Speaker-Independent Isolated Word Recognition for a Moderate Size (54 Word) Vocabulary

LAWRENCE R. RABINER, FELLOW, IEEE, AND JAY G. WILPON

Abstract—Recent work at Bell Laboratories has shown that statistical clustering techniques could be used to provide a reliable set of reference templates for a speaker-independent isolated-word recognition system. The vocabulary on which the system was tested consisted of the 26 letters of the alphabet, the 10 digits (0 to 9), and 3 command words. Since this vocabulary consisted of a large number of acoustically similar words (e.g., b, c, d, e, g, p, t, v, z), the recognition accuracy on the top candidate was only about 80 percent. In this paper results are presented using a considerably less difficult 54 word vocabulary of computer terms. Recognition accuracies from 95-98 percent were obtained across a wide variety of talkers. These results tend to support the hypothesis that carefully trained speaker-independent word recognizers can perform essentially as well as casually trained speaker-independent systems.

I. INTRODUCTION

ISOLATED word recognition systems may be either speaker-trained or speaker-independent. Word templates for speaker-trained systems are generally obtained via a sequential sampling technique, i.e., the talker (for whom the recognition system is trained) repeats each of the vocabulary words from 1 to 10 times in a single (or possibly two) short training session(s) [1]-[6]. For multiple repetitions of the vocabulary words, reference templates are obtained either by suitably averaging the replications of each word [1], [2], or by retaining each training word as a separate template [3]-[6]. For speaker-independent systems such training methods are generally not applicable because of the high interspeaker variability in individual words [7]-[10]. As such, a wide variety of statistical techniques have been developed either for characterizing the inherent acoustical properties of each vocabulary word [8], or for clustering multiple repetitions of each word (by different talkers) [9]-[13].

The clustering methods, because of their inherent independence of the vocabulary and the speech model, show great promise for providing a set of reference templates for speaker-independent, isolated word recognition. In an earlier study Rabiner, Levinson, Rosenberg, and Wilpon showed that a supervised statistical approach (a human interacting in the clustering loop) could cluster 100 repetitions of each word of a 39 word vocabulary into from 6 to 12 groups (clusters) which would essentially encompass from 90 to 98 percent of the replications [12]. The vocabulary used in this study consisted of the 26 letters of the alphabet (A to Z), the 10 digits (0 to 9), and 3 command words (STOP, ERROR, and REPEAT). This vocabulary was intended as a basis for the spoken, spelled name recognizer of Rosenberg and Schmidt [6]. Since this vocabulary, both trained and tested over conventional dialed-up telephone lines, was so difficult (i.e., many acoustically similar words such as b, d, p, t, v, z, e, etc.) the recognition accuracy for the first candidate was about 80 percent. The correct word was in the top 5 candidates about 98 percent of the time. These recognition accuracies were comparable to those obtained by Rosenberg and Schmidt for a speaker-trained system. Thus, the indications were that the clustering procedure was providing a fairly representative set of word templates.

The first improvement on the recognition system was to replace the supervised clustering procedure by a fully automatic one [13]. Experimentation with several automatic procedures showed that recognition accuracies (and cluster quality ratios) comparable to a supervised procedure could readily be obtained. It was also found that reference templates obtained from clusters by averaging the tokens in the cluster, rather than choosing the cluster minimax center, gave small but consistent improvements in recognition accuracy.

In order to provide a benchmark for how well the fully automatic recognition system worked, a new vocabulary was used to train and test the system. The vocabulary chosen was the 54 word vocabulary of computer terms originally proposed by Gold [8], and also used by Rabiner [7]. Fig. 1 shows the words in the vocabulary. It is seen that half the words are monosyllables, and 19 of the remaining 27 are 2 syllable words. Furthermore, some fairly close sounding words are included in the list, such as four, core and store, add and end, etc. As such, this vocabulary has a moderate degree of difficulty.

