Speaker-Independent Recognition of Isolated Words Using Clustering Techniques

LAWRENCE R. RABINER, FELLOW, IEEE, STEPHEN E. LEVINSON, MEMBER, IEEE, AARON E. ROSENBERG, MEMBER, IEEE, AND JAY G. WILPON

Abstract—A speaker-independent isolated word recognition system is described which is based on the use of multiple templates for each word in the vocabulary. The word templates are obtained from a statistical clustering analysis of a large database consisting of 100 replications of each word (i.e., once by each of 100 talkers). The recognition system, which accepts telephone quality speech input, is based on an LPC analysis of the unknown word, dynamic time warping of each reference template to the unknown word (using the Itakura LPC distance measure), and the application of a K-nearest neighbor (KNN) decision rule. Results for several test sets of data are presented. They show error rates that are comparable to, or better than, those obtained with speaker-trained isolated word recognition systems.

I. Introduction

ALTHOUGH there are a large number of factors which influence the implementation of a discrete word recognizer, perhaps two of the most important ones are vocabulary size and degree of speaker dependence. These factors are illustrated in Fig. 1 which also shows the areas in which the most current word recognition research is being performed. For speaker-dependent systems, vocabulary sizes of from 40 to 1000 words have been investigated [1]-[4], with vocabulary sizes of from 100 to 200 words being most typical of modern systems. For speaker-independent systems, vocabulary sizes of from 2 to 50 words have been used with varying degrees of success [5]-[10]. Error rates associated with such systems range from 20 percent (for the larger vocabulary sizes or the more difficult vocabularies) to less than 1 percent (for the easier or smaller vocabularies).

Although most word recognition systems are either speaker-dependent or speaker-independent the dichotomy between these two categories is more one of implementation than of structure. For statistical pattern recognition systems, a speaker-dependent word recognizer can be used as a speaker-independent word recognizer (and vice versa) by interchanging the set of reference templates. Recently, several attempts have been made at "bridging the gap" between completely speaker-dependent and completely speaker-independent word recognizers by using multiple templates per word instead of the usual single template per word [8]-[10]. In addition, increasingly sophisticated pattern recognition or clustering algorithms have been used to aid in the optimal selection of the word templates [11]. As such, these word recognizers fall midway between the classic speaker-independent and speaker-dependent recognizers, as illustrated in Fig. 1.

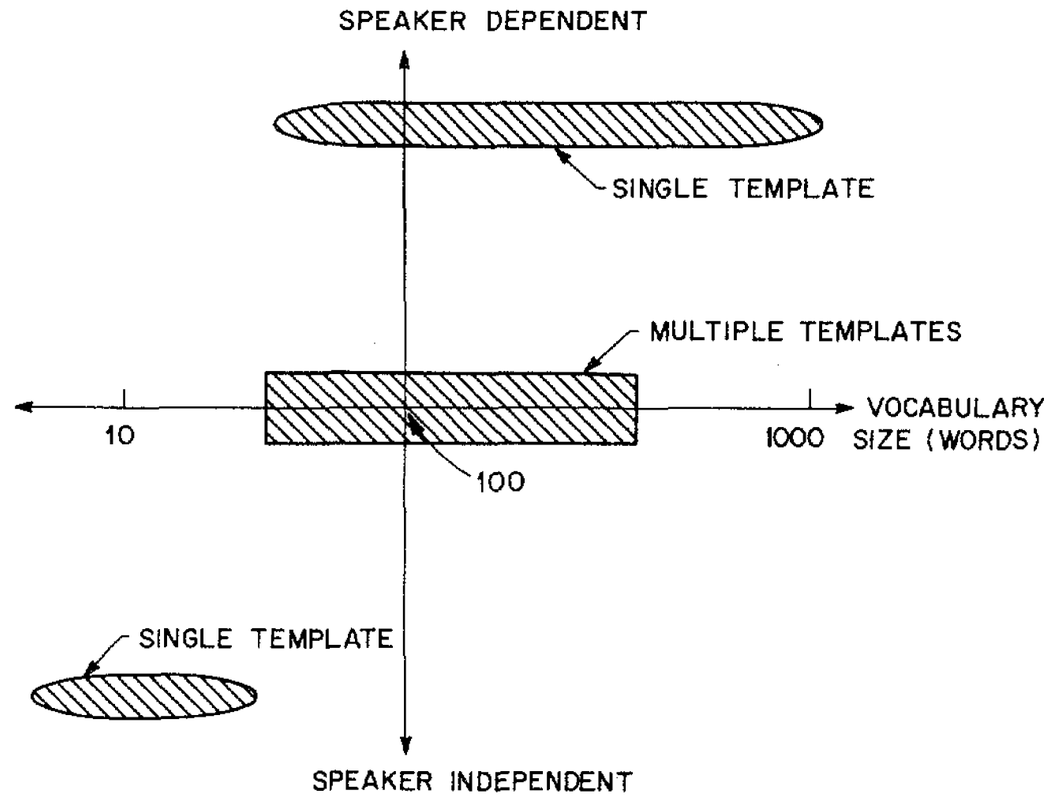

Fig. 1. Illustration of the range of recent isolated word recognition systems as a function of vocabulary size and degree of speaker dependence.

The purpose of this paper is to describe some recent results on speaker-independent recognition of isolated words based on word templates obtained from a statistical clustering analysis. In building the recognition system, many aspects were considered, such as time alignment procedures, endpoint robustness by means of a backup frame, rejection thresholding, and statistical decision rules. The theory of these procedures is described in Section II. The cluster analysis which is central to these investigations is described briefly in Section III. Extensive test results are provided in Section IV. In particular, we have evaluated each of the aspects discussed in Section II independently to observe its contribution to the overall system performance. Finally, we summarize in Section V by presenting an error analysis, a comparison of our results with those of other researchers and directions for further studies.

