W. E. Dickinson

A Character-Recognition Study

Abstract: A study of the single-gap-scan approach to character recognition, using an IBM 650 for simulation, is reported. Ten specially designed digits were used in this study. Character recognition is discussed in terms of some simple concepts from *n*-dimensional geometry. The main contribution is an effective method for using a computer to aid in the design of the type font. This procedure is a natural development of the vector approach. Experimental results show the sensitivity of the system to phasing. An expression is given for a "quality factor." The relationship of this factor to errors and to ink density is illustrated.

Introduction

This article is a summary of a study made on a particular method of character recognition. This method was considered as a possible low-cost character-recognition system which could recognize a limited number of characters. The character set was limited to the ten digits in this study; in most applications a few special characters would be required also.

In this character-recognition system, the characters are scanned through a slit. The signal obtained from this scan is then compared with a set of "compare" signals and the best match chosen as the "machine-read" character.[1] The advantage of such an approach is its simplicity and low cost. Since considerable information is lost during the scanning process, this system has its limitations, not the least of which is the requirement that the characters be reasonably different in terms of their scanned signals. This limitation usually requires a set of characters especially designed for this system. Other systems which extract more information during the scan have a greater recognition potential, but also require more equipment.

In general, it can be said that, with normal printing size and quality, this approach is limited to a small number of characters. With suitable "reject" control, it should be possible to keep the "substitution-error" rate low for this system. A specific rate is difficult to determine because of such subjective factors as printing quality and character appearance. The system appears to represent the lower performance limit of useful character-recognition systems.

This paper will discuss various aspects of this slit-scan approach, present certain experimental data and conclude with a brief mathematical appendix. With the exception of a scanner system which punched the scanned information into a punched card, all experimental work was done on a simulation basis on an IBM 650 computer.

Various aspects of the slit-scan method

• 1. *The general method*

In all character-recognition systems, a signal obtained by scanning a character is compared with signals representing each of the possible characters; in general, the compare-character signal with the best match to the scanned signal is selected as the machine-read character.

In the approach discussed here, the character to be read is scanned by passing it under a narrow slit. The slit may be passed over the character in any direction (in all instances the relative motion of the slit is perpendicular to its length). The slit height is such that it includes the character being read and excludes any other character. These restrictions impose some constraints on printing location, but the constraints appear reasonable. With the slit-scan technique, information about the location of ink along the length of the slit is lost. In exchange for this information loss, a positioning tolerance of the slit relative to the character is obtained.

The signal obtained from the slit-scan is an analog waveform. The specific shape of the waveform will depend upon the transducer used and the character scanned. The waveform obtained with an optical transducer is shown in Fig. 1. The waveform was obtained with a vertical slit moved horizontally across a zero; the peaks are shown to correspond to the sides of the zero, with a lower-level amplitude between the peaks resulting from the two horizontal bars of the zero. With printing variations, the height of the waveform will vary, but the center-to-center distance between peaks will be relatively constant.

After linear amplification, the scanned signal is compared with each of the possible signals. In the method described here, the comparison is made using a set of "matched filters." Since the term "matched filter" may be unfamiliar to some, it will be discussed in greater detail

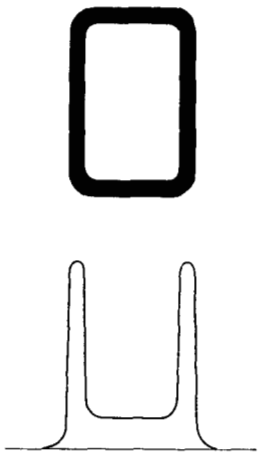


Figure 1 **Waveform obtained with optical transducer for the character zero.**

later. At this point it is sufficient to say that this filter will yield a maximum output for an input waveform of a specific shape.

Recognition, in this system, consists in choosing the character whose filter gives the largest output. To avoid substitution errors the choice may be qualified to require that the ratio of largest to second-largest output exceed a certain set value.

The operation involved in comparing the scanned waveform against all possible waveforms and choosing the best match is performed with analog circuitry. Inherent in this type of circuitry is a degree-of-match indication which may be useful in other comparing applications not requiring a high percentage of accuracy.

Returning now to the matched filter, some of its aspects will be discussed in a general manner. A detailed discussion may be found in References 2 and 3, where it is shown that for a signal contaminated by white noise (noise whose frequency components are uniformly distributed over the frequency range of interest), the best possible linear filter is a matched filter. Although the noise encountered in character recognition is unlikely to be white noise, and the possibility of nonlinear filters need not be excluded, the matched filter is likely to prove to be a good filter for this application.[5] (Section 5, *Choice of compare vectors,* discusses some modifications.) A "matched filter" is matched with respect to a specific signal in such a way that the impulse response of the filter is the time-inverse of the signal. That is,

$$h(t) = As(b-t),$$

where

$h(t)$ is the filter impulse response,

$s(t)$ is the signal being matched, and

A and b are constants.

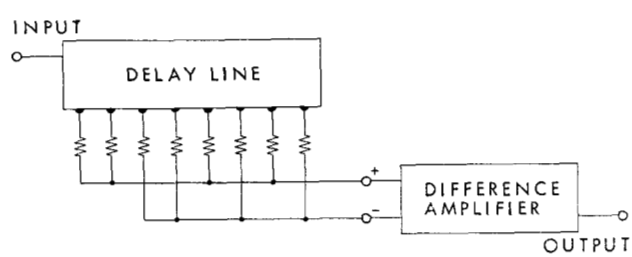
The matched filter for electrical signals can be constructed using conventional filter theory, provided the desired impulse response is simple enough. For frequency-band-limited signals, an approach using a tapped delay line is preferable since it is much more flexible, and many matched filters can be obtained from one delay line by adding inexpensive components. The tapped-delay-line matched filter is shown schematically in Fig. 2. The introduction of an impulse to the input results in a pulse travelling down the delay line (passing the taps in succession). From each tap the signal is attenuated by an amount determined by the resistor connected to that particular tap. To allow positive and negative polarities, a polarity inversion is accomplished in the difference amplifier. All positive taps of the delay line are added together (resistor adder), and all the negative taps are added together to form the two inputs to the difference amplifier. If only positive outputs are desired, the difference amplifier may be eliminated. In character recognition, the signals may be frequency-band limited (put through a low-pass filter) with little loss in information content since the frequency components above ten times the character scan rate are very minor. Thus, for a scan rate of 1 kc (1 millisecond per character), a cutoff of 10 kc may be used. The use of sample-data points rather than a continuous waveform does not lose further information provided the sample points are spaced at $1/2W$ seconds or less. (W is highest frequency present in sampled waveform.) For these reasons the use of a tapped-delay-line matched filter appears suitable for this approach to character recognition.

After introducing some relationships between geometry and waveforms, the general approach considered here will be continued.

• 2. *Geometric interpretation of waveforms*

In discussing certain aspects of our character-recognition study it is desirable to use a geometrical viewpoint. Before doing this, certain essential aspects of the geometric approach will be discussed briefly. In general, n-dimensional geometry is involved where n is greater than 3, and the geometric space cannot be visualized. It is paradoxical that, under certain conditions, a geometric space which cannot be visualized allows a greater insight into understanding the problem. A quotation from Shannon,[4] somewhat paraphrased, presents the essence of this approach: "We replace a complex entity (the waveform) in a simple environment (the waveform requires only a plane for its representation as a function of time) by a simple entity (a point) in a complex environment (n-dimensional