To train the system 100 talkers (50 male, 50 female) each recited the vocabulary of Fig. 1 once (in a random order) directly into a standard dialed-up telephone connection. A real-time analysis program (using the high-speed CSP MAP-200 array processor) recorded the features of each word for each talker and put them into a large word store for use by the automatic clustering package. On-line editing was used to eliminate spurious mouth clicks, pops, and breath noise, as well as any spurious sounds recorded off the telephone line itself. After training and clustering, the system was tested by 40 talkers, 10 of which were in the original training set of 100 talkers, 30 of which were not. Average recognition accuracies of close to 97 percent were obtained on 38 of the 40 talkers. The remaining two talkers (one foreign born, and one with a strong dialectal accent) had significantly poorer results.

Manuscript received April 10, 1979; revised June 27, 1979.

The authors are with the Acoustics Research Department, Bell Laboratories, Murray Hill, NJ 07974.

0096-3518/1200-0583$00.75

1. INSERT (2)	28. NAME (1)
2. DELETE (2)	29. END (1)
3. REPLACE (2)	30. SCALE (1)
4. MOVE (1)	31. CYCLE (2)
5. READ (1)	32. SKIP (1)
6. BINARY (3)	33. JUMP (1)
7. SAVE (1)	34. ADDRESS (2)
8. CORE (1)	35. OVERFLOW (3)
9. DIRECTIVE (3)	36. POINT (1)
10. LIST (1)	37. CONTROL (2)
11. LOAD (1)	38. REGISTER (3)
12. STORE (1)	39. WORD (1)
13. ADD (1)	40. EXCHANGE (2)
14. SUBTRACT (2)	41. INPUT (2)
15. ZERO (2)	42. OUTPUT (2)
16. ONE (1)	43. MAKE (1)
17. TWO (1)	44. INTERSECT (3)
18. THREE (1)	45. COMPARE (2)
19. FOUR (1)	46. ACCUMULATE (4)
20. FIVE (1)	47. MEMORY (2)
21. SIX (1)	48. BITE (1)
22. SEVEN (2)	49. QUARTER (2)
23. EIGHT (1)	50. HALF (1)
24. NINE (1)	51. WHOLE (1)
25. MULTIPLY (3)	52. UNITE (2)
26. DIVIDE (2)	53. DECIMAL (3)
27. NUMBER (2)	54. OCTAL (2)

Fig. 1. Words in the vocabulary. The number in parentheses is the number of syllables in the word.

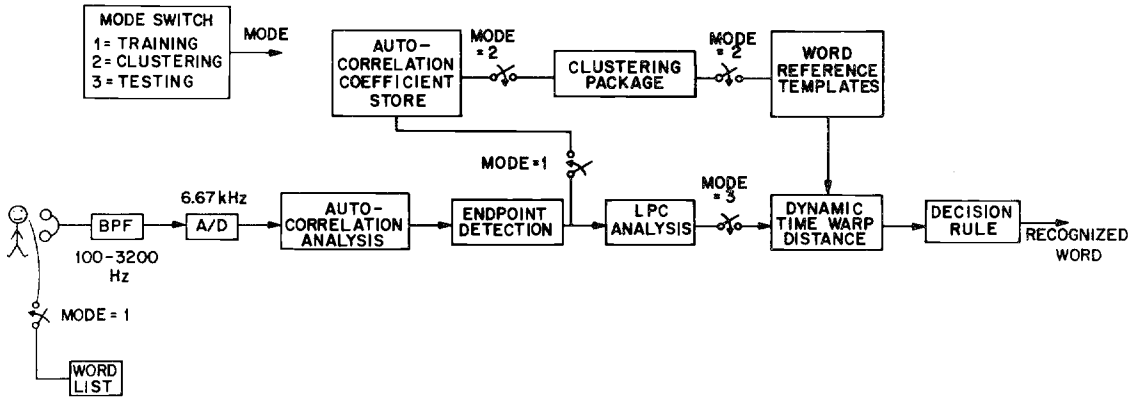

Fig. 2. Overall block diagram of the recognition system showing training (Mode 1), clustering (Mode 2), and testing (Mode 3).

In Section II of this paper a discussion of the experimental techniques is given in which we review briefly the recognition system and the training methods used. Section III presents the experimental results and an analysis of the errors, etc. Finally, in Section IV a discussion of the results and their relation to previous results is given.

