THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 58, No. 10, December 1979
Printed in U.S.A.

Application of Clustering Techniques to Speaker-Trained Isolated Word Recognition

By L. R. RABINER and J. G. WILPON

(Manuscript received July 25, 1979)

Speaker-trained, isolated word recognizers have achieved notable success in a wide variety of applications. The training for such systems generally involves a single (or sometimes two) replication(s) of each word of the vocabulary by the designated talker. Word reference templates are then formed directly from these replications. In recent work on speaker-independent word recognition, it has been shown that statistical clustering procedures provided an effective way for determining the structure in multiple replications of a word by different talkers. Such techniques were then used to provide a set of reference templates based on the clustering results. In this paper we discuss the application of clustering techniques to speaker-trained word recognizers. It is shown that significant improvements in recognition accuracy are obtained when using templates obtained from a clustering analysis of multiple replications of a word by the designated talker. It is also shown that recognition accuracy did not change with time (over a 6-month period) for any of the subjects tested, thereby indicating that the reference templates were reasonably stable.

I. INTRODUCTION

Although a great deal has been learned about isolated word speech recognition systems,[1-14] several key issues are not as well understood as others. One such issue is the manner in which the word reference templates for such a system are obtained. To date, there have been at least three distinct ways of obtaining templates, including:

(*i*) Casual training in which the designated talker (for a speaker-trained system) speaks each word of the vocabulary (one or more times) and a reference template is created for each spoken word.[3,4] Thus, for casual training, there is a direct correspondence between a spoken token of the word and the reference template.

(*ii*) Averaging methods in which the designated talker (for a speaker-trained system) or a set of talkers (for a speaker-independent system) speaks the word a number of times and a weighted, time-normalized average of the feature sets for that word is used as the reference template.[1,7,15]

(*iii*) Statistical clustering methods in which a set of talkers speak the word and a statistical pattern recognition algorithm is used to group the feature sets of the tokens into a set of clusters.[14,16] The similarity of tokens within a cluster is high (small intratoken distances), whereas the similarity of tokens in different clusters is low (large intertoken distances). Reference templates are obtained by representing each cluster by a single template (either using a minimax approach,[14] or via averaging techniques[17]). Thus, a word is generally represented by a *set* of templates rather than one or two templates.

The third method above, the statistical approach, has been successfully applied to a speaker-independent word recognizer for a variety of vocabularies.[14,17,18] It is the purpose of this paper to show how this technique can be applied in a speaker-trained system to further increase their accuracy and robustness over systems in which the reference templates are obtained by casual training.

The organization of this paper is as follows. In Section II we review the operation of the basic word recognizer and the clustering procedures. In Section III we present the experimental procedures used to obtain the data for training and testing the system. The statistics of the clustering for each of three talkers are presented in Section IV, and the recognition accuracy as a function of key system parameters is given in Section V. Finally, Section VI discusses the results and their implications for practical implementations of word recognition systems.

II. REVIEW OF THE WORD RECOGNITION SYSTEM

The word recognizer, shown in Fig. 1, is similar to the one originally proposed by Itakura,[3] and has been used in a variety of applications.[4,13,14,16–18] Telephone line input signals (100- to 3200-Hz bandwidth) are digitized at a 6.67-kHz rate, and a $p = 8$th-order autocorrelation analysis is performed on overlapping frames of $N = 300$ samples (45 ms), with an overlap of 200 samples between frames. Prior to the autocorrelation analysis, each frame of data is preemphasized with a first-order digital network with transfer function $(1 - 0.96\ z^{-1})$ and windowed by a 300-sample Hamming window. If we denote the lth preemphasized, windowed frame of speech as $\hat{x}_l(n)$, $0 \le n \le N-1$, then

$$\hat{x}_l(n) = \hat{x}(l \cdot S + n) \cdot w(n) \qquad 0 \le n \le N-1, \quad 0 \le l \le L-1, \quad (1)$$