Figure 2 **Tapped-delay-line matched filter.**



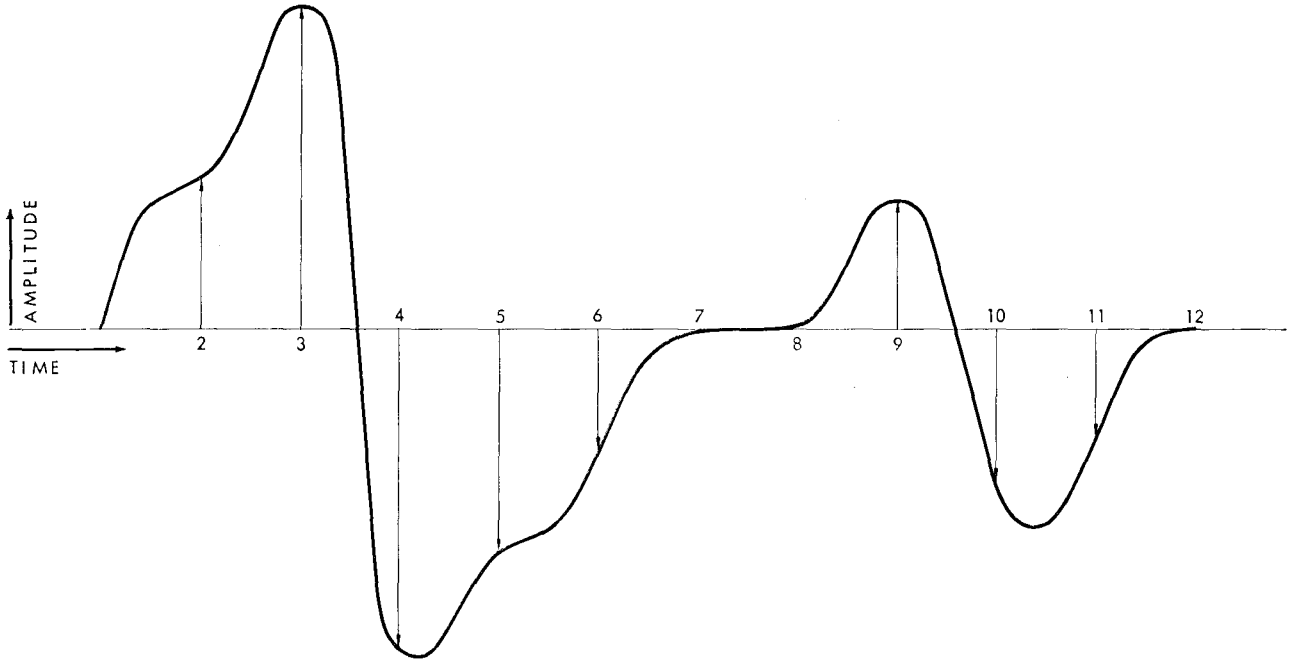


Figure 3 **Waveform to be represented as a vector (see text).**

space)." In using the geometric approach, only three points are normally considered at one time, one point representing the origin and the other two points representing the two waveforms under consideration. Since three points require only a plane for their representation, a two-dimensional diagram can be used to *display* them, although it is understood that a high dimensional space is involved. The two waveforms can be considered in this representation as two vectors, a vector being the line connecting the origin and a waveform point. As a waveform is increased or decreased uniformly in amplitude, the corresponding vector length increases or decreases in length, but does not change its direction relative to the n-space.

If we consider the n coordinate axes to be at right angles to each other, then the length of a vector from the origin is defined as in one-, two-, or three-dimensional geometry:

$$L=\sqrt{x_1^2+x_2^2+x_3^2+\ldots.+x_n^2}=\sqrt{\sum_i^N x_i^2}\,.$$

A vector can be made into a unit-length vector by dividing each component by the vector's total length.

The product of two vectors will be defined in the vector, innerproduct sense. The resultant of the product will be a scalar number. Thus:

$$\mathbf{X}\cdot\mathbf{Y}=x_1y_1+x_2y_2+\ldots.+x_ny_n=\sum_i^n x_iy_i\,.$$

If **X** and **Y** are unit vectors (of unit length) then their scalar product must be between $+1$ and -1, and will be $+1$ only when **X** equals **Y**. A further point of interest is the angle between two vectors. Here, an extension from two- and three-dimensional geometry is also used, namely:

$$\mathbf{X}\cdot\mathbf{Y}=|\mathbf{X}|\,|\mathbf{Y}|\cos\theta_{xy}\,,$$

where: θ_{xy} is the angle between **X** and **Y**, and **X** is the absolute value (or length) of **X**.

Thus, where **X** and **Y** are unit vectors, the product of the two vectors is equal to the cosine of the angle between them.

Returning now to the waveforms, the means of representing a waveform as a vector is as follows. Consider the waveform shown in Fig. 3. Here 12 sample points are shown along the waveform. If the waveform contains no frequencies above W, then the sampling theorem[4] states that if the time between sample points is equal to or less than $1/2W$, no information is lost. The formation of a vector from this sampled waveform is accomplished by considering the amplitude of the "i^{th}" sample point as the "i^{th}" component of the vector. Because of this one-to-one correspondence between waveform and "vector," the two terms will be used interchangeably in the remainder of this paper.

• 3. *The geometry of waveform comparison*

The matched filter, in the geometric terms we have discussed, provides a means for obtaining the scalar product of an unknown vector (the scanned waveform) and a compare vector (the waveform the matched-filter matches). It is clear that, if the scalar products of the unknown vector with a set of unit-length compare vectors are obtained, the compare vector with the smallest angular separation from the unknown vector will yield the largest scalar product. If the unknown vector increases in length, all scalar products increase proportionately. Thus, by

taking the ratio between any two scalar products, a means of eliminating the effect of amplitude variations of the unknown vector is obtained.

An alternative approach might be to compare the differences between the unknown vector and each of the compare vectors. It can be shown that this approach leads to essentially the same result as the scalar-product approach. With some ways of mechanization it might even be preferable. Consider Fig. 4. The unknown vector, **X**, has been normalized to the same length as the compare vector, **C**. The difference vector is identified as **D** and the angular separation of **X** and **C** is identified as θ. An expression for **D** is:

$$|\mathbf{D}| = \sqrt{\mathbf{X}^2 + \mathbf{C}^2 - 2\mathbf{X}\mathbf{C}\cos\theta}\,.$$

If **X** and **C** are considered to be of unit length, then

$$|\mathbf{D}| = \sqrt{2(1-\cos\theta)}\,.$$

In terms of the components of **X** and **C**:

$$|\mathbf{D}| = \sqrt{\sum_i^n (x_i - c_i)^2}\,, \quad \text{and so}$$

$$\sum_i^n (x_i - c_i)^2 = 2(1-\cos\theta) = 2(1-R)\,,$$

where R = scalar product of **X** and **C**.

Thus it is seen that the two approaches to expressing the separation of vectors (waveforms) are directly related by the above expression. In general, any expedient which can be used to improve one approach can be revised to fit the other approach.

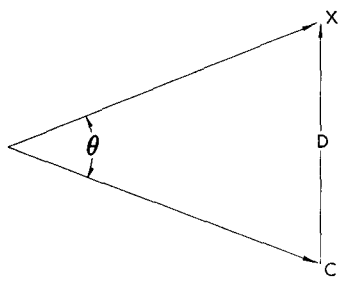
Up to this point one aspect of comparing vectors has not been mentioned. This is the requirement to properly phase the scanned waveform against the compare vectors. A system which is able to perform a continuous scalar product (such as the matched filter approach) offers a few possibilities not available readily where the scalar product is obtained serially. Two such possibilities which might be of interest are:

a) Sometime prior to the expected "in-phase" time, start to obtain the scalar product. Continue beyond the expected "in-phase" time. During this phasing period hold the maximum scalar product obtained for each compare vector.

b) Do a similar approach to method (a), but monitor the highest scalar product. When it begins to decrease, hold all scalar products as of that time. If later, during the phasing period, the highest scalar-product exceeds the previous maximum, the process is repeated until the end of the phasing period.

Either of these methods could be mechanized in a system which permits continuous monitoring of the scalar products. Method (a) is simpler, but will yield a poorer ratio between the closest and the second-closest matches. That this ratio will be poorer follows from the reasoning that the closest match will be the same for the two approaches, whereas the second closest in method (a) will be *at least* as close as it is in method (b).

Figure 4 **Vectors (waveforms) may be compared using the difference (D) between the unknown vector and the compare vectors (see text).**



Less complex methods of phasing could be used, either to indicate roughly the phasing period for the methods just discussed, or to give the phase time directly. Such methods would be based upon a search for some distinctive characteristic of the waveform which is essentially invariant for all the characters scanned. Such characteristics as the following might be used:

a) Amplitude exceeds a certain level.

b) First part of waveform with negative slope.

c) First zero crossing of waveform for waveforms having both positive and negative polarities.

d) Comparison of leading edge of waveform with a pulse shape.

Others might be suggested, but these are adequate to illustrate the general idea. In all of these it will be necessary to detect that the scan slit has just passed into the leading edge; that is, the passage of a between-character space must be detected. With typefont spacings having equal center-to-center or edge-to-edge distances, certain historical information can be used to reduce errors in detecting the between-character space. The performance of the above methods of phasing will be discussed further in examining simulation results; however, the performance of the first method is likely to be quite poor in any case because of noise and will not be considered further. The performance of these phasing methods will depend to a large extent upon how distinctive the leading edge (or trailing edge, if it is used) is for the typefont used. It is evident that none of these simple methods can give the accuracy of phasing possible with the second of the two methods where the phasing is continuously monitored.

