Copyright © 1982 American Telephone and Telegraph Company
THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 10, December 1982
Printed in U.S.A.

Isolated Word Recognition for Large Vocabularies

By L. R. RABINER, A. E. ROSENBERG, J. G. WILPON, and W. J. KEILIN*

(Manuscript received June 2, 1982)

It has long been known that one of the key factors in determining the accuracy of isolated word recognition systems is the size and/or complexity of the vocabulary. Although most practical isolated word recognizers use small vocabularies (on the order of 10 to 50 words), there are many applications that require medium- to large-size vocabularies (e.g., airlines reservation and information, data retrieval, etc). This paper discusses the problems associated with speaker-trained recognition of a large vocabulary (1109 words) of words. It is shown that the practicability of using large vocabularies for isolated word vocabularies is doubtful, both because of the problems in training the system, and because of the difficulty the user has in learning and remembering the vocabulary words for any significant size vocabulary. The importance of studying large word vocabularies for recognition lies in the flexibility it provides for understanding the effects of vocabulary size and complexity on recognition accuracy for both small- and medium-size vocabularies. By constructing subsets of the total vocabulary for recognition, we show that a judicious choice of words can lead to significantly better recognition accuracy than a poor choice of the words in the subset. We show that for each doubling of the size of the vocabulary, the recognition accuracy tends to decrease by a fixed amount, which is different for each talker.

I. INTRODUCTION

In the field of automatic speech recognition, the only type of system to date that has proven useful and practical is the isolated word recognizer. Isolated word recognizers have been in use commercially for a number of years,[1-5] and have been extensively studied in several

* Work performed on BLESP summer assignment at Bell Laboratories.

major research laboratories throughout the world.[6-13] For the most part, applications of isolated word recognizers have limited themselves to vocabulary sizes ranging from small (10 to 30 words) to moderate (30 to 200 words). There are several reasons why there are no commercially available systems that can recognize words from large vocabularies (greater than 200 words). These include:

(*i*) The difficulty of training the system on large vocabularies

(*ii*) The storage required for word templates for large vocabularies

(*iii*) The processing required to recognize words from large vocabularies

(*iv*) The difficulty of accurately recognizing word vocabularies that would be useful in a variety of applications.

The computational problems associated with reasons *ii* and *iii* above (i.e., large storage and large amounts of computation) are rapidly becoming less important as memory and processing costs decrease, and should continue to do so for the forseeable future. The problems in training are very real ones, and will be discussed further in this paper. The problems associated with choice of vocabulary words and accuracy of word recognition are the main topics of this paper.*

Although the practicability of large vocabularies for isolated word recognition is doubtful, the experimental use of large vocabularies provides the opportunity to examine significant issues in automatic word recognition that cannot be examined with small vocabularies. If the vocabulary is sufficiently general, in some sense, it is possible to choose several smaller partitions from the vocabulary, of a given size or complexity, and thereby better understand the effects of vocabulary size, or complexity, on word recognition accuracies.

At the present time it is not even known how currently available isolated word recognizers would perform on large vocabularies—i.e., what factors would most influence accuracy. For small- and medium-size vocabularies there is a wide body of experimental data that indicates that vocabulary complexity (not size) is the key indicator of accuracy.[8,12,14] Furthermore, most experimental studies have shown that speaker-independent word recognizers can (and do) perform as well as speaker-trained recognizers; however, they require an order of magnitude more computation.[15]

A brief summary of recent experimental results on isolated word recognition is given in Table I. The results given in this table illustrate the complex relationship between accuracy and vocabulary size and complexity. In Section II of this paper we give a simple model that helps to explain this relationship in terms of the relationships between

* For any practical system, using a large vocabulary of isolated words, syntactic constraints of the recognition task would effectively reduce the vocabulary size and speed up the processing most of the time.

Table I—Accuracies of isolated word recognizers on several vocabularies

Vocabulary	Speaker Trained (%)	Speaker Independent (%)
10 digits [1, 12]	99	98
26 letters of alphabet [16, 17]	80	70
39 alphadigits [8, 12]	87	80
54 computer terms [18, 19]	99	96
91 North American States [14]	99	—
129 Airline terms [20, 21]	88	91
561 Words and Phrases [18]	92	—

words in the vocabulary. In Section III we describe an experiment designed to measure word recognition accuracy for an 1109-word vocabulary. The recognizer was run in a speaker-trained mode on six talkers (three male, three female), in which each talker used a robust training procedure to give individual word templates. In Section IV we discuss results on recognition of subsets of the 1109-word vocabulary. These results illustrate the degree to which choice of vocabulary words can influence word accuracy for given vocabulary sizes. Finally, in Section V we summarize our findings and discuss their implications for practical systems.

II. MODEL FOR ISOLATED WORD RECOGNITION ACCURACY AND COMPLEXITY

Assume we have a specified vocabulary, V, of Q words, i.e.,

$$V = \{v_1, v_2, \cdots, v_Q\}. \tag{1}$$

We define a word similarity index as $D(v_i, v_j)$, which measures the distance (in whatever units are desirable) between pairs of vocabulary words, v_i and v_j. The distance can be an acoustic one (e.g., the average distance of the time-aligned words) or a phonetic one [e.g., the average number of phonemes (syllables, demisyllables) that are different in the words]. We next define a word overlap index, q_i, for the ith vocabulary word as

$$q_i = C\{j{:}s.t.\ D(v_i, v_j) \leq T\}, \tag{2}$$

where C is the cardinality of the set of indices j such that the pairwise word distance score, $D(v_i, v_j)$* falls below a threshold T. Basically, q_i is a count of the number of words in the vocabulary similar to word v_i.

