16.01.071.003

December 15, 1959

THE APPLICATION OF DECISION THEORY TO VOICE RECOGNITION MACHINES, WITH AN ILLUSTRATED EXAMPLE

N. M. Abramson
W. E. Dickinson
F. B. Wood

ABSTRACT

This report is a revision of IBM Research Report RJ 158, "The Application of Decision Theory to Voice Recognition Machines," by N. M. Abramson, W. E. Dickinson, and F. B. Wood, March 5, 1959.

The difficult problem of the automatic recognition of spoken words is discussed. The mathematical solution to this problem (and related problems of pattern recognition) is obtained through the use of statistical decision theory. The main result of this paper is not in obtaining this solution--which is relatively trivial--but in showing that the implementation of this solution is possible in two and only two ways.

Each of these methods of implementation is discussed extensively. The first (or experimental) approach is the most powerful and leads to a probability computer. The second (or perfect signal) approach may be preferable in a few voice recognition problems. This approach leads to a correlation type of voice recognition machine.

The final section, which may be read independently, deals with three voice recognition problems and indicates how they might best be solved. It is concluded that a general-purpose, large-vocabulary, voice recognition machine must be self-programming. That is, it must be capable of selecting and automatically adjusting the parameters of a probability computer to conform to the experimentally determined statistics of the vocabulary.

A numerical illustration of the use of the decision-theory approach to word recognition is given in the Appendix.

CONTENTS

I. INTRODUCTION

The machine recognition of spoken words is a good example of a broad class of problems which we may call problems of pattern recognition. This class also includes automatic recognition of printed or written characters, automatic recognition of digital or analog information sent over a communication channel, and retrieval of stored information. The common characteristic of this class is that in each case the machine must be capable of assigning any one of a large number of possible inputs to one of a (smaller) number of outputs. In this paper we shall discuss the problem of machine recognition of spoken words, but we wish to emphasize that the techniques presented and the conclusions drawn are applicable to a wide group of problems.

The primary input to a voice recognition machine (VRM) may be taken as a voltage wave out of a microphone. We wish to identify any one of these voltage waves with a particular output. Because of the large number of possible inputs it is clearly not feasible to build what might be called a deterministic machine--that is, a machine which merely lists the response required for every possible different input. The input to a VRM must be considered statistical in nature, and consequently its operation must be based on sound statistical principles. This is not to say that it is impossible to construct a VRM of limited performance using an intuitive approach to the recognition problem. This may have been the approach used in developing Bell Telephone Laboratory's Audrey,[1] whose outputs are the ten digits. It should be possible, however, to demonstrate the correspondence of the products of the intuition with the solution based upon statistical principles.

An excellent illustration of this point is in the use of various waveform correlation techniques[2, 3] in pattern recognition. In Part III we shall show where these techniques correspond to a rigorous statistical approach and what assumptions are necessary to insure this correspondence. But if intuition appears to be a reasonable guide to the very simplest problems in automatic voice recognition, its effectiveness decreases radically with the complexity of the problem. Indeed, it is not difficult to find people whose intuition tells them that complex voice recognition problems are insoluble!

The statistical framework into which the problem of automatic voice recognition falls is the theory of statistical decisions as formulated by Wald.[4] Chow has developed a more detailed application of statistical decision theory to the waveforms encountered in a printed character recognition system.[5] Statistical decision theory allows one to solve rigorously the problem of recognizing the various inputs to a VRM as one of a number of words so as to minimize the probability of misrecognition. It is important to emphasize that the problem of automatic voice recognition can best be attacked as a problem in statistical decision theory. For any specified criterion (such as minimum probability of misrecognition), decision theory shows how to obtain the best choice as to which word was spoken based on the statistical properties of the input. The solution to this decision theory problem is quite simple to obtain, and the problem of the design of a VRM is easily resolved into the problem of implementing this solution or finding the decision probability densities.*

This last point is of the greatest importance. It should be realized that the problems in the design of a VRM arise not in obtaining the solution but in obtaining the decision probability densities in order to realize the solution. The problem of automatic voice recognition is nothing more than the problem of finding an adequate representation of the statistics of the required vocabulary.

Since the solution to the voice recognition problem hinges directly upon the statistics of the vocabulary, it is clear that we have but two choices. Either we must experimentally determine these statistics or we must assume that we know them a priori. The first method is discussed in Part II while the second method is discussed in Part III.

In Part II it is shown that the experimental determination of the statistics of the complete input waveform is not feasible. It is necessary, therefore, to focus our attention on some set of "secondary inputs" and to obtain the statistics of these secondary inputs. We then decide which word was spoken on the basis of the secondary inputs.

In Part III we make the assumption that the inputs to the VRM may be represented by perfect signals corrupted by additive white Gaussian noise. This assumption leads to a correlation type of VRM. A VRM based on this assumption will work only when the inputs conform closely to some standard saying of each word--i.e., differences in accent and pitch should not be allowed if a machine of this sort is to perform satisfactorily.

We will assume that the spoken units to be recognized are words. It is also possible to take the phoneme as the unit to be recognized. This course may have the advantage of requiring a smaller vocabulary in many cases of automatic voice recognition. Thus we may replace a vocabulary of 100, 1,000, or 10,000 words by a smaller vocabulary of, say, 80 phonemes. The choice between the use of phonemes or the use of words as the basic recognition unit is not always clear and, in fact, the choice may sometimes be considered

* These decision probabilities are sometimes called the "channel probability function."

academic. Where the word vocabulary is smaller than the number of phonemes, then the use of words as the unit appears preferable. However, because of the nature of the secondary inputs we may actually be determining and combining phonemes to recognize the words. Seeking out the invariants of the phonemes is thus useful, as it can be helpful in telling us what secondary inputs to consider.

It is clear that in many cases the sound of a word by itself is not sufficient to determine which word is being spoken. This is true not only of words which sound alike when pronounced clearly (two, to, too), but also of words which sound alike only when the pronunciation is poor (road, wrote). Thus, if a VRM is to recognize ordinary English sentences it must contain, in addition to the probability-density computer previously described, additional components to take into account the information available from context. These components might consist of logical circuitry based on the structure of English sentences. Whatever its nature, we see that it is possible to trade complexity in that part of the VRM operating on the waveform information for complexity in that part of the VRM operating on the context information.

We shall now outline, in broad terms, the manner in which statistical decision theory is applied to this particular problem. Further discussions along this line may be found in Reference 6. In the remainder of this paper we shall make two assumptions in order to simplify the discussion. The first is that we want a VRM which will minimize the probability of misrecognition of the spoken words. In doing this we will be assuming that all errors in recognition are equally costly and that to identify a "yes" as a "no" is no worse than to identify a "yes" as a "perhaps." The second assumption we shall make is that a recognition is always made. That is, in the simplified discussion to follow we do not admit the possibility of an "I don't know" response for the VRM. Neither of these assumptions is necessary for mechanization, although removal of the first assumption may cause practical difficulties in the construction of VRM's having a large vocabulary. The removal of the second assumption would require a decision feedback system to request the transmitter (message source) to restate a word or phrase in a more redundant form.

In the problem of voice recognition as already outlined, we are required to decide, on the basis of a received signal, which word (or phoneme, syllable, or sentence) was spoken. That is, we receive a voltage waveform x_j (t) from a microphone, and, given that x_j (t) is received, we calculate the probability that Word 1 was spoken, the probability that Word 2 was spoken, and so on. We shall refer to these probabilities as the a posteriori probabilities. Since we assume all errors are equally costly and we must pick one of the words, statistical decision theory (and our intuition) tells us to pick the word with the highest a posteriori probability.

How do we actually calculate these probabilities, however? The calculation of the a posteriori probabilities is based directly upon one of the simplest equations of probability theory--known as Bayes' rule.

Let $\Pr\{W_i\}$ = probability that Word i was spoken,

$\Pr\{x_j\}$ = probability that $x_j(t)$ is received,

$\Pr\{x_j/W_i\}$ = probability that x_j will be received if W_i is spoken,

$\Pr\{W_i/x_j\}$ = probability that W_i was spoken given that x_j is received--an a posteriori probability.

Then Bayes' rule tells us that the following equation may be written for the a posteriori probabilities:*

$$\Pr\{W_i/x_j\} = \frac{\Pr\{x_j/W_i\}\ \Pr\{W_i\}}{\Pr\{x_j\}} \tag{1}$$

Equation (1) contains the key to a large portion of the pattern recognition problem. We receive some signal $x_j(t)$ and using Eq. (1) we obtain the a posteriori probabilities for $x_j(t)$:

$$\begin{gathered}\Pr\{W_1/x_j\}\ ,\\ \Pr\{W_2/x_j\}\ ,\\ \vdots\\ \Pr\{W_n/x_j\}\ .\end{gathered} \tag{1b}$$

Now, remembering that to minimize the probability of misrecognition we must choose the word W_o which yields the highest a posteriori probability, we merely scan the above list and choose W_o, where

$$\Pr\{W_o/x_j\} \geq \Pr\{W_i/x_j\} \quad \text{for any } W_i \quad . \tag{2}$$

* In the above definitions we have tacitly assumed that there are only a finite or at most a countably infinite number of possible $x_j(t)$. This, of course, is not ordinarily the case, but the assumption allows us to speak in terms of probabilities rather than probability densities. Later we shall withdraw this assumption.