• 4. *Character font design techniques*

In the special font of type required for the single slit-scan system, an attempt should be made to make each character as distinctive as possible. Because the font must be humanly readable, some compromise in machine readability is required. Furthermore, the usage may require that the font be esthetically satisfying as well as capable

of being read with accuracy by humans. Where the esthetic requirement is important, it is difficult to decide when restyling should stop. Where accuracy is the only requirement for human readability, objective tests can be made to evaluate the font based upon human and machine readability. The discussion here will be confined to a technique used to restyle the characters to improve machine readability.

The basic approach used is to take the vectors representing a set of characters and to restyle one of the vectors at a time until it is sufficiently well separated from all the other vectors, the appearance is unsatisfactory, or the vector is no longer physically realizable (i.e., negative ink is required).

An iterative approach has been used on the computer. In this approach a vector is chosen which is modified gradually until it is satisfactorily separated from the other vectors. The chosen vector is successively compared with each of the vectors of the set. If the scalar product indicates less separation than is desired, a negative fraction of the interfering vector is added to the chosen vector. If the scalar product indicates satisfactory separation, no modification is performed. This process is continued, using the modified vector each time, until the scalar product with all of the vectors has been accomplished. Then a printout of the modified vector along with its scalar products with all vectors is obtained. If the scalar products indicate the desired separation has been achieved, the program stops; if not, the process is repeated.

The comparison of waveforms may not be made in the amplitude domain in some systems. Consider Fig. 5. Here the character is scanned so as to give a derivative-like waveform. To restyle a character for a system such as this, it may be desirable to separate the vectors in the domain where the comparisons (scalar products) are made,[6] then to convert the modified vector to an amplitude pattern and finally to a character shape. The steps for doing this are indicated in Fig. 6. The character is converted to a "black-pattern" by passing under a slit. Then by use of the convolution integral the black pattern is converted to the derivative-like waveform. After separation by the method described above, the resulting waveform is put through an inverse response to obtain a restyled black pattern. At this point human intervention is required to make a character out of the black pattern using, as an aid, an abacus-like device or bead matrix. (The human is here doing the inverse of what the slit did in Fig. 5.)

Unless numerical value can be assigned to the appearance of the restyled digits, an optimum set of digits can never be obtained since the criterion is subjective. Any set of digits may be further separated in the machine-readability sense.

The principal source of noise in character recognition is caused by printing variations. Since the printing noise will depend upon the shape of the printed symbol, the noise is not "white" noise. A better character design technique would consider the noise. This study did not consider the noise in designing the characters.

• 5. *Choice of compare vectors*

Given a font of type to be read by the single-slit system, the question arises as to what compare vectors should be used. A natural choice would be to use an average vector for each of the compares. This average vector would be obtained by scanning a large number of printed samples of each symbol. Another possibility investigated was to find a compare vector which had a large angular separation from all nonmatching, average-value vectors and a minimum angular separation from its matching average-value vector. Some of the aspects of this latter approach will now be discussed.

The computer program previously described to redesign characters can be used to obtain a set of vectors which has the characteristics desired. The constraints are relaxed, however, since the compare vector is used only in the circuitry.

What length should be used for these compare vectors? With compares which are *average* vectors, making all of the compare vectors the same length is a good first choice. (The effect of this choice will be considered later.) With the modified compare vectors, which we are discussing, the angle between the compare vector and its matching-character mean-value vector will not always be the same for all characters. For this reason, in our tests a compare-vector length was used which made the scalar product of the compare vector and a normalized, matching-character average vector the same for all characters. In other words, we normalized the *scalar product* of these two vectors by

Figure 5 **Example of system that produces derivative-like waveform.**

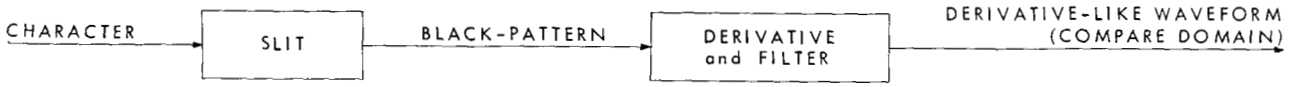
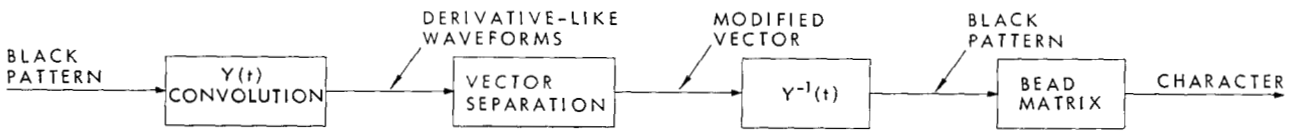


Figure 6 **Steps required to restyle a character.**



choosing the compare-vector length. If the minimum angle between the chosen average-value vector and all other average-value vectors was greater than the desired separation, the compare vector was identical in length and direction to the chosen average-value vector. When the desired angle was increased beyond this minimum, the compare vector had to be lengthened to maintain a constant scalar product.

When properly phased, this modified compare set gave much better separation ratios between the largest and second largest scalar products. The better ratio would be advantageous from a mechanization standpoint since component drift becomes less critical. Using these modified compares leads to a serious difficulty; namely, phasing becomes more difficult. The average-value compares led to a peak scalar product of the scanned vector with the matching compare vector at a predictable time, and all nonmatching compares gave lesser peaks (within printing tolerance). These modified compares yield peaks with the matching compare at a different time, and nonmatching compares may yield scalar products which exceed the scalar product with the matching compares. Thus, although at *the time of phasing which is best for average compares* a large ratio exists between the matching and second-highest scalar products, it is difficult to determine, with these modified compares, this sample time. Primarily for this reason, the use of this version of modified compares does not look attractive.

Again consider the average compare vectors. When equal length compares are used, the hyper-plane which falls (angularly) half-way between two compare vectors separates the selection region. If a compare vector is increased in length, it will tend to push the dividing hyper-planes further away from the compare vector. Because the noise will not be equal for all characters, it will probably be desirable to experimentally adjust the compare-vector lengths. Another modification of the average-value vectors which might be helpful is to remove completely any components of the average-value vectors which are common to all these vectors. If this is done, these components must be eliminated in the scanned vector as well, which may prove to be very difficult.

6. *The reject problem*

One of two decisions must be made after a character has been scanned and compared. Either a character is selected, or a reject indication is given. The rejects would indicate that uncertainty about the correct character is too great, and the document would be processed manually. If a character is selected incorrectly, a substitution error is made. The cost of a substitution error in any system is high compared to the cost of a reject. For a useful system, the substitution errors must be kept to a very small fraction of the characters read. The reject rate is dependent upon how well the printing falls within the system capabilities, the method used to determine rejection, and the level at which rejection is made. If, for example, the rejection level is set so that no rejects can occur, then a substitution error will occur whenever the printing is poor enough. The other extreme in rejection level would reject everything. The optimum, but unattainable, rejection level would be one which rejects only those characters which, if not rejected, would result in substitution errors.

The rejection method is, therefore, an important part of the system. The method which has been mentioned earlier is to reject if the ratio between the highest and second highest scalar products does not exceed a certain level. This method has some weaknesses which become apparent when poorly printed characters are tested. In a vector sense a poorly printed character has a large angular separation from its matching-compare vector.

Under these conditions the ratio of scalar products (highest to second highest) is an unreliable indication of whether rejection should be made or not. What is needed, in fact, is some measure of the quality of the printing. When the printing quality is good, the ratio of scalar products can be much lower than when the printing is poor. Thus, it would appear desirable to make the level of the reject ratio a function of the print-quality measure or to reject poor printing entirely.

A possible way to measure the print quality for the system discussed here is to obtain a ratio, Q, as follows:

$$Q = \frac{\sum_i \mathbf{C}_i \mathbf{X}_i}{\sqrt{\sum_i \mathbf{X}_i^2}}.$$

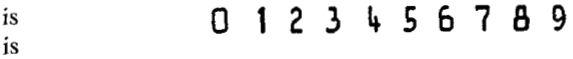
With the **C** vectors normalized in length, this expression for Q is equivalent to normalizing the scan vector, **X**, as well. Q is the cosine of the angle between **C** and **X**, and thus is a measure of the print quality which we desired. The compare vector with the largest Q would be selected only if its Q were near enough to one and sufficiently greater than the second-highest Q.