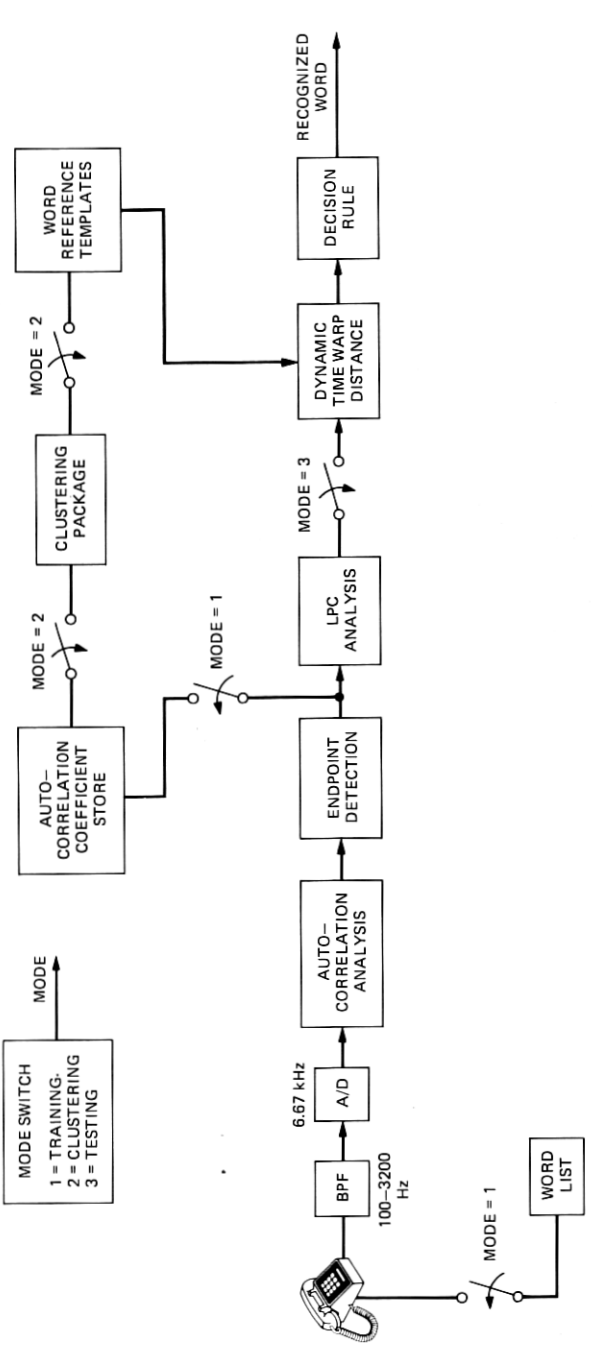

Fig. 1—Block diagram of the word recognizer.

where $\hat{x}(n)$ is the preemphasized speech, $w(n)$ is a Hamming window, $S = 100$ samples (15 ms) is the shift in samples between adjacent frames, and L is the number of frames in the recording interval. The autocorrelation coefficients of the lth frame, $R_l(m)$ are given by

$$R_l(m) = \sum_{n=0}^{N-l} \hat{x}_l(n)\hat{x}_l(n+m) \quad 0 \le m \le p \tag{2}$$

$$= \sum_{n=0}^{N-1} \hat{x}(lS+n)\hat{x}(lS+n+m). \tag{3}$$

The zeroth autocorrelation coefficient of each frame ($R_l(0)$) is the energy in the frame. The time pattern of $R_l(0)$ (i.e., $R_l(0)$ vs l) is used to determine the end-point boundaries of the spoken, isolated word in a simple manner based on the measured energy of the background noise in the recording environment.[14]

As noted in Fig. 1, there are three modes of operation of the word recognizer, namely, training, clustering, and testing (normal usage). As discussed earlier, for many speaker-trained systems, the training mode is simply a recording of each word of the vocabulary and the clustering (i.e., formation of word reference templates) is a direct conversion from stored autocorrelation coefficients to the format required for the word reference template. (For this recognizer, the word reference templates are stored as frames of autocorrelated linear predictive coding (LPC) coefficients. This is explained later in this section.) For a statistical clustering approach, however, training consists of storing sets of autocorrelation coefficients for a number of replications of each word of the vocabulary, and clustering consists of grouping the replications of each word into clusters and creating a single word reference template for each cluster.

For the third mode of the system, namely the testing or normal usage mode, following end-point detection, an LPC analysis of each frame is performed (the autocorrelation method), and each autocorrelation frame is converted to a normalized form as follows. If we denote the frames of autocorrelations in the test word as $R_i(m)$, $i = 1, 2, \cdots, NT$, and the LPC prediction residual of each frame as E_i, then the test frame parameters are given as

$$V_i(m) = \frac{R_i(m)}{E_i} \qquad 0 \le m \le p, \quad 1 \le i \le NT. \tag{4}$$

For the word reference templates if we denote the jth frame of LPC coefficients (derived from the autocorrelation coefficients) as $a_j(k)$, $0 \le k \le p$, then the jth reference frame parameter set is given as

$$P_j(m) = 2 \sum_{k=0}^{p} a_j(k)a_j(k+m) \qquad 1 \le m \le p \tag{5a}$$

$$= \sum_{k=0}^{p} [a_j(k)]^2 \qquad m = 0. \tag{5b}$$

A distance can now be defined between the ith test frame and the jth reference frame as [3,19]

$$d(i, j) = d(V_i, P_j) = \log \left[\sum_{m=0}^{p} V_i(m) P_j(m) \right]. \tag{6}$$

The distance measure of eq. (6) has been shown to be an effective measure for comparing sets of LPC coefficients in a variety of applications,[3,19–22] and it can be computed with $(p + 1)$ multiplications and additions and one logarithm.