We can now define an average probability of error as

* For simplicity we assume that work distances, D, are symmetric, i.e., $D(v_i, v_j) = D(v_j, v_i)$. In practice, for nonsymmetric distances we use the average pairwise word distance, i.e., $[D(v_i, v_j) + D(v_j, v_i)]/2$.

$$P(E_Q) = \sum_{i=1}^{Q} P(v_i)P(E|v_i), \tag{3}$$

where $P(v_i)$ is the a priori probability word v_i is spoken, $P(E|v_i)$ is the probability of error given word v_i is spoken.[†] Since we assume all words are equiprobable, we have

$$P(v_i) = \frac{1}{Q}. \tag{4}$$

We now make the simplistic assumption that the probability of error given word v_i is spoken can be written as

$$P(E|v_i) = 1 - \frac{1}{q_i}, \tag{5}$$

i.e., we assume a random choice is made among the q_i similar versions of word v_i. Clearly the resulting error rate based on this assumption is an overbound on the true probability of error. Combining eqs. (2) through (5) we get

$$P(E_Q) = \frac{1}{Q} \sum_{i=1}^{Q} \left(1 - \frac{1}{q_i}\right). \tag{6}$$

To illustrate the interpretation of eq. (6) consider calculating the average value of q_i as

$$\bar{q} = \frac{1}{Q} \sum_{i=1}^{Q} q_i. \tag{7}$$

The quantity $\bar{q}$, which we call the average vocabulary complexity, is a measure of the average number of candidates in the vocabulary similar to any word. Since q_i satisfies the constraint

$$1 \le q_i \le Q \tag{8a}$$

then $\bar{q}$ satisfies the constraint

$$1 \le \bar{q} \le Q. \tag{8b}$$

Consider now a $Q = 10$ word vocabulary. We can define various possible sets of q_i and compute $P(E_Q)$ and $\bar{q}$ for each set. For example, if we have

$$\{q_i\} = \{3, 3, 3, 3, 3, 3, 2, 2, 2, 2\} \tag{9a}$$

then $\bar{q} = 2.6$ and $P(E_Q) = 0.6$. Similarly, if

$$\{q_i\} = \{7, 7, 7, 7, 7, 7, 7, 1, 1, 1\} \tag{9b}$$

[†] Technically, eq. (3) should contain a small residual error term that accounts for errors owing to improper recordings, mispronunciations, etc. We will omit this term for simplicity.

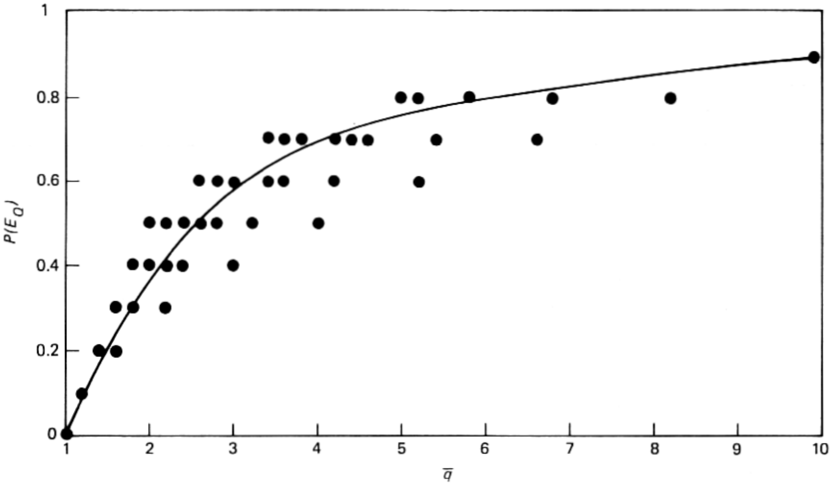

Fig. 1—Plot of average word error rate as a function of average word complexity for all possible combinations of a 10-word vocabulary. The smooth curve is a hand-drawn curve, which approximates the average behavior of the data.

then $\bar{q} = 5.2$ and $P(E_Q) = 0.6$. The vocabulary of eq. (9a) consists of four subsets, two of which have three confusable words, and two of which have two confusable words. The vocabulary of eq. (9b) consists of one subset with seven confusable words, and three distinct words. Both vocabularies, however, yield identical error probabilities using the simple model given above.

If we consider all possible subsets of the 10-word vocabulary, and plot the values of $P(E_Q)$ versus $\bar{q}$ for each such subset, the resulting plot would be as shown in Fig. 1. This figure shows that for a given probability of error a wide range of vocabulary complexities can often be found. It also shows that as the probability of error goes to the residual value, the choice of vocabularies becomes sparse—i.e., only well-designed vocabularies will achieve the lowest error rates.

This simple word recognition model could also be described in information theoretic terms based on channel models.[22] From such models one could derive plots equivalent to the one of Fig. 1.