A more elegant reject criterion has been described by Chow.[5] The essence of his method is to determine the conditional probability of each of the characters having been sent, given the unknown waveform. Then, with a knowledge of the loss for rejecting a character, the loss for selecting a wrong character, and these conditional probabilities, the action which has the least risk (in a cost sense) can be chosen.

Experimental work

1. *Description of digits studied*

The experimental work described in this section was all done on a set of digits which is shown below.

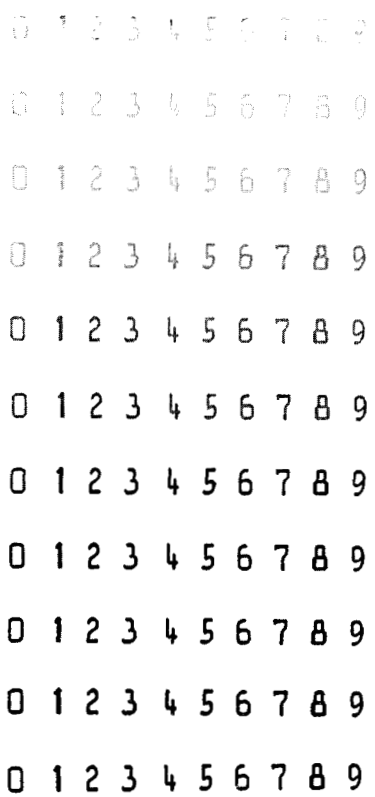


This set was designed to be scanned in a horizontal direction by the method described in the section entitled *Character font design techniques*. The set is such that the mean-value vectors of the digits are separated by amounts varying from about 45 to 100 degrees (correlation range of 0.69 to −0.16 respectively). It should be understood that these separations are for the noise-free characters.

These digits could be further improved in the machine readability sense. Since the decision as to when to stop the design of the characters is subjective, an optimum font is undefined.[7] Further separation can be achieved with the computer program whenever desired.

In light of the test results, it now appears that the phasing could be simplified by modifying the right-hand side of these digits to make the vertical edge more nearly alike for all the digits.

To force errors, with a limited amount of scanning, a set of eleven letterpress-printed samples was produced which had a wide range of inking. These samples were scanned both optically and magnetically. The range of output from the magnetic head was somewhat greater than 20 to 1 in going from the heaviest- to the lightest-inked samples. The lightest printed sample was too light to be handled by our circuitry with magnetic scan. Our optical-scan circuitry could accommodate all eleven samples. The range of printing is illustrated in the samples below.



An experimental scanner was used to scan the printed samples and reproduce the data on punched cards for computer processing. Briefly, it consists of a pair of scanning heads: one optical, the other magnetic. The amplified voltage from the heads is fed through an analog-to-digital converter and the digital amplitude, for each sample point on the waveform, is punched into a card. These cards then serve as the data-input source for the computer programs.

• 2. *Illustration of typical waveforms*

A sketch of the average-value, derivative waveforms for the set of digits is shown in Fig. 7.

• 3. *Spectral characteristics of the scan waveforms*

Low-pass filtering of scan signal — With heavy or light inking there is an increase in the amount of higher frequency "noise" present in the scanned signals. This is illustrated in Fig. 8, where f_c is determined by the sample-point interval. For this reason a low-pass filter which essentially removed the frequencies above the normalized frequency of 0.5 was used for most of the tests. This filter reduced the number of substitution and reject errors slightly.

• 4. *Correlation functions of compare waveforms*

The correlation functions for the mean-value waveforms near the in-phase time are shown in Figs. 9a through 9k. The solid line is a plot of the auto-correlation function while the dotted lines show the cross-correlation function. Thirty sample points per waveform were used for these curves. In many instances an error of one sample point would be serious (especially for the digits 2, 3, 4, 5, 6, 8 and 9). To illustrate the critical nature of the phase problem when other than mean-value compares are used, Fig. 10 shows a set of compares which are approximately orthogonal to the mean-value vectors at in-phase time and with lengths to normalize the in-phase, matching character scalar product to one. It is seen that although good separation occurs at the correct in-phase instant, a very poor situation occurs just before and after this time.

• 5. *Print quality versus error*

Figure 11 shows a scatter diagram which relates the print-quality measure, Q, and the ratio of the highest nonmatching-character scalar product to the matching-character scalar product. When the print quality is high, the likelihood of a substitution error (points above the dotted line) decreases.

• 6. *Print quality versus ink density*

Figure 12 shows a scatter diagram which indicates the relationship between print quality measure, Q, and ink density for ten samples each of ten printing densities.

• 7. *Performance as related to inherent separation*

Figure 13 is a scatter diagram. The ordinate represents the scalar product of a compare vector and its closest neighboring compare vector. The abscissa represents the ratio between the largest scalar product of a nonmatching compare vector and the scanned vector, and the scalar product of the scanned vector and its matching compare. The results are not surprising. Compare vectors which are well separated from other compare vectors have better performance than those which are poorly separated.

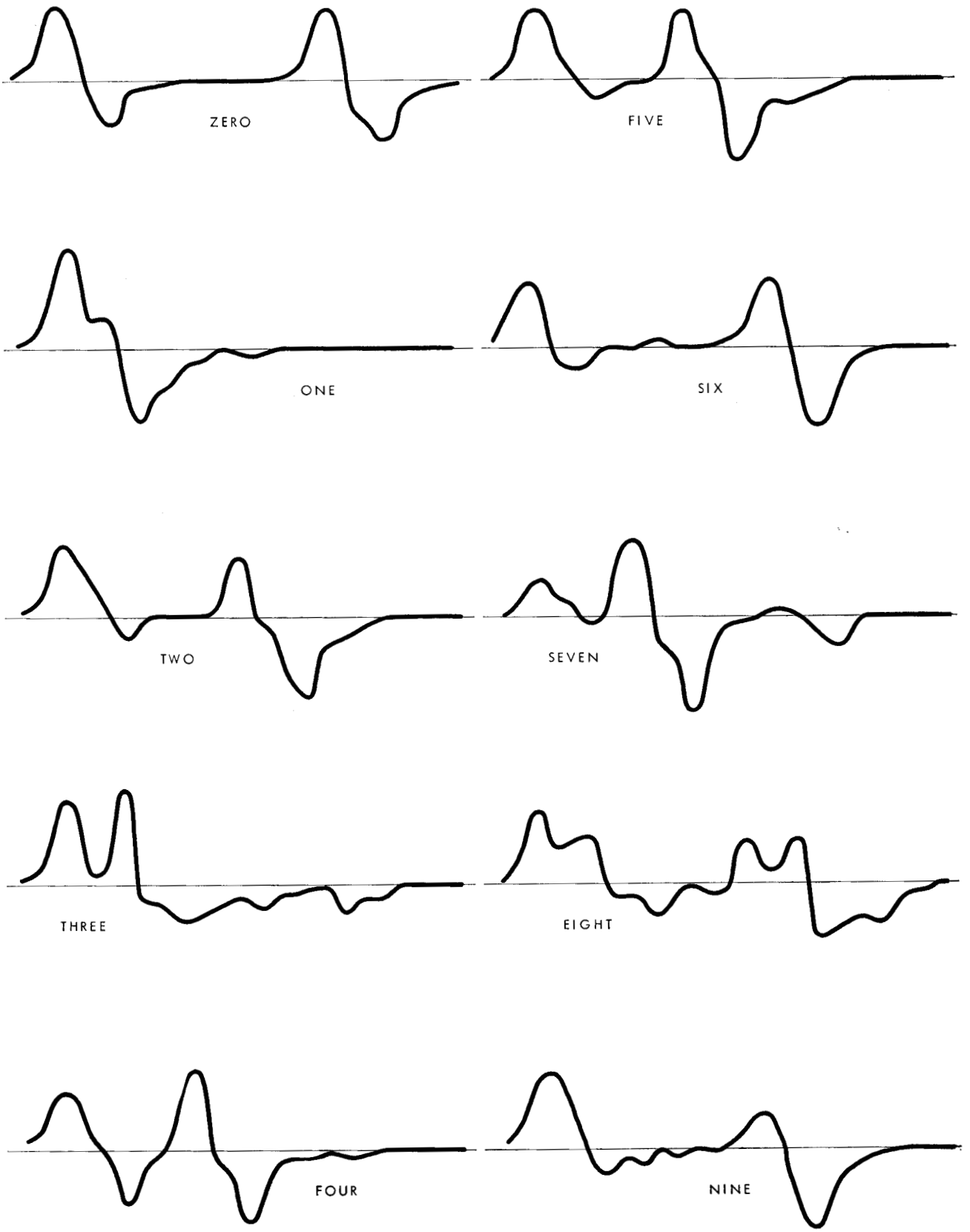


Figure 7 **Sketch of average-value, derivative-waveforms for digits zero to nine.**