By Eq. (1) we see that W_o also has the property that

$$\mathrm{Pr}\,\{x_j/W_o\} \;\; \mathrm{Pr}\,\{W_o\} \geq \mathrm{Pr}\,\{x_j/W_i\} \;\; \mathrm{Pr}\,\{W_i\} \quad \text{for any } W_i. \tag{3a}$$

If we are considering catalog numbers, the digits 0 through 9 occur with equal probability. This is shown in Fig. 1, where the probability of any digit's occurring approaches one-tenth as we go to large sets of numbers. Under these conditions,

$$\mathrm{Pr}\,\{W_i\} = \mathrm{Pr}\,\{W_j\} \quad \text{for all i and j} . \tag{4a}$$

When we are dealing with words that are not numbers, the relative frequency of words plotted against rank is a distribution similar to Zipf's law (see also J. B. Estoup), refined by Mandelbrot as shown in Fig. 2. Let f_i be the frequency of occurrence of word W_i of rank r_i in a sample of N words. Then Pr $\{W_i\}$ is estimated by f_i/N for large N:

$$\mathrm{Pr}\,\{W_i\} \cong \frac{f_i}{N} . \tag{4b}$$

For example,

$$\mathrm{Pr}\,\{W_1\} \cong 100/2000 = 0.05 ,$$

$$\mathrm{Pr}\,\{W_{10}\} \cong 50/2000 = 0.025 ,$$

$$\mathrm{Pr}\,\{W_{100}\} \cong 5/2000 = 0.0025 ;$$

$$\sum_{i=1}^{r_{max} \sim 200} \mathrm{Pr}\,\{W_i\} = \sum_{i=1}^{r_{max}} \frac{f_i}{N} = 1.0 . \tag{4c}$$

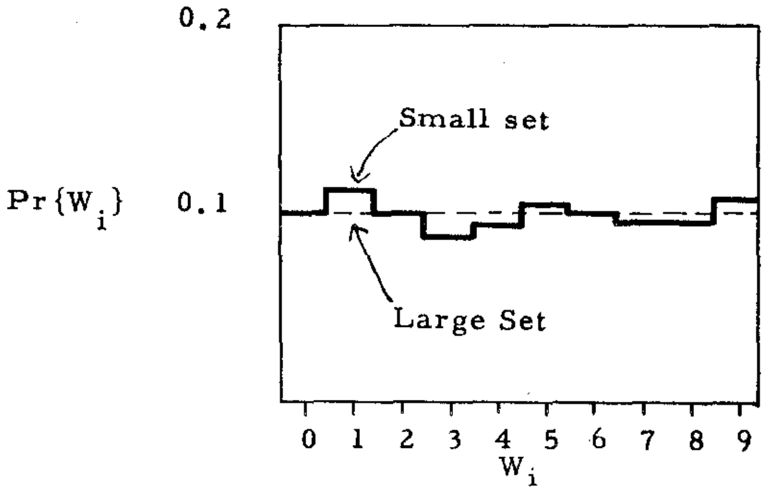


Fig. 1. Sample Distribution of Probability of Occurrence of Word "W_i" in a Catalog Number System

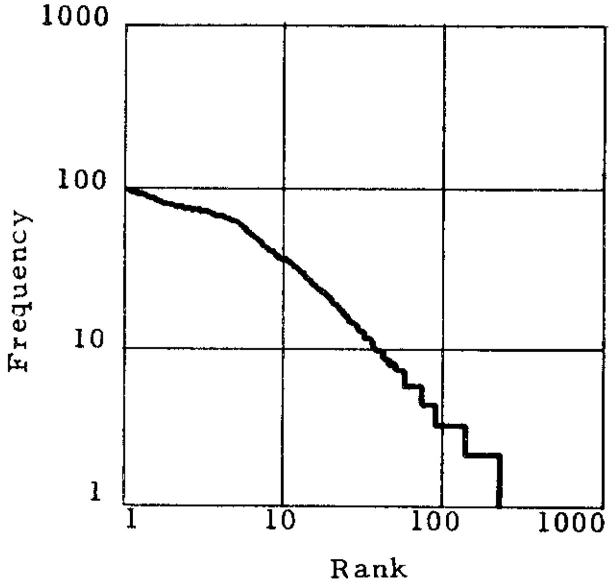


Fig. 2. Rank-Frequency Distribution of Words in a Sample of 2000 Words[7]

If Eq. (4b) holds, we must include the probability of the word occurring and use Eq. (3a) to find the most probable word to choose. If Eq. (4a) holds, then Eq. (3a) becomes

$$\Pr\{x_j/W_o\} \geq \Pr\{x_j/W_i\} \quad \text{for any } W_i \quad . \tag{3b}$$

Note the difference between Eq. (2) and Eq. (3b). Equation (3b) tells us that if we wish to identify W_o, we need not calculate the a posteriori probabilities at all! If the assumption summarized in Eq. (4a) holds, when we receive signal x_j (t) we merely select the W_o satisfying Eq. (3b). This procedure is simpler than selecting W_o from inspection of the a posteriori probabilities, since the probabilities involved in Eq. (3b) are more easily obtained than the a posteriori probabilities.

Because of their central importance in the selection of the word spoken we shall call probabilities of the form $\Pr\{x_j/W_i\}$, as in Eq. (3b), decision probabilities.

At this point let us summarize. It has been assumed that:

1. We wish to minimize the probability of misrecognition.
2. Some choice of which word was spoken must be made.
3. Any word is as likely to be spoken as any other word.

Given these assumptions, we have shown that if we receive a signal x_j (t) we must choose that word W_o which has the largest a posteriori probability, Pr $\{W_o/x_j\}$. Furthermore, we have shown that if W_o has the largest a posteriori probability it will also have the largest decision probability.

Now in many problems of statistical decision theory the decision probabilities are known a priori. If this were true of the speech recognition problem, the results summarized in the previous paragraph would contain the complete solution. Because of the nature of the speech recognition problem (and most other pattern recognition problems), however, we do not know the decision probabilities a priori. In Part II and Part III we shall derive two methods of obtaining such decision probabilities and discuss the circumstances under which each method is applicable.

Up to this point, we have assumed that the x_j (t)'s--the possible voltage waveforms out of the microphone--were discrete in nature, so that their statistics could be described by probabilities. Now we shall have to generalize a bit, to a more realistic model where the x_j (t)'s are continuous waveforms whose statistics are given by the joint probability densities, $p(\vec{x}_j/W_i)$. $\vec{x}_j$ is a vector whose components are the values of x_j (t) at certain sample times. The number of components necessary for $\vec{x}_j$ is just 2TW where T is the time duration of x_j (t) and W is the effective bandwidth of x_j (t).* Then, instead of finding the decision probabilities and selecting the largest, we must find the $p(\vec{x}_j/W_i)$, which we shall call the decision probability densities, and select the largest. This generalization does not lead to any significant change in the preceding material.

We have presented above the solution of all parts of the voice recognition problem except the calculation of the decision probability densities, $p(\vec{x}_j/W_i)$. Neither of the two methods which we shall present for the determination of

* This is strictly true only for functions which are "series-bandlimited" as defined in Reference (9). For any physically realizable signal, however, the treatment given here is quite adequate.

these quantities is wholly satisfactory in all situations. A large factor in the determination of the effectiveness of any voice recognition scheme, therefore, will be the selection of the proper method for the problem being considered. Consequently, we shall attempt to emphasize the assumptions and restrictions (both theoretical and practical) inherent in these two methods.

II. EXPERIMENTAL METHOD OF OBTAINING PROBABILITY DENSITIES

When considered as a function of $\vec{x}_j$, the term $p(\vec{x}_j / W_i)$ may be interpreted as a plot of the relative frequency of occurrence of the different $\vec{x}_j$ as shown in Fig. 3. This interpretation leads directly to our first method of obtaining the decision probability densities.

