AT&T BELL LABORATORIES TECHNICAL JOURNAL
Vol. 63, No. 1, January 1984
Printed in U.S.A.

A Probabilistic Model for the Performance of Word Recognizers

By A. E. ROSENBERG*

(Manuscript received February 7, 1983)

This paper develops a probabilistic model to account for the error-rate behavior of isolated-word speech-recognition systems. It examines two kinds of errors, confusion error, an a priori characterization of a recognizer, which measures differences between words, and recognition (rank) error, an a posteriori characterization, which, in addition to taking into account differences between words, accounts for differences between different tokens of the same word. It is shown that these kinds of errors can be modeled by describing recognition trials as Bernoulli trials. Good models of error-rate behavior as a function of vocabulary size can be obtained if the distributions of confusion and recognition (rank) number are considered to be mixtures of binomial distributions. The data obtained from a recent experiment in isolated-word recognition with a large vocabulary (1109 words) are used to evaluate the model. Experimental error-rate functions obtained from each of six talkers and three partitions of the vocabulary are fit by means of an optimization algorithm to model functions based on mixture distributions. The results indicate that two-way mixture distributions account quite well for the experimental performance results.

I. INTRODUCTION

A critical concern in the study and development of automatic speech-recognition systems is specification of their performance. Performance is typically specified by recognition error rate, which is the

* AT&T Bell Laboratories.

fraction of trials in a test of the system in which incorrect decisions are obtained. This specification should be accompanied by a description of the test vocabulary, the talker population, the talking environment, and other pertinent conditions relating to both the training and testing of the system. The interaction among these factors and their effect on performance are not well understood. Indeed, altering a variable associated with any of these factors can change the performance of a system in often unpredictable and drastic ways.

A more general specifier of recognizer performance is the rate at which the best n choices offered by the recognizer contain the correct word. More recently, specifiers measuring "complexity"[1] and efficiency[2] have been introduced. The relationship among specifiers is another aspect of recognizer performance that is not well understood.

It is the purpose of this paper to examine and establish probabilistic models to describe the performance of isolated-word speech-recognition systems and to relate various performance specifiers. The distinction will be made between performance specifiers that characterize systems through the training phases of the system and those that characterize the overall behavior in test use of the system. We will focus on modeling performance behavior as a function of vocabulary size for a given recognizer, over a small population of talkers, and three types of vocabularies. Some speculation will be offered on the relation between model parameters and the recognizer, talker, and vocabulary.

The paper is organized as follows. In Section II, performance measures are defined and the probabilistic models, which form the basis for describing the behavior of these measures as a function of vocabulary size, are introduced. In Section III we make use of data obtained in an experiment with a speaker-dependent isolated-word recognizer using a large vocabulary to illustrate the behavior of some of the performance measures and evaluate how well the probabilistic models account for the behavior. Section IV presents a discussion that offers some speculation regarding the significance of the parameters that specify the models. Section V presents some conclusions.

II. DEFINITIONS AND PROBABILISTIC MODELS

2.1 Bernoulli trials as the basis for confusion and rank

Suppose we have a vocabulary of N words, $V = \{v_1, v_2, \cdots, v_N\}$. Let d_{ij} be a distance measure between a token of word v_i and a token of word v_j. The source for this distance might be some perceptual experiment, a phonetic or linguistic measurement, or the output of an automatic recognizer. In what follows the distance is considered to be the output of a recognizer. As the output of a recognizer it will normally be assumed that the first index, i, refers to an input test word while the second index, j, refers to a (single) prototype word.

The experiment underlying the formulations that follow is the comparison of a token of an input word v_I with the prototype for each of the remaining $N - 1$ words in the vocabulary.

Suppose we are concerned with a particular word, v_I, in the vocabulary. Consider two events

$$d_{Ij} < T, \qquad j \neq I \tag{1}$$

and

$$d_{Ij} < d_{II}, \qquad j \neq I, \tag{2}$$

where T is some preassigned distance threshold and d_{II} is a "self-distance". Note that self-distance, d_{II}, generally represents the distance between two different tokens of a word, v_I, and therefore, must be greater than zero.

Now consider the number of occurrences of these events in the underlying experiment, defined as follows:

$$q_I(T) = |\{j \neq I\colon d_{Ij} < T\}| \tag{3}$$

and

$$r_I = |\{j \neq I\colon d_{Ij} < d_{II}\}|, \tag{4}$$

where $|\{\quad\}|$ is the cardinality or count of the events in the brackets. Note that $q_I(T)$ is the basis for the notion of confusability or complexity introduced in Rabiner et al.[1] whereas $r_I + 1$ is the rank of the correct word input to a recognizer. When $r_I = 0$, the best matching reference prototype corresponds to the correct word v_I.

If, in eq. (3), both I and j represent reference word tokens, $q_I(T)$ can be considered to characterize a recognizer through the training phase of the system, in other words, an a priori characterization. In eq. (4), however, the self-distance, d_{II}, specifically represents the distance between a test word input and the reference prototype for that word. Thus r_I characterizes a system in its test or use phase, and is therefore an a posteriori characterization.

Consider now a probabilistic formulation that can be applied to either of the events defined in (1) and (2). Define a random variable X_{Ij} such that

$$X_{Ij} = \begin{cases} 1 & \text{if the event occurs} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

with

$$\text{Prob}\{X_{Ij} = 1\} = p_I \tag{6}$$

and

$$\text{Prob}\{X_{Ij} = 0\} = 1 - p_I. \tag{7}$$

Each event [(1) or (2)] is thus considered to be a Bernoulli trial. We assume that the Bernoulli probability is independent of the reference word j. When referring to event (1), p_I depends on T, but this is omitted to keep the notation simple. The counts defined in eqs. (3) and (4) are thus sums over j of the random variable X_{Ij}. We denote this sum generically as s_I, it being understood that $s_I = q_I(T)$ when referring to events of type (1), and $s_I = r_I$ when referring to events of type (2). We can see that

$$s_I = \sum_{j \neq I} X_{Ij}, \tag{8}$$

from which it follows that $0 \leq s_I \leq N - 1$. Given eqs. (6), (7), and (8) we obtain

$$\mathscr{E}\{s_I\} = (N - 1)p_I \tag{9}$$

and

$$\text{Var}\{s_I\} = (N - 1)p_I(1 - p_I). \tag{10}$$

Also, the probability that s_I assumes a particular value k, $0 \leq k \leq N - 1$, obeys the binomial probability law

$$\text{Prob}\{s_I = k\} = \binom{N-1}{k} p_I^k(1 - p_I)^{N-1-k}, \tag{11}$$

where $\binom{N-1}{k}$ is a binomial coefficient.

Two kinds of error measures are introduced. An error measure is a monotonically increasing function of s_I that equals 0 when s_I equals 0, and approaches or equals 1 when s_I equals $N - 1$.

The first error measure, e_I, is defined as

$$e_I = 1 - \frac{1}{1 + s_I}, \tag{12}$$

from which it follows that $0 \leq e_I \leq 1 - 1/N$. When s_I is understood to be $q_I(T)$, e_I is similar to confusability or complexity error as defined in Rabiner et al.[1] When $s_I = r_I$, e_I is related to the notion of "efficiency" introduced by Smith and Erman[2] to characterize recognizer performance. Using eqs. (11) and (12) it can be shown that

$$\mathscr{E}\{e_I\} = \sum_{k=0}^{N-1} P\{s_I = k\}\left(1 - \frac{1}{1+k}\right)$$

$$= 1 - \frac{1 - (1 - p_I)^N}{p_I N}. \tag{13}$$

The second error measure is associated with the occurrence of any nonzero value of s_I, that is, any confusion, $q_I(T)$, at all, or any rank, $r_I + 1$, other than 1. Define

$$E_I = \begin{cases} 1 & \text{if} \quad s_I > 0 \\ 0 & \text{if} \quad s_I = 0. \end{cases} \tag{14}$$

This is the conventional or standard recognition error generally used to characterize the performance of automatic recognizers. From eq. (11) we have

$$\text{Prob}\{s_I > 0\} = 1 - \text{Prob}\{s_I = 0\} = 1 - (1 - p_I)^{N-1}. \tag{15}$$

Then it follows that

$$\mathscr{E}\{E_I\} = 1 - (1 - p_I)^{N-1}. \tag{16}$$

Often recognition-error rates are calculated to provide the frequency with which the correct word is included among the top c choices provided by the recognizer $s_I = 0, 1, 2, \cdots, c - 1$. This represents a generalization of the preceding definition for E_I which is expressed by

$$E_I(c) = \begin{cases} 1 & \text{if} \quad s_I \geq c \\ 0 & \text{if} \quad s_I < c. \end{cases} \tag{17}$$

Since

$$\text{Prob}\{s_I \geq c\} = \sum_{k=c}^{N-1} p_I^k (1 - p_I)^{N-1-k} \tag{18}$$

we have

$$\mathscr{E}\{E_I(c)\} = \sum_{k=c}^{N-1} p_I^k (1 - p_I)^{N-1-k}. \tag{19}$$

Although the models which are derived in this paper could easily include this generalization, we restrict our attention to the case for which $c = 1$, referred to as standard error.

2.2 Mixture models

The foregoing formulas pertain to a single word v_I in a vocabulary V. Our object is to model behavior of confusability or rank over an entire vocabulary V. It is therefore necessary to make some assumptions about the behavior of p_I over all the words in V. The simplest possible assumption is that p_I is constant over V, i.e., $p_I = p$ for $I = 1, 2, \cdots, N$. It will be shown in the following sections that this assumption leads to very poor models of the actual experimental behavior.