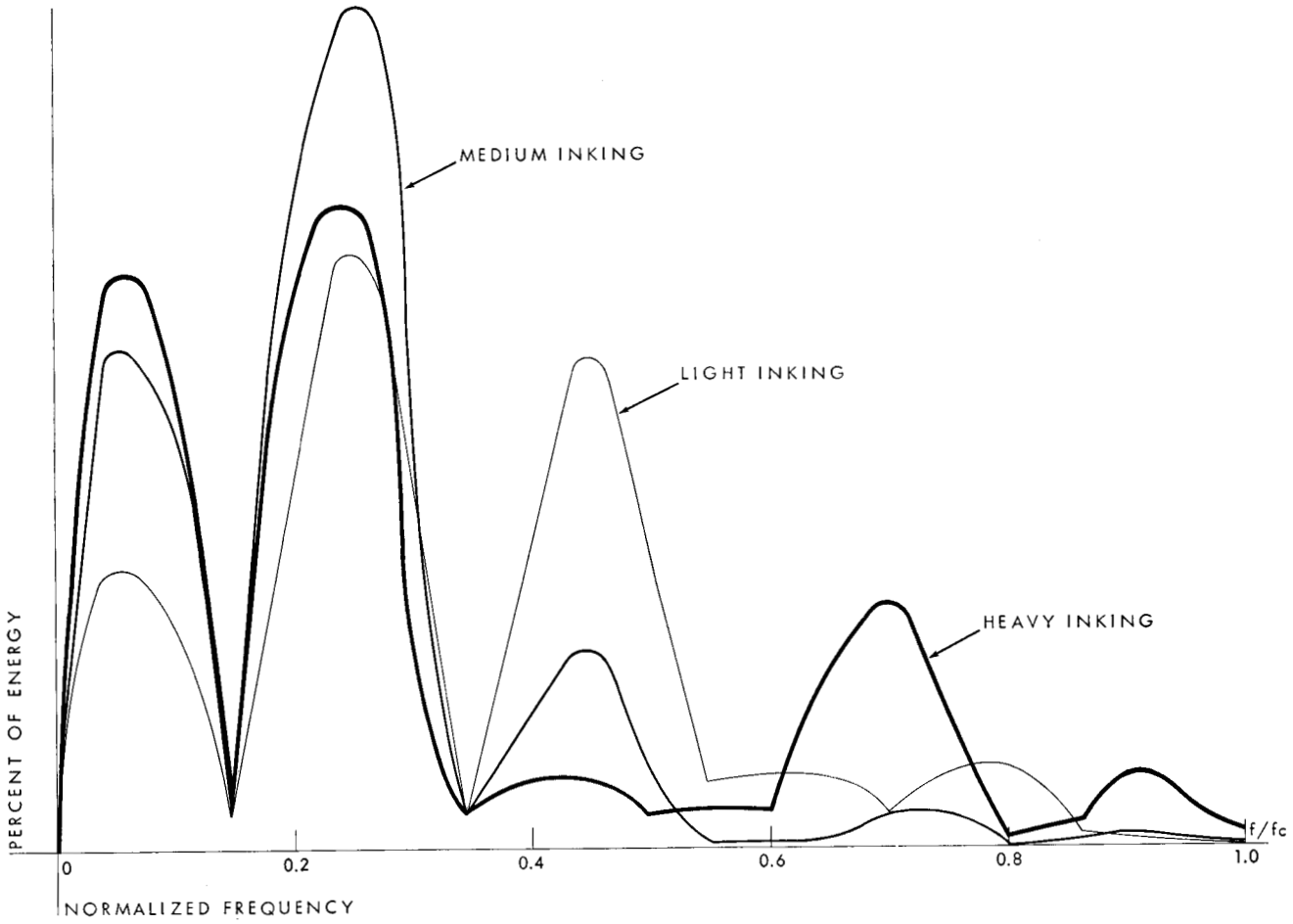


Figure 8 **Energy spectrums resulting from light, medium, and heavy inking of the printed character "five."**