The next step in the recognizer is to compare the test word against each stored word reference template. A dynamic time-warping algorithm is used to optimally align in time the test and reference patterns and to give the average distance associated with the optimal warping path. The average distance for the qth template of the rth reference word is

$$\bar{D}_{r,q} = \frac{1}{NT} \left[\min_{w_{r,q}(i)} \sum_{i=1}^{NT} d(i, w_{r,q}(i)) \right], \tag{7}$$

where $w_{r,q}(i)$ is the optimally determined warping path. The final step in the process is to choose the recognized word based on the set of average distances $\bar{D}_{r,q}$. The most common decision rule is the minimum distance rule which chooses the word r^* such that

$$\bar{D}_{r^*,q^*} \leq \bar{D}_{r,q} \quad \text{all} \quad r, q \tag{8}$$

for some value q^*. An alternative and more powerful decision rule (for the case of multiple reference templates) is the K-nearest neighbor rule (KNN), which says that for each word r, the distances $\bar{D}_{r,q}$ are reordered according to average distance so that

$$\bar{D}_{r,[1]} \leq \bar{D}_{r,[2]} \leq \cdots \leq \bar{D}_{r,[Q]}, \tag{9}$$

where Q is the number of templates for the rth word, and the KNN rule says to choose the word r^* such that

$$\sum_{k=1}^{K} \bar{D}_{r^*,[k]} \leq \sum_{k=1}^{K} \bar{D}_{r,[k]} \quad \text{all} \quad r. \tag{10}$$

For $K = 1$ the KNN rule is the minimum distance rule. Unless otherwise noted, a value of $K = 2$ was used in the recognition tests in this paper.

2.1 The clustering procedure

The clustering analysis is based on the fully automatic technique (unsupervised with averaging—UWA) described in Ref. 17. It was

assumed that we begin with M replications of each word in the vocabulary and, based on the pairwise dynamic time-warped average distance between words, the M tokens are grouped into P disjoint clusters, ω_i, such that

$$\Omega = [t_1, t_2, \cdots, t_M] = \bigcup_{i=1}^{P} \omega_i, \tag{11}$$

where $t_1, t_2, \cdots, t_M$ are the M tokens in the set. The total number of clusters, P, is determined automatically by the clustering procedure. Each cluster, ω_i, is represented by a prototype $\hat{x}_i$. Based on the work of Rabiner and Wilpon,[17] the tokens within cluster ω_i are averaged (using dynamic time warping for time alignment) to give the prototype $\hat{x}_i$. Word reference templates are determined as the prototypes $\hat{x}_i$ corresponding to the $\hat{P}$ largest clusters, i.e., for a single template we choose the prototype of the cluster with the most tokens; for a two-template representation, we choose the prototypes of the two clusters with the largest number of tokens, etc.

The grouping of the M tokens into P clusters is based on splitting of the set Ω by iteratively determining cluster centers (based on a minimax criterion) and cluster points based on a given distance threshold. Ultimately, all M tokens are assigned to one of the clusters. A cluster may consist of a single outlier token whose distance to all other tokens in Ω is greater than the distance threshold of the procedure. The final set of P clusters is ordered based on size of the clusters, and the averaged centers of the $\hat{P}$ largest clusters are retained as the $\hat{P}$ word reference templates.

III. EXPERIMENTAL PROCEDURES

To test the effectiveness of the clustering analysis for a speaker-trained system, three talkers trained the recognizer of Fig. 1. One of the three talkers was the first author of this paper. The other two talkers were experienced workers in the area of speech processing. All three talkers were instructed to speak the words naturally, but in an isolated format. No specific motivation for good performance was employed, as the talkers' interest in the area was considered sufficient. The vocabulary for these tests consisted of the letters A to Z, the digits 0 (zero) to 9, and the command words STOP, ERROR, and REPEAT for a total of 39 words. This vocabulary is an extremely difficult one (especially when recorded over telephone lines, as was done here), but one which has utility in a wide range of practical applications.[23]

Each talker spoke the 39-word vocabulary (in a random order) three times per session over a one-month period for a total of 50 replications for each word in the vocabulary. A total of 17 sessions was used, with only two recordings in the last session.

A clustering analysis was performed for each talker, and a set of word reference templates was obtained. For speaker-independent systems, a total of $\hat{Q} = 12$ reference templates per word was used. For comparison purposes, the same number of templates was obtained for the speaker-trained vocabulary. However, results are also given for a variable number of templates per word.

To test the system, each of the three talkers spoke the 39-word vocabulary five times per session for a total of 10 sessions. Each session was at least two weeks after the preceding one; thus, a total of at least 20 weeks was used to obtain the 10 test sets.

Additional analyses were performed to show the effects of reduced training on the recognition accuracy. To do this, we simply used fewer training runs in the clustering analysis. As such, results are presented for cluster sets based on 24, 12, and 6 replications of the word vocabulary during the training phase.

IV. CLUSTER STATISTICS

Based on the clustering analysis, a set of objective statistics on the clusters can be given which indicates how the tokens cluster. In accordance with past experience with these clustering algorithms, the following statistics appear to be most meaningful:

(*i*) Number of clusters per word. A cluster is defined as a set with at least two tokens.

(*ii*) Number of outliers per word. An outlier is a token that does not fall into one of the clusters above, i.e., its distance to all other tokens in the training set exceeded a threshold.

(*iii*) Quality ratio, σ, defined as the ratio of the average intercluster distance (as defined between cluster prototypes) to the average intracluster distance (as defined between cluster tokens).