A more general assumption is the following. Assume that the Bernoulli probability defined in eq. (6) is itself a random variable, p_V, which may assume different values from word to word in a vocabulary, or indeed, from trial to trial of the same word. Suppose there are M values p_V can assume, p_m, $m = 1, 2, \cdots, M$,* such that

$$\text{Prob}\{p_V = p_m\} = h_m, \qquad m = 1, 2, \cdots, M \tag{20}$$

with

$$\sum_{m=1}^{M} h_m = 1, \tag{21}$$

where h_m is the probability that p_V assumes the value p_m. (It is possible to generalize still further by assuming p_V to be continuously distributed.) It is now possible to generalize s_I to s_V over the entire vocabulary V. From eqs. (11) and (16) we obtain

$$\text{Prob}\{s_V = k\} = \sum_{m=1}^{M} h_m \binom{N-1}{k} p_m^k (1 - p_m)^{N-1-k}. \tag{22}$$

This expression represents a so-called compound binomial distribution or mixture of binomial distributions.[3,4] With this interpretation, each time we perform the underlying experiment represented by (1) or (2) the probability assumed one of the M values p_m over all the $N - 1$ comparisons with the words in V. (This is in contrast to the situation in which the probability may assume different values for each comparison in the underlying experiment.) Using eq. (22), general expressions can be obtained for the mean and variance of s_V and for the two generalized error formulations e_V and E_V. All of these have the same form as eq. (22), that is, $\mathscr{E}\{\cdot\} = \sum_{m=1}^{M} h_m \mathscr{E}\{\cdot \mid m\}$. Thus,

$$\mathscr{E}\{s_V\} = (N - 1)\bar{p}_V \tag{23}$$

and

$$\begin{aligned}\text{Var}\{s_V\} = (N - 1) \sum_{m=1}^{M} h_m p_m (1 - p_m)& \\ + (N - 1)^2 \sum_{m=1}^{M} h_m (p_m - \bar{p}_V)^2,&\end{aligned} \tag{24}$$

* Note that the index m on p no longer refers in general to individual words in the vocabulary as in eq. (6).

where

$$\bar{p}_V = \sum_{m=1}^{M} h_m p_m. \tag{25}$$

Also,

$$\begin{aligned}\mathscr{E}\{e_V\} &= \sum_{m=1}^{M} h_m \left(1 - \frac{1 - (1 - p_m)^N}{p_m N}\right)\\ &= 1 - \sum_{m=1}^{M} h_m \left(\frac{1 - (1 - p_m)^N}{p_m N}\right),\end{aligned} \tag{26}$$

and

$$\mathscr{E}\{E_V\} = 1 - \sum_{m=1}^{M} h_m (1 - p_m)^{N-1}. \tag{27}$$

With M set to 1, these expressions revert to the form of the earlier expressions for a single word I.

III. EXPERIMENTAL EVALUATION

The experimental data used in this study were obtained using the AT&T Bell Laboratories Linear Predictive Coefficient (LPC) based isolated-word recognition system.[5,6] The vocabulary was the 1109-word so-called Basic English vocabulary of Ogden.[7] The recognizer was used in a speaker-dependent mode. Six native American talkers, three male, three female, participated in the experiment. Each talker trained the system using the robust training procedure of Rabiner and Wilpon,[8] giving a single reference prototype for each word in the vocabulary. In addition, four sets of test utterances were obtained from each talker over a four-week period. Both the training and test utterances were collected over dialed-up telephone lines using an ordinary telephone handset with the talker seated in a sound booth. For each talker, each test word was input to the recognizer and compared with every reference prototype word for that talker. For each such comparison the recognizer provides a distance figure measuring how closely the test word matches a prototype word. In a typical recognition trial the word recognized is associated with the best matching prototype word, that is, the one with the smallest distance. The raw experimental data consist of four sets of 1109 × 1109 distance matrices for each of the six talkers.

The large size of the vocabulary provides an opportunity to investigate recognition performance over a variety of experimental conditions related to vocabulary size and content by choosing appropriate subsets of the whole vocabulary. A series of such experiments using this experimental database has been described in a previous report.[1]

In the present experiment we focus on three partitions of the 1109-word vocabulary, the 605 monosyllabic words contained in the vocabulary, the remaining 504 polysyllabic words, and the entire vocabularly itself. For each of these, randomly selected subsets of various sizes are chosen. The subset sizes chosen are

$$N = 10,\ 20,\ 50,\ 100,\ 200,\ 400,\ (800),\ PARTSIZ,$$

where $PARTSIZ = 605, 504$, or 1109 for the monosyllabic, polysyllabic, and whole vocabulary partitions, respectively. For each subset size N, a total of $MT = \min[50, PARTSIZ/N]$ subsets of words selected at random without replacement from each partition are specified. The same subsets are specified over all test sets and talkers. Thus, in the aggregate, for each subset size N, results are obtained over N^*MT different words, where $500 \leq N^*MT \leq PARTSIZ$.

The experimental performance data that are presented in this paper are generally given as functions of subset size for each talker and vocabulary type, and represent an average over all the talker's four test sets and vocabulary subsets for each subset size.

3.1 Experimental performance measures

This section presents experimental examples of the performance measures introduced in Section II. To recapitulate, confusion number and rank number are defined as follows:

1. $q_I(T)$: confusion number for a given word $v_I \epsilon V$, is the number of words (other than v_I) in a given vocabulary subset V whose distance to the given word is less than some threshold T [from eq. (3)].

2. r_I: rank number for a given word $v_I \epsilon V$, is the number of words (other than v_I) in a given vocabulary subset V whose distance is less than the self-distance for v_I [from eq. (4)].

Experimental averages are obtained as follows. Suppose word v_I is included in $V_m(N)$, where $V_m(N)$ is the mth vocabulary subset of size N, $m = 1, 2, \cdots, MT$, and MT is the total number of subsets of size N. The words in each subset $V_m(N)$ are selected at random without replacement from a vocabulary V of total size $Q \geq N$. Then, given the confusion number and rank number, $q_{I,V_m(N),s,t}(T)$ and $r_{I,V_m(N),s,t}$, respectively, for word v_I from subset $V_m(N)$, test set t, and talker s, the experimental averages are

$$\bar{q}_{s,V}(N,\ T) = \frac{1}{4}\sum_{t=1}^{4}\frac{1}{MT}\sum_{m=1}^{MT}\frac{1}{N}\sum_{I\epsilon V_m(N)} q_{I,V_m(N),s,t}(T) \tag{28}$$

and

$$\bar{r}_{s,V}(N) = \frac{1}{4}\sum_{t=1}^{4}\frac{1}{MT}\sum_{m=1}^{MT}\frac{1}{N}\sum_{I\epsilon V_m(N)} r_{I,V_m(N),s,t}\,, \tag{29}$$

respectively. Similarly, for the two error measures that were introduced in Section II, the experimental averages are

$$\bar{e}_{q,s,V}(N,\ T) = 1 - \frac{1}{4}\sum_{t=1}^{4}\frac{1}{MT}\sum_{m=1}^{MT}\frac{1}{N}\sum_{I\epsilon V_m(N)}\frac{1}{1+q_{I,V_m(N),s,t}(T)}, \tag{30}$$

and

$$\bar{e}_{r,s,V}(N) = 1 - \frac{1}{4}\sum_{t=1}^{4}\frac{1}{MT}\sum_{m=1}^{MT}\frac{1}{N}\sum_{I\epsilon V_m(N)}\frac{1}{1+r_{I,V_m(N),s,t}} \tag{31}$$

for the efficiency errors of confusion and rank, respectively, and

$$\bar{E}_{q,s,V}(N,\ T) = \frac{1}{4}\sum_{t=1}^{4}\frac{1}{MT}\sum_{m=1}^{MT}\frac{1}{N}\ |\{I\colon I\ \epsilon\ V_m(N),\ q_{I,V_m(N),s,t}(T) \geq 0\}| \tag{32}$$

and

$$\bar{E}_{r,s,V}(N) = \frac{1}{4}\sum_{t=1}^{4}\frac{1}{MT}\sum_{m=1}^{MT}\frac{1}{N}\ |\{I\colon I\ \epsilon\ V_m(N),\ r_{I,V_m(N),s,t} \geq 0\}| \tag{33}$$

for the standard errors of confusion and rank, respectively.

Note that both experimental confusion and rank or recognition data are obtained by averaging over four test utterances. In the previous section we noted that confusion could be considered an a priori, training characterization of system performance if it is based on distances between prototypes. Since this is not the case here, the description of confusion does not strictly hold. However, we do not expect distances between test utterances of different words to be significantly different from distances between prototypes of the same words. This point is discussed again in Section IV.

Shown plotted in Fig. 1a as a function of N are experimental averages for confusion number and rank number for talker 3 and the whole vocabulary V_W, $\bar{q}_{3,V_W}(N,\ T)$ and $\bar{r}_{3,V_W}(N)$. Average confusion number is plotted for five threshold values, $T = 0.20$, 0.25, 0.30, 0.35, and 0.40. For each of these plots, straight-line fits are obtained by least-squares regression. It can be seen that straight-line fits are quite good, each one having a correlation coefficient of better than 0.9995 with the data, with the exception of $\bar{q}_{3,V_W}(N,\ 0.20)$ and $\bar{r}_{3,V_W}(N)$, whose coefficients of fit are both 0.998. The linear trend is predicted by the model as expressed in eq. (23). The interpretation of the linearity is quite natural. Simply, as the size of the vocabulary grows, the number of confusable words, or the number of words better than the input word (the rank number of the input word), grows proportionately.