Consider applying the word recognition model to some of the vocabularies of Table I. For the 10 digits we get (using $q_i = 1$, all i) $P(E_{10}) = 0$, $\bar{q} = 1$. For the 26 letters of the alphabet (ordered alphabetically), using*

$$\{q_i\} = \{1, 5, 2, 5, 5, 2, 5, 1, 2, 2, 2, 1, 2, 2, 1, 2, 1, 1, 2, 2, 1, 5, 1, 1, 2, 2\}$$

we get $\bar{q} = 2.2$, $P(E_{26}) = 0.385$. For the 39-word alphadigit vocabulary[16]

* The values of q_i for the alphabet were obtained from the letter confusion matrix in Ref. 16.

we get $\bar{q} = 2$ and $P(E_{39}) = 0.28$. For the 54-word computer terms vocabulary[18,19] we get $\bar{q} = 1.1$ and $P(E_{54}) = 0.04$. For the 129-word airline terms vocabulary[20,21] we get $\bar{q} \approx 1.1$ and $P(E_{129}) = 0.08$. By comparing the error probabilities from the model with those given in Table I it can be seen that a reasonable match to all vocabularies can be obtained using this simple recognition model.

The major purpose of this section has been to illustrate the range of variability in error rate associated with a vocabulary of fixed-size Q words, and to roughly explain the source of variation. The key point is to keep in mind that judicious choice of vocabulary items can lead to considerably higher word recognition accuracies than can a poor choice of vocabulary items. We will illustrate this key point further in later sections of this paper.

III. WORD RECOGNITION ON AN 1109-WORD VOCABULARY

To evaluate the performance of an isolated word recognizer on large vocabularies, the linear predictive coefficient (LPC) based recognizer developed at Bell Laboratories was tested on a vocabulary of 1109 words from the Basic English vocabulary of Ogden.[23] The recognizer was tested in a speaker-trained mode with six talkers (three male, three female) each training the recognizer.

Before presenting results of the evaluation tests, we briefly review the techniques used for recognition and training.

3.1 The LPC-based word recognizer

Figure 2 shows a block diagram of the LPC-based word recognizer. The input speech signal, $s(n)$, recorded off a standard dialed-up telephone line, is bandpass-filtered between 100 and 3200 Hz, and digitized at a 6.67-kHz rate. The first step in the processing is the preprocessing and blocking step, which consists of a simple first-order preemphasis network. The preemphasized signal is blocked into frames of 45 ms $(N = 300)$ with each consecutive frame spaced 15 ms apart $(L = 100)$. An 8-pole LPC analysis (autocorrelation method) is performed on each frame of the word (which has presumably been located by an endpoint

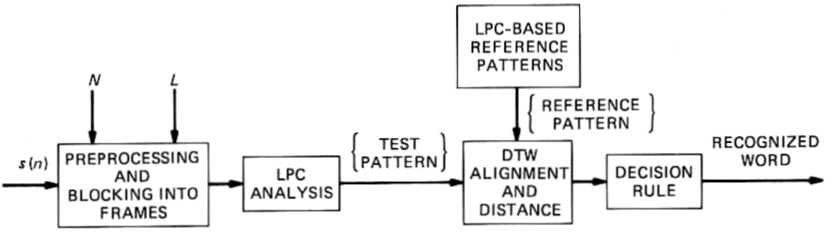
LPC-BASED REFERENCE PATTERNS
N
L
REFERENCE PATTERN
s(n)
PREPROCESSING AND BLOCKING INTO FRAMES
LPC ANALYSIS
TEST PATTERN
DTW ALIGNMENT AND DISTANCE
DECISION RULE
RECOGNIZED WORD

Fig. 2—Block diagram of isolated word recognizer.

detector), creating the test pattern T. This test pattern is compared with each reference pattern using a dynamic time warping (DTW) alignment algorithm, which simultaneously provides a distance score associated with the alignment. The distance scores for all the reference patterns are sent to a decision rule, which provides an estimate as to the spoken word, and possibly an ordered (by distance) set of the best n candidates.

3.2 The robust training procedure

The procedure used to obtain speaker-dependent reference patterns is the robust training procedure of Rabiner and Wilpon.[24] For this method each talker speaks each vocabulary word repetitively (up to six times) until a pair of word tokens are deemed sufficiently similar (based on a DTW distance score). The word reference pattern is created by averaging the two time-aligned versions of the word (the autocorrelation coefficients of each frame are averaged). This procedure yields robust reference patterns since a tight similarity threshold is used to guarantee that the two tokens of the word are free of artifacts by either the talker (e.g., lip smacks, pops, heavy breathing), or from the transmission environment. In this manner it is essentially guaranteed that each of the two tokens being averaged represents a valid pronunciation of the word.

There are two points worth noting about the robust training procedure. The first is that each word is not spoken repetitively until the robust reference pattern can be created. To maximally separate in time the repetitions of each vocabulary word, the talker training the system speaks the entire vocabulary in a random sequence once each pass through the training. The disadvantage of such a procedure is that a considerable amount of storage is required to save the multiple versions of each word that may be required before a robust reference pattern can be obtained. The big advantage of this method is that each word token tends to be an independent pronunciation of the word; hence, word variability is easily and readily measured.