II. Experimental Techniques

A block diagram of the word recognizer is shown in Fig. 2. There are 3 modes for the system. Mode 1 is a training mode in which talkers say each of the words of the vocabulary directly over a standard dialed-up telephone line, a $p = 8$ pole autocorrelation analysis is performed, and the beginning and end of each word is located. The frames of autocorrelation coefficients for the word are then stored for use later by the clustering package.

When all of the training data has been collected, the system enters Mode 2 in which the clustering package creates a set of word reference templates for each word in the vocabulary by grouping the replications of each word into a suitable set of clusters. The algorithm used for clustering was the UWA (unsupervised without averaging) method described in [13]. This method finds a set of clusters from the matrix of pairwise distances between the N tokens in the set by iteratively finding the minimax center of the current observation set (the token whose maximum distance to all remaining tokens is minimum), forming a cluster of all tokens within a prescribed distance of the minimax center, and checking for stability of the minimax center. The final cluster template is obtained as either the minimax center of the cluster, or an appropriately obtained average center [13].

The 100 replications of the 54 word vocabulary were clustered by the UWA algorithm. Table I shows the results of the clustering. Included in the table are 4 cluster statistics, namely:

1) the number of clusters per word, NC, where a cluster was a group containing at least 2 tokens;

2) the number of outliers (i.e., clusters with a single token) per word, NO. The number of outliers represents the number of tokens which do not fall into any of the NC clusters;

3) the size (in tokens) of the largest cluster SL; and

4) the quality ratio defined as the ratio of the average intercluster distance to the average intracluster distance ([11], [13]) σ.

The average minimum and maximum values of NC, NO, SL, and σ are given in Table I. The quantities σ_{MM} and σ_{AV} of Table I are the σ ratios for minimax centers and for average

TABLE I
STATISTICS OF THE CLUSTERS FOR THE 54 WORD VOCABULARY

	NC	NO	SL	σ_{MM}	σ_{AV}
Average	12	17	26	2.82	3.78
Minimum	4	5	13	2.49	3.00
Maximum	22	25	43	3.50	4.57

cluster centers. The statistics of the clusters for this vocabulary are comparable to the statistics of the 39 word vocabulary [13]. Furthermore, we see from Table I that the σ_{AV} values are significantly higher than the σ_{MM}, again indicating the improvements obtained from using the averaged center over the minimax center to represent the template.

In addition to the clustered templates, a template set obtained by randomly selecting tokens from the training data was used in testing the recognition system. Based on previous experience, the number of templates used (in all template sets) for each word in the vocabulary was 12; however, tests were run using from 1 to 12 templates per word. In the clustered case the templates were chosen based on the size of the clusters they represented; thus the largest clusters were represented before the smaller ones.

The final mode of the system of Fig. 3 (Mode 3) is the testing mode in which the unknown (test) word was compared to each of the word reference templates (using a dynamic time-warping alignment procedure), and a distance was assigned to the reference. The decision rule ordered the distance values according to an appropriate K-nearest neighbor decision rule [12], and gave an ordered list of candidates and their associated distances. In the next section recognition accuracy results are given for several test sets of data.

III. RECOGNITION RESULTS

A series of 4 recognition tests were used to evaluate the accuracy of the system. The test sets, denoted as *TS*1 to *TS*4, consisted of the following:

*TS*1—10 talkers (5 male, 5 female), each of which had been part of the original training set.

*TS*2—10 talkers (5 male, 5 female), none of whom had been part of the original training set. Talkers were all native American with no strong accent, i.e., of the same composition as the training set talkers.

*TS*3—20 talkers (10 male, 10 female), none of whom had been part of the original training set. No restrictions on talkers in this set.

*TS*3′—18 talkers (8 male, 10 female). This set is a subset of *TS*3 with 2 of the male talkers omitted. (This will be explained later.)

*TS*4—1 talker, originally part of training set; 6 replications of the entire 54 word vocabulary. All the recordings were made automatically, on-line, over dialed-up telephone lines. Each talker except the one in *TS*4 spoke the entire vocabulary one time.