(*iv*) Size of largest cluster—i.e., the number of tokens in the largest set.

This set of cluster statistics gives an excellent picture of how the M tokens are distributed in the feature space of the problem being studied.

Table I gives the statistics of the word clusters for the three talkers used in this investigation. Included in the table are averaged, minimum, and maximum values of the cluster statistics for each of the three talkers. The statistics in Table I were obtained from clustering the 50 replications of each word for each talker. It is seen that the average values of all statistics are about the same for all three talkers. Typically, about 6 clusters per word were sufficient to include all nonoutlier tokens. Included in the six clusters were, on the average, 38 of the 50 tokens, with about 20 of the 50 tokens in the biggest cluster. The

Table I—Statistics of the word clusters for the three talkers

	Subjects								
	LRR			AER			SWC		
	Avg	Min	Max	Avg	Min	Max	Avg	Min	Max
Number of clusters per word	6	1	10	6	3	10	6	3	10
Number of outliers per word	13	3	17	12	5	17	11	5	18
Quality ratio (σ)	2.90	1.99	4.06	2.92	2.52	3.86	2.68	2.12	3.41
Size of largest cluster	20	9	46	20	7	35	21	9	31

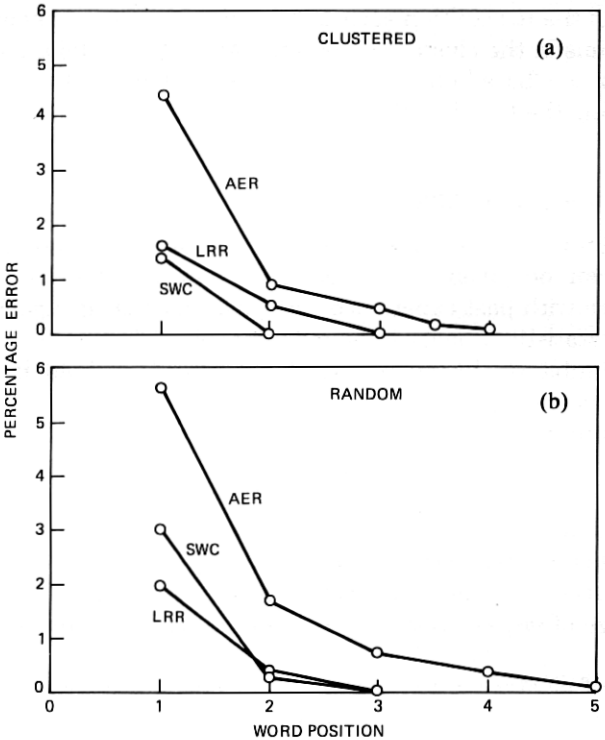

Fig. 2—Recognition error as a function of word position for the three talkers for (a) clustered templates and (b) randomly chosen templates.

quality ratios of between 2.6 and 2.9 indicate good cluster separation for each of the talkers.[16]

V. RECOGNITION RESULTS

Recognition results on the total of 1950 words (50 replications of the 39-word vocabulary) for each of the three talkers (LRR, AER were male, SWC was female) are presented in Figs. 2 and 3. Figure 2a shows a

series of plots of the percentage errors as a function of word position for the three talkers for reference templates obtained from the clustering analysis. Word error rate for the kth word position is the percentage of words which were not within the top k candidates. A total of 12 templates per word was used in these tests. Overall error rates of 1.4

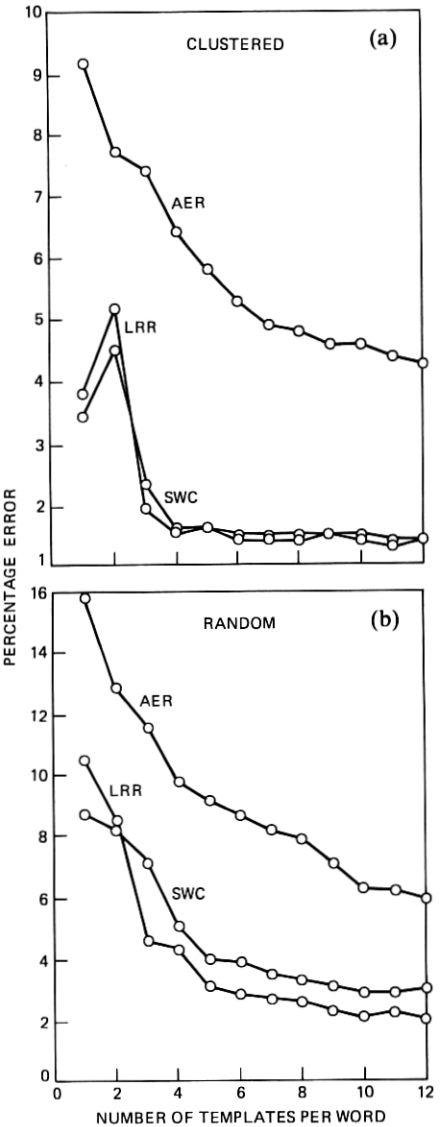

Fig. 3—Recognition error as a function of the number of templates per word (top choice candidate) for the three talkers for (a) clustered templates and (b) randomly chosen templates.

percent (SWC), 1.6 percent (LRR), and 4.4 percent (AER) were obtained for the three talkers for the top recognition candidate (i.e., the first-choice candidate). The error rate was below 1 percent for the top two candidates (word position 2) for all three talkers.