For the same talker and vocabulary, and for the same set of thresholds, experimental averages for efficiency error, $\bar{e}_{q,3,V_W}(N,\ T)$ and

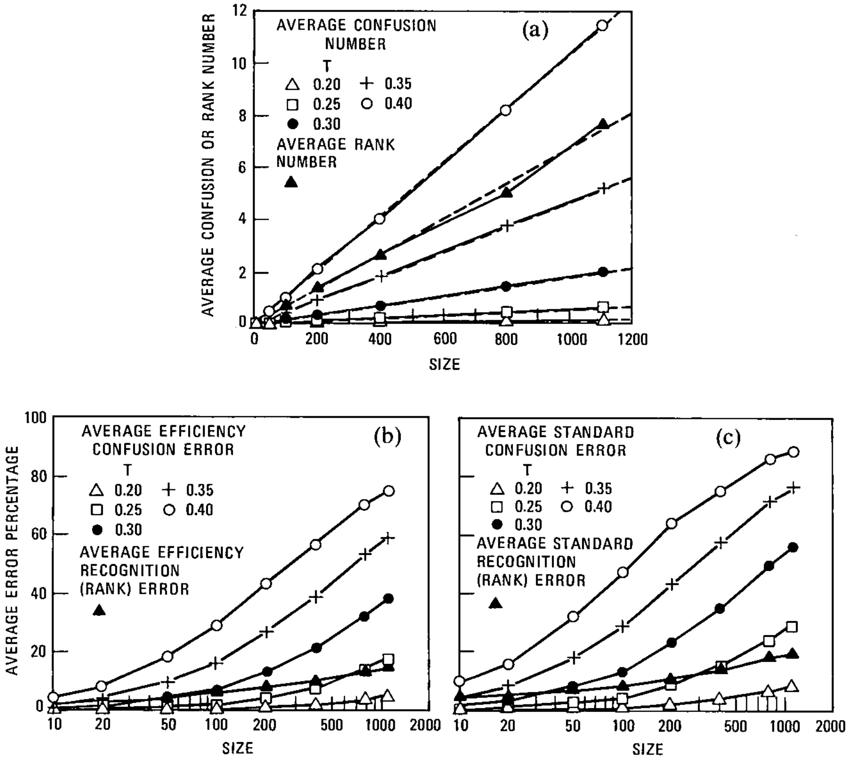

Fig. 1—(a) Average confusion and rank number, (b) average efficiency confusion and rank error, and (c) average standard confusion and rank error as functions of vocabulary size, for talker 3, vocabulary type V_W, and five values of threshold, T (for confusion).

$\bar{e}_{r,3,V_W}(N)$, are plotted in Fig. 1b and standard error $\bar{E}_{q,3,V_W}(N, T)$ and $\bar{E}_{r,3,V_W}(N)$ are plotted in Fig. 1c, both as a function of N scaled logarithmically. Both the efficiency- and standard-error curves assume the same trends as a function of vocabulary size, increasing monotonically and approaching one asymptotically. The efficiency-error curves assume uniformly smaller values than their standard error counterparts for each value of N. For small N each efficiency-error curve has approximately half the value of its standard-error counterpart. For confusion error, error increases monotonically as the distance threshold, T, is increased or relaxed.

The standard recognition- or rank-error curve plotted in Figure 1c is representative of results presented in the earlier report.[1] In the present study, additional data values are presented that extend the curve to vocabulary sizes less than 100. Standard error rate for this talker is approximately 4 percent for 10-word vocabularies, 9 percent for size 100, and 20 percent for the full vocabulary of 1109 words. (The same curve is shown with an expanded error scale in Fig. 2.) In the earlier report it was suggested that for vocabulary sizes greater than 100, doubling the size increases error by a constant amount, a linearly increasing trend with N scaled logarithmically. It can be seen here that with the extension to smaller vocabulary sizes the linear trend is restricted and approximate.

3.2 *The relation between recognition and confusion error*

The difference in form between the rank- or recognition-error curves and the confusion-error curves, for any threshold value, is quite marked. The relation between recognition error and confusion error as a function of vocabulary size is rather complex.

The following are two hypotheses for relating recognition and confusion error. First, we might examine average confusion number as a function of distance threshold T for any given vocabulary size and find that value of T for which average confusion number is equal to average rank number. For the results shown in Fig. 1a for talker 3 and vocabulary V_W, this value of T lies between 0.35 and 0.40. We could reason that confusion-error rates ought to be the same as recognition-error rates for a value of T, which on the average includes as many confusable words as the rank of the correct word. However, examining the error rates in Figs. 1b and 1c we find that this hypothesis holds only for the very smallest vocabulary size, $N = 10$. The threshold suggested by this hypothesis leads to confusion-error rates much greater than the recognition-error rates actually observed for larger values of N.

The second hypothesis suggests that the appropriate threshold for which confusion-error rates ought to be the same as recognition-error rates can be found by associating the threshold with the average self-distance. We have carried out the calculation of average self-distance for this talker and vocabulary in the same way as the other calculated experimental averages,

$$\overline{DSLF}_{3,V_W}(N) = \frac{1}{4}\sum_{t=1}^{4}\frac{1}{MT}\sum_{m=1}^{MT}\frac{1}{N}\sum_{I\epsilon V_m(N)} d_{II}. \qquad (34)$$

We obtain $\overline{DSLF}_{3,V_W}(N) \approx 0.235$ for all values of N. Examining the confusion-error rates in Figs. 1b and 1c, only for the very largest value of N, where recognition error is bracketed by confusion error rates for $T = 0.20$ and $T = 0.25$, do we find agreement with this hypothesis.

It is not surprising that average self-distance is independent of N, since the similarity between two tokens of the same word is not dependent on the size of the vocabulary from which the words are taken. Nor is it surprising, as we have seen earlier, that the rate of increase in average rank number is independent of N. But neither of the associated thresholds is adequate to relate confusion and recognition for all values of N. The explanation lies in the following observations. Referring back to the basic definitions expressed in (1) and (2), if the individual self-distance, d_{II}, were absolutely constant from trial to trial and from word to word, there would be a threshold equal to this constant for which confusion-error rates would be equal to recognition-error rates. However, even though average self-distance is constant for all N, individual self-distances fluctuate widely from trial to trial resulting in rank number distributions which are also quite wide. Thus for small subset sizes, these fluctuations produce an average rank number that is significantly greater than the average confusion number associated with a threshold equal to the average self-distance. However, when vocabulary size grows, so does word density, that is, the average distance between different words decreases. As this occurs the self-distance fluctuations become less important compared with errors attributed to the increasing density. Thus for large vocabularies a threshold equal to average self-distance relates confusion error to recognition error.

The interpretation of the relation between confusion and recognition in the light of model parameters will be brought up in the discussion, Section IV.

3.3 *Estimation of model parameters and fits to experimental results*

It has already been shown that average confusion number and average rank number grow linearly with subset size in agreement with the model as expressed in eq. (23). The slope of this linear function is an estimate of $\bar{p}_V$ given in eq. (25), the average model Bernoulli probability of an error or confusion. Estimates of $\bar{p}_V$ are obtained as the slope estimates for the regression-line fits shown in Fig. 1a. Table I shows these estimates along with the coefficients of fit (correlation coefficients). Since these are linear relations, they present no information concerning individual mixture probabilities, nor, indeed, whether there are any mixtures at all.

The effect of mixtures becomes apparent when we attempt to model

Table I—Linear growth of average confusion and average rank numbers

T Values	$\bar{p}_V$ Estimates	r Coefficients
	Average Confusion Number	
0.20	1.18×10^{-4}	0.99784
0.25	5.43×10^{-4}	0.99980
0.30	1.77×10^{-3}	0.99995
0.35	4.63×10^{-3}	0.99993
0.40	1.03×10^{-2}	0.99994
	Average Rank Number	
—	6.67×10^{-3}	0.99801

error-rate behavior as a function of vocabulary size. To illustrate the effect, the standard-rank error-rate curve shown in Fig. 1b is displayed once more on an expanded error-rate scale in Fig. 2a. (We refer to rank and recognition error rate interchangeably.) Along with this curve, we have plotted the function for expected standard error rate given in eq. (27) with M set to 1 for four different values of p. These

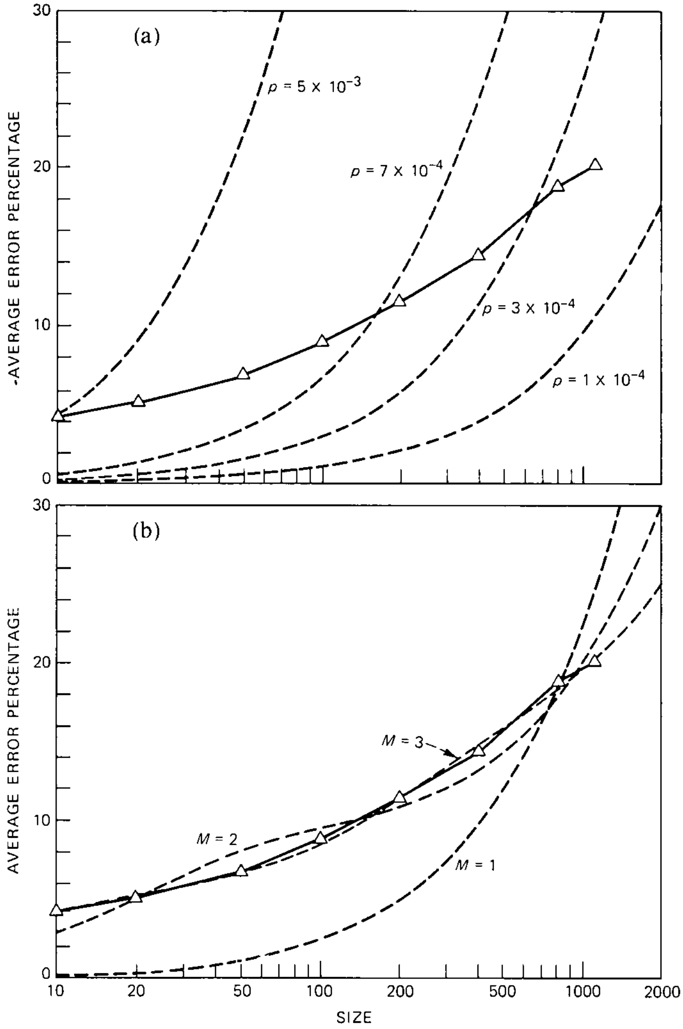

Fig. 2—Average standard recognition (rank) error as a function of vocabulary size for talker 3, and vocabulary type V_W with (a) four different simple (one-way) models of standard error [eq. (27) with $M = 1$], and (b) one-, two-, and three-way mixture-model fits of standard error.

four values of p bracket the range of the values of $\bar{p}_V$ given in Table I. It is quite evident that the simple Bernoulli trial model, obtained with M set to 1 in eq. (27), cannot provide a good fit to the experimental results.