The second point about the robust training concerns the validity of the reference pattern that is obtained. For words with stop releases at the end, e.g., act, back, stop, etc., a speaker will often vary the pronunciation (almost at random). Thus, on some occurrences of these words the speaker will release the stop consonant (leading to a burst at the end of the word) and on other occasions will not release the stop consonant. For such words it should be clear that the robust training procedure cannot adequately represent this dichotomous method of speaking the word and will instead lock onto one of the two variations. For such words a high probability of error is introduced, not by alternate competing words in the vocabulary, as discussed in Section

II, but instead by alternate competing word pronunciations. We see no simple or obvious way of handling this problem.

The robust training procedure trades training time (on the part of the talker) for the ability to obtain robust reference patterns for each word in the vocabulary. For a large vocabulary, notably the 1109-word vocabulary, the average time to train all 1109 words was 5-1/2 hours for each talker! It became imminently clear to the authors that, in practice, one could never consider training speech recognizers for large vocabularies in such a manner. If the need ever arose for word recognition for large vocabularies, an automatic template generation procedure would be required in lieu of the robust training that was used here. Such automatic-template-generation techniques have been used by Mermelstein[25] and by Rosenberg et al.[26] for equivalent size vocabularies using syllable and demisyllable representations of words.

3.3 *Isolated word recognition experiments*

To evaluate the isolated word recognizer on various size vocabularies, a series of word recognition experiments were run. For each of the six talkers, four complete test sets, each consisting of one token each of the entire 1109-word vocabulary, were recorded. The recording took place over four weeks in time, and required about eight hours of recording time for each talker.

The entire data base was used in the first experiment, which consisted of measuring the error rate, E_{1109}, as a function of talker (i), replication (j), and candidate position (n). This experiment provides the absolute performance measure of the word recognizer on the largest vocabulary tested to date.

The next series of experiments basically considered subsets of the 1109-word vocabulary for both training and testing. The Q-word subset of the vocabulary was chosen in several ways to study the influence of means of vocabulary choice on the error rate. The ways in which vocabulary entries were chosen for the Q-word vocabulary included:

(i) Random without replacement—i.e., each of the Q vocabulary words was chosen at random from the 1109-word vocabulary. For each replication of this experiment, the Q words were chosen from the candidates not selected on previous trials. Clearly, a maximum number of trials, $MT = 1109/Q$, is possible with this selection procedure. Since we considered values of Q of 100, 200, 400, and 800, values of MT of 11, 5, 2, and 1 were used, respectively, for the different values of Q.

(ii) Random with replacement—i.e., each of the Q vocabulary words was chosen at random from the 1109-word vocabulary. On subsequent replications a new set of Q words was chosen at random, again from the complete set of 1109 words. For this method of word selection, the same vocabulary word could appear in several replications of the vocabulary. To compare the results of this experiment with those of the one above, the same values of Q and MT were used.

(*iii*) Vocabulary chosen based on best training tokens—i.e., the Q words of the vocabulary, for each talker, were chosen as the Q words (of the 1109) that required the fewest training tokens before the robust reference pattern was obtained. Such words represent the "easiest words to train on," and were expected to be least affected by inherent variability in word pronunciations. Values of Q of 100, 200, 400, and 800 were used.

(*iv*) Vocabulary chosen based on worst training tokens—i.e., the Q words of the vocabulary, for each talker, were chosen as the Q words that required the most training tokens before the robust reference pattern was obtained. Such words represent the "hardest words to train on," and were expected to be most affected by inherent variability in word pronunciations. Values of Q of 100, 200, 400, and 800 were used.

(*v*) Vocabulary with proportional training statistics—i.e., the Q words of the vocabulary, for each talker, were chosen on an equal proportion with their statistics on training. Thus, if a talker had P_2 training words requiring two replications, P_3 training words requiring three replications, etc., then in the test set a total of $(P_j/1109) \cdot Q$ words were chosen at random from the set of words requiring j training replications. In this manner a vocabulary with statistics representative of the training difficulty was obtained. Values of Q of 100, 200, 400, and 800 were used.

(*vi*) Vocabulary with all monosyllabic words. A separate score was obtained using only the $Q = 605$ monosyllabic words in the 1109-word vocabulary.

(*vii*) Vocabulary with all polysyllabic words. A separate score was obtained using only the $Q = 504$ polysyllabic words in the 1109-word vocabulary.

The results of these word recognition experiments are given in the next section.

3.4 Recognition test results

The results of the first experiment, using all 1109 words in the vocabulary, are given in Table II and shown graphically in Figs. 3 and 4. Table II shows values of $E_{1109}(i, j, n)$ for values of i $(i = 1, 2, \cdots, 6)$, j $(j = 1, 2, 3, 4)$, and n $(n = 1, 2, 3, 4, 5)$. Figure 3 shows plots of $E_{1109}(i, j, n)$ versus n for each talker, i, and each replication, j. Figure 4a shows plots of $\bar{E}_{1109}(i, n)$ versus n, where

$$\bar{E}_{1109}(i, n) = \frac{1}{4} \sum_{j=1}^{4} E_{1109}(i, j, n), \tag{10a}$$