For comparison purposes, a set of word templates was created by randomly choosing tokens from the 50 replications of each word and creating one reference template directly from each token. Again, a total of 12 templates per word was used. Figure 2b shows the error scores for the random set of word templates for the three talkers. Overall error rates of 3 percent (SWC), 2 percent (LRR), and 5.6 percent (AER) were obtained for the top recognition candidate. Although these error rates were somewhat higher than the scores obtained from the clustered template set, the differences are reasonably small and indicate that the clustering analysis is unnecessary if we are using 12 templates per word. In such a case, a random selection of word templates is essentially equivalent.

It is shown in Fig. 3, however, that the results given in Fig. 2 are not a complete picture of the effectiveness of the clustering analysis. Figure 3a shows plots of percentage error for the top recognition candidate as a function of the number of templates per word for the clustered template set, and Fig. 3b shows similar plots for randomly chosen templates. For talkers LRR and SWC, it is readily seen that the error rate does not change for more than four templates per word for the clustered data. For talker AER, the error rate decreases by about 0.6 percent as the number of templates per word increases from 6 to 12. Thus, Fig. 3a indicates that from 4 to 6 templates per word obtained via clustering give comparable recognition accuracies to 12 templates per word obtained in the same manner.

A totally different picture emerges from Fig. 3b for the case of randomly chosen templates. (Note that the vertical scale of Fig. 3b is different from the vertical scale of Fig. 3a.) It is seen that the percentage error decreases steadily as the number of (random) templates per word increases until about 10 templates per word. Thus for randomly chosen templates a substantially larger number of templates per word are required than for templates obtained from a clustering analysis. An alternative way of stating this is that recognition accuracies from 3 to 4 templates per word (obtained from the clustering analysis) are comparable to those obtained from 10 to 12 randomly chosen templates per word.

5.1 Confusions among the words

The spoken letters of the alphabet form one of the most difficult of word recognition vocabularies because of the high confusability among sets of the letters.[14,23] A major advantage of the clustering analysis is

that confusions among many of the subsets are entirely eliminated. The major confusions, for all three talkers, were in the equivalence class of letters containing B, C, D, E, G, P, T, V, and Z. The confusion matrices for this class for the three talkers (for 12 templates per word) are shown in Table II. For talker SWC, 26 of the 27 errors were within this equivalence set; for talker LRR all 31 errors occurred within the equivalence set; for talker AER, 72 of the 85 recognition errors occurred within the equivalence class—however, one confusion was with a word outside the set. (Nine of the remaining errors were A, K confusions.)

Table II shows that each talker had one or more letters in the major equivalence class which were hard to reliably recognize; however, for all three talkers most letters in the hard equivalence class were correctly recognized. This result again demonstrates the power of the clustering analysis in determining the structure of each word in the vocabulary.

5.2 Recognition accuracy vs time

An important aspect of a speaker-trained word recognizer is the stability of the reference templates as a function of time. For casually

Table II—Confusion matrices of the equivalence class with B, C, D, E, G, P, T, V, Z for the three talkers

		Recognized Word										
		B	C	D	E	G	P	T	V	Z	Other	
Spoken Word	B	46		1			1	1			1	LRR
	C		50									
	D	4		35			4	4	3			
	E				50							
	G					50						
	P						44	6				
	T						5	45				
	V	1							49			
	Z									50		
Spoken Word	B	36		3			6	4	1			AER
	C		48							2		
	D			31				17	2			
	E	3			45		2					
	G					50						
	P						39	10			1	
	T			3				46			1	
	V	1	1				1		47			
	Z		12						1	36	1	
Spoken Word	B	47		2	1							SWC
	C		50									
	D	1		48	1							
	E			1	49							
	G					48		2				
	P						49	1				
	T					3		47				
	V								50			
	Z		4						10	36		

trained, nonadaptive systems, the reference templates often degrade with time and the system must be retrained.[1] Since training is so simple for these systems, this generally does not pose a problem. However, some mechanism must be provided for detecting the degradation of the reference templates and retraining the system.

For a clustering analysis method of obtaining reference templates, it is imperative that the templates be robust in time, i.e., that no degradation in recognition accuracy occurs, since the training procedure is a long and involved one. To demonstrate that the reference templates from this system are indeed robust, Fig. 4 shows plots of the error rate vs time for each of the three talkers. It is seen that over the 20-week period of testing, only small changes occur in the recognition accuracy.

5.3 The effects of reduced training

Since the amount of training used to obtain the accuracies reported here was quite extensive (50 repetitions of each word), the effects of reduced training on the recognition scores are important to understand. Thus, the clustering analysis was redone using subsets of the 50 replication training data. The subsets included the first 24, 12, and 6 replications. Before discussing the results, two points should be noted. Each recording session consisted of three consecutive replications of the word list. Thus the three subsets constitute eight, four, and two recording sessions. This is important since it was found that a high degree of correlation existed between tokens within a given recording session.