The recognition error-rate curve is plotted once again in Fig. 2b along with the function of eq. (27) for three values of M, M = 1, 2, and 3. Table II shows the h and p parameters selected for these functions along with coefficients of fit. The parameters are obtained by using an optimization routine[9] to provide a best fit to the experimental data. We employ the convention that $p_1 \geq p_2 \geq, \cdots, \geq p_M$ and $h_M = 1 - \sum_{m=1}^{M-1} h_m$ to satisfy eq. (21). Note that the fits obtained improve progressively with increasing M and the values of $\bar{p}_V$ obtained for $M = 2$ and $M = 3$ bracket the observed value for the slope estimate for $\bar{r}_{3,V_W}(N)$ given in Table I.

Although it is clear that the three-way mixture model, $M = 3$, provides a superior fit, hereafter we will provide only two-way mixture model fits, $M = 2$. It seems reasonable to expect that models with larger M's should provide better fits because it is reasonable to expect such models to accommodate and discriminate better among all the effects that contribute to the error-rate functions. However, it is also true that there are only eight data points, which is a small number of points to support the number of parameters associated with such models. In addition, there is a certain appeal of parsimony in using two-way models, since it may lead to simpler or more direct interpretations of the parameters. Some suggestions for interpretations are discussed in Section IV. Also, although the two-way model fit is somewhat deficient for the case shown here, in most of the other experimental results that are presented the fits are quite adequate.

For the two-way model, we refer to p_1 and p_2, $p_1 \geq p_2$, as the type 1 and type 2 population probabilities, respectively, and h (dropping the subscript), as the mixing coefficient for the two populations.

The optimization-fitting procedure is described briefly. The function that is minimized is the sum over the subset sizes of the squared differences between the observed values and the calculated model function value. This function, the gradient of the function, initial values for the parameters, and some convergence constants are supplied to the optimization routine. Usually the routine is run several

Table II—Model parameter estimates for average standard recognition error

M	$h_1, h_2, \ldots h_{M-1}$	$p_1, p_2, \ldots p_M$	$\bar{p}_V$	r
1	—	2.61×10^{-4}	2.61×10^{-4}	0.9607
2	0.085	4.32×10^{-2}, 1.37×10^{-4}	3.82×10^{-3}	0.9856
3	0.046, 0.093	1.71×10^{-1}, 4.72×10^{-3}, 7.26×10^{-5}	8.35×10^{-3}	0.9990

times for each set of experimental points with different sets of initial values to ensure that the optimized parameters represent a global rather than a local minimization. In some cases, particularly for overall low error rates, the minimization is relatively insensitive to variation of the h parameter.

3.4 Two-way mixture model fits to experimental data

This section presents a variety of experimental confusion- and recognition-error results as a function of vocabulary size, together with two-way mixture model fits. The object is to demonstrate that the model represents the error-rate behavior as a function of vocabulary size quite well, and to point out the effects of talker, vocabulary type, etc., on the parameter estimates obtained from the model.

3.4.1 Model fits to recognition error as a function of talker and vocabulary type

Recognition error rate results as a function of vocabulary size, both efficiency and standard error, are displayed in Figs. 3 and 4, together with model fits for each example. Figure 3 shows results for the three vocabulary types, the whole vocabulary, V_W, monosyllables, V_M, and polysyllables, V_P, for three talkers. Figures 3a through 3c show efficiency-error results for the three talkers while Figs. 3d through 3f show standard-error results. Figure 4 presents results for all six talkers for a single vocabulary type, V_W. Fig. 4a presents efficiency-error results and Fig. 4b presents standard-error results. The three talkers selected for Fig. 3 are associated with median performances in Fig. 4. The performance trends of these three talkers for the three vocabulary types in Fig. 3 are representative of all six talkers.

Recall once again the distinction between standard error and efficiency error. Standard error is based on a count of trials with nonzero rank, while efficiency error accounts for the distribution of all rank numbers over all the trials, and is therefore, in some sense, a finer characterization of error performance. The differences between the two are generally predictable, as pointed out in the previous section. Both kinds of error results are shown, principally, to compare the parameter estimates obtained for each model.

The trend in error-rate performance as a function of vocabulary type for individual talkers presented in Fig. 3 is a familiar one. That is, performance degrades for any vocabulary size from the more redundant to the less redundant vocabulary types, from V_P to V_W to V_M.

The performance of individual talkers for a single vocabulary type presented in Fig. 4 shows considerable variability. The performance of one talker, talker 4, is distinctly poorer than the others. The best

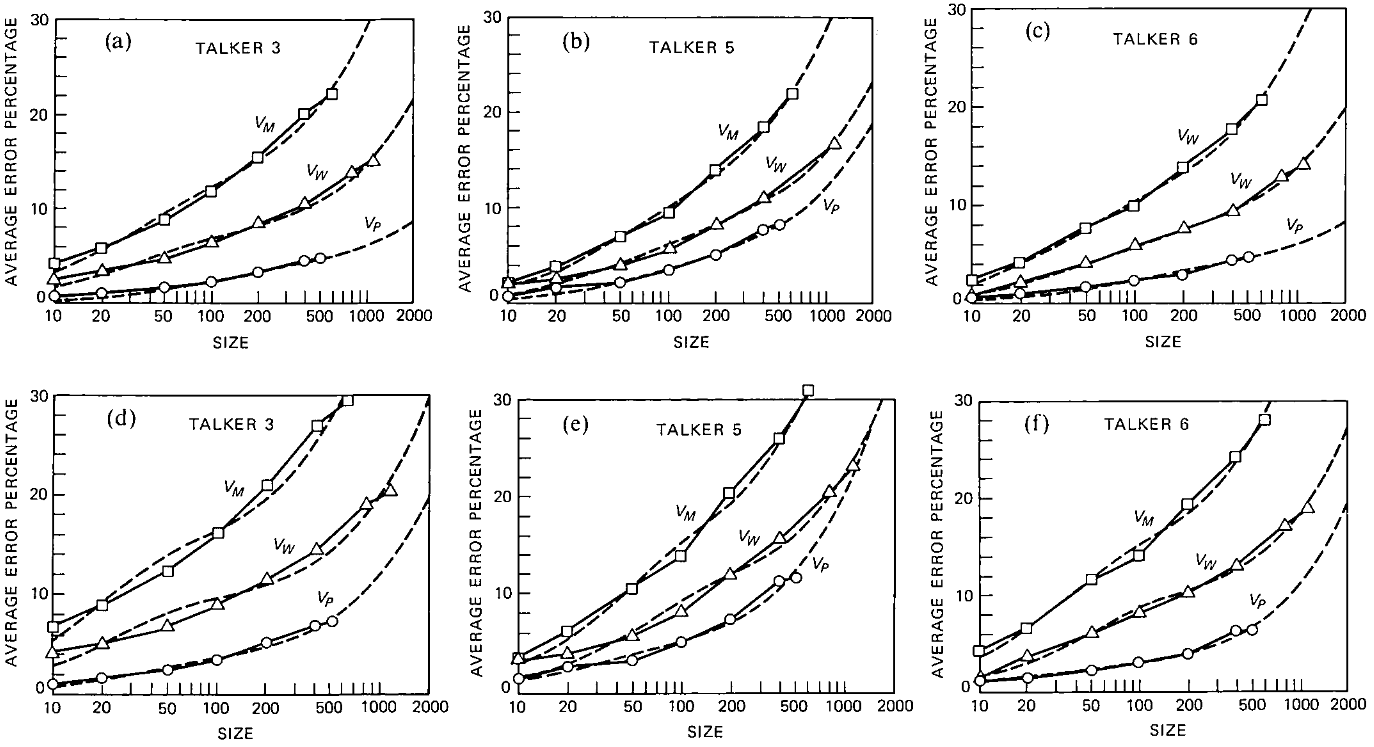

Fig. 3—(a), (b), (c) Average efficiency recognition (rank) error, and (d), (e), (f) average standard recognition (rank) error as a function of vocabulary size, with two-way mixture model fits, for three vocabulary types.

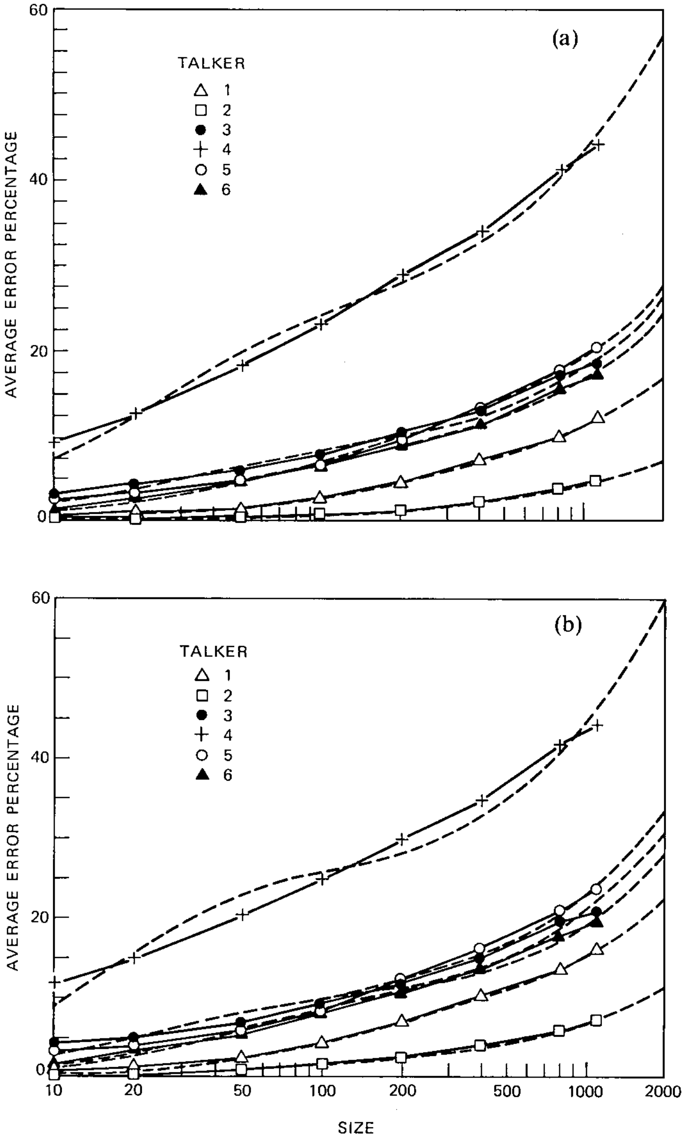

Fig. 4—(a) Average efficiency recognition (rank) error, and (b) average standard recognition (rank) error as a function of vocabulary size, with two-way mixture model fits, for six talkers and vocabulary type V_W.

performances are obtained for talkers 1 and 2, while the remaining three are grouped together in an intermediate range of performance.