As in the earlier investigations [12], [13], the recognition variables that were studied included:

1) *KNN* (K-nearest neighbor) decision rule. Values of *KNN* from 1 to 4 were used.

TABLE II(a)
RECOGNITION ACCURACIES USING AVERAGED CENTER TEMPLATES

KNN	TS1	TS2	TS3	TS3′	TS4
1	98	95.3	92.1	95.3	99.4
2	98.3	95.3	92.3	95.9	99.4
3	98.3	95.3	92.3	95.1	98.5
4	97.8	94.4	91.6	94.5	97.8

TABLE II(b)
RECOGNITION ACCURACIES USING RANDOM TEMPLATES

KNN	TS1	TS2	TS3	TS3′	TS4
1	94.4	93.4	86.6	89.7	98.1
2	95.3	93.8	88.7	92.0	96.6
3	96	93.6	89.3	92.4	96.9
4	95.6	93.4	88.6	91.9	96

TABLE II(c)
RECOGNITION ACCURACIES USING MINIMAX CENTER TEMPLATES

KNN	TS1	TS2	TS3	TS3′	TS4
1	95.1	91.6	89.4	93.0	97.8
2	97.5	94.0	90.6	93.8	98.5
3	97.5	93.2	91.2	94.2	97.8
4	97.3	93.2	89.9	93.1	97.2

2) L, number of templates per word. Values of L from 1 to 12 were used.

3) C, number of ordered candidates (from the decision rule output) that were considered. Values of C from 1 to 5 were used.

4) Set of templates. Templates from the clustered analysis both with averaged centers (*CLAV*) and with minimax centers (*CLMM*), as well as randomly chosen templates (*RAN*) were used.

A summary of the average recognition scores for each of the 4 test sets is given in Table II. Results are given for the top candidate ($C = 1$), using 12 templates per word ($L = 12$), for various values of *KNN* and for each of the 3 template sets. Results are also given for a test set *TS*3′ which is *TS*3 without 2 of the 20 talkers. The reason for this will be explained later.

Based on the average scores given in Table II, the following observations can be made:

1) Recognition accuracies using clustered templates with averaged centers are consistently higher than the results with minimax centers, and significantly (based on measured standard deviations of the data) better than results using random templates.

2) Differences in recognition accuracies using different *KNN* rules were fairly small. However, in general, slightly but consistently better recognition scores were obtained for $KNN = 2$ and/or 3 than for $KNN = 1$ and/or 4.

3) Recognition accuracies for talkers in the original training set (*TS*1 and *TS*4) were significantly higher than recognition accuracies for talkers not in the original training set.

4) Recognition accuracies for *TS*1 and *TS*4 were 98.3 percent and 99.4 percent, respectively, for $KNN = 2$ and clustered templates with average centers.

5) Recognition accuracies for *TS*2 and *TS*3 were 95.3 percent and 92.3 percent, respectively, for $KNN = 2$ and clustered templates with average centers.

Because of the significantly reduced recognition accuracies of the *TS*3 data, the individual scores of each of the talkers

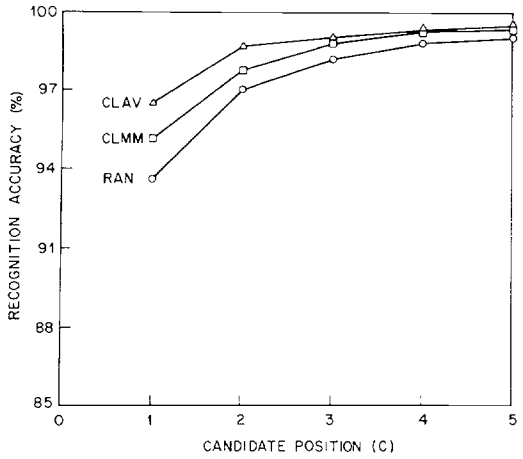

Fig. 3. Average word recognition accuracy as a function of the candidate position (C) for each of the template sets and for $KNN = 2$.