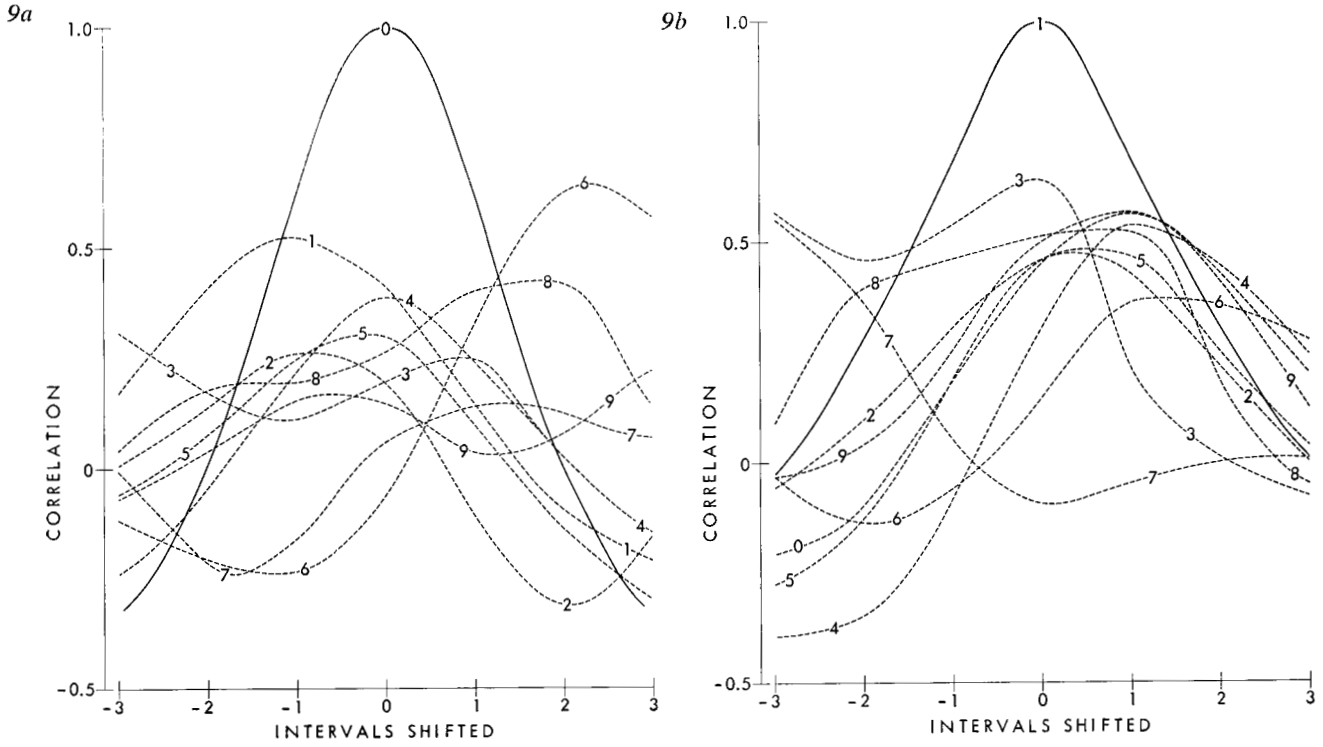
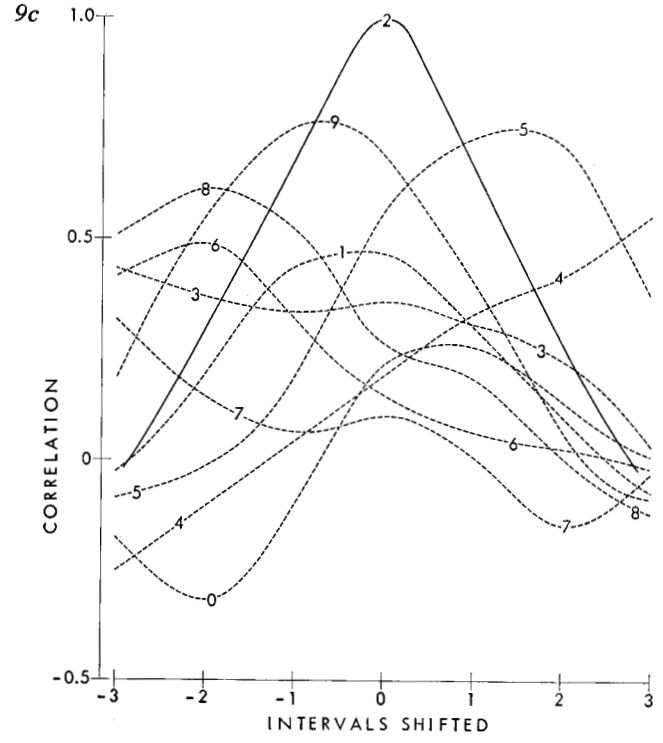
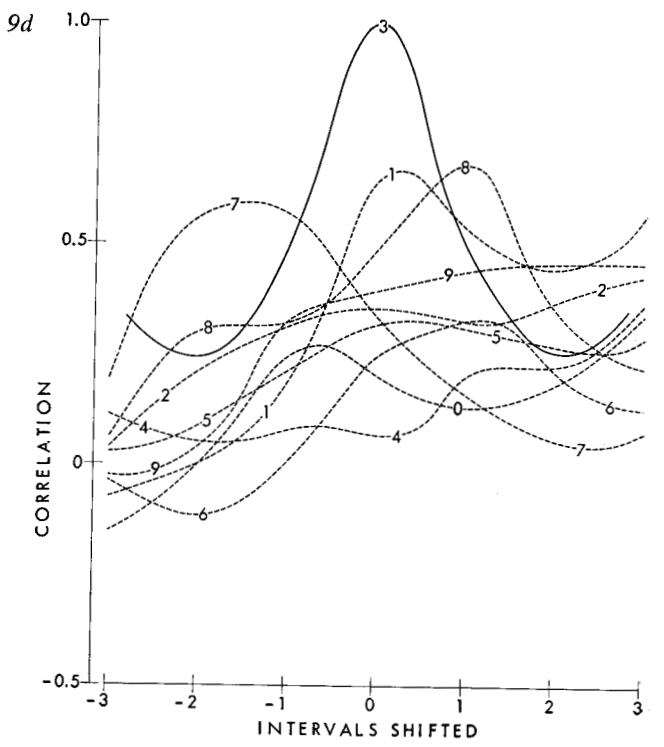
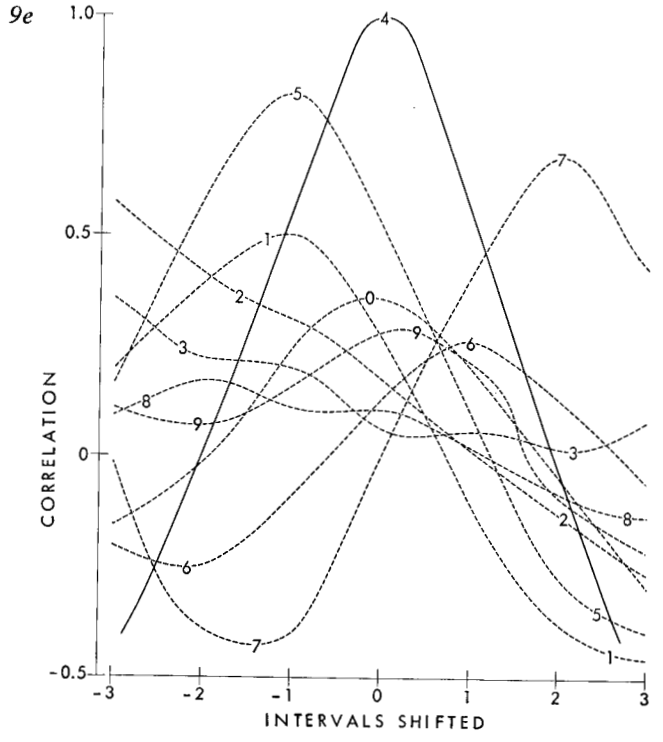
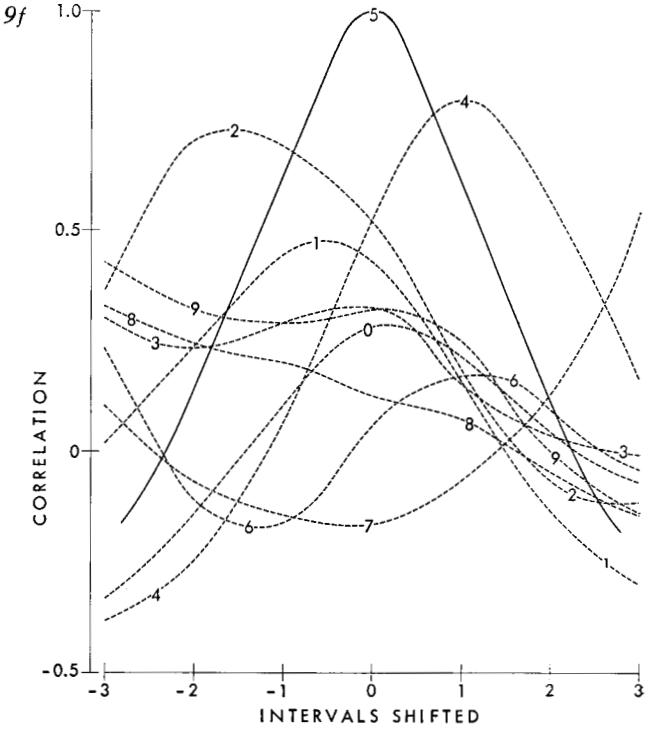


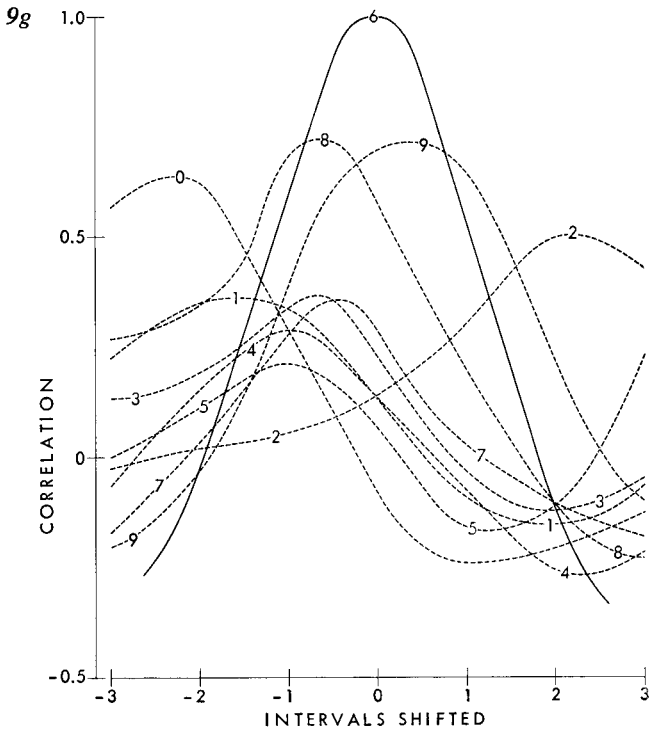
Figure 9a-k **Correlation functions for mean-value waveforms near the in-phase time.**