Model fits have been carried out, as described previously, for both the efficiency and standard-error results for each of the six talkers

Table III—Model parameter estimates for average efficiency recognition (rank) error

	V_P			V_W			V_M		
Talker	h	p_1	p_2	h	p_1	p_2	h	p_1	p_2
1	0.037	7.85×10^{-3}	4.92×10^{-5}	0.055	9.49×10^{-3}	9.80×10^{-5}	0.105	8.29×10^{-3}	1.96×10^{-4}
2	0.020	9.99×10^{-4}	2.53×10^{-5}	0.020	5.24×10^{-3}	4.17×10^{-5}	0.040	9.04×10^{-3}	9.31×10^{-5}
3	0.040	1.99×10^{-2}	5.24×10^{-5}	0.076	5.47×10^{-2}	1.68×10^{-4}	0.126	6.22×10^{-2}	4.36×10^{-4}
4	0.119	1.09×10^{-1}	5.07×10^{-4}	0.208	7.36×10^{-2}	4.21×10^{-4}	0.277	8.16×10^{-2}	9.40×10^{-4}
5	0.046	2.23×10^{-2}	1.74×10^{-4}	0.076	3.04×10^{-2}	1.99×10^{-4}	0.109	3.71×10^{-2}	4.79×10^{-4}
6	0.038	1.69×10^{-2}	5.07×10^{-5}	0.072	3.28×10^{-2}	1.52×10^{-4}	0.109	4.66×10^{-2}	4.16×10^{-4}
Mean	0.050	2.95×10^{-2}	1.43×10^{-4}	0.085	3.44×10^{-2}	1.75×10^{-4}	0.128	4.08×10^{-2}	4.27×10^{-4}
Standard deviation	0.035	3.98×10^{-2}	1.86×10^{-4}	0.064	2.62×10^{-2}	1.30×10^{-4}	0.079	2.91×10^{-2}	2.94×10^{-4}
Without Talker 4									
Mean	0.036	1.36×10^{-2}	7.03×10^{-5}	0.060	2.65×10^{-2}	1.32×10^{-4}	0.098	3.26×10^{-2}	3.24×10^{-4}
Standard deviation	0.010	8.92×10^{-3}	5.90×10^{-5}	0.024	1.99×10^{-2}	6.22×10^{-5}	0.033	2.37×10^{-2}	1.69×10^{-4}

	Mean			Standard Deviation		
Talker	h	p_1	p_2	h	p_1	p_2
1	0.066	8.54×10^{-3}	1.14×10^{-4}	0.035	8.49×10^{-4}	7.48×10^{-5}
2	0.027	5.09×10^{-3}	5.34×10^{-5}	0.012	4.02×10^{-3}	3.54×10^{-5}
3	0.081	4.56×10^{-2}	2.19×10^{-4}	0.043	2.26×10^{-2}	1.97×10^{-4}
4	0.201	8.81×10^{-2}	6.23×10^{-4}	0.079	1.86×10^{-2}	2.78×10^{-4}
5	0.077	2.99×10^{-2}	2.84×10^{-4}	0.032	7.41×10^{-3}	1.69×10^{-4}
6	0.073	3.21×10^{-2}	2.06×10^{-4}	0.036	1.49×10^{-2}	1.89×10^{-4}

and each of the three vocabulary types. This includes all the results shown in Figs. 3 and 4 plus the V_M and V_P results for talkers 1, 2, and 4. These are two-way mixture fits obtained by setting M to 2 in eq. (26) for efficiency error and eq. (27) for standard error. Model parameter estimates for efficiency error are presented in Table III. The parameter estimates for standard error are not shown, but comparing them to the efficiency-error estimates indicates general, if not necessarily close, agreement. This agreement reinforces our assumption that the models developed to account for both kinds of error functions are substantially correct since the same experimental data underlay both performance measures.

Table IV compares model fits for average rank, efficiency recognition error, and standard recognition error for each of the six talkers for V_W. The table presents estimates for $\bar{p}_V$ and coefficients of fit, r. As in Table I, the estimates of $\bar{p}_V$ for average rank are obtained from slope estimates for least-squares regression lines. For efficiency and standard error, the $\bar{p}_V$ estimates are reconstructed using eq. (25). The $\bar{p}_V$ estimates obtained from efficiency and standard-error parameters are in fairly good agreement with each other, but are generally less than half the values of the estimates obtained for average rank. This discrepancy was pointed out in the previous section, where it was implied that it is related to the extent that two-way mixtures model the data compared with models with higher specified values of M. An examination of the model function fits in Figs. 3 and 4 and the coefficients of fit in Table IV indicates generally close agreement with the experimental results. The possible exceptions are associated with high error-rate performances (for example, for talkers 3 and 4). For these cases the fits are poorer for the standard-error functions than for the corresponding efficiency-error functions.

As an aid to improve interpretations for the model parameters, it would be useful in an examination of the parameter estimates to detect significant trends associated with the variation of experimental con-

Table IV—Comparison of model fits for average rank, efficiency recognition error, and standard recognition error

	Average Rank		Efficiency Recognition Error		Standard Recognition Error	
Talker	$\bar{p}_V$	r	$\bar{p}_V$	r	$\bar{p}_V$	r
1	1.13×10^{-3}	0.99705	6.11×10^{-4}	0.99905	5.54×10^{-4}	0.99942
2	3.49×10^{-4}	0.97985	1.46×10^{-4}	0.99877	2.49×10^{-4}	0.99891
3	6.67×10^{-3}	0.99801	4.31×10^{-3}	0.99454	3.28×10^{-3}	0.98563
4	2.69×10^{-2}	0.99994	1.57×10^{-2}	0.99522	1.26×10^{-2}	0.98681
5	6.05×10^{-3}	0.99905	2.49×10^{-3}	0.99675	1.79×10^{-3}	0.99350
6	4.86×10^{-3}	0.99997	2.49×10^{-3}	0.99881	1.97×10^{-3}	0.99742

ditions. In particular, it would be useful to identify those parameters that remain more or less constant over a particular set of conditions. General trends are apparent. As performance degrades, either from one talker to another or from one vocabulary type to another, the estimates of each of the parameters, h, p_1, and p_2, generally increase. Differential trends are harder to detect. No definite conclusions are provided in this set of estimates, but some of it is suggestive.

Table III provides means and standard deviations for each parameter estimate across vocabulary types for each talker and across talkers for each vocabulary type. Since there are only three vocabulary types, caution should be exercised with respect to statistics over this variable. If we use the ratio of the standard deviation to the mean for each parameter as an indicator of variability, we find that p_1, across vocabulary types, has consistently less variability than h or p_2, with ratios generally less than 0.5 for both efficiency and standard error. Across talkers, suggestions are somewhat vaguer, chiefly because of the especially large variability provided by talker 4. If we disregard the estimates for talker 4, which may be justified by the fact that the two-way mixture fits are rather poor for this talker, then low variability is indicated for the h parameter, and to a lesser extent, for p_2, as shown by the second set of means and standard deviations in the table.

Another general observation that can be made is that the ratio of p_1 to p_2 is of the order of 100 and generally decreases across vocabulary types from V_P to V_W to V_M.

3.4.2 Model fits to confusion error as a function of threshold, talker, and vocabulary type

We turn now to two-way mixture models of confusion error and estimates of the model parameters. Experimental confusion error results are shown plotted as a function of vocabulary size in Fig. 5 for talker 6, vocabulary type V_W, and seven threshold values. Efficiency error is plotted in Fig. 5a and standard error in Fig. 5b. Accompanying each curve is a two-way mixture model fit based on eqs. (26) and (27). Parameter estimates for efficiency error are presented in Table V. As threshold increases so does confusion error, as well as all the model parameters, h, p_1, and p_2. As with recognition error, parameter estimates for standard error are omitted. However, there is reasonable agreement between the parameter estimates obtained from efficiency error and standard error with the exception of the lowest threshold value, where the data are too sparse for reliable estimation. Above the lowest threshold the ratio of p_1 to p_2 remains fairly constant at approximately nine.