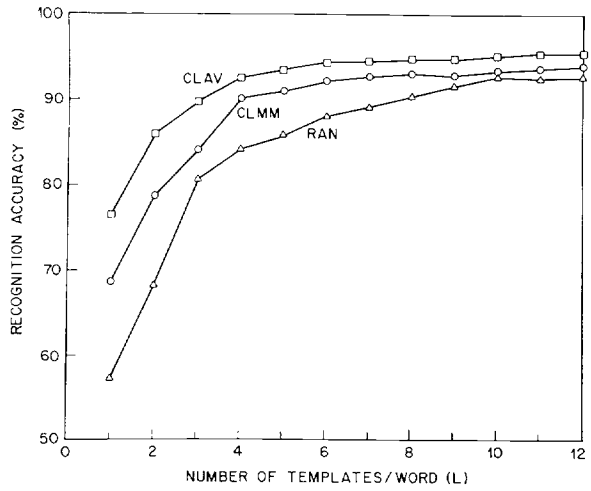

Fig. 4. Average word recognition accuracy as a function of the number of templates per word (L) for each of the template sets and for $KNN = 2$.

were studied to find the cause of the problem. It was found that 18 of the 20 talkers had recognition scores comparable to those of *TS*2. However, 2 talkers had scores of 50 percent and 79.6 percent. One of the talkers was both of Indian origin and a professional actor who was trained to overarticulate words to be best understood in a theatrical environment. This extreme emphasis on initial and final parts of each isolated word caused severe time alignment problems and caused the incorrect recognition of such words. The other talker with a low recognition had a strong, dialectal accent, again leading to severe problems in time alignment with the given template set. To show the effect of these two talkers on the recognition score, they were eliminated and the remaining 18 talkers' results are given as *TS*3′ in Table II. It is seen that an average recognition accuracy of 95.9 percent is obtained on the remaining 18 talkers, a score that is comparable to the score of the *TS*2 talkers.

To demonstrate the effect of ordered candidate position on the recognition score, Fig. 3 shows plots of average recognition accuracy (over all 4 test sets of data, including the 2 talkers with the poor scores) versus C for the three sets of templates using $KNN = 2$. This figure shows that the *CLAV* results were consistently better than the *CLMM* or *RAN* results for all values of C. The average recognition accuracy for $C = 1$ was 96.5 percent, and rose to a score of 99.5 percent for $C = 5$ for the *CLAV* templates. For this vocabulary, the $C = 1$ results are most significant; however, there are many applications in which information in the ordered list of candidates can be utilized to correct errors [5], [6]. For such applications, the $C = 2$ or $C = 5$ scores are often a better indicator of the overall performance of the system than the $C = 1$ score [6].

Fig. 4 demonstates the effect of the number of templates per word on the recognition score. Results are given here for the average recognition accuracy (over *TS*1, *TS*2 and *TS*3 data) versus L for the three sets of templates using $KNN = 2$ ($KNN = 1$ rule was used when $L = 1$). It can be seen that the recognition accuracies steadily increase as L increases starting at about 76.6 percent for $L = 1$ for the *CLAV* templates, and flattening off at about 95 percent for $L = 6$ and above. As such we see that about 6-8 templates per word would give essentially the same recognition accuracy (to within 1 percent) as 12 templates per word for the *CLAV* templates. For *RAN* templates, the accuracy steadily increases until $L = 10$, showing that we would need 10 to 12 templates per word for the best accuracy with the random templates.

IV. Discussion

The main purpose of this investigation was to verify that a recently proposed speaker-independent isolated word recognizer was capable of providing recognition accuracies comparable to such systems trained to the talker. In addition the goal was to verify the utility of a fully automatic clustering algorithm and the use of the K-nearest neighbor rule in the recognition system for a less difficult vocabulary than the one used initially to test the system.

The recognition results presented in the previous section lead to the following general conclusions:

1) The fully automatic clustering procedure with averaged center templates provided consistently high recognition accuracies for all test sets of data.

2) Templates obtained from clustering procedures gave consistently higher recognition scores than templates obtained randomly, i.e., clustering is an efficient method for determining the structure (similarities) among a group of tokens. This was especially the case when the number of templates per word was small (1 to 6).