Table II—Word error rates as a function of talker (*i*), replication (*k*), and word position (*n*)

	n					
Talker	1	2	3	4	5	k
$i = 1$	12.3	6.3	4.2	3.7	2.9	1
	18.8	9.8	7.2	5.0	4.2	2
	15.6	9.4	6.0	4.4	3.8	3
	12.1	6.4	4.1	2.4	2.1	4
$i = 2$	6.1	2.5	1.4	1.0	0.8	1
	5.7	2.4	1.5	1.4	1.2	2
	6.0	2.1	1.3	1.0	0.9	3
	6.4	3.1	1.9	1.4	1.3	4
$i = 3$	18.4	12.2	10.2	9.2	8.6	1
	19.9	13.1	10.9	9.2	8.7	2
	23.0	15.7	12.3	10.6	10.0	3
	17.8	12.9	10.1	8.9	8.4	4
$i = 4$	40.2	31.2	27.1	24.3	23.4	1
	38.0	29.9	26.1	24.1	22.4	2
	46.7	36.9	32.2	29.8	27.5	3
	48.3	39.9	35.2	32.3	30.1	4
$i = 5$	17.7	10.9	8.7	7.2	6.3	1
	24.8	16.5	12.9	11.0	9.8	2
	20.9	23.5	9.9	7.6	6.9	3
	27.3	17.9	14.2	13.0	11.6	4
$i = 6$	17.0	12.0	9.7	8.3	7.3	1
	17.3	12.4	9.9	8.6	7.7	2
	21.1	15.0	11.8	10.3	9.6	3
	18.6	12.4	9.9	8.8	7.8	4

where $\bar{E}_{1109}(i, n)$ is the error rate averaged over replications; Fig. 4b shows the grand average plot $\hat{E}_{1109}(n)$ versus n, where

$$\hat{E}_{1109}(n) = \frac{1}{6}\sum_{i=1}^{6} \bar{E}_{1109}(i, n) \tag{10b}$$

$$= \frac{1}{24}\sum_{i=1}^{6}\sum_{j=1}^{4} E_{1109}(i, j, n). \tag{10c}$$

Two points in the results are worth noting. It can clearly be seen that within the four replications of a single talker, the error rate scores for a given value of n do not vary a great deal (relative to the absolute error rates). However, across talkers a large amount of variation in error scores is seen for all values of n (see Fig. 4a). Thus, talker 4 has an average error rate of 43.3 percent for $n = 1$, whereas talker 2 has an average error rate of 6.0 percent, a range of over 7 to 1 in error rates.

The grand average (over talkers and replications) error rate curve shows an average error rate of 20.8 percent of the top candidate, and the error rate falls to 9.3 percent for the top five candidates. Although these absolute scores are highly biased by the talker with the high error rate (talker 4), the curves of Fig. 4 show that the error rate for all

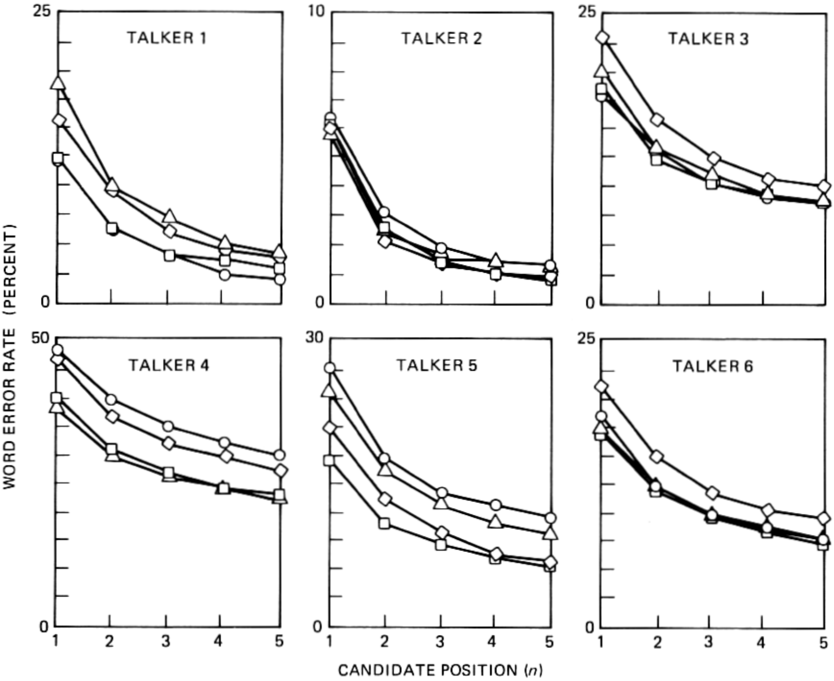

Fig. 3—Word error rate scores for each talker and for each replication as a function of word candidate positions.

talkers displayed similar trends, and hence the grand average curve is representative of the overall behavior of the isolated word recognizer for this vocabulary.

The results of the tests using subsets of the 1109-word vocabulary are given in Table III. For each talker and for each vocabulary partition size Q, this table gives the average error rate (averaged over the four replications) for the top candidate as a function of the subset condition (1 to 7 as described previously). An examination of the data in this table shows the following:

(*i*) Conditions 1 and 2 (random selection without and with replacement) lead to essentially the same error scores on all subsets of the vocabulary for all talkers.