Estimates of $\bar{p}_V$ derived from average confusion number data and

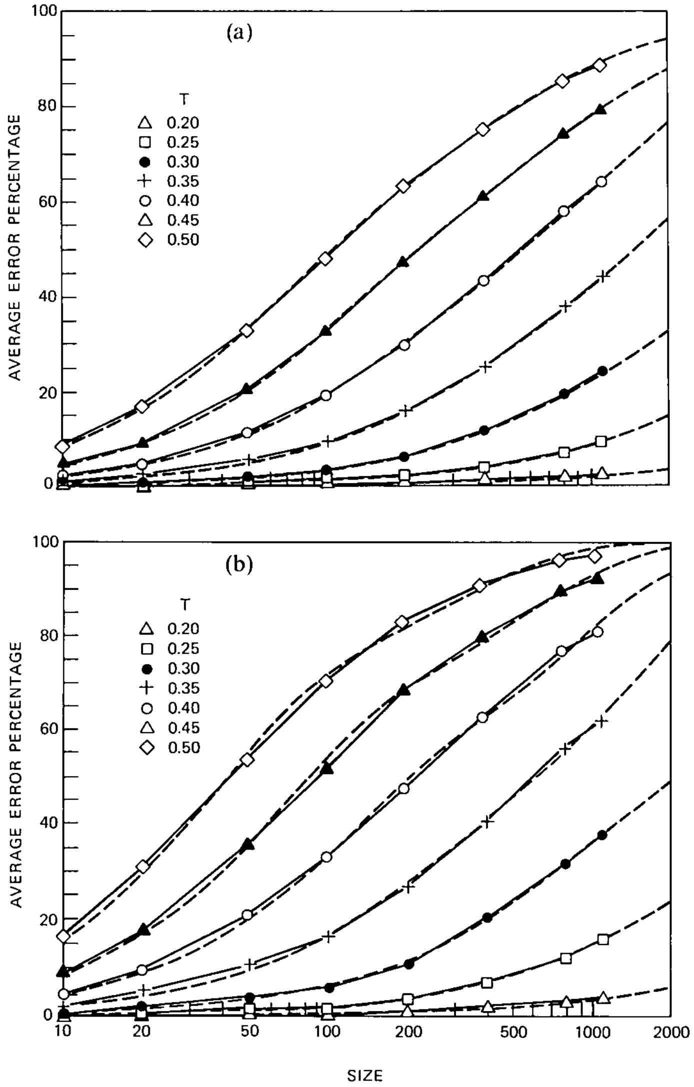

Fig. 5—(a) Average efficiency confusion error, and (b) average standard confusion error as a function of vocabulary size, with two-way mixture model fits for talker 6, vocabulary type V_W, and seven values of threshold, *T*.

from estimates of h, p_1, and p_2 for efficiency and standard confusion error are presented in Table VI together with coefficients of fit. Note that compared to recognition error results shown in Table IV, the coefficients of fit indicate better fits for the confusion models, sustaining a subjective impression gained by examining the figures. In addition there is much better agreement in estimates of $\bar{p}_V$ between

Table V—Model parameter estimates for average efficiency confusion error

T	h	p_1	p_2
0.20	0.050	4.62×10^{-4}	1.46×10^{-5}
0.25	0.250	5.70×10^{-4}	6.04×10^{-5}
0.30	0.350	1.73×10^{-3}	1.14×10^{-4}
0.35	0.310	5.27×10^{-3}	5.67×10^{-4}
0.40	0.379	1.15×10^{-2}	1.26×10^{-3}
0.45	0.527	1.84×10^{-2}	2.02×10^{-3}
0.50	0.620	3.08×10^{-2}	3.47×10^{-3}

Table VI—Comparison of model fits for average confusion number, efficiency confusion number, and standard confusion number

	Average Confusion Number		Efficiency Confusion Error		Standard Confusion Error	
T	$\bar{p}_V$	r	$\bar{p}_V$	r	$\bar{p}_V$	r
0.20	3.73×10^{-5}	0.99521	3.70×10^{-5}	0.99539	4.67×10^{-5}	0.99670
0.25	2.03×10^{-4}	0.99952	1.88×10^{-4}	0.99910	1.97×10^{-4}	0.99895
0.30	7.17×10^{-4}	0.99984	6.81×10^{-4}	0.99971	6.62×10^{-4}	0.99950
0.35	2.13×10^{-3}	0.99994	2.02×10^{-3}	0.99981	2.05×10^{-3}	0.99951
0.40	5.26×10^{-3}	0.99999	5.15×10^{-3}	0.99989	5.00×10^{-3}	0.99956
0.45	1.14×10^{-2}	0.99999	1.06×10^{-2}	0.99995	1.01×10^{-2}	0.99964
0.50	2.24×10^{-2}	0.99998	2.04×10^{-2}	0.99986	1.97×10^{-2}	0.99930

efficiency and standard error, and between these estimates and the estimates obtained from the regression line fits for average confusion-number results. This improved agreement is attributed to the fact that the ratio of p_1 to p_2 is much smaller for confusion than for recognition results. The size of this ratio is a good indicator of the disparity among the underlying populations that we are attempting to model with two-way mixtures. The smaller the disparity, the better is the model. In the limit when p_1 equals p_2, indicating a uniform population, a simple model containing no mixtures is appropriate.

Figures 6 and 7 and Table VII present some additional aspects of confusion-error models. Figure 6 shows confusion error results for the three vocabulary types as a function of vocabulary size, together with model fits, with the threshold, T, set to 0.3. Results are shown individually for each of three talkers. Efficiency error results are shown in Fig. 6a though 6c and standard error results in Fig. 6d through 6f. The usual degradation in performance is found passing from V_P to V_W to V_M. Model parameter estimates for efficiency error for all six talkers and three vocabulary types are presented in Table VII. It can be noted that the increase in error rate across these vocabulary types is not consistently accompanied by an increase in the value of the parameter estimates, as was obtained for recognition error.

Figure 7 shows confusion error results with model fits for the six

Table VII—Model parameter estimates for average efficiency confusion error

	V_P			V_W			V_M		
Talker	h	p_1	p_2	h	p_1	p_2	h	p_1	p_2
1	0.350	3.59×10^{-3}	1.16×10^{-4}	0.209	8.34×10^{-3}	7.29×10^{-4}	0.723	6.65×10^{-3}	3.63×10^{-5}
2	0.204	2.84×10^{-3}	4.02×10^{-5}	0.200	4.00×10^{-3}	3.64×10^{-4}	0.661	4.34×10^{-3}	3.11×10^{-5}
3	0.450	1.97×10^{-3}	1.10×10^{-5}	0.200	6.45×10^{-3}	5.90×10^{-4}	0.700	5.74×10^{-3}	9.55×10^{-5}
4	0.060	1.94×10^{-3}	1.12×10^{-4}	0.300	9.87×10^{-4}	4.89×10^{-5}	0.353	2.60×10^{-3}	6.52×10^{-5}
5	0.161	2.50×10^{-3}	1.67×10^{-4}	0.400	1.96×10^{-3}	3.07×10^{-5}	0.551	3.66×10^{-3}	8.43×10^{-5}
6	0.070	4.18×10^{-3}	2.41×10^{-4}	0.350	1.73×10^{-3}	1.14×10^{-4}	0.450	3.68×10^{-3}	3.00×10^{-4}
Mean	0.216	2.84×10^{-3}	1.15×10^{-4}	0.277	3.91×10^{-3}	3.13×10^{-4}	0.573	4.45×10^{-3}	1.02×1^{-4}
Standard deviation	0.156	8.99×10^{-4}	8.73×10^{-5}	0.087	2.94×10^{-3}	2.97×10^{-4}	0.149	1.49×10^{-3}	1.00×10^{-4}

	Mean			Standard Deviation		
Talker	h	p_1	p_2	h	p_1	p_2
1	0.427	6.19×10^{-3}	2.94×10^{-4}	0.266	2.41×10^{-3}	3.79×10^{-4}
2	0.355	3.73×10^{-3}	1.45×10^{-4}	0.265	7.87×10^{-4}	1.90×10^{-4}
3	0.450	4.72×10^{-3}	2.32×10^{-4}	0.250	2.41×10^{-3}	2.13×10^{-4}
4	0.238	1.84×10^{-3}	7.54×10^{-5}	0.156	8.11×10^{-4}	3.28×10^{-5}
5	0.371	2.71×10^{-3}	9.40×10^{-5}	0.197	8.69×10^{-4}	6.87×10^{-5}
6	0.290	3.20×10^{-3}	2.18×10^{-4}	0.197	1.29×10^{-3}	9.50×10^{-5}