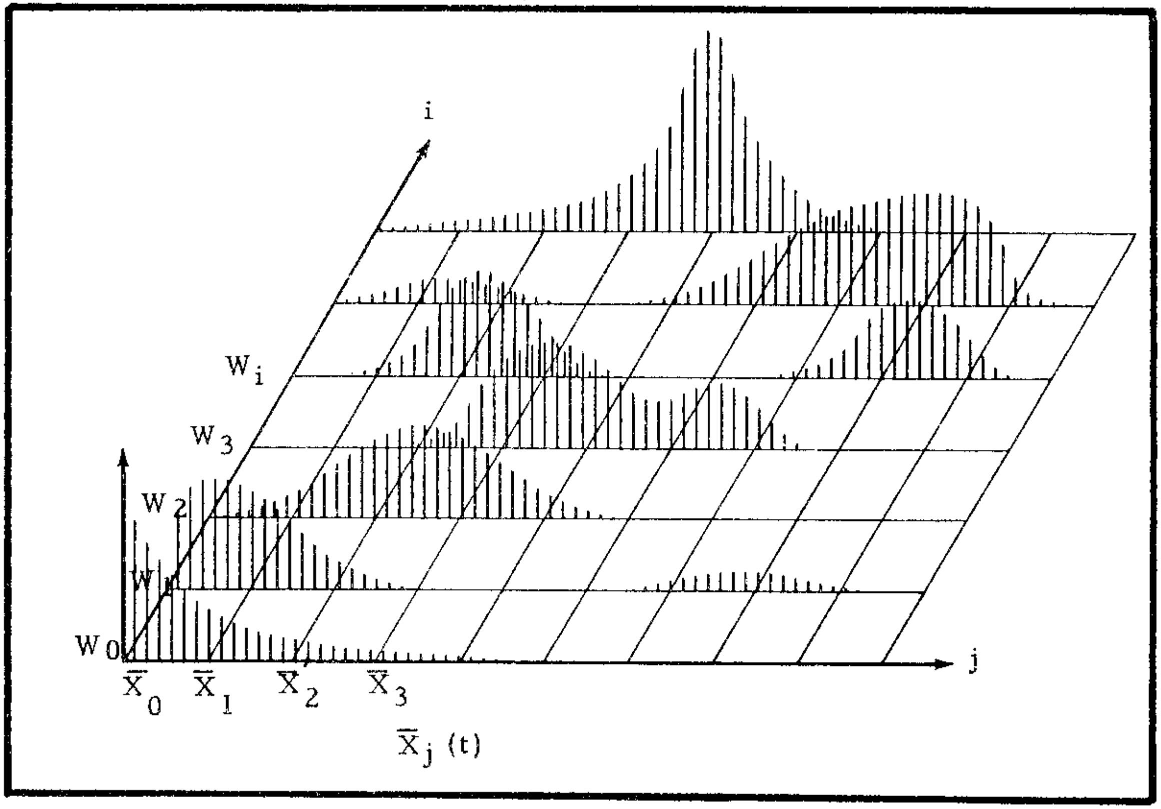


Fig. 3. A Sample Distribution of Pr $\{\vec{X}_j/W_i\}$

We might consider the possibility of obtaining $p(\vec{x}_j / W_i)$ experimentally. That is, we might have a large number of people speak word W_i into a VRM and observe the various $\vec{x}_j$ which occur. From this data the construction of $p(\vec{x}_j / W_i)$ could be attempted. However, three factors conspire to make this method highly impractical:

1. $\vec{x}_j$ will ordinarily require a large number of components--about 5,000. Our illustration in Fig. 4 shows but 24 sampling points making a part of the vector $\vec{x}_j\,(t)$.

2. The components of $\vec{x}_j$ will ordinarily not be independent random variables.

3. The components of $\vec{x}_j$ may assume one of a continuous range of values as is illustrated in Fig. 4.

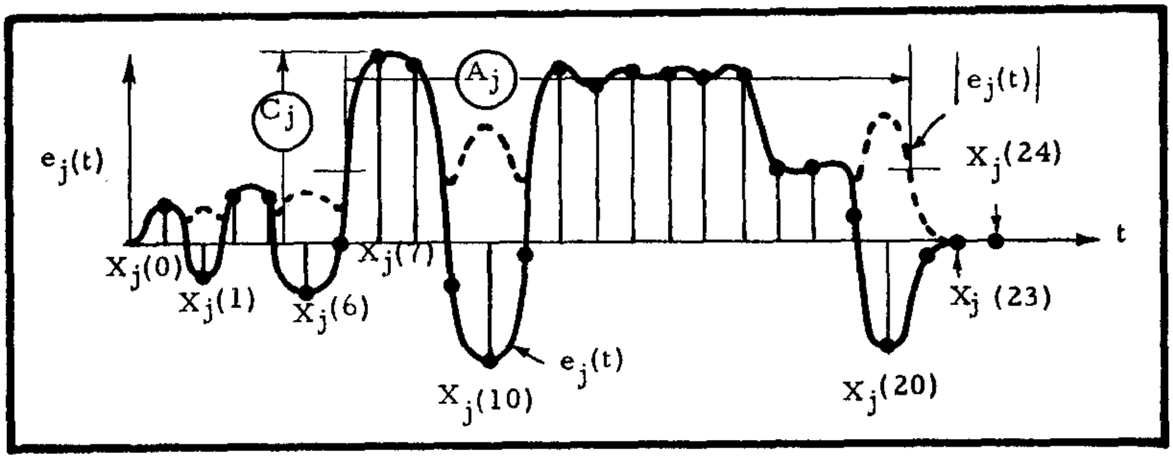


Fig. 4. Expansion of Vector $\vec{X}_j(t)$ in Sampling Points

As a matter of practical necessity, therefore, if we are to use this method of experimental determination of the statistics of the input, we must restrict ourselves to the consideration of only a small part of the input. That is, we must find some method of operating on the inputs, the random variables $x_j\,(t)$, to produce, say, the random variables A_j, B_j and C_j where the statistics* of A_j, B_j and C_j are relatively simple to obtain and where A_j, B_j and C_j retain most of the information originally contained in $x_j\,(t)$.

We shall refer to the quantities A_j, etc., as secondary inputs, in contrast to $x_j\,(t)$, which we may call a primary input. Note that any secondary input is merely a function of the primary input.

As an example of a particular set of secondary inputs, we might take A_j equal to the time duration of $x_j\,(t)$, B_j equal to the bandwidth occupied by $x_j\,(t)$, and C_j equal to the maximum value of $x_j\,(t)$.

* The subsets A_j, B_j, etc., should be chosen from the more complete set $\vec{X}$ such that there is minimum dependence and maximum discrimination ability.

These secondary inputs are shown in Figs. 4 and 5. The time duration A_j shown in Fig. 4 is derived by setting a threshold level and recording the time during which the signal exceeds the threshold. The secondary input B_j is the bandwidth of the Fourier transform G_j (t), which can be obtained from a bank of parallel filters.

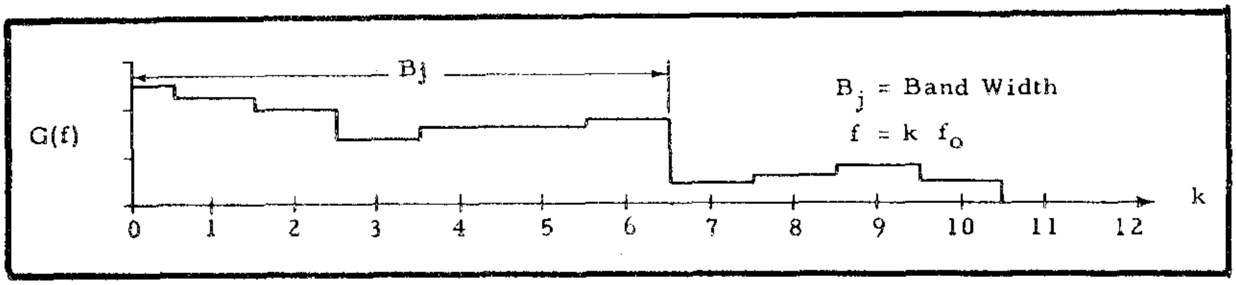


Fig. 5. Bandwidth from Fourier Transform of $\vec{X}_j(t)$

The remaining secondary input C_j --the maximum amplitude--can be obtained from a vacuum tube voltmeter that is discharged once each word-time.