The second point of note is that, for the 12 replications training set, the maximum number of clusters was limited to six (including outliers), and for the six replication training set the maximum number of clusters was limited to four. The reason for this limitation is that more than six (or four) meaningful clusters cannot be obtained from the reduced

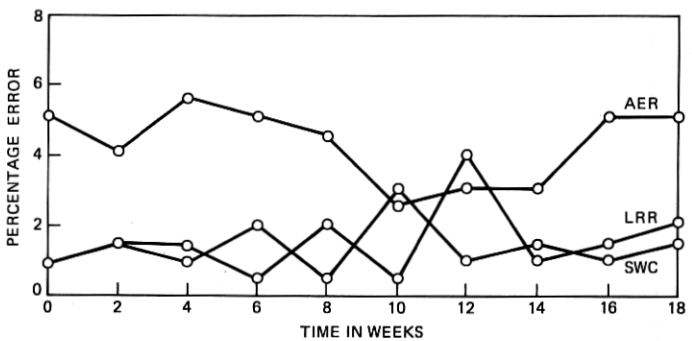

Fig. 4—Recognition error as a function of time for the three talkers.

number of tokens. For recognition purposes, the KNN = 1 rule was used for the 12 and 6 replication template sets.

The recognition results for the reduced training sets are shown in Figs. 5 and 6. Figure 5 shows plots of percentage error as a function of word position for each training set and for each talker. Figure 6 shows plots of percentage error vs the number of templates per word for

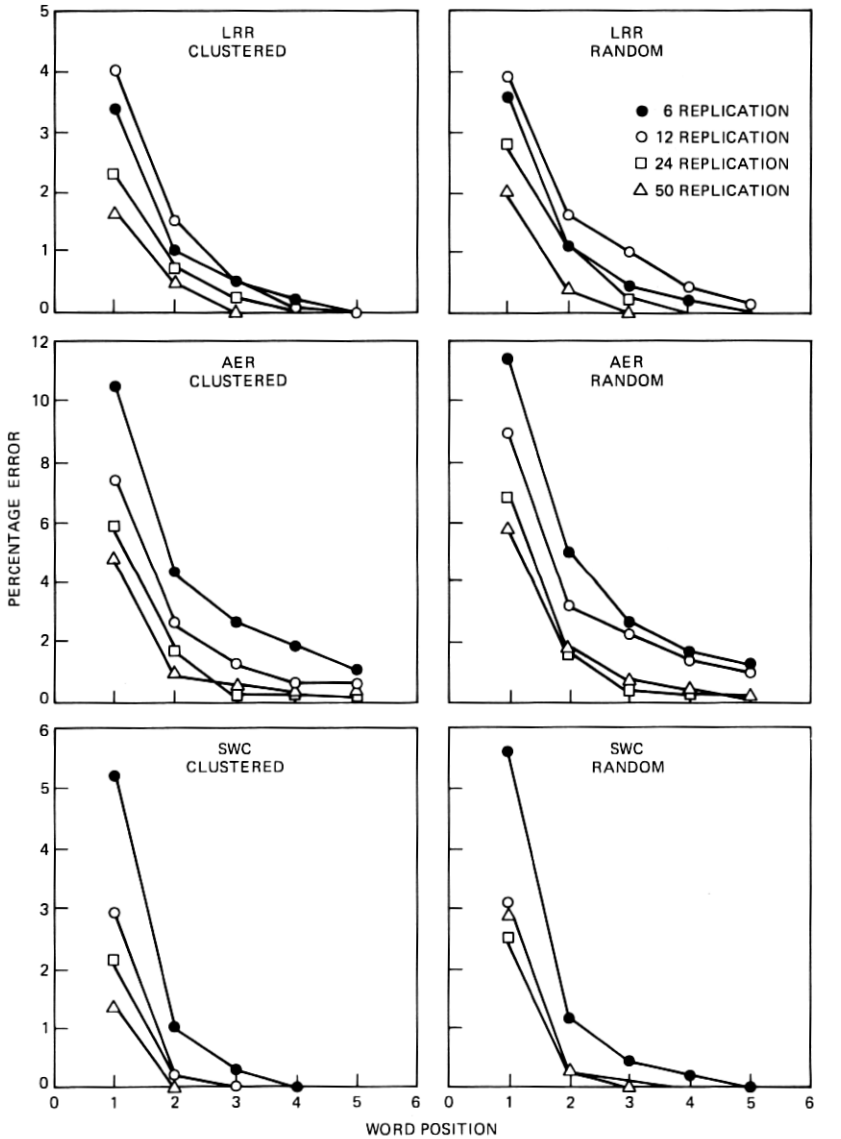

Fig. 5—Recognition error as a function of the word position for different numbers of training sets for the three talkers for both clustered and random templates.