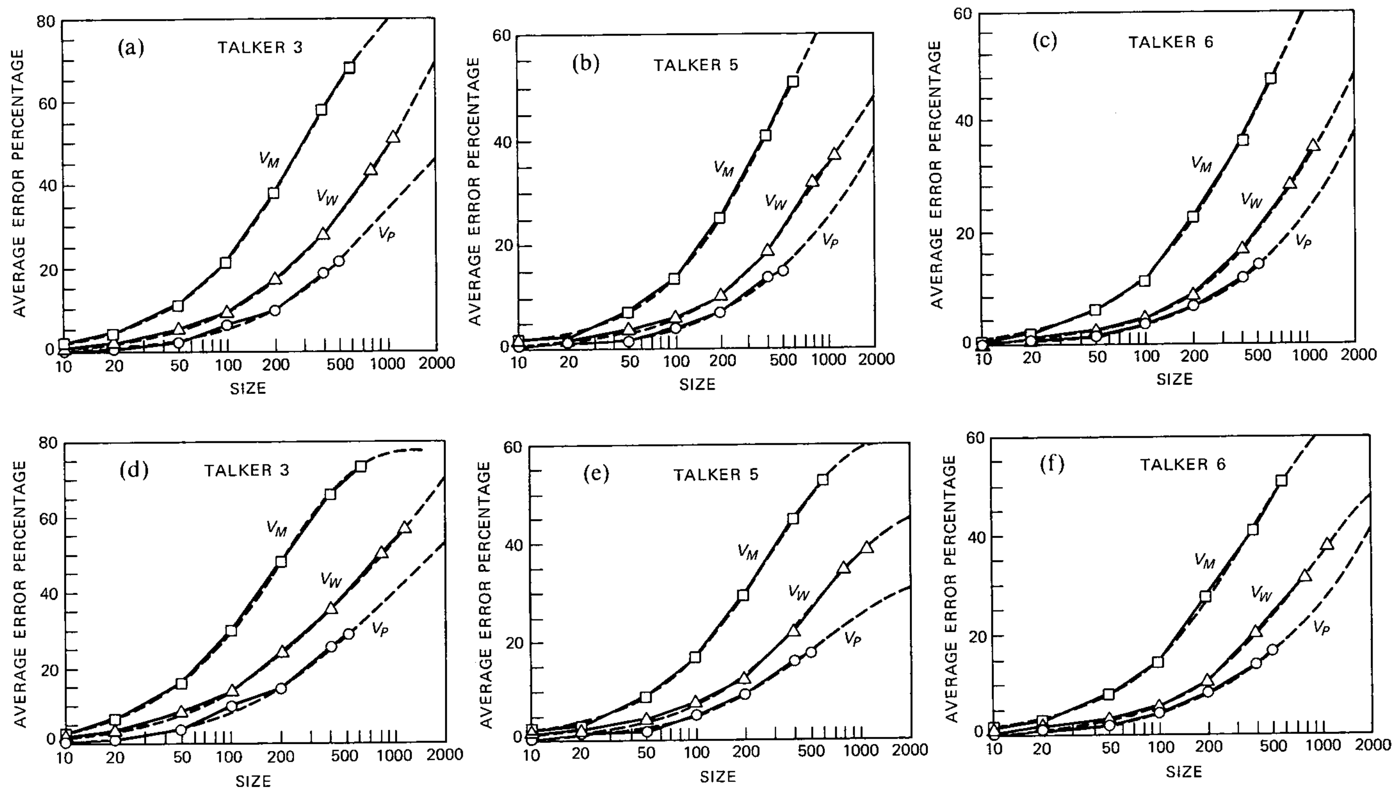

Fig. 6—(a), (b), (c) Average efficiency confusion error, and (d), (e), (f) average standard confusion error as a function of vocabulary size, with two-way mixture model fits, for three vocabulary types.

talkers and the single vocabulary type V_W with the threshold, T, set at 0.3. Efficiency error results are shown in Fig. 7a and standard error results in Fig. 7b. As with recognition error there is considerable variability across talkers, although for confusion error there are no prominent extreme individual performances. There is also an apparent greater tendency for the error rates to converge for small vocabulary

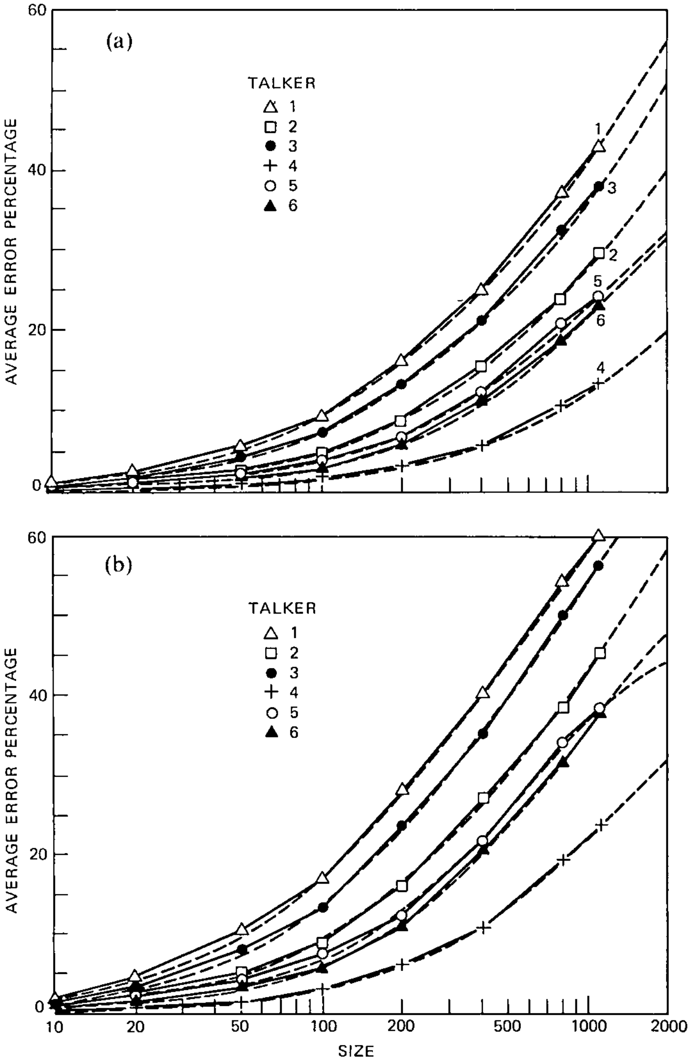

Fig. 7—(a) Average efficiency confusion error, and (b) average standard confusion error as a function of vocabulary size, with two-way mixture model fits, for six talkers, vocabulary type V_W, and threshold value, $T = 0.30$.

sizes. From Table VII, it is apparent that there is a fairly consistent tendency for the parameters p_1 and p_2, but not h, to increase with increasing error rate. Means and standard deviations are calculated across talkers and across vocabulary types in the tables. Using these as indicators of variability, as was done for recognition-error parameters, it appears that estimates of p_1 across vocabulary types, and h across talkers, have relatively small variability, the same as for recognition error. There is also some suggestion that p_1 across talkers also has low variability.

The foregoing observations, together with similar ones made for recognition error, will be discussed in the following section in connection with interpretation of the parameters.

IV. DISCUSSION

In the preceding section we have shown that both the confusion- and recognition-error performance of a recognition system can be modeled quite closely by assuming that there is a mixture of types of recognition or confusion trials, each type associated with a distinct probability for the occurrence of a recognition or confusion error. We have shown that, in most cases, assuming a mixture of two population types is quite adequate to represent the experimental behavior that has been presented, although more than two types might very well underlie this behavior.

A substantial dichotomy of population types is evidenced by a large ratio of p_1 to p_2, the probability estimates of the two populations. It has been found that large ratios, of the order of 100 or 200, are generally associated with recognition error, while smaller ratios, from 10 to 30, generally characterize confusion error.

Where substantial dichotomies exist, it would be most interesting and useful to relate the different population types to actual phenomena associated with the speech-recognition process. Unfortunately, the experimental results incorporate and merge various sources of recognition and confusion error, including the talker, the talking environment, various aspects of the vocabulary, and the recognizer itself. It is difficult, if not impossible, to uniquely characterize these sources in the model parameters. Moreover, in these experimental results, trials are averaged over all four repetitions of each word and all the words in a subset [see eqs. (28) to (33)]. It is therefore not possible to uniquely attribute the different types of behavior to phenomena associated with repeated utterances of the same word on one hand, or to different word types in a subset, on the other. It is reasonable to believe, however, that a dichotomy of types for repeated utterances of the same word would be more significant for rank or recognition performance than for confusion performance. This is because an atypical pronunciation of a word should perturb the self-distance distribution, and consequently the rank distribution, far greater than the distribution of distances to other words in the vocabulary. In fact, in some results not presented here, confusion error was calculated from the distances between reference prototypes alone. Only small differences were obtained with the confusion error results calculated from distances between test utterances and prototypes, which have been presented in the previous section. We are therefore led to believe that differences in populations of trials for confusion error are mostly associated with different types of words, rather than different types of pronunciations of words. So we speculate that for confusion error there are two (or more) populations of words with confusion probabilities differing by ratios of 10 to 30; whereas, for recognition error there are two (or more) populations of trials with recognition probabilities differing by ratios of 100 to 200, where large discrepancies in self-distance distributions for repeated utterances of words are superimposed on population differences among words.

There are two other possibilities for making educated speculations associating model parameters with various phenomena in the recognition process. First, we can examine trends as an experimental parameter such as vocabulary type or distance threshold varies. For example, in the previous section note was taken of which model parameters vary little over the range of some experimental variables. Second, some clues might be obtained if small or large subset approximations of the model error formulations isolate one model parameter from the others. In anticipation of this possibility, some derivations of small and large subset approximations are provided.

4.1 Small and large subset size approximations

Small subset size approximations for the error formulations are obtained by assuming that for small p and N, $(1 - p)^N$ can be represented by the first few terms in the binomial expansion. Rewriting eq. (26) for $M = 2$ we obtain

$$\mathscr{P}\{e_V\} = 1 - h\,\frac{1 - (1 - p_1)^N}{p_1 N} - (1 - h)\,\frac{1 - (1 - p_2)^N}{p_2 N}, \quad (35)$$

where, as in Section 3.3, we assume $p_1 > p_2$. Using the approximation

$$(1 - p)^N \approx 1 - Np + \frac{N(N - 1)}{2}\,p^2 \quad (36)$$

for small N, we obtain

$$\mathscr{P}\{e_V\} \approx \frac{N - 1}{2}\,(hp_1 + (1 - h)p_2) = \frac{N - 1}{2}\,p_V. \quad (37)$$

Similarly, eq. (27) rewritten for $M = 2$ is given by

$$\mathscr{E}\{E_V\} = 1 - h(1 - p_1)^{N-1} - (1 - h)(1 - p_2)^{N-1}. \qquad (38)$$

Approximating $(1 - p)^{N-1}$ by $1 - (N - 1)p$, we obtain

$$\mathscr{E}\{E_V\} \approx (N - 1)[hp_1 + (1 - h)p_2] = (N - 1)p_V. \qquad (39)$$

Thus, for small N, expected error grows linearly with N just as expected rank or confusion number does for all N [see eq. (23)]. In fact, the approximation for standard error is identical to eq. (23). Also, for these small N formulations, the expected value of efficiency error is just one half the expected value of standard error. This can be observed in the experimental results as pointed out in Section 3.1. It is easily verified from the basic expressions for efficiency and standard error found in eqs. (12) and (14) for $N = 2$.

For large subset size approximations we might assume that $(1 - p)^N \approx 0$ for large N. Then we can approximate the efficiency error formulation, eq. (35), by

$$\mathscr{E}\{e_V\} \approx 1 - \frac{1}{N}\left(\frac{h}{p_1} + \frac{1 - h}{p_2}\right). \qquad (40)$$

A comparable approximation does not exist for standard error. However, it is possible to approximate $(1 - p)^N$ by e^{-pN} for small p and moderate pN. This can be introduced in eq. (39) to obtain

$$\mathscr{E}\{E_V\} \approx 1 - he^{-(N-1)p_1} - (1 - h)e^{-(N-1)p_2}, \qquad (41)$$

which is a potentially useful approximation.