We would allow only discrete values for these three variables. Then, instead of trying to obtain the complicated decision probability densities $p(\vec{x}_j/W_i)$ based on the complete signal x_j (t), we would obtain the relatively simple decision probabilities $\Pr\{A_j, B_j, C_j/W_i\}$ based on the secondary inputs A_j, B_j and C_j. Our selection of the word spoken would then depend only on the functions A_j, B_j and C_j of x_j (t) and not on the complete x_j (t). Using Bayes' rule, it is a simple matter to show that if we receive signal x_j (t) and extract from this signal A_j, B_j and C_j, then (under the three assumptions stated in Part I) we minimize the probability of misrecognition if we select word W_o where

$$\Pr\{A_j, B_j, C_j/W_o\} \geq \Pr\{A_j, B_j, C_j/W_i\}.^* \qquad (3c)$$

Note the correspondence of Eq. (3c) with Eq. (3b). In general, of course, we cannot expect the performance of a VRM using merely secondary

[*] In practice even these probabilities are difficult to obtain if A_j, B_j and C_j are not independent random variables. What is ordinarily done is to try to select A_j, B_j and C_j so that they are approximately independent. Then we may make the simplification

$$\Pr\{A_j, B_j, C_j/W_i\} = \Pr\{A_j/W_i\} \; \Pr\{B_j/W_i\} \; \Pr\{C_j/W_i\} .$$

inputs to equal the performance of a VRM using the total signal x_j (t) -- if we could build such a VRM. A question of prime importance, therefore, is how to evaluate the performance of a VRM using some given set of secondary inputs. This question is answered in the Appendix, where we show how to calculate the probability of recognition given any set of secondary inputs and the corresponding decision probabilities. In order to make this calculation, it is necessary to obtain the decision probabilities experimentally. Practically speaking, this means that we do not wish to have to test too many sets of possible secondary inputs in order to obtain a set which produces satisfactory results in a VRM.

We might sum up this experimental method of calculating the statistics of the input by saying that it ducks the problem completely. That is, instead of obtaining what is really needed to do the best possible job, we admit that it is impossible (in a practical sense) to do the best possible job and start looking for an "almost-as-good" method. Instead of trying to use all of the information presented to us in x_j (t), we throw away some of the information in order that we may simplify the handling of the information which remains, hoping that enough information remains to do an adequate job.

From the preceding discussion and the Appendix it seems probable that a search for a suitable set of secondary inputs will produce a satisfactory voice recognition scheme only in a restricted class of problems. That is, we must restrict the words which the VRM should recognize so that no two words "look alike" in terms of the secondary input. This, of course, depends to a great extent on which properties of the primary input are chosen as the secondary inputs. If we are not to require a highly complicated set of secondary inputs, however, it appears that either the words to be recognized must be carefully chosen or their number must be quite small. This solution, then, is suitable for two types of automatic voice recognition problems:

1. The recognition of a small number of "natural" words (i. e., words as they are ordinarily spoken) when the speaker may be any one of a large number of people. It is comparatively simple for a VRM of this sort to ignore problems arising from different accents, different pitches (male and female), etc., by ignoring these features as much as possible in the selection of the secondary inputs. That is, the search for satisfactory secondary inputs in this sort of problem may be viewed as the search for what have been called the "statistical invariants" of speech--those properties of speech which, in a statistical sense, convey the meaning of the word.

2. The recognition of a large number of "artificial" words (i. e., words composed of sounds designed to be machine-recognized). This type

of scheme may take the form of a machine which recognizes a large number of nonsense-words, or the less radical form of a machine which recognizes ordinary words with certain constraints put upon their pronunciation; for example, we might specify that the word "five" be pronounced "fie-yuv." The essence of this solution is the designing of the words for the secondary inputs rather than vice versa.

III. A PRIORI METHOD OF OBTAINING PROBABILITY DENSITIES

The decision probability densities are essential to the solution of the voice recognition problem. These quantities contain the clues available to us from the primary input. We must obtain these quantities somehow, and if we do not determine them experimentally as in Part II, our only alternative is to assume that we know these quantities a priori. This alternative approach forms the body of the discussions of Part III.

The assumption that the decision probability densities are known a priori may not be as unrealistic as it at first appears. In some restricted voice recognition situations, it is indeed true that we have an idea of the form of these densities, and intuitively it is quite plausible that the assumptions we make are close enough to the actual situation to yield an adequate voice recognition scheme. This view is further buttressed by the fact that, as previously mentioned, the assumption that the decision-probability densities are known a priori leads to a waveform correlation VRM--an intuitively satisfying solution.

This is the approach used by ERMA for character recognition. Its use here is based upon the assumption that the spoken words may be represented by a "true" signal disturbed by Gaussian noise. Although this approach can work well with a few printed symbols, it is not likely to be very good for a vocabulary of spoken words large enough to be useful. A discussion of the method, in some detail, is included for completeness.

Historically, the a priori method is derived from the solution to the problem of detecting a known radar signal immersed in noise. The critical assumption which it is necessary to make is that there exists a "perfect" way to speak each word which we wish the VRM to recognize.* This is not

* Earlier assumptions, that are used here also, are restated; namely,

1. All words are equally likely to occur;
2. All misrecognitions are equally costly;
3. An "I don't know" response is not permitted.

to say that the "perfect" words are always (or ever) spoken into the VRM. What is necessary is merely that we may view the input waveforms, $x_j(t)$, as a signal (the "perfect" word) corrupted by some additive noise, $n_j(t)$. We assume, therefore, that we wish to build a VRM to recognize the words W_1, W_2, ... W_n. To each word we assign a "perfect" input - $s_1(t)$ to W_1, $s_2(t)$ to W_2, ... $s_n(t)$ to W_n.

When the actual input to the VRM is $x_j(t)$, then, we say that $x_j(t)$ is composed of one of the $s_i(t)$'s plus noise voltage. That is, we say that one of the following equations must be true:

$$x_j(t) = s_1(t) + n_1(t) \quad , \tag{5a}$$

$$x_j(t) = s_2(t) + n_2(t) \quad , \tag{5b}$$

$$\vdots$$

$$x_j(t) = s_n(t) + n_n(t) \quad . \tag{5n}$$

Now what we are after is $p(x_j/W_i)$, the probability density of x_j given that word W_i is spoken, or equivalently the probability density of $x_j(t)$ given that signal $s_i(t)$ was sent. From Eq. (5) above we see that if $s_i(t)$ is sent and $x_j(t)$ is the input received, then the noise must make up the difference; in other words, the noise must equal $x_j(t) - s_i(t)$. The probability density of $x_j(t)$ given that $s_i(t)$ is sent must, therefore, equal the probability density of the noise evaluated at $[x_j(t) - s_i(t)]$. If we represent these time functions by the usual 2TW dimensional vectors and let $p(\vec{n})$ be the probability density of the noise, we may write the decision probability densities as

$$p(\vec{x}_j/W_1) = \left[p(\vec{n})\right]_{\vec{n} = \vec{x}_j - \vec{s}_1} \quad , \tag{6a}$$

$$p(\vec{x}_j/W_2) = \left[p(\vec{n})\right]_{\vec{n} = \vec{x}_j - \vec{s}_2} \quad , \tag{6b}$$

$$\vdots$$

$$p(\vec{x}_j/W_n) = \left[p(\vec{n})\right]_{\vec{n} = \vec{x}_j - \vec{s}_n} \quad , \tag{6n}$$

where the notation indicates that the noise densities are to be evaluated at $(\vec{x}_j - \vec{s}_i)$, $(\vec{x}_j - \vec{s}_2)$, ..., etc.

At this point we find ourselves in the dilemma of having expressed our solution in terms of the probability density of a "noise" which does not really exist! This has occurred, of course, because at the outset we insisted upon viewing the primary input as a corrupted version of a signal, or "perfect word," which also did not exist. Physically, then, this "noise" corresponds to a statistical spread about the "perfect" words which we must represent by some probability density. If we can assume that our fictitious noise vector has components which are independent normal random variables with their means all zero and their variances all equal to σ^2, then we can express the noise probability density as:

$$p(\vec{n}) = \frac{1}{(\sqrt{2\pi}\,\sigma)^k} \exp\left[- \frac{\vec{n}\,.\,\vec{n}}{2\sigma^2} \right], \tag{7}$$

where $\vec{n} = \bar{a}_1\alpha + \bar{a}_2\beta + \ldots + \bar{a}_k K$, or there are k orthogonal components of n.

