Speaker Trained Recognition of Large Vocabularies of Isolated Words

A. E. Rosenberg

L. R. Rabiner

J. G. Wilpon

Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT 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 which require medium to large size vocabularies (e.g. airlines reservation and information, data retrieval etc). It is the purpose of this paper to discuss 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 for the user to learn and remember 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 by 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 which has proven useful and practical is the isolated word recognizer. Isolated word recognizers have been in use commercially for a number of years [1], and have been extensively studied in several major research laboratories throughout the world [2]. 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).

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. This is because 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. 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 [3].

II. Model for Isolated Word Recognition Accuracy and Complexity

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

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

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 i^{th} vocabulary word as

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

where C is the cardinality of the set of indices j such that the pairwise word distance score 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

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

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) - (5) we get

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

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

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

CH 1746-7/82/0000 - 2018 $ 00.75 © 1982 IEEE

then $\bar{q}$ satisfies the constraint

$$1 \leq \bar{q} \leq Q \tag{8b}$$

If we consider all possible subsets of a 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 Figure 1. This figure shows tbat 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.

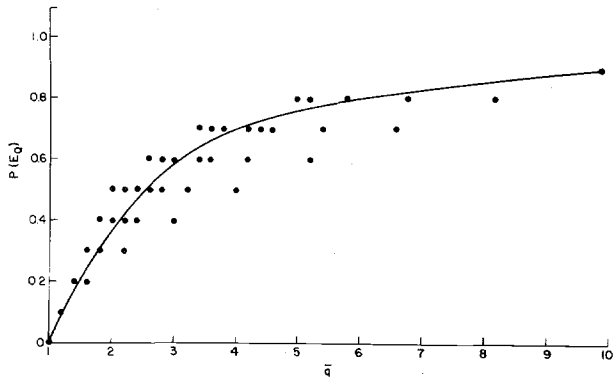

Figure 1 Plot of average word error rate as a function of average word complexity for all possible combinations of a 10-word vocabulary. The smooth curve is a hand drawn curve, which approximates the average behavior of the data.

III. Word Recognition on an 1109 Word Vocabulary

In order to evaluate the performance of an isolated word recognizer on large vocabularies, the LPC based recognizer developed at Bell Laboratories was tested on a vocabulary of 1109 words from the Basic English vocabulary of Ogden [4]. The recognizer was tested in a speaker trained mode with 6 talkers (3 male, 3 female) each training the recognizer using the robust training procedure of Rabiner and Wilpon [5]. For each of the 6 talkers, 4 complete test sets, each consisting of 1 token each of the entire 1109 word vocabulary, were recorded. The recording took place over 4 weeks in time, and required about 8 hours of recording time for each talker.

Using the entire data base the first experiment 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:

1. 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.
2. 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, tbe same vocabulary word could appear in several replications of the vocabulary. In order to compare the results of this experiment with those of the one above, the same values of Q and MT were used.
3. Vocabulary Chosen Based on Best Training Tokens — i.e. the Q words of the vocabulary, for eacb talker, were chosen as the Q words (of the 1109) which 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.
4. Vocabulary Chosen Based on Worst Training Tokens — i.e. the Q words of the vocabulary, for each talker, were chosen as the Q words which 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.
5. 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 2 replications, P_3 training words requiring 3 replications etc, then in tbe test set a total of $(P_j / \sum_{i=2}^{6} P_i) \cdot Q$ words were chosen at random from the 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.
6. Vocabulary With All Monosyllabic Words. A separate score was obtained using only the $Q = 605$ monosyllabic words in the 1109 word vocabulary.
7. Vocabulary With all Polysyllabic Words. A separate score was obtained using only the $Q = 504$ polysyllabic words in the 1109 word vocabulary.

3.1 Recognition Test Results

The results of the first experiment, using all 1109 words in the vocabulary, are shown graphically in Figure 2. Figure 2a shows plots of $\bar{E}_{1109}(i,n)$ versus n, where

$$\bar{E}_{1109}(i,n) = \frac{1}{4} \sum_{j=1}^{4} E_{1109}(i,j,n) \tag{9a}$$

where $\bar{E}_{1109}(i,n)$ is the error rate averaged over replications, and Figure 2b shows the grand average plot $\hat{E}_{1109}(n)$ versus n, where

$$\hat{E}_{1109}(n) = \frac{1}{6} \sum_{i=1}^{6} \bar{E}_{1109}(i,n) \tag{9b}$$

Two points are worth noting about the results. Within the 4 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. 2a). The grand average (over talkers and replications) error rate curve shows an average error rate of 20.8% for the top candidate, and the error rate falls to 9.3% for the top 5 candidates.