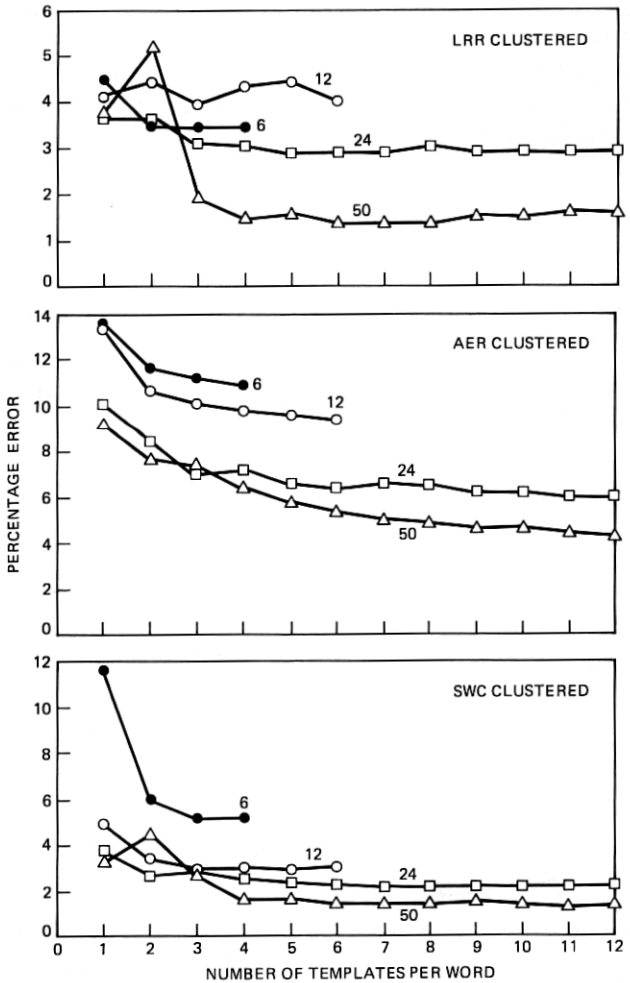

Fig. 6—Recognition error as a function of the number of templates per word and the number of training sets for the three talkers for the clustered templates.

these cases. It is seen that, in all cases, the reduced training set leads to increased error in recognition. In reducing the size of the original training set (from 50 to 24 replications), the error increased about 1.5 percent, on the average, for the three talkers. In going from 50 to 12 replications for training, the error increased by about 2.5 percent for the three talkers, and in going from 50 to 6 replications, the error increased by about 3.3 percent.

The results given above indicate that increased training always gave better templates from the clustering analysis and reduced the recognition error rate.

5.4 Comparisons to casually trained systems

The recognition system of Fig. 1 was casually trained to each of the three talkers by having them speak the vocabulary twice and creating reference templates directly from the spoken words. The recognition tests were then rerun using the casually obtained word templates. The average error rates (over the 50 replications) for the three talkers were 6.5 percent for talker LRR, 12.9 percent for talker AER, and 12.5 percent for talker SWC. Since the overall error rates for the clustered data were 1.6, 4.3, and 1.4 percent for these talkers, respectively, reductions in error rate of 3.9, 8.6, and 11.1 percent were obtained. These error reductions represent a substantial improvement in the recognition.

VI. DISCUSSION

The main result of this paper is the demonstration that statistical clustering techniques can be applied equally well to speaker-trained word recognizers as they have been to speaker-independent ones. It was shown that, with sufficient training and through the use of well-developed clustering algorithms, extremely high recognition scores can be obtained, even with vocabularies as difficult as the letters of the alphabet. This result indicates that, if a user is sufficiently motivated to spend the time necessary to train a word recognizer, he can reliably use the recognizer with a range of vocabularies in a wide variety of applications.[1,23]

An important consideration in a practical implementation of a system like the one described in this paper is to keep the number of reference templates as small as possible. It was shown that about 4 to 6 templates per word were sufficient for the given vocabulary. It is anticipated that for alternative, less complex vocabularies even fewer templates per word would be required. The templates themselves appear to be stable with time as the recognition scores did not change appreciably through the 20 weeks of testing.

One point in question about this work is that only three (experienced) talkers were used. We can only speculate on what the results would be for a larger set of talkers. It is believed that the clustering approach would be highly effective for any talker. (It should be especially good for an inexperienced one who has a lot of replication-to-replication variability in the way he says the words.) As such, the conjecture is that even larger improvements in recognition accuracy over casual training would be obtained by using this system for a wide range of talkers.

Finally, it was shown that the clustering analysis could be bypassed with sufficient training, if a large number of randomly chosen templates (10 to 12) were used to represent each word in the vocabulary. If

computational complexity was not an issue, this result could be useful for some applications.

VII. SUMMARY

We have shown that statistical clustering techniques can be applied to a speaker-trained, isolated word recognition system to provide significant improvements in recognition accuracy over casually trained systems. The amount of training required for such a system is fairly extensive. Thus, this method would probably be limited to applications requiring extremely difficult vocabularies (e.g., the letters of the alphabet), or those in which very high recognition accuracies are required.