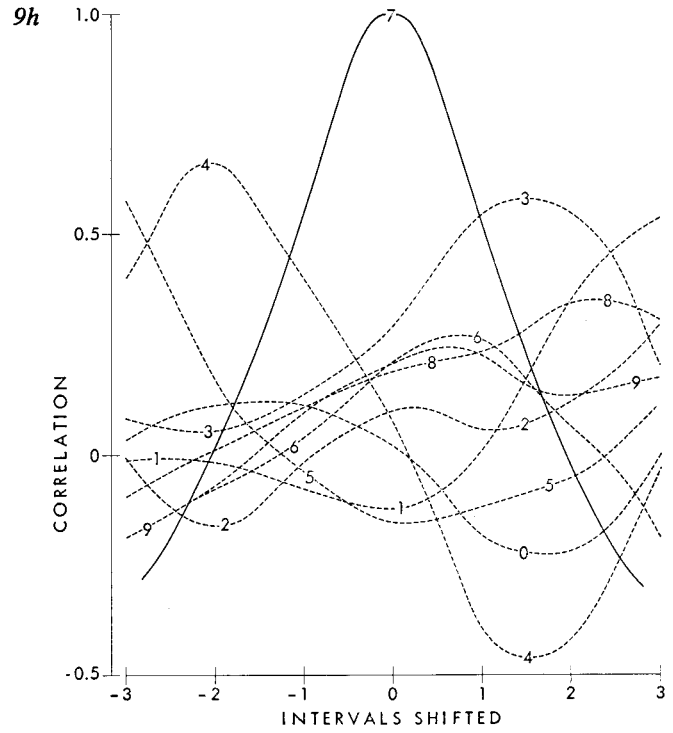

9c
1.0
0.5
0
-0.5
CORRELATION
-3
-2
-1
0
1
2
3
INTERVALS SHIFTED

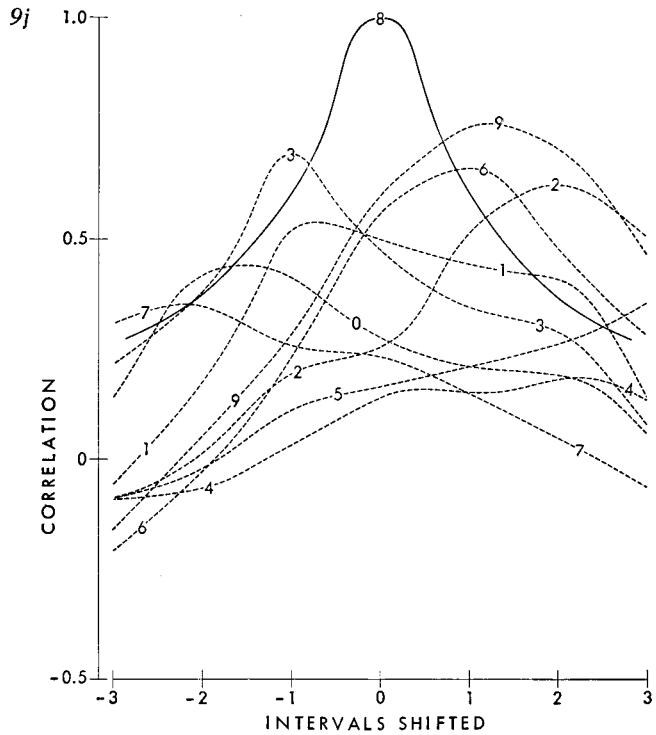

9d
1.0
0.5
0
-0.5
CORRELATION
-3
-2
-1
0
1
2
3
INTERVALS SHIFTED

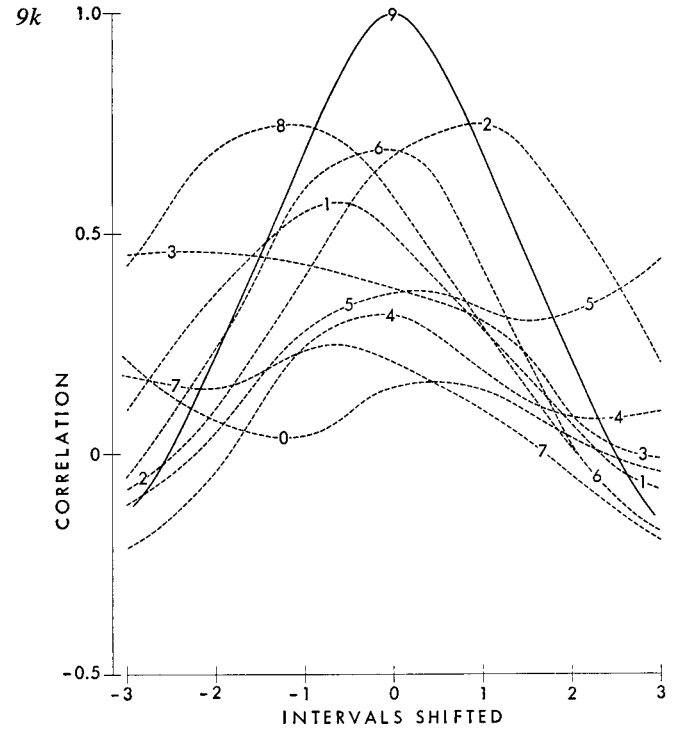

9e
1.0
0.5
0
-0.5
CORRELATION
-3
-2
-1
0
1
2
3
INTERVALS SHIFTED


9f
1.0
0.5
0
-0.5
CORRELATION
-3
-2
-1
0
1
2
3
INTERVALS SHIFTED


9g
1.0
0.5
0
-0.5
CORRELATION
-3
-2
-1
0
1
2
3
INTERVALS SHIFTED


9h
1.0
0.5
0
-0.5
CORRELATION
-3
-2
-1
0
1
2
3
INTERVALS SHIFTED


9j
1.0
0.5
0
-0.5
CORRELATION
-3
-2
-1
0
1
2
3
INTERVALS SHIFTED


9k
1.0
0.5
0
-0.5
CORRELATION
-3
-2
-1
0
1
2
3
INTERVALS SHIFTED

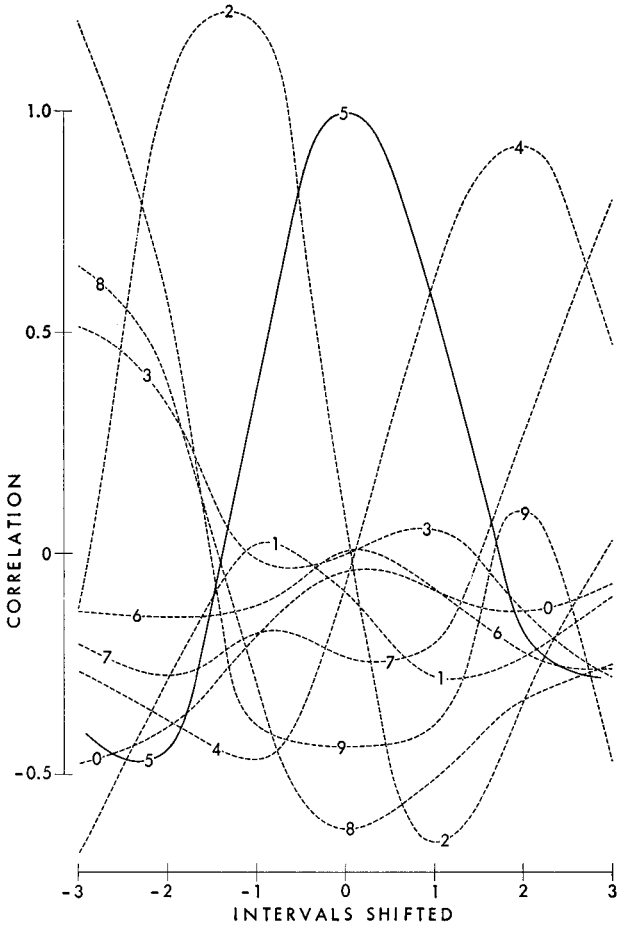


Figure 10 **Set of compares orthagonal to the mean-value vectors illustrates critical nature of the phase problem.**

• 8. *A simulation difficulty*

As a consequence of the Sampling Theorem, one can show that for the band-limited waveforms considered in this article a certain number of sample-data points are adequate to describe the waveforms. In general about two or three times the minimum number were used for the computer studies. Since a half-sample-point phasing error is possible, a brief investigation was made with four times as many sample points (120) as were used for the majority of the tests (30). The results indicated considerable improvement. The improvement is attributed to two causes:

a) The compare waveforms used in the computer tests were formed by averaging a group of scan waveforms. In the process of obtaining the average, it was necessary to phase each waveform to the partial compare waveform already formed. The phasing error was broadened and distorted relative to the waveforms from which it originated. The compare waveforms made with a larger number of sample-data points were sharper and more representative.