3) The *KNN* rule with $KNN = 2$ or 3 gave higher recognition scores than the $KNN = 1$ or 4 rules for clustered templates.

The above conclusions essentially provide a strong confirmation that the methods and procedures described previously are applicable to almost any set of vocabulary words.

In regard to the question of how the absolute recognition accuracies compared to those obtained in earlier studies [7], [8] it is seen that the average accuracy of 95 percent or higher for each of the test sets (*TS*3′ instead of *TS*3) is considerably higher than the score of 86 percent obtained by Gold using high-quality input speech, or 85 percent obtained by Rabiner using the same recognizer but considerably less training data and a different decision rule. As such this system represents

a substantial improvement over alternative recognizers with the same vocabulary.

In comparison to word recognizers trained to the talker, recognition accuracies over 98 percent have been reported [1]-[3], [14], but with considerably different vocabularies. However, even accounting for vocabulary differences, it seems clear that systems that are carefully trained to the individual talker should and will perform more accurately and in a more robust manner than systems which are speaker-independent. The robustness comes in when we realize that there will be talkers, such as the 2 in *TS*3 of our data, who are poorly if at all represented in the training set. For such talkers, as seen here, the recognition process can and sometimes will break down entirely. Thus, even though we have made a large step to bridge the gap between systems that are trained to the talker and those that are talker-independent, there still remain some real obstacles to a universal word recognizer.

V. Summary

We have presented results on the recognition of a 54 word vocabulary of computer terms using a fully automatic clustering technique to obtain speaker-independent work templates. Recognition accuracies of about 95 percent have been obtained for different sets of talkers. These results show considerable improvement over earlier speaker-independent recognizers using the same vocabulary.

References

[1] T. B. Martin, "Practical applications of voice input to machines," *Proc. IEEE*, vol. 64, pp. 487–501, Apr. 1976.

[2] M. B. Herscher and R. B. Cox, "An adaptive isolated-word speech recognition system," in *Conf. Rec., 1972 Conf. Speech Commun. and Processing*, pp. 89–92, Newton, MA, April 1972.

[3] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-23, pp. 67–72, Feb. 1975.

[4] A. E. Rosenberg and F. Itakura, "Evaluation of an automatic word recognition system over dialed-up telephone lines," *J. Acoust. Soc. Amer.*, vol. 60, suppl. 1, p. S12 (abstract), Nov. 1976.

[5] S. E. Levinson, A. E. Rosenberg, and J. L. Flanagan, "Evaluation of a word recognition system using syntax analysis," *Bell Syst. Tech. J.*, vol. 57, pp. 1619–1626, May–June 1978.

[6] A. E. Rosenberg and C. E. Schmidt, "Recognition of spoken spelled names applied to directory assistance," *J. Acoust. Soc. Amer.*, vol. 62, suppl. 1, p. S63, Dec. 1977.

[7] L. R. Rabiner, "On creating reference templates for speaker independent recognition of words," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 34–42, Feb. 1978.

[8] B. Gold, "Word recognition computer program," Massachusetts Inst. Tech., Cambridge, RLE Tech. Rep. 452, June 1966.

[9] P. B. Scott, "VICI–A speaker independent word recognition system," in *Conf. Rec., 1976 IEEE Int. Conf. Acoust., Speech, Signal Processing*, Philadelphia, PA, pp. 210–213, Apr. 1976.

[10] V. N. Gupta, J. K. Bryan, and J. N. Gowdy, "A speaker-independent speech recognition system based on linear prediction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 27–33, Feb. 1978.

[11] S. E. Levinson, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "Interactive clustering techniques for selecting speaker-independent reference templates for isolated word recognition," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. ASSP-27, pp. 134–141, Apr. 1979.

[12] —, "Speaker-independent recognition of isolated words using clustering techniques," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 336–349, Aug. 1979.

[13] L. R. Rabiner and J. G. Wilpon, "Considerations in applying clustering techniques to speaker-independent word recognition," *J. Acoust. Soc. Amer.*, 1979.

[14] S. Tsuruta, "DP-100 voice recognition system achieves high efficiency," *JEE*, pp. 50–54, July 1978.