The results of the tests using subsets of the 1109 word vocabulary are given in Table I. This table gives, for each talker and for each vocabulary partition size Q, the average error rate (averaged over tbe 4 replications) for the top candidate as a function of the subset condition (1-7 as described previously). An examination of the data in tbis table shows the following:

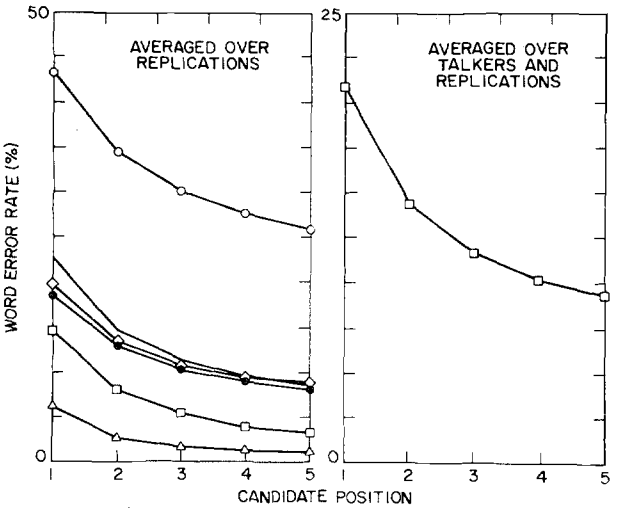

Figure 2 a) Average word error rate for each talker as a function of word position; b) grand average word error rate as a function of word position.

Condition		Q					
		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		
	6					29.5	
	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

Table I

Average Word Rates as a Function of the Partitioning of the Vocabulary for Each Talker

1. 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.
2. 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 3 (best training words) were significantly lower than the error rate scores for condition 4 (worst training words). The error rate scores for condition 5 (equal proportions) were essentially comparable to those of conditions 1 and 2 and somewhere between those of conditions 3 and 4.
3. For the larger vocabulary partitions (Q=400,800) the effects of choosing vocabulary words based on training statistics on the error rate were small.
4. The error rates for monosyllabic words alone (condition 6) 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 7) were significantly smaller than for any other subset of the vocabulary.

Figure 3 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 condition 1 and condition 2 data of Table I. It can be seen that remarkably good fits to the data are obtained, for all talkers, by the least squares regression line.

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 Figure 3 is the fact that, for each talker, a doubling in the vocabulary size leads to a constant (talker dependent) increase in error rate. This effect has been noted previously by Smith and Erman [6] 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

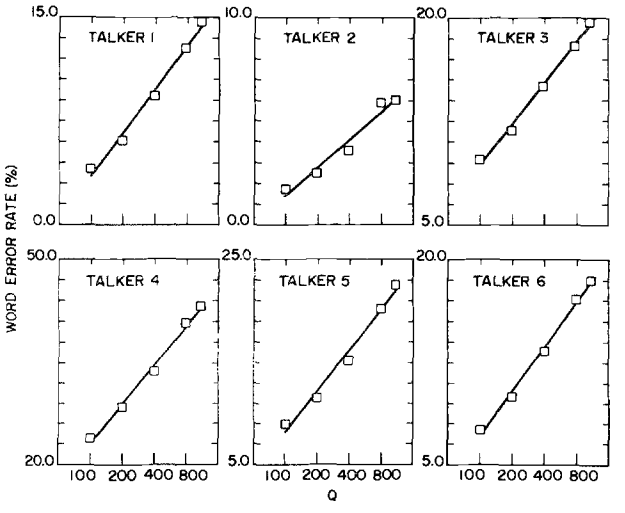

Figure 3 Average word error rate as a function of vocabulary size for each talker. The straight line is the least squares linear regression fit to the data.

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_l values greater than 1. For other talkers (e.g. Talker #4) the threshold is set very high and therefore most vocabulary words have q_l 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 3-7 in Section III) 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.

The average error rates for monosyllables versus polysyllables vividly drives home the point as to 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; hence it has 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 an 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% on the top candidate and 9.3% on the top 5 candidates was obtained. These scores represent the anticipated average performance of the recognizer across different talkers. The best talker achieved a 6.0% error rate on the first candidate, whereas the worst talker achieved a 43.3% 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*, Vol. 64, pp. 487-501, Apr. 1976.

[2] W. Lea, Ed., *Trends in Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1980.

[3] L. R. Rabiner and S. E. Levinson, "Isolated and Connected Word Recognition — Theory and Selected Applications," *IEEE Trans. on Communications* Vol. COM-29, No. 5, pp. 621-659, May 1981.

[4] C. K. Ogden, *Basic English: International Second Language*, Harcourt, Brace and World Inc., 1968.

[5] L. R. Rabiner and J. G. Wilpon, "A Simplified Robust Training Procedure for Speaker Trained, Isolated Word Recognition Systems," *J. Acoust. Soc. Am.*, Vol. 68, No. 5, pp. 1271-1276, Nov. 1980.

[6] A. R. Smith and L. D. Erman, "NOAH — A Bottom-Up Word Hypothesizer for Large Vocabulary Speech Understanding Systems," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. PAMI-3, No. 1, pp. 41-51, Jan. 1981.