For these small and large vocabulary size approximations to be useful in providing interpretations for the model parameters, conditions must exist for one or the other of the population types to dominate. For small N, since we have assumed $p_1 > p_2$, for type 1 populations to dominate in eqs. (37) and (39) we should have

$$\frac{h}{1 - h}\frac{p_1}{p_2} \gg 1. \qquad (42)$$

Conversely, for type 2 populations to dominate in eq. (40), we should have

$$\frac{1 - h}{h}\frac{p_1}{p_2} \gg 1. \qquad (43)$$

4.2 Small and large subset approximations and the relation between rank and confusion error

These small and large N formulations for error coincide with our earlier discussion in Section 3.1 on the relation between confusion and rank or recognition error. There we found that for small N, recognition error is approximately equal to confusion error at a threshold for which average confusion number equals average rank number. For large N we found that recognition error approximates confusion error at a threshold equal to average self-distance. Examining these relationships once more with respect to model parameters, we see that for small N, from eq. (37) or (39), the threshold for equality should be set such that $\bar{p}_{q_{V}(T)} = \bar{p}_{r_V}$. (The subscripts are used to differentiate between confusion and rank.) From eq. (23) we know that $\bar{p}_V$ is the slope coefficient for expected rank or confusion number, so our earlier hypothesis for small N is confirmed. For the example of talker 3 and vocabulary V_W used in Section III, $\bar{p}_{r_{V_W}}$ is 6.67×10^{-3} from Table I. From the same table, we see that for $\bar{p}_{q_{V_W}(T)}$ to have this value, T should be between 0.35 and 0.40.

For large N, from eq. (40), assuming eq. (43) holds, equal errors should be obtained if $1 - h/p_2$ is the same for both confusion and rank. Again for the same example, for rank or recognition, $1 - h/p_2 \approx 6 \times 10^3$. Although confusion model parameter estimates are not shown for talker 3 as a function of threshold, for a threshold of 0.235, corresponding to average self-distance, they are approximately 0.04, 5.3×10^{-3}, and 1.8×10^{-4}, for h, p_1, and p_2, respectively. Thus, $1 - h/p_2 \approx 5.3 \times 10^3$, which agrees well with the value obtained for rank. Thus, the model formulations and parameter estimates support the original hypothesis for large N as well. For large N the density of words for a given vocabulary type is great enough so that even though the distribution of rank numbers and confusion numbers for a threshold set to average self-distance is not the same, the proportion of zero and nonzero rank and confusion numbers, which correlates well with both kinds of error, is about the same. In the discussion that follows, we will conjecture that the parameters of the large N formulation, h and p_2, are largely associated with vocabulary type which should be the major factor controlling density.

4.3 Small and large subset size approximations and dominance of population types

Now let us examine some of the experimental results to determine to what extent eq. (42) or (43) holds. For recognition error, in Table III, we find that generally high ratios of p_1 to p_2 are to a large extent offset by small values of $h/1 - h$. Thus, although the value of the expression in eq. (42) is nearly always greater than one, it is not consistently greater than 10, a value for which we could say unequivocally that type 1 populations dominate small vocabulary size behavior. Those instances in which the expression assumes values less than 10 are associated with low error rates, for example, for talkers 1 and 2. Quite the opposite is true for large size behavior, since in eq. (43), $1 - h/h$ is always greater than one and the ratio of p_1 to p_2 remains large. Thus large size behavior is consistently dominated by type 2 populations in eq. (40), and also in eq. (41), as is easily verified.

For confusion error, with the threshold fixed at 0.3, the results in Table VII indicate that although the ratio of p_1 to p_2 is smaller than for recognition error, the value of $h/1 - h$ is generally greater. Consequently, overall, the expression in eq. (42) assumes about the same range of values as for recognition error. Similar observations are made for large size behavior in both confusion error and recognition error.

To the extent that type 1 populations control small vocabulary behavior and type 2 populations control large vocabulary behavior, it is natural to associate type 1 populations with trials or words that are chronically "bad" in some sense, and type 2 populations with the "natural" density of a particular vocabulary type. Thus, type 1 errors persist when alternate choices are few and the vocabulary size is small, while the natural density of words in the vocabulary must be important when the vocabulary size is large. By this hypothesis we should expect that good performance associated with low error rates should have only a weak dominance of type 1 trials or words, since there should be fewer bad words or trials. This is, in fact, what is observed. The hypothesis is also in agreement with the observations made earlier in this section on the dichotomy of populations for recognition and confusion error.

4.4 Experimental variability of model parameter estimates and parameter origins

We can now stretch further our speculations on the origins of model parameters by recalling our earlier observations of their relative variability across the experimental variables we observed. We assume three sources of error, the talker, the vocabulary, and everything else which we lump into the recognition system. For confusion error, p_1 was observed to have low variability across vocabulary types, and to a lesser extent, across talkers. Therefore, we could associate p_1, the type 1 probability, largely with the system. For recognition or rank performance, p_1 has low variability only across vocabulary types, and is therefore associated with both talker and the system. This reflects the effect of self-distance distribution, which is clearly talker dependent. The type 2 probability, p_2, was observed to have low variability across talkers for recognition or rank performance (except one talker). Although a similar observation was not made for confusion performance, it is natural to associate p_2 with vocabulary type and the system. This hypothesis is compatible with the vocabulary density role associated with p_2 earlier in our discussion. Finally, h, the mixing coefficient for the two types of populations, was observed to have low variability across talkers for both confusion and recognition performance. We are therefore led to believe that h is largely a function of vocabulary type, with the role of the system unclear.

V. CONCLUSION

The data extracted from a series of isolated word recognition experiments with large vocabularies have enabled us to hypothesize and verify a simple probabilistic model underlying performance of recognizers. Essentially, we have attempted to model the distributions of confusion number, an a priori characterization of a recognizer, and rank number, an a posteriori characterization. Expressions have been derived for three confusion or rank number functions, average confusion or rank number, and two error functions, standard error and efficiency error. Models have been evaluated and interpreted using experimental values of these functions. The difference between standard error and efficiency error has been described and an attempt has been made to describe and interpret the difference between confusion and rank performance.

It is significant that good models for performance are obtained only by assuming a mixture of probability distributions as the basis. The reduction of the performance of a recognition system over a large range of vocabulary sizes to as little as three parameters enhances our understanding of the processes involved and has some potential practical utility in the evaluation of systems. Over the range of experimental variables available in this series of experiments we have been able to speculate on associations of the model parameters with variables in the recognition process. To place these suggestions on firmer footing will require additional experimental data. For example, useful data can be obtained from a large number of repeated utterances for a given talker and vocabulary in order to attribute behavior differences uniquely to the different words in a vocabulary or to repetitions of the same words. Examining results obtained by passing the same utterances through different recognizers, or systematic variations of the same recognizer or recording environment, will also be revealing. The use of a larger number and more sharply distinct vocabulary types will also provide useful information. In addition it is important to devise experiments to evaluate the predictive power of the models. Thus, once the parameters of a model have been estimated, new experimental data obtained with controlled variation of experimental variables should be consistent with the model.

VI. ACKNOWLEDGMENTS

The author wishes to thank J. G. Wilpon for maintaining the experimental database, L. R. Rabiner, S. E. Levinson, J. L. Hall, and M. M. Sondhi for stimulating discussions, and L. A. Shepp for valuable suggestions on the use of probability mixtures.

REFERENCES

1. L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon, and W. J. Keilin, "Isolated Word Recognition for Large Vocabularies," B.S.T.J., *61,* No. 10 (December 1982), pp. 2989–3005.
2. A. R. Smith and L. D. Erman, "Noah—A Bottom-up Word Hypothesizer for Large-Vocabulary Speech Understanding System," IEEE Trans. on Pattern Analysis and Machine Intelligence, *PAMI-3* (January 1981), pp. 41–51.
3. W. Feller, *An Introduction to Probability Theory and Its Applications, Vol. II,* New York: Wiley, pp. 52–7.
4. W. L. Johnson and S. Kotz, *Distributions in Statistics: Discrete Distributions,* Boston: Houghton-Mifflin, 1969, pp. 76–9.
5. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. on Acoust., Speech, and Signal Processing, *ASSP-23* (February 1975), pp. 67–72.
6. L. R. Rabiner and S. E. Levinson, "Isolated and Connected Word Recognition—Theory and Selected Applications," IEEE Trans. Commun., *COM-29* (May 1981), pp. 621–59.
7. C. K. Ogden, *Basic English: International Second Language,* New York: Harcourt, Brace and World, Inc., 1968.
8. L. R. Rabiner and J. G. Wilpon, "A Simplified Robust Training Procedure for Speaker Trained, Isolated Word Recognition Systems," J. Acoust. Soc. Amer., *68* (November 1980), pp. 1271–6.
9. J. E. Dennis, Jr. and H. H. W. Mei, "An Unconstrained Optimization Algorithm Which Uses Function and Gradient Values," unpublished report based on M. J. D. Powell's, "A New Algorithm for Unconstrained Optimization," contained in J. B. Rosen et al., eds., *First Symposium on Nonlinear Programming,* New York: Academic Press, 1970.

AUTHOR

Aaron E. Rosenberg, S.B. and S.M. (Electrical Engineering), 1960, Massachusetts Institute of Technology; Ph.D. (Electrical Engineering), 1964, University of Pennsylvania; AT&T Bell Laboratories, 1964—. Mr. Rosenberg is a Member of Technical Staff in the Acoustics Research Department. Since joining AT&T Bell Laboratories his research interests have included auditory psychophysics, speech perception, and currently, speech and speaker recognition. He has authored or coauthored over 35 papers in these fields. Former chairman, IEEE Acoustics, Speech, and Signal Processing Society Technical Committee on Speech Communication; secretary, ASSP Conference Board; associate editor for speech communication, ASSP Transactions. Member, IEEE, Sigma Xi. Fellow, Acoustical Society of America.