(*ii*) For small vocabulary sizes (Q = 100, 200) selection of vocabulary items based on training statistics leads to very different error rates, depending on the exact set of training statistics used. The error rate scores for condition (*iii*) (best training words) were significantly lower than the error rate scores for condition (*iv*) (worst training words). The error rate scores for condition (*v*) (equal proportions) were essentially comparable to those of conditions (*i*) and (*ii*) and somewhere between those of conditions (*iii*) and (*iv*).

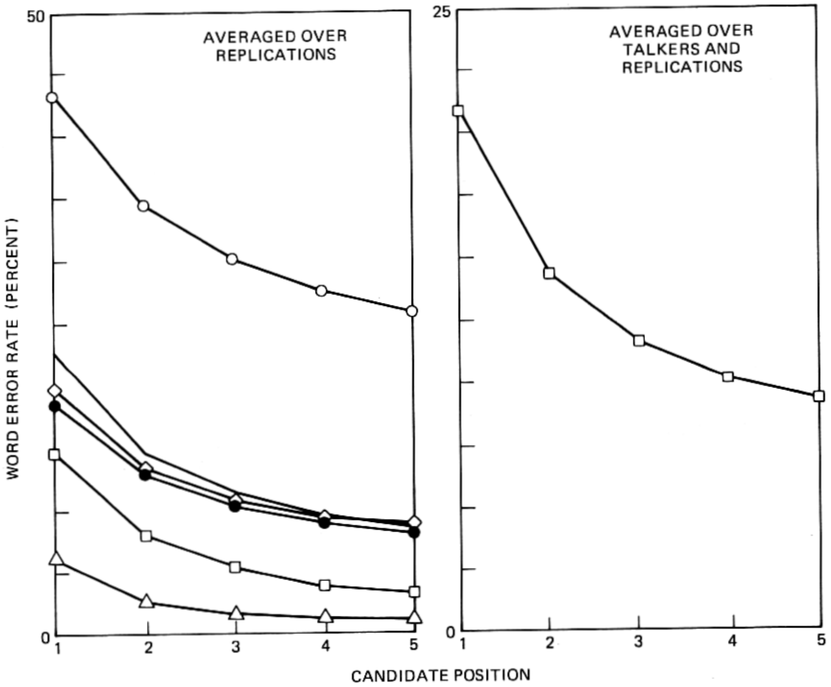

Fig. 4—(a) Average word error rate for each talker as a function of word position. (b) Grand average word error rate as a function of word position.

(*iii*) For the larger vocabulary partitions (Q = 400, 800) the effects of choosing vocabulary words based on training statistics on the error rate were small.

(*iv*) The error rates for monosyllabic words alone [condition (*vi*)] were always significantly larger than for any other subset (or even the whole vocabulary) of the vocabulary; similarly, the error rate scores for polysyllabic words alone [condition (*vii*)] were significantly smaller than for any other subset of the vocabulary.

Figure 5 shows a summary plot of the average error rate for each talker as a function of the logarithm of the vocabulary size, and a least squares regression fit to the data points. The data points represent averages of conditions (*i*) and (*ii*) data of Table II. It can be seen that remarkably good fits to the data are obtained, for all talkers, by the least squares regression line. It should be noted that the scales for each talker are different, reflecting the differences in absolute error rates. Similarly, the slopes of the linear fits are different for each talker. In particular the slopes for the individual talkers are 3.1, 1.3, 2.9, 5.7, 4.0, and 3.2, respectively. A slope of α means that for each doubling of vocabulary size, the predicted error rate increases by α percent.

Table III—Average word rates as a function of the partitioning of the vocabulary for each talker

		Q					
Condition		100	200	400	800	605	504
Talker 1	1	3.9	6.4	9.1	13.1		
	2	4.2	5.7	9.4	12.3		
	3	2.5	7.4	7.7	12.2		
	4	5.3	8.0	10.6	12.9		
	5	2.5	6.2	9.2	12.9		
	6					20.8	
	7						6.1
Talker 2	1	1.9	2.6	4.0	6.1		
	2	1.5	2.4	3.2	5.6		
	3	1.5	2.4	2.8	4.2		
	4	3.5	4.6	4.8	5.7		
	5	3.0	2.1	3.8	4.6		
	6					10.0	
	7						2.0
Talker 3	1	10.2	12.9	15.3	18.1		
	2	9.3	10.8	14.7	17.7		
	3	9.0	10.9	12.1	15.2		
	4	25.7	21.6	18.9	18.9		
	5	9.0	11.1	12.7	16.1	29.5	
	6						
	7						7.4
Talker 4	1	23.2	28.0	33.3	40.9		
	2	24.6	29.0	34.6	40.8		
	3	19.8	25.0	29.8	37.5		
	4	31.7	37.2	40.5	43.3		
	5	23.7	24.6	34.2	38.2		
	6					53.4	
	7						28.0
Talker 5	1	8.9	12.0	15.3	20.3		
	2	9.2	11.5	15.5	20.4		
	3	8.0	9.0	13.0	18.6		
	4	18.7	19.5	18.2	21.5		
	5	9.0	11.1	14.2	19.3		
	6					30.9	
	7						11.8
Talker 6	1	7.5	10.0	13.5	17.0		
	2	7.8	10.3	13.4	17.3		
	3	4.7	7.4	10.2	14.3		
	4	15.2	17.1	18.8	18.5		
	5	7.5	8.2	12.6	14.6		
	6					28.2	
	7						6.6

IV. DISCUSSION

The results presented in the previous section demonstrate clearly the effects of vocabulary complexity on error rate for isolated word recognizers. They also show the high degree of variability among talkers in the error rates for almost any size vocabulary.

