exceeds $\frac{1}{2}$. The effective range of t for problems of two classes is, therefore, from 0 to $\frac{1}{2}$. With $n = 2$ and $0 \leq t \leq \frac{1}{2}$ it can be readily verified that the condition for rejection (4) is equivalent to

$$\frac{t}{1-t} \leq \frac{p_1F(v \mid 1)}{p_2F(v \mid 2)} \leq \frac{1-t}{t} \tag{24}$$

A. Normal Distributions of Equal Variance

Consider two normal distributions with means μ_1 and μ_2 and equal covariance σ^2, as shown in Fig. 4. Take $\mu_1 > \mu_2$. The density function is, $i = 1$ or 2,

$$F(v \mid i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(v-\mu_i)^2}{2\sigma^2}\right]. \tag{25}$$

With (25) and some algebraic manipulations, (24) can be transformed to

$$|v - \tfrac{1}{2}(\mu_1 + \mu_2)| \leq \frac{\sigma^2}{\mu_1 - \mu_2} \ln\left(\frac{1}{t} - 1\right), \tag{26}$$

i.e., the optimum rule is to reject whenever the pattern lies within a certain distance of the midpoint between the two means. The corresponding error and reject rates are

$$E(t) = \Phi(a) \tag{27}$$

$$R(t) = \Phi(b) - \Phi(a) \tag{28}$$

where Φ is the normal cumulative distribution function, namely,

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-(x^2/2)}\,dx \tag{29}$$

and the parameters are

$$a = -\tfrac{1}{2}s - \frac{1}{s}\ln\left(\frac{1}{t} - 1\right) \tag{30a}$$

$$b = -\tfrac{1}{2}s + \frac{1}{s}\ln\left(\frac{1}{t} - 1\right) \tag{30b}$$

$$s = \frac{\mu_1 - \mu_2}{\sigma}. \tag{30c}$$

The parameter s is the (normalized) separation between the means of the distributions and is the only (composite) parameter of the distributions that $R(t)$ and $E(t)$ depend upon. It is straightforward to verify (14) for this example. A set of the error, reject, and tradeoff curves, for $s = 1$, 2, 3, and 4, is depicted in Fig. 5.

B. Uniform Distributions

Consider two uniform distributions:

$$F(v \mid 1) = \begin{cases} 1 & \text{when} \quad 0 \le v \le 1 \\ 0 & \text{elsewhere} \end{cases} \tag{31a}$$

$$F(v \mid 2) = \begin{cases} \tfrac{1}{2} & \text{when} \quad \tfrac{1}{2} \le v \le \tfrac{5}{2} \\ 0 & \text{elsewhere} \end{cases} \tag{31b}$$

as shown in Fig. 6.

The reject function $R(t)$ is simply

$$R(t) = \begin{cases} \tfrac{3}{8} & \text{when} \quad 0 \le t \le \tfrac{1}{3} \\ 0 & \text{when} \quad \tfrac{1}{3} < t \le 1, \end{cases} \tag{32}$$

which is discontinuous [Fig. 7(a)], and the integral of (13) is evaluated to

$$E(t) = \begin{cases} 0 & \text{when} \quad 0 \le t \le \tfrac{1}{3} \\ \tfrac{1}{8} & \text{when} \quad \tfrac{1}{3} < t \le 1, \end{cases} \tag{33}$$

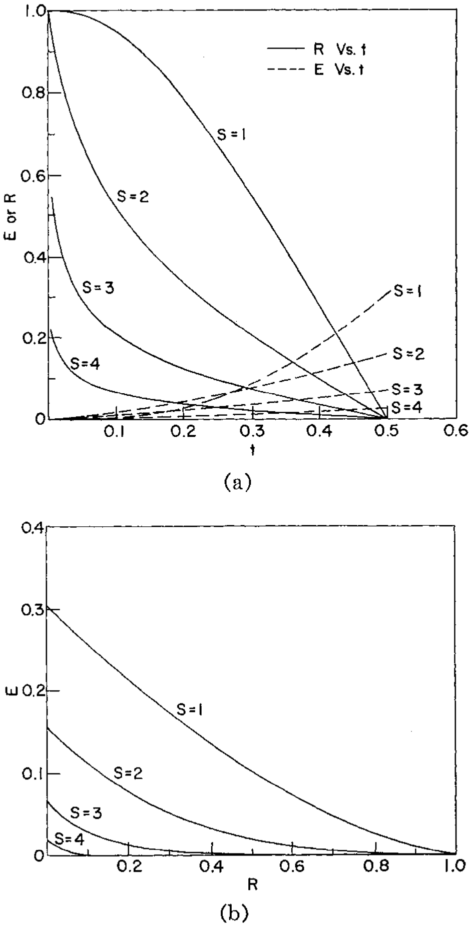
which is shown in Fig. 7(b) and the tradeoff can assume only two values, namely $(E, R) = (\tfrac{1}{8}, 0)$ or $(0, \tfrac{3}{8})$ [Fig. 7(c)]. However, if a randomized scheme is used in the range $\tfrac{1}{2} \le v \le 1$, R may vary continuously from $\tfrac{1}{8}$ to 0 as shown by the dotted line in Fig. 7(c).