Assuming that Eq. (7) holds, we may use Eq. (6) to obtain

$$p(\vec{x}_j/W_i) = \frac{1}{\left(\sqrt{2\pi}\;\sigma\right)^k} \exp\left[- \frac{(\vec{x}_j - \vec{s}_1)\,.\,(\vec{x}_j - \vec{s}_1)}{2\,\sigma^2} \right], \tag{8a}$$

$$p(\vec{x}_j/W_2) = \frac{1}{\left(\sqrt{2\pi}\;\sigma\right)^k} \exp\left[- \frac{(\vec{x}_j - \vec{s}_2)\,.\,(\vec{x}_j - \vec{s}_2)}{2\,\sigma^2} \right], \tag{8b}$$

$$\vdots$$

$$p(x_j/W_N) = \frac{1}{\left(\sqrt{2\pi}\;\sigma\right)^k} \exp\left[- \frac{(\vec{x}_j - \vec{s}_N)\,.\,(\vec{x}_j - \vec{s}_N)}{2\,\sigma^2} \right]. \tag{8c}$$

Now, recall that we are not really interested in the values of these decision probability densities. We merely want to identify the word W_o such that

$$p(\vec{x}_j/W_o) \geq p(\vec{x}_j/W_i) \quad \text{for all } W_i \quad . \tag{9}$$

From Eq. (8) we see that Eq. (9) may be written as

$$\exp\left[\frac{2\,\vec{x}_j\,.\,\vec{s}_o - \vec{s}_o\,.\,\vec{s}_o}{2\,\sigma^2} \right] \geq \exp\left[\frac{2\,\vec{x}_2\,.\,\vec{s}_i - \vec{s}_i\,.\,\vec{s}_i}{2\,\sigma^2} \right], \text{ for all } s_i, \tag{10a}$$

or,

$$2\,\vec{x}_j\,.\,\vec{s}_o - \vec{s}_o\,.\,\vec{s}_o \geq 2\,\vec{x}_j\,.\,\vec{s}_i - \vec{s}_i\,.\,\vec{s}_i \quad \text{for all } \vec{s}_i \quad . \tag{10b}$$

From Reference 9 we see that each of the dot products above may be written as an integral involving the original time functions, so that Eq. (10b) becomes

$$2\int_{0}^{T} x_j(t)\, s_o(t)\, dt - \int_{0}^{T} s_o^2\, (t)\, dt \geq 2\int_{0}^{T} x_j(t)\, s_i(t)\, dt - \int_{0}^{T} s_i^2(t)\, dt \text{ for all } s_i(t) \qquad (10c)$$

Equation (10c) above indicates the operation which must be instrumented for this type of VRM. Despite the imposing appearance of this equation, it is possible to instrument it either electronically or optically. However, it is impractical to instrument this equation for the complete speech waveform. The waveform of a word ordinarily consists of several bursts of sound with the energy concentrated in a small number of relatively narrow frequency bands. The construction of an electric filter or a small photographic transparency thus becomes impractical if $x_j(t)$ is the pure waveform of the word. Also, it is quite unlikely that the variations in these waveforms can be described as Gaussian noise. For these reasons, therefore, we might resort to a procedure analagous to that of selecting a "secondary input" as discussed in Part II. Instead of taking $x_j(t)$ and $s_i(t)$ as the waveforms of the received word and the "perfect words," respectively, we would take $x_j(t)$ and $s_i(t)$ as the envelopes of these waveforms, or perhaps as the envelopes of the waveforms after they have been passed through a bandpass filter. These envelope waveforms have a better chance of being useful, but, here too, a weakness is apparent. Many spoken sounds have considerable variation in duration while others do not, and these variations cause parts of the waveform to be out-of-phase even though the received and perfect waveforms are in phase at the start. This phasing difficulty makes it unlikely that the independent Gaussian assumption will hold.

This correlation approach is best suited to signals that are well timed and that have nearly Gaussian disturbances.* Spoken-word signals do not have these attributes. We know, for example, that ERMA-type character recognition requires a special type font to be useful. When its scanned signal is out-of-phase with the compare signals by about five percent, the recognition accuracy becomes very poor. We can require the equivalent of a special font by calling upon the speaker to use certain words and pronunciations. We would have great difficulty in controlling his single-sound rate. For these reasons the approach described in Part II appears to be the better one for all but very-small-vocabulary VRM's.

* An alternative is to remove the "time" element from the speech wave, but this may be considered to be one approach to the secondary inputs covered in Part II.

IV. SUMMARY

In this final section we shall select several specific voice recognition problems and outline what we feel are reasonable methods of attacking these problems in view of the preceding material. We shall take a good deal more liberty in Part IV in presenting conclusions which depend to some extent on the subjective opinions of the writers. This is necessary if any significant conclusions are to be extracted from the technical (and, therefore, we hope, objective) data presented in Parts I, II and III.

First consider a relatively simple problem of automatic voice recognition. Let us say we are interested in operating a three-position switch by spoken commands. We wish the switch to be capable of operation by people with radically different accents. Clearly this is a case where we can design the input language of the VRM. We do not want to use the perfect signal technique since there are likely to be wide variations of any given command when spoken by the different operators. We shall, therefore, employ the experimental technique as given in Part II. We select an easily instrumented secondary input, say the duration of the command, and then design three command words which differ radically in duration. The design of the VRM is then trivial--the VRM is merely a timing device.

If we wished to increase the number of possible positions of the switch (i. e., the vocabulary of the VRM) then we might have to select an additional secondary input (say the frequency band of maximum energy), design our words more carefully, and include a simple probability computer as part of the VRM.

Next, we take the problem of the recognition of a small set of words as spoken by one person. The perfect signal approach of Part III applies in this case, and either electronic or optical correlation techniques can be used. Because the number of phonemes in this set of words is greater than the number of words, we can consider the word as the unit of input.

A considerably more difficult problem than the two considered above is the recognition of sentences consisting of a small number of words (perhaps 500), selected so that no two sound alike. We assume that the VRM is required to recognize these words as spoken by several different people. Because we shall probably be unable to define a reasonable set of "perfect words," the problem calls, in this case, for an experimental determination of the decision probabilities of a set of secondary inputs as described in Part II. The machine will be made to select a small number of words (five or less) that best "fit" the unknown waveform. The final selection of an output word from among these five might then be made by logical circuitry using the context.

It is perhaps unnecessary to remark that a VRM as described above would be quite difficult to construct. In the first place, the selection of an adequate set of secondary inputs is a problem for which no analytic methods of attack are known. It is true that the intuition can be of great help in the selection of the secondary inputs, but the primary reliance must be on trial-and-error methods. At this point it is necessary to point out that the law of diminishing returns conspires to make large vocabularies for the VRM difficult to achieve. We assumed in Part II that it was possible to choose secondary inputs which were independent of each other. As we increase the vocabulary size, however, it is necessary to increase the number of secondary inputs; the more secondary inputs we select, the harder it is to insure that they are independent. In effect this means that we do not obtain as much information from the last few secondary inputs as we do from the first few. As we increase the size of the vocabulary required of the VRM, we find that the number of secondary inputs necessary for satisfactory recognition probabilities increases rapidly. The experimental determination of the decision probabilities is perhaps the greatest single deterrent to the design of VRM's of large vocabulary.

Finally, we note that the design of logical circuitry (taking into account the context) which might operate upon the most probable (taking into account only the secondary inputs) words also presents formidable difficulties.

The conclusion to be drawn from the gloomy picture presented above is, we feel, inescapable. It is undoubtedly possible to design a VRM to recognize English sentences composed from a small vocabulary with certain confusing words deleted. The labor involved in obtaining the necessary decision probabilities and the sheer mass of equipment necessary to instrument the probability computations and logical operations, however, seem to indicate that if such a machine were built it could be nothing more than a scientific toy.

If a practical VRM of the type discussed above (or a more versatile VRM) is to be built, then, it will not be built by techniques presently available. It is important to note at this point that we are talking about the techniques of implementing the solution to the voice recognition problem and not the solution itself. The solution of the voice recognition problem lies in the decision probabilities and there exist two and only two ways of obtaining these probabilities--either determine them experimentally or obtain them from a priori knowledge.*

* It is, of course, possible to conceive of mixtures of these two basic methods.

For general purpose large-vocabulary voice recognition we must use the experimental method because we do not have the necessary a priori information. The stumbling block in all presently available techniques, then, is clearly the vast amount of work necessary to obtain the statistics of the words of the vocabulary and, after they are obtained, to program these statistics into the probability computer part of the VRM.

The solution to this problem too, we feel, can be clearly defined (although it is by no means simple to accomplish). The tremendous volume of work in obtaining the statistics and programming the probability computer must be done automatically. That is, for each word of the required vocabulary, the VRM must be capable of reading the statistics contained in a number of samples of this word and (after being instructed as to the word from which the samples are derived) adjust parameters of the probability computer accordingly. What we are describing, then, is a method by which the VRM acts as a self-programming (or learning) machine. If the VRM were able to obtain the necessary statistics and program itself automatically, we would be able to use a large number of secondary inputs upon which to base the decision probabilities.*

The design of a general-purpose large vocabulary VRM is merely the design of a practical method of obtaining and utilizing the statistics of the different inputs. The solution of this voice recognition problem is almost trivial--it is the implementation of the known solution in the manner described above which presents the difficulties.

* It is interesting to note that some elementary forms of this type of self-programming have been studied.[10, 11, 12] One of these references (11) even mentions the possibility of using a particular type of learning machine (the Perceptron) to recognize spoken words.

REFERENCES

1. K. H. Davis, R. Biddulph and S. Balashek, "Automatic Recognition of Spoken Digits," Journal Acoustical Society of America, Vol. 24, November 1952, pp. 637-642.

2. K. R. Eldredge, F. J. Kamphoefner and D. H. Wendt, "Teaching Machines to Read," SRI Journal, First Quarter 1957, pp. 18-23.