Perhaps the most startling observation from the data of Fig. 5 is the fact that, for each talker, a doubling in the vocabulary size leads to a

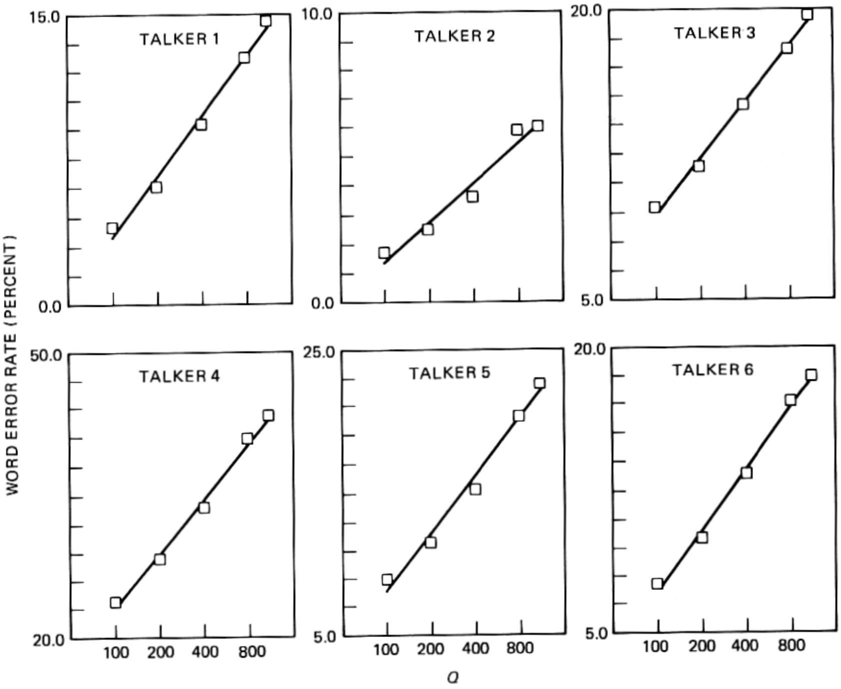

Fig. 5—Average word error rate as a function of vocabulary size for each talker. The straight line is the least squares linear regression fit to the data.

constant (talker-dependent) increase in error rate. This effect has been noted previously by Smith and Erman[27] in their work on word hypothesizing for large vocabulary recognizers. The explanation for this effect is that the error rate is essentially proportional to the density of words in the pattern space [e.g., the factor $(1 - 1/q_i)$ in eq. (6)]. As the number of words in the vocabulary doubles (by random selection), the density increases a constant amount, thereby leading to a constant increase in error rate.

The fact that different talkers have different absolute error rates and different slopes for the same vocabulary sets can be explained by the model of Section II as follows. We postulate that the word similarity threshold, T, of eq. (2) is a talker-dependent threshold in that it is a function of the inherent variability of a talker in repeating a given vocabulary word. For some talkers (e.g., Talker 2) the threshold is set very low and hence very few vocabulary words have q_i values greater than 1. For other talkers (e.g., Talker 4) the threshold is set very high and therefore most vocabulary words have q_i values greater than 1. Thus, the absolute error rate [eq. (6)] will be much higher for talkers with high variability in their word pronunciations than for talkers with low variability in their word pronunciations. Similarly, the increase in error rate for a doubling of vocabulary size is a function (to first order) of the absolute error rate since the density of words in pattern space increases more rapidly for talkers with high word variability than for talkers with low variability.

If the words in the vocabulary are not chosen at random [e.g., conditions (*iii*) to (*vii*) in Section 3.4] then the above analysis is not correct. For example, by choosing words with poor training statistics the average word density is higher than expected, leading to higher word error rates. Similarly, by choosing words with good training statistics, the average word density is lower than expected. Since most words had good training statistics, the effect on the error rate of choosing good training words is generally much smaller than the effect of choosing poor training words. For values of Q approaching 1109 (e.g., $Q = 800$ and $Q = 400$), both effects are smaller since there are only a small number of words (for each talker) whose training statistics were poor.

The average error rates for monosyllables versus polysyllables vividly point out the strong effects of vocabulary complexity. The monosyllable vocabulary of 605 words has a much higher complexity than the total 1109-word vocabulary; hence, it has a much higher error rate for all talkers. Similarly, the 504-word polysyllable vocabulary has a much lower complexity than the 1109-word vocabulary and therefore a much smaller error rate.

V. SUMMARY

In this paper we have presented results of a series of speaker-trained, isolated word recognition tests on a 1109-word vocabulary, and various subsets of the vocabulary. We have shown that although a great deal of variability in error scores was noted across talkers, a fairly good consistency in error scores across replications by the same talker was attained. On the total vocabulary an average (over talkers) error rate of 20.8 percent on the top candidate and 9.3 percent on the top five candidates was obtained. These scores represent the anticipated average performance of the recognizer across different talkers. The best talker achieved a 6.0-percent error rate on the first candidate, whereas the worst talker achieved a 43.3-percent error rate on the first candidate.