VII. Some Practical Implications

Most of our results on the error–reject tradeoff seem consistent with our intuition, although the simple integral relation between the error and reject rates is somewhat unexpected.

Since the slope of the error–reject tradeoff curve (Fig. 2) is the value of the rejection threshold, the tradeoff is most effective initially (i.e., at the low level of rejection) and it gets more difficult as the error rate is lower. This is certainly common in our practical experience; excessive rejection is generally required to reduce residual errors.

Practical applications of the present results are in the areas of system design and performance evaluation of the recognition systems. The general characteristics of the error–reject tradeoff curve provide the system de-



(a)

(b)

Fig. 5. Normal distributions. (a) Reject and error curves. (b) Tradeoff curve.

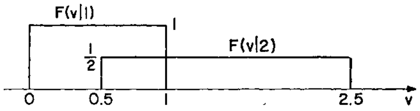
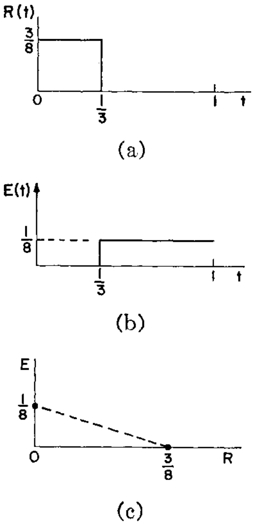


Fig. 6. Example in uniform distributions.



(a)

(b)

(c)

Fig. 7. Uniform distributions. (a) Reject curve. (b) Error curve. (c) Tradeoff curve.

signer with a convenient means of verifying the basic assumption on the underlying probability distributions. The integral of (13) makes it possible to calculate error rates and consequently, the tradeoff curve from the empirically observed reject rates. No class identification of the sample patterns are required in obtaining the empirical rejection curve. Or equivalently, one can just obtain an empirical density function of the maximum of the a posteriori probabilities, and then calculate the error and reject rates.

In most recognition tasks, the underlying probability distributions of the patterns are not completely known and the design of the recognition systems is generally based on empirical data. A common design procedure is to assume, on the basis of available (usually limited) a priori information and the designer's intuition, some functional forms of the distributions, to derive the system structure based on these assumptions, and to adjust the system parameters by using the empirical data. It is not always a simple matter to verify the validity of the assumptions on which the system structure is based. However, one can always, though laboriously, obtain the empirical error–reject tradeoff curve and compute the theoretical one from the basic assumptions. A comparison of the empirical and theoretical tradeoff curves can quickly reveal how well the theoretical model agrees with the empirical data, and it can serve as a checkpoint for initiating the process of revising and improving the theoretical model.

The data used in any meaningful evaluation of a recognition system are usually large, and it is extremely costly and time consuming to detect the recognition errors. To identify a recognition error additional information, usually human inspection, at some stage is required. On the other hand, the rejection is the explicit result of a definite decision, and the rejects can be readily recorded and tallied. Equation (13) provides a simple means of calculating the error rate from the reject curve without actually identifying the errors.

VIII. Conclusion

A general error and reject tradeoff relation is derived for the (Bayes) optimum recognition system with an option to reject. The error probability is a Stieltjes integral of the rejection threshold with respect to the reject probability. The error function can be directly evaluated from the reject function. Hence, the reject function determines the recognition error and reject tradeoff and completely characterizes the performance of the optimum recognition system. The error–reject integral provides a simple means of calculating the error rate from the empirical reject curve without actually identifying the recognition errors.

References

[1] C. K. Chow, "An optimum character recognition system using decision functions," *IRE Trans. Electronic Computers*, vol. EC-6, pp. 247–254, December 1957.

[2] C. K. Chow and C. N. Liu, "An approach to structure adaptation in pattern recognition," *IEEE Trans. Systems Science and Cybernetics*, vol. SSC-2, pp. 73–80, December 1966.

[3] R. Bakis, N. Herbst, and G. Nagy, "An experimental study of machine recognition of hand-printed numerals," *IEEE Trans. Systems Science and Cybernetics*, vol. SSC-4, pp. 119–132, July 1968.

[4] R. B. Hennis, "The IBM 1975 optical page reader, Part I: System design," *IBM J. Res. and Develop.*, vol. 12, pp. 346–353, September 1968.