3. W. E. Dickinson, "A Character-Recognition Study" (to be published in IBM Journal of Research and Development).

4. A. Wald, Statistical Decision Functions, John Wiley and Sons, New York, 1950, 179 pp.

5. C. K. Chow, "An Optimum Character Recognition System Using Decision Functions," IRE Transactions on Electronic Computers, Vol. EC-6, December 1957, pp. 247-254.

6. N. M. Abramson, "Seminar on Decision Theory," IBM Research Memorandum RJ-MR-8, San Jose, California, January 1958, 45 pp.

7. B. Mandelbrot, "An Informational Theory of the Statistical Structure of Language," pp. 486-502 in Communication Theory (London Symposium, September 1952), W. Jackson, Ed., Butterworths, London, 1953, 532 pp.

8. C. E. Shannon, "Communication in the Presence of Noise," Proceedings IRE, Vol. 37, January 1949, pp. 10-21.

9. W. Peterson, T. Birdsall and W. Fox, "The Theory of Signal Detectability," Transactions IRE, PGIT, Vol. 4, September 1954, pp. 171-212.

10. R. M. Friedberg, "A Learning Machine," IBM Journal of Research and Development, Vol. 2, January 1958, pp. 2-13.

11. F. Rosenblatt, "The Perceptron," Cornell Aeronautical Laboratory Report No. VG - 1196 - G - 1 (Office of Technical Services, U.S. Dept. Commerce, PB 151247), January 1958, 268 pp.

12. F. Rosenblatt, "Two Theorems on Statistical Separability in the Perceptron," Cornell Aeronautical Laboratory Report No. VG - 1196 - G - 2 (Office of Technical Services, U.S. Dept. Commerce, PB 151247S), September 1, 1958, 42 pp.

APPENDIX

A. The Measurements

In this Appendix we will show how measurements of the waveform of a spoken word are used to decide which word was spoken. The example used to illustrate the method is a simple one, but one which may be useful in its own right. Only simple mathematics is used. The method is that of statistical decision theory as covered in the main body of the report.

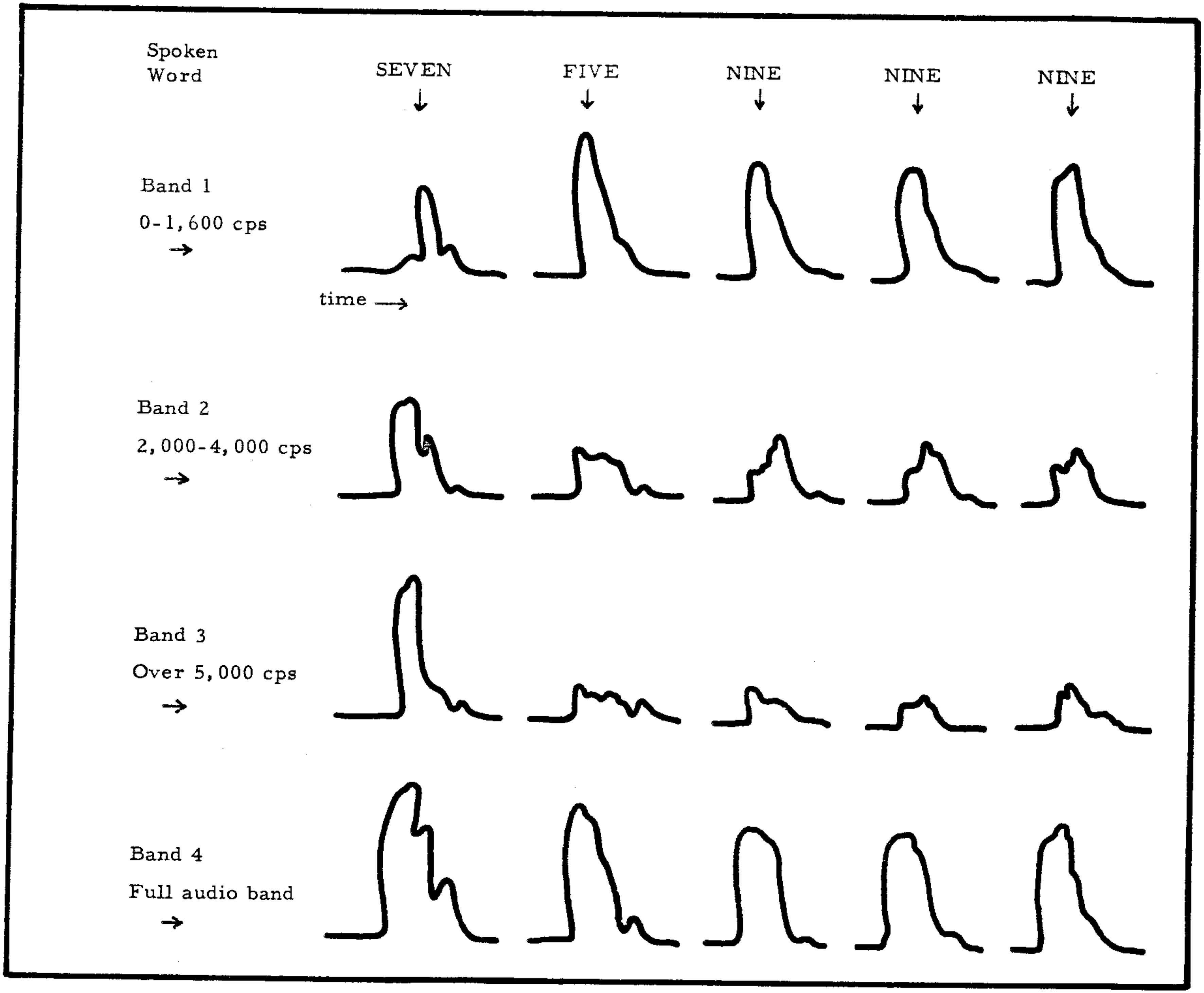
The set of spoken words or the vocabulary considered here consists of the digits ZERO through NINE. All the words were spoken by one speaker during one recording session. This person had considerable experience with recording words on the equipment used. Thus, the amount of variation in the words was likely to be less than for an "average" speaker or even for the same speaker on different days. The number of words in the sample was small; only 25 repetitions of each of the ten words were recorded. These 250 spoken samples were recorded randomly.

After recording the words on magnetic tape, the recorded signal was played back and fed through three filters. These filters separated the signal into its low, medium, and high frequency components. The specific frequency bands were: 0-1,600, 2,000 - 4,000 and above 5,000 cycles per second. The envelope of the rectified voltage at each filter output was obtained. These three envelopes were recorded using a Brush recorder. The envelope of the full-frequency band was recorded along with each of these three bands. This gave us four envelopes, or waveforms, per word. The measurements used in the calculations we will describe were taken from these waveforms. For identification, the low-frequency band is called Band 1; the middle, Band 2; the high, Band 3, and the full-frequency, Band 4.

The general appearance of these four waveforms is shown in Fig. A1.* The words shown are a SEVEN, a FIVE and three NINES. As the FIVE and NINE give similar waveforms, three NINE'S are given to show some of its variations.

Altogether we had 1,000 waveforms to make measurements upon.* The measurements, called "secondary inputs" in Part II of this report, were made

* The time scale, with amplitude, is distorted because the recorder-pen swings on an arc.


Spoken
Word
SEVEN
FIVE
NINE
NINE
NINE
Band 1
0-1,600 cps
time
Band 2
2,000-4,000 cps
Band 3
Over 5,000 cps
Band 4
Full audio band

on each waveform and are shown in Fig. A2. A few words about these measurements. The different measurements were chosen to describe the waveform quantitatively. Other, better, measurements might have been selected.

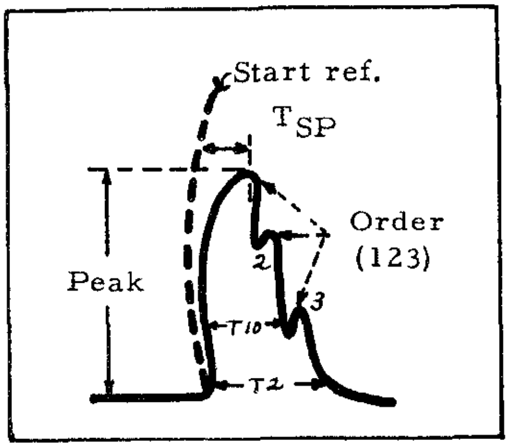


Fig. A2. Measurements Made on Waveforms

The measurement called PEAK gives the height of the highest peak in the waveform. The measurement TSP gives a measure of when the peak occurs relative to the start of the signal. The measurement T10 gives the time during which the waveform exceeds a ten-unit level of amplitude. If the waveform dips below the ten-unit level and later exceeds it again, this is noted. If this information is used as part of the measurement it is called a T10S measurement. A T2 or T2S measurement is the same as the T10 or T10S except that a two-unit level is used. The ORDER measurement is described as follows. The peaks are numbered in chronological order and these numbers are then arranged in order of their peak's amplitude (e.g., a "312" ORDER means the 3rd peak to occur is largest, the 1st peak second largest, and the 2nd peak smallest.) All five of these measurements, plus identification of word, frequency band, and recording sequence, were punched in a card for each waveform. This was a tedious job and about one man-week was required to take the measurements and punch the 1,000 cards.