By considering various subsets of the 1109-word vocabulary we were able to show that the method of selection of the words within the vocabulary had a strong effect on the word error rate achieved. However, when we used randomly chosen vocabulary subsets all talkers had error rates that increased by a constant percentage for each doubling in the vocabulary size. A simple explanation for this effect was given.

REFERENCES

1. T. B. Martin, "Practical Applications of Voice Input to Machines," Proc. IEEE, *64* (April 1976), pp. 487–501.
2. S. Moshier, "Talker Independent Speech Recognition in Commercial Environments," in Speech Commun. Papers, 97th ASA Meeting, June 1979, pp. 551–3.
3. Interstate Electronics Corp., Voice Data Entry System, unpublished technical descriptions.
4. Centigram Corp., Mike, unpublished technical descriptions.
5. Heuristics Corp., Speechlab, unpublished technical description.
6. W. Lea, ed., *Trends in Speech Recognition*, Englewood Cliffs, NJ: Prentice-Hall, 1980.
7. D. R. Reddy, ed., *Speech Recognition*, New York: Academic, 1974.
8. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-23* (February 1975), pp. 67–72.
9. A. E. Rosenberg and F. Itakura, "Evaluation of an Automatic Word Recognition System Over Dialed-up Telephone Lines," J. Acoust. Soc. Amer., suppl. 1, *60* (1976), p. S12.
10. M. R. Sambur and L. R. Rabiner, "A Speaker-Independent Digit-Recognition System," B.S.T.J., *54* (January 1975), pp. 81–102.
11. L. R. Rabiner, "On Creating Reference Templates for Speaker Independent Recognition of Isolated Words," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-26* (February 1978), pp. 34–42.
12. L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon, "Speaker-Independent Recognition of Isolated Words Using Clustering Techniques," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-27* (August 1979), pp. 336–49.
13. J. S. Bridle and M. D. Brown, "Connected Word Recognition Using Whole Word Templates," in Proc. Inst. Acoust., Autumn 1979.
14. G. M. White and R. B. Neely, "Speech Recognition Experiments With Linear Prediction, Bandpass Filtering, and Dynamic Programming," IEEE Trans. Acoust., Speech, Signal Processing, *ASSP-24* (April 1976), pp. 183–8.
15. L. R. Rabiner and S. E. Levinson, "Isolated and Connected Word Recognition—Theory and Selected Applications," IEEE Trans. Commun., *COM-29*, No. 5 (May 1981), pp. 621–59.
16. B. Aldefeld, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "Automated Directory Listing Retrieval System Based on Isolated Word Recognition," Proc. IEEE, *68*, No. 11 (November 1980), pp. 1364–79.
17. A. E. Rosenberg and C. E. Schmidt, "Automatic Recognition of Spoken Spelled Names for Obtaining Directory Listings," B.S.T.J., *58*, No. 8 (October 1979), pp. 1797–1823.
18. P. Vicens, "Aspects of Speech Recognition by Computer," Ph.D dissertation, Stanford University, April 1969.
19. L. R. Rabiner and J. G. Wilpon, "Speaker-Independent Isolated Word Recognition for a Moderate Size (54-Word) Vocabulary," IEEE Trans. Acoustics, Speech, and Signal Processing, *ASSP-27*, No. 6 (December 1979), pp. 583–7.
20. S. E. Levinson and A. E. Rosenberg, "A New System for Continuous Speech Recognition—Preliminary Results," Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (April 1979), pp. 239–43.
21. J. G. Wilpon, L. R. Rabiner, and A. Bergh, "Speaker-Independent Isolated Word Recognition Using a 129-Word Airline Vocabulary," J. Acoust. Soc. Am., *72,* No. 2 (August 1982), pp. 390–6.
22. R. M. Fano, *Transmission of Information, A Statistical Theory of Communications*, Cambridge, MA: MIT Press, Wiley, 1961, pp. 186ff.
23. C. K. Ogden, *Basic English*: *International Second Language*, New York: Harcourt, Brace and World Inc., 1968.
24. L. R. Rabiner and J. G. Wilpon, "A Simplified Robust Training Procedure for Speaker Trained, Isolated Word Recognition Systems," J. Acoust. Soc. Am., *68*, No. 5 (November 1980), pp. 1271–6.
25. M. J. Hunt, M. Lennig, and P. Mermelstein, "Experiments in Syllable-Based Recognition of Continuous Speech," Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, Denver, Colorado (April 1980), pp. 880–3.
26. A. E. Rosenberg, L. R. Rabiner, S. E. Levinson, and J. G. Wilpon, "A Preliminary Study on the Use of Demisyllables in Automatic Speech Recognition," Proc. Int.

Conf. on Acoustics, Speech, and Signal Processing, Atlanta, Georgia (March 1981), pp. 967–70.

27. A. R. Smith and L. D. Erman, "NOAH—A Bottom-Up Word Hypothesizer for Large Vocabulary Speech Understanding Systems," IEEE Trans. Pattern Analysis and Machine Intelligence, *PAMI-3*, No. 1 (January 1981), pp. 41–51.