VIII. ACKNOWLEDGMENTS

The authors acknowledge the diligence and time of Susan Christiansson (SWC) and Aaron Rosenberg (AER) of Bell Laboratories for participating as subjects in an experiment that lasted the better part of six months.

REFERENCES

1. T. B. Martin, "Practical Applications of Voice Input to Machines," Proc. IEEE, *64*, No. 4 (April 1976), pp. 487–501.
2. P. Vicens, "Aspects of Speech Recognition by Computer," Ph.D. Thesis, Stanford Univ., April 1969.
3. F. Itakura, "Minimum Prediction Residual Applied to Speech Recognition," IEEE Trans. Acoustics, Speech, and Signal Proc., *ASSP-23*, No. 1 (February 1975), pp. 67–72.
4. A. E. Rosenberg and F. Itakura, "Evaluation of an Automatic Word Recognition System Over Dialed-Up Telephone Lines," J. Acoust. Soc. Am., *60*, Supplement No. 1 (November 1976), p. S12 (Abstract).
5. S. R. Hyde, "Automatic Speech Recognition: A Critical Survey and Discussion of the Literature," in *Human Communication: A Unified View*, E. E. David, Jr., and P. B. Denes, eds., New York: McGraw-Hill, 1972, pp. 399–438.
6. G. M. White and R. B. Neely, "Speech Recognition Experiments With Linear Prediction, Bandpass Filtering, and Dynamic Programming," IEEE Trans. Acoustics, Speech, and Signal Proc., *ASSP-24*, No. 2 (April 1976), pp. 183–188.
7. M. B. Herscher and R. B. Cox, "An Adaptive Isolated-Word Speech Recognition System," Conf. Rec. 1972 Conf. Speech Comm. and Proc., *AD-742236*, (April 1972), pp. 89–92.
8. B. Gold, "Word Recognition Computer Program," MIT Res. Lab of Electronics, Tech. Report 452, June 1966.
9. J. M. Shearme and P. F. Leach, "Some Experiments With a Simple Word Recognition System," IEEE Trans. Audio and Electroacoustics, *AU-16*, No. 2 (June 1968), pp. 256–261.
10. V. M. Velichiko and N. G. Zagoruiko, "Automatic Recognition of 200 Words," Int. J. Man-Machine Studies, *2* (1970), p. 23.
11. C. F. Teacher, H. G. Kellet, and L. R. Rocht, "Experimental, Limited Vocabulary Speech Recognizer," IEEE Trans. Audio and Electroacoustics, *AU-15*, No. 3 (September 1967), pp. 127–130.
12. A. Ichikawa, Y. Nakamo, and K. Nakata, "Evaluation of Various Parameter Sets in Spoken Digits Recognition," IEEE Trans. Audio and Electroacoustics, *AU-21*, No. 3 (June 1973), pp. 202–209.
13. L. R. Rabiner, "On Creating Reference Templates for Speaker Independent Recognition of Isolated Words," IEEE Trans. Acoustics, Speech, and Signal Proc., *ASSP-26*, No. 1 (February 1978), pp. 34–42.
14. L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon, "Speaker Independent Recognition of Isolated Words Using Clustering Techniques," IEEE

Trans. Acoustics, Speech, and Signal Proc., *ASSP-27*, No. 4 (August 1979), pp. 336–349.
15. M. R. Sambur and L. R. Rabiner, "A Statistical Decision Approach to the Recognition of Connected Digits," IEEE Trans. Acoustics, Speech, and Signal Proc., *ASSP-24*, No. 6 (December 1976), pp. 550–558.
16. 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. Acoustics, Speech, and Signal Proc., *ASSP-27*, No. 2 (April 1979), pp. 134–141.
17. L. R. Rabiner and J. G. Wilpon, "Considerations Applying Clustering Techniques to Speaker Independent Word Recognition," J. Acoust. Soc. Amer., *66* (September 1979), pp. 663–673.
18. 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 Proc., 1979.
19. J. M. Tribolet, L. R. Rabiner, and M. M. Sondhi, "Statistical Properties of an LPC Distance Measure," IEEE Trans. Acoustics, Speech, and Signal Proc., 1979.
20. J. Makhoul, R. Viswanathan, L. Cosell, and W. Russell, "Natural Communication with Computers," BBN Rep. No. 2976, December 1974.
21. M. R. Sambur and N. S. Jayant, "LPC Analysis/Synthesis from Speech Inputs Containing Quantizing Noise or Additive White Noise," IEEE Trans. Acoustics, Speech, and Signal Proc., *ASSP-24*, No. 6 (December 1976), pp. 488–494.
22. D. J. Goodman, C. Scagliola, R. E. Crochiere, L. R. Rabiner, and J. Goodman, "Objective and Subjective Performance of Tandem Connections of Waveform Coders with a LPC Vocoder," B.S.T.J., *58*, No. 5, (March 1979), pp. 601–629.
23. A. E. Rosenberg and C. E. Schmidt, "Recognition of Spoken Spelled Names Applied to Directory Assistance," J. Acoust. Soc. Amer., *62*, Supplement No. 1 (December 1977), p. 563 (Abstract).