Generally, the number of different values and the ranges of these values were not the same for the five measurements. For convenience in calculation it was desirable that the values of all measurements be of the same range. This was done by a simple transformation, as follows. The 1,000 cards were first sorted into the four frequency, or waveform, bands. Within each band the cards were then sorted on each measurement to arrange them in increasing value. Ten value ranges were chosen from a listing of each of these "sorts" that put about the same number of waveforms in each range. A new deck of cards was then punched which contained the word and measurement identification as well as a converted value ranging from zero to nine depending upon which range the original measurement value fell into. These cards, a 250-card deck for each measurement, formed the basis for all further calculations.

To recapitulate, the recorded spoken-word signals were filtered into three frequency bands. The envelopes of these bands and the full-frequency band were recorded. These recorded waveforms were measured and the values obtained on each measurement assigned to one of ten ranges.

We can now describe the method by which the probability of correct recognition is obtained from these measurements. Throughout we will assume that all our spoken words will be used with equal frequency.

For each word and each measurement we will calculate the probabilities of the ten values. The probabilities are obtained by the expression*

$$P_i = \frac{n_i}{\sum_{i=o}^{9} n_i} = \frac{n_i}{n_o + n_1 + \ldots . n_9} , \qquad \text{(A1)}$$

where (for a given WORD and MEASUREMENT) P_i is the probability that the value i will be measured; n_i is the number of times the value i has been measured; and $\sum n_i$ is the number of measurements that have been taken.

These probabilities represent the statistics of a word when measured a certain way. We can combine measurements. A set of these probabilities will be put in matrix form where the matrix has a column for each of the ten values and a row for each measurement. We will call this a "test matrix." An an example, the test matrix for the word ZERO with the first measurement, PEAK-Band 1, and the second measurement, T10-Band 3, is shown below:

	values									
	0	1	2	3	4	5	6	7	8	9
PEAK-1	0.0	0.0	0.29	0.38	0.25	0.04	0.0	0.04	0.0	0.0
T10-3	0.0	0.13	0.35	0.43	0.09	0.0	0.0	0.0	0.0	0.0

If the value read by the first measurement is independent (i.e., does not affect the value read by the second measurement), then the probability that the word ZERO would give rise to a value 2 on the first measurement and a value 4

* In Section C of this Appendix another possibility is considered.

on the second measurement is the product of the two probabilities: 0.29 x 0.90 = 0.0261. For more measurements than two, all combinations need to be considered.

In the same way, if we make test matrices for these measurements for all the words, we can obtain the probability that this "reading" (i.e., 2,4) came from each of the words.

The ten probabilities, so obtained, form one of the columns of a second matrix we will call the "experiment matrix." This "experiment matrix" will have a row for each word, and a column for each "reading." The readings for this example run from 0,0 to 9,9. The idea is illustrated below. Here the matrix is filled out only for reading "2,4." The measurements used are PEAK, Band 1 and T10, Band 3.

				Readings			
Word	0,0	0,1	. . . .	2,4	. . . .	9,8	9,9
ZERO				0.0261			
ONE				0.0			
TWO				0.005			
THREE				0.0			
FOUR				0.0			
FIVE				0.0			
SIX				0.0			
SEVEN				0.1872			
EIGHT				0.0840			
NINE				0.0			

A completed row of this matrix would list a word's probability of causing each of the "readings." Since a word must give rise to one and only one "reading," the sum of each <u>row</u> is one. (The readings are mutually exclusive and the probability of <u>some</u> reading is a certainty.) The sum of <u>all</u> columns is ten, equal to the number of words. Thus, it follows that the sum of a column is ten times the probability of the occurrence of that column (reading). The sum of the column in the above matrix is 0.3023; the probability of reading "2,4" is, therefore, about 0.03. The most likely word and the only logical choice for reading "2,4" is SEVEN. If a SEVEN had been spoken, the reading "2,4" would occur with probability 0.1872. The probability that a SEVEN would be spoken from our vocabulary is 0.1 (all words equally likely). Thus, the probability that a spoken word from our vocabulary would be a SEVEN and the reading would be "2,4" is 0.1 x 0.1872 = 0.01872. The probability that SEVEN is the correct word for this reading is 0.01872/0.03023 or 0.62. This little calculation is according to Bayes' rule. We can summarize it in equation form as:

$$\text{Pr (SEVEN/Reading 2,4)} = \frac{\text{Pr (SEVEN) Pr (Reading 2,4/SEVEN)}}{\text{Pr (Reading 2,4)}}$$

The notation Pr (A/B) is read as the probability of A given B. Returning to our example--on the average, we will be right 62% of the time and wrong 38% of the time, if we always choose SEVEN when reading "2,4" occurs. If we had the "experiment matrix" completely filled out we could perform the same calculation for each reading. The probability of correct recognition for the set of measurements would be the weighted average of the probability of correct recognition for each reading. We have seen that the probability of correct recognition for a reading is the ratio

$$\frac{\text{Maximum probability in the column}}{\text{Sum of all probabilities in the column}} .$$

The probability of the occurrence of a reading is also a ratio,

$$\frac{\text{Sum of all probabilities in the (reading) column}}{\text{Sum of all columns}}$$

Combining these (sum of all the products) we find the probability of correct recognition for a set of measurements is the ratio of the sum of the column maxima to the total of all columns (here equal to 10).

The method of calculation now appears clear:

1. Choose a set of measurements and form the ten "test matrices."
2. Calculate the "experiment matrix" for these measurements.
3. Obtain the "probability of correct recognition" from the experiment matrix.

Unfortunately, there are two complications. The first problem is that we must choose our set of measurements so that they are an independent set. If they are not independent of each other, we will get less information from a measurement than our numbers indicate. Our calculated probability of correct recognition will be too high. The second problem is that working out the entire experiment matrix may not be feasible. For example, to calculate an experiment matrix for 4 measurements on the IBM 650, would require about 12 hours; for 5 measurements, about 20 working days. The solution to the first problem was to choose measurements thought to be reasonably independent. For the second problem we chose "readings" according to their probability. Thus, we evaluated only the most probable columns of the experiment matrix. We have found that such results rapidly approach the results obtained for the complete matrix.

When we find the probability of correct recognition for the set of words, we also work out the probability of correct recognition for each of the words. This is helpful when we try to decide which measurements to add or subtract from the set.

B. Calculations and Results

After the original values had been converted into the ten value ranges and the probability of these values had been calculated for each digit, the probabilities were charted and are shown in Fig. A3. The measurement headings are the same as those used in Fig. A2. The best single measurement is T10S-Band 4. Its probability of correct recognition is 56.8%. It is perfect* on the words ZERO, ONE, and EIGHT; does well on SIX and FIVE; but does not recognize THREE, FOUR, or SEVEN at all.

Single measurements are unable to recognize our words well enough. We must add other measurements. Because the measurements should be as independent as possible, certain measurements should never be combined. T10 and T10S for the same band are merely reclassifications of the same data. The same is true for T2 and T2S. There will be high dependency between T2 and T10 in the same band, since they both express the time during which the waveform exceeds a certain level. Dependency will exist between other measurements to a lesser degree. Since frequency Bands 1 and 3 are best separated, they are likely to be less dependent; hence, if a type of measurement must be used twice, we probably should use these two bands. It is possible to test the dependency of two measurements, but we did not do so and will use the rough guides given here to combine measurements.

Assuming independence, a choice between sets of measurements can be made by choosing the set that gives the better probability of correct recognition. For example, T10S-Band 4 and ORDER-Band 4 give 77% recognition; T10S-Band 4 and TSP-Band 2 give 73% recognition. The first set is better. Although both of these sets are better than T10S-Band 4 alone (56.8%), the

* For these data. A discussion of the data will follow later.