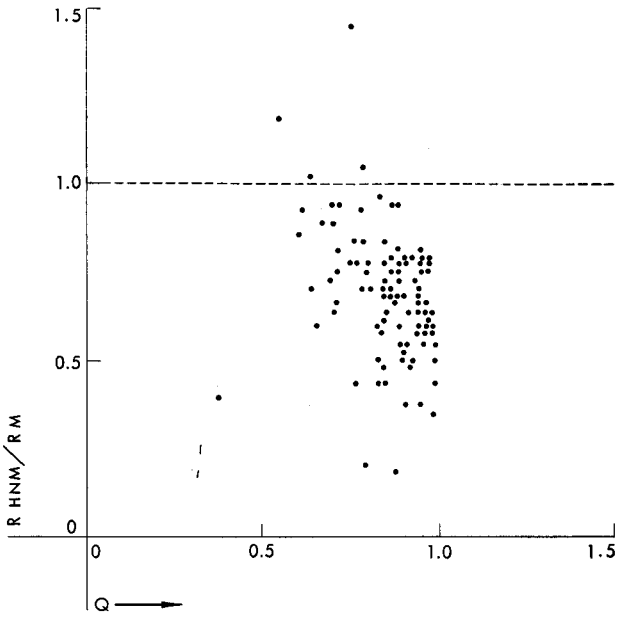


Figure 11 **Print quality variation with error.**

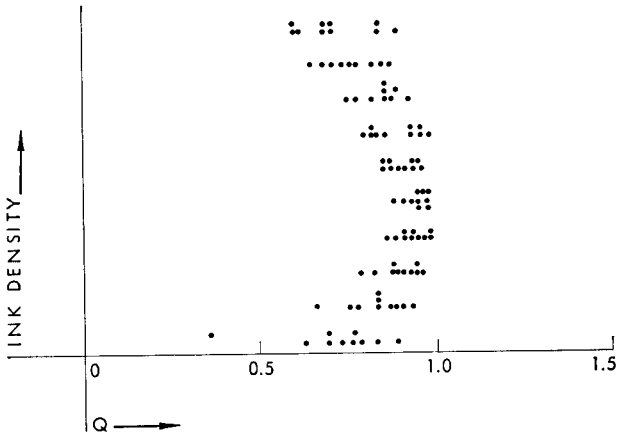


Figure 12 **Print quality variation with ink density.**

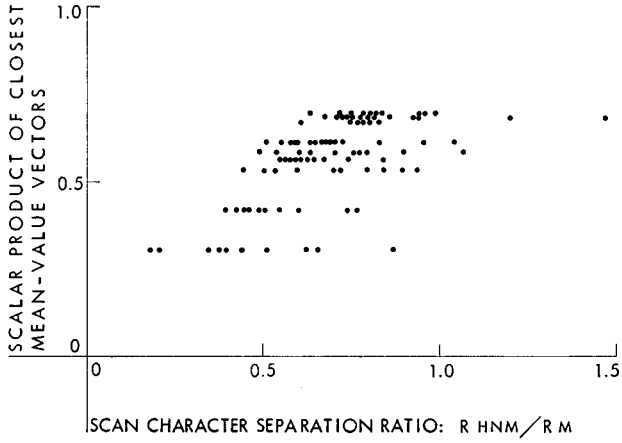


Figure 13 **Performance related to inherent separation.**

b) The larger number of sample points led to a more accurate phasing when the set of scalar products were obtained. Since the auto-correlation function changes more rapidly near the correct phase point than do the cross-correlation functions, a significant improvement results with more nearly correct phasing. This characteristic can be seen by referring to Fig. 9.

Taking more sample-data points for simulation does not mean that more taps on the delay line are required. It does mean that choosing the correct phase is important. Based on these tests, the method which "holds" the peak scalar product, as suggested earlier in this article, appears to be increasingly important.

A comparison between 30 and 120 sample-data points is shown in Table 1 for a set of ten "4's" which were scanned optically and part of which were misrecognized as "5's" with 30 sample-data points. (The "4's" gave the poorest performance in this particular test.)

No great significance should be attached to the high scalar products in this tabulation. This is a consequence of the choice of origin of the coordinate system used and the transformations involved. (Discussed in section entitled *Geometric interpretation of waveforms.*) In this instance, the scan was done optically and measured the amount of ink seen by the scanning slit; all earlier data were from magnetic scanning which involves taking a derivative of the ink presented to the slit.

Conclusion

The use of n-dimensional geometry not only has provided a means to better visualize the compare process but also has resulted in a straightforward procedure for designing the characters on a computer. Unfortunately, the design process is not completely resolved because of the conflict between machine, machine readability, human readability and appearance.

The experimental work has shown that the recognition accuracy is good with good printing. By means of a "quality factor," an indication of the quality of the printing may be determined. This is not a guarantee that errors will not occur, but only a step toward greater reliability.

In simulation, such as this, it was found desirable to use sample-data points much closer together than would be indicated by the sampling theorem. The increased number of sample points resulted in more accurate "compare" waveforms and more accurate phasing of the unknown waveform.

The single-slit scan system can be used for a limited number of characters provided a special font of characters is used. The extension of this system to reliably read alphanumeric characters of normal size that are humanly readable does not appear feasible.

Appendix A: Relation of matched filter to statistical decision theory

It is of interest to see the relationship[8] between the matched-filter approach we have discussed in this article and an approach based upon statistical-decision theory.[9] The relationship will be shown for a special case, namely where the noise is white, gaussian and band-limited to W.

Consider the case of two characters, C_1 and C_2 an rms noise, N, and a scanned signal, X. The probability that C_1 was sent, given X, is:

$$p(C_1/X) = K \exp \frac{-\sum_i (C_{1i} - X_i)^2}{2N^2}$$

Similarly for C_2:

$$p(C_2/X) = K \exp \frac{-\sum_i (C_{2i} - X_i)^2}{2N^2}$$

With X, C_1 and C_2 normalized so that:

$$\sum C_{1i}^2 = \sum C_{2i}^2 = \sum X^2 = 1 \ .$$

Table 1 **Comparison of sample-data points.**

Sample	*30-Sample Points*		*120-Sample Points*	
	"4" Compare	*Highest Other*	*"4" Compare*	*Highest Other*
1	0.960	0.932	0.975	0.925
2	.957	.940	.984	.926
3	.957	.941	.987	.921
4	.968	.947	.987	.935
5	.954	.963	.996	.960
6	.945	.958	.996	.972
7	.937	.946	.991	.966
8	.931	.949	.989	.971
9	.922	.958	.988	.977
10	.919	.971	.987	.983

Then:

$$\frac{p(C_1/X)}{p(C_2/X)} = \exp - \left\{ \frac{\sum(C_{1i}-X_i)^2 - \sum(C_{2i}-X_i)^2}{2N^2} \right\}$$

$$= \exp \left\{ \frac{\sum C_{1i}X_i - \sum C_2X_i}{N^2} \right\}$$

Taking the log of both sides:

$$N^2\{\ln p(C_1/X) - \ln p(C_2/X)\} = \sum C_{1i}X_i - \sum C_{2i}X_i .$$

Thus the noise power times the difference between the logarithm of the probabilities is equal to the difference between the scalar products.

Appendix B: Maximum possible separation of characters

In designing characters it is desirable to separate the characters from each other as far as possible. Obviously there is a limit to the amount of separation possible. In terms of vectors the question can be stated as: what is the maximum separation possible for m vectors in n-space?

A few special cases can be readily examined. When $m=2n$, the vectors can be placed along the plus and minus axes and thus give scalar products of zero (90° separation). When $n=2$, a separation angle of $360/m$ can be obtained. In general as n increases for a given number of vectors, m, the separation increases.

When $n=m-1$, the m vectors can be chosen so the vectors are symmetric and the scalar product of every product of every vector pair equal. The scalar product which is obtained is $-1/(m-1)$. Furthermore, increasing the space dimension does not lead to greater separation. Thus, where $n \geqq m-1$, a maximum scalar-product separation of $-1/(m-1)$ can be obtained.[10] Where $m/2 \leqq n < m-1$, at least 90° separation is possible; and, when $n <$ m/2, less than 90° separation can be achieved.

Footnotes and References

1. This is the approach used, for example, by Stanford Research Institute's ERMA. See K. R. Eldredge, et al., "Teaching Machines to Read," *S. R. I. Journal,* p. 18, First Quarter 1957.
2. D. W. Lytle, "On the Properties of Matched Filter," Tech. Report No. 17, Stanford Electronics Laboratory, June 10, 1957.
3. L. A. Zadeb and J. R. Ragazzini, "Optimum Filters for the Detection of Signals in Noise," *Proc. IRE,* **40,** 1223 (1952).
4. C. E. Shannon, "Communication in the Presence of Noise," *Proc. IRE,* **37,** 10 (1949).
5. C. K. Chow, "An Optimum Character Recognition System Using Decision Functions," *IRE Trans. On Electronic Computers,* **EC-6,** 247 (1957).
6. The separation can also be done in the amplitude domain. See section entitled "Choice of scan waveform."
7. The set which was tested had sufficient separation to yield meaningful results.
8. Pointed out by N. M. Abramson.
9. N. M. Abramson, "The Application of 'Comparison of Experiments' to Detection Problems," *IRE National Convention Record,* Pt. 4, p. 22 (1958).
10. A mathematical proof has been provided by Dr. R. C. Wrede.

Revised manuscript received August 17, 1959