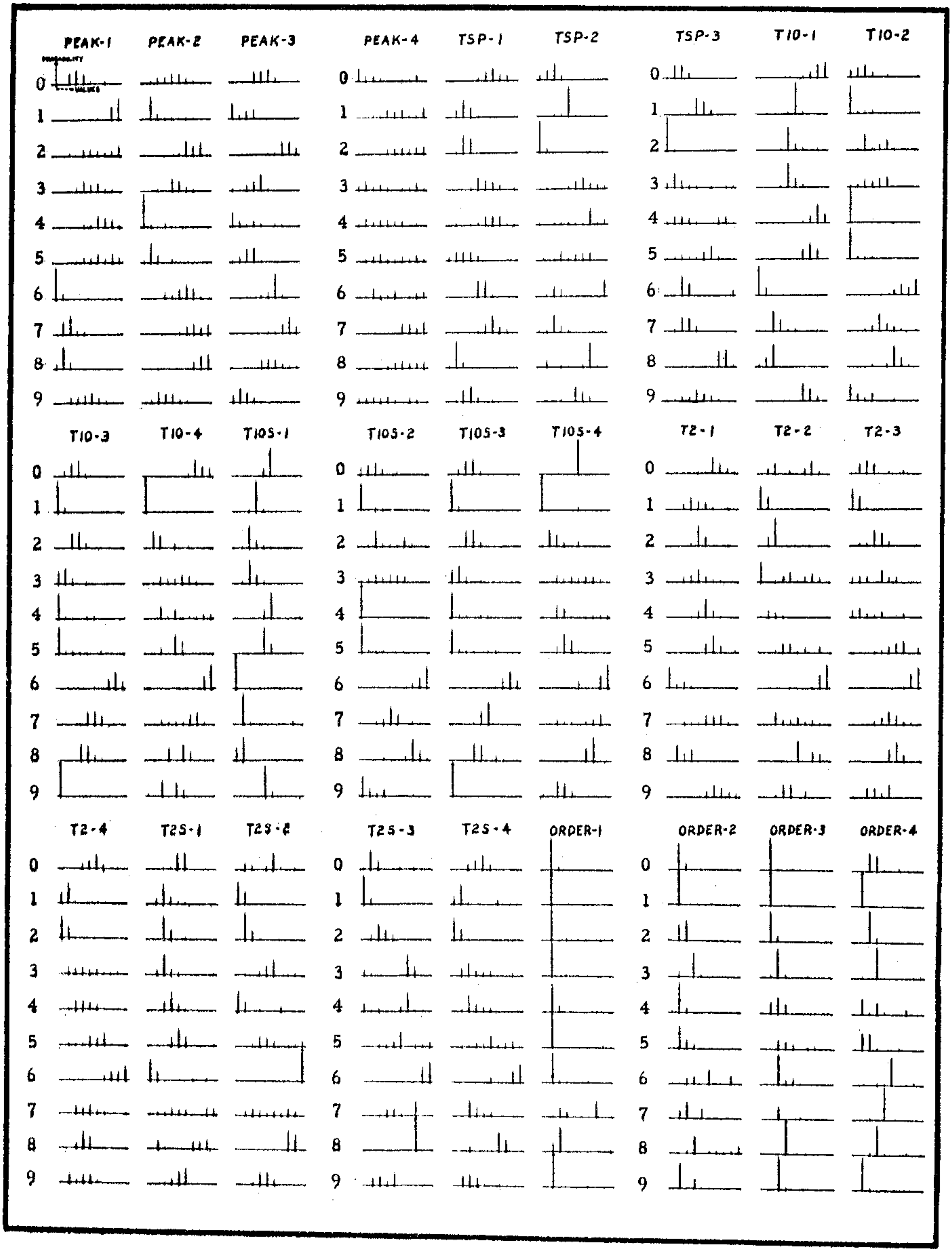


Fig. A3. Statistics of Waveform Measurements on Spoken Digits

recognition of certain words may be worse. We would like to find the best two-measurement set, the best three-measurement set, the best n-measurement set. We could find these best sets by trying all sets, but this is not practical. In the example just given ORDER-Band 4 is a somewhat better single-measurement performer (.480) than is TSP-Band 2 (.472). We find that combining T2S-Band 3 (.551) with TS10-Band 4 gives even better recognition---79.6%; combining T2S-Band 2 (.499) with T10S-Band 4 gives 75.7%. Thus, we find that the better the performance of a measurement by itself the better it usually is when combined with another. This is not surprising. An intuitive guide that did not work is the following: If one measurement is weak in recognizing certain words, W_w, and strong in recognizing the other words, W_s, then choose as a second measurement one that is strongest for the words W_w, when the words W_s are ignored. For example, T10S-Band 4 does a good job on words ZERO, ONE, EIGHT, SIX, and FIVE (i.e., W_s-words) and does poorly on the other words (W_w). If we choose a second measurement that is best for these W_w; namely, TWO, THREE, FOUR, SEVEN, and NINE by themselves, then we would hope this two-measurement set is the best over-all performer. This guide has been found to be a poor one, although it may be helpful for a few stubborn words.

The results of various combinations of six measurements are shown in Table A1.

The performance of the measurements used singly is given in Table A2.

Other calculations, not listed here, gave up to 80% recognition for two measurements, up to 89% for three measurements, up to 95% for four measurements, and nearly 98% for eight readings. It becomes harder to improve the performance as the recognition percentage increases. The more measurements used, the greater the chance of dependency between measurements. Perfection cannot be achieved for the measurements made because the statistics of different words overlap.

C. General Discussion of the Method

Dr. Abramson has shown that we can form our test matrices in a way different from that which we have used. [6] We have used the ratio of the number of times a value occurs to the number of times all values occur for a measurement as the probability of the value.

He indicates that another estimate might be

$$P_i = \frac{1 + n_i}{k + \sum_{i=0}^{9} n_i} \quad , \qquad (2A)$$

T2S-3	T10-1	ORDER-4	TSP-2	PEAK-1	ORDER-2	RECOGNITION (%)
X						55
	X					54
		X				48
			X			47
				X		45
					X	35
X	X					84
X	X	X				94
X	X	X	X			97
X	X	X	X	X		98
X	X	X	X	X	X	97
	X	X	X	X	X	94
X		X	X	X	X	96
X	X		X	X	X	96
X	X	X		X	X	97
X	X	X	X		X	97

Table A1

Recognition Performance by Various Combinations of Six Measurements

T10S-4	.57
T2S-3	.55
T10-1	.54
T2S-2	.50
ORDER-4	.48
TSP-2	.47
PEAK-1	.45
TSP-3	.45
T2S-4	.44
T10S-3	.44
T10S-2	.44
PEAK-2	.40
PEAK-3	.40
T2S-1	.39
ORDER-2	.35
TSP-1	.32
ORDER-3	.32
ORDER-1	.25
PEAK-4	.25

(The better of T10 or T10S in each band is listed. This is true also of T2 or T2S.)

Table A2

Probability of Correct Recognition-Measurements Used Singly
(In Decreasing Performance Order)

where n_i is the number of times the value i has been measured; $\sum_{i=o}^{9} n_i$ is the number of waveforms that have been measured; k is the number of different values equal to 10; and p_i is the probability of value i.

Before any data are taken the probability of each value is 1/k. When the number of word samples is very large, adding 1 to n and k to $\sum n_i$ is insignificant. Thus, in the two extremes the correct probabilities are obtained. For our sample the effect of k (= 10) and the 1 is not insignificant. The theoretical basis for this equation (only for k = 2) is given in Reference 6; a practical reason for using it is given here. If we do not use Eq. A2 we can get value probabilities equal to zero for some words. Furthermore, if these values do occur when words are measured during machine operation, we will eliminate these words from consideration even though all measurements but this one "fit" with high probability.

The probability of correct recognition may be calculated using the following steps:

1. Form the ten "test matrices" for the set of measurements chosen.
2. Choose the probable "reading" using random numbers and these test matrices.
3. Obtain the reading-column of the "experiment matrix" and then form the ratio of the column maximum to the column sum.
4. Repeat steps 2 and 3 to form the average ratio or probability of correct recognition.

An alternative approach would be to replace step 2 with a set of measurements on a spoken word (not used in the statistics of step 1) to get the "reading." This reading could then be applied in step 3. Although there are advantages to both methods, we have not used this approach here.

The method employed had the advantage that it made use of the statistics of all the words. But the second method, in retrospect preferable, has the advantage that dependency between measurements is no longer a question since, if dependence is present, the probability of correct recognition will be less.

In conclusion, we have shown, by example, how to calculate the probability of correct recognition from a set of measurements on spoken words. We have been able to cover much more ground than would have been possible experimentally. However, the computer-time saved by filtering the waveforms before our calculations began should not be overlooked. It has not been feasible to perform an exhaustive test of these data, but we feel that some important sets of measurements have been touched upon. Since these data were from but one speaker, we must emphasize that the results are not likely to be general (i. e., other sets of measurements may be better for another speaker). These data could have been handled in other ways. For example, many other measurements could have been taken from the recorded waveforms, and the value-bracketing could have been done differently.

Our aim has been to outline the decision-theory approach to recognition problems, and to illustrate this approach with a specific example. We do not intend that the preceding pages be considered a complete solution, but rather that they serve as a useful guide.