II. Word Recognition System

Fig. 2 shows a block diagram of the word recognition system. The speech signal was recorded using a standard telephone line, bandpass filtered from 100 to 3200 Hz, and sampled at a 6.67 kHz rate. The first step in the digital processing of Fig. 2 is endpoint detection to determine points in time at which the unknown word begins and ends. The major causes of errors

Manuscript received September 8, 1978; revised January 15, 1979.

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

0096-3518/79/0800-0336$00.75

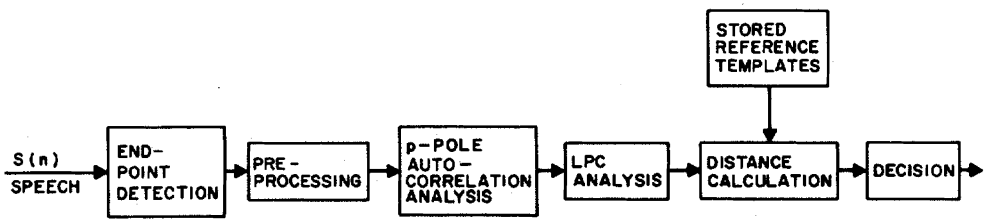

Fig. 2. Block diagram of the word recognition system.

in endpoint detection are clicks on the lines and heavy breathing at the ends of words. Special care was taken to minimize the possibility of endpoint errors. In particular, a backup ending point was calculated to account for mouth clicks, breath noise, etc. at the end of the utterance. The calculation for the backup point was as follows. We denote the detected frames of the word as having indices $i = 1, 2, \cdots, N_{END}$. For each frame we denote the zeroth autocorrelation coefficient as $R_i(0)$. We then define the function $g(i)$ as

$$g(i) = \sum_{j=1}^{i} \frac{R_j(0)}{R_{\max}} \tag{1}$$

where

$$R_{\max} = \max_i [R_i(0)]. \tag{2}$$

The backup frame N_{BU} is calculated as the largest index i satisfying the constraint

$$g(i) < g(N_{END}) - S \tag{3}$$

where S is an empirically determined threshold. (It was set to 0.001 in our simulations.) The backup frame is used as an alternative endpoint for the distance calculations to be described later.

Following endpoint detection, the speech is preemphasized using a simple first-order digital filter with z transform

$$H(z) = 1 - az^{-1} \tag{4}$$

where a value of $a = 0.95$ was used in our simulations. Extensive experimental evidence has shown that preemphasis serves to reduce the variance of the distance calculations used in the recognition system when LPC parameters are used as the feature set, and the autocorrelation method of analysis is used [12].

The next step in the recognition system is to perform a p-pole autocorrelation analysis of the word. A value of $p = 8$ was used for the telephone quality speech. The autocorrelation coefficients were calculated from overlapping frames of length $N = 300$ samples (45 ms) using a Hamming window on the data. A total of 67 frames/s (i.e., every 15 ms) were calculated. Each frame of autocorrelation coefficients was then converted to linear prediction coefficients (LPC), using the autocorrelation method, for subsequent processing and/or storage as reference patterns.

A. Dynamic Time Warping

The recognition phase is essentially a matching process in which an unknown sample pattern of autocorrelation coefficients is compared with an ensemble of stored reference patterns (templates). The reference patterns may be from a designated speaker (for speaker-dependent systems) or a "universal" set (for speaker-independent systems). In the comparison, a frame-by-frame scan of the sample pattern is carried out against each reference pattern. A distance score (or measure of dissimilarity) is calculated and accumulated using a dynamic programming technique [1], [13]-[15] as the scan proceeds. A simple decision rule which is often used designates the vocabulary item corresponding to the reference pattern with the lowest accumulated distance as the recognized word. A somewhat more powerful decision rule is discussed later in this section.

The use of a dynamic programming algorithm provides an efficient procedure for obtaining a nonlinear time alignment between each reference pattern and the unknown sample. By means of a simple recursion formula, a sequence of frames through each reference pattern is generated, associated with a minimum accumulation of distance from beginning to end. The use of a nonlinear time alignment has been shown to be a significant factor in the performance of the recognizer, especially for polysyllabic words [1].

B. Variants of the Time-Warping Algorithm

Recently, several variants on the basic dynamic time-warping algorithm have been proposed [14]-[15]. These modified dynamic time-warping algorithms account for misregistrations between the unknown sample and the reference patterns due to errors in the word endpoints. In addition, Sakoe and Chiba [14] have proposed a modified version of the algorithm which is symmetrical in the time alignment procedure, i.e., neither the unknown sample nor the reference guides the frame-by-frame matching process, but instead a parametric index (which is a function of both time scales) is used.

The three versions of the dynamic time warping algorithm proposed by Rabiner *et al.* [15] have been studied in the context of the recognition system of Fig. 2. These three algorithms and their properties are as follows.

1) CE2-1—Constrained Endpoints, 2-to-1 Range of Slopes: This algorithm is the one proposed by Itakura [1] in which the starting and ending points are assumed to be in perfect registration, and the dynamic path is assumed to be in a fixed parallelogram whose slopes are 2 and $\frac{1}{2}$ at the edges.

2) UE2-1—Unconstrained Endpoints, 2-to-1 Range of Durations: For this version, the boundary conditions are relaxed and it is assumed that a region of width of δ frames exists in which the initial and final frames could be mapped. (A value of δ of 5 frames (75 ms) was used in our implementation.) The dynamic path was again assumed to lie within a fixed parallelogram whose slopes are 2 and $\frac{1}{2}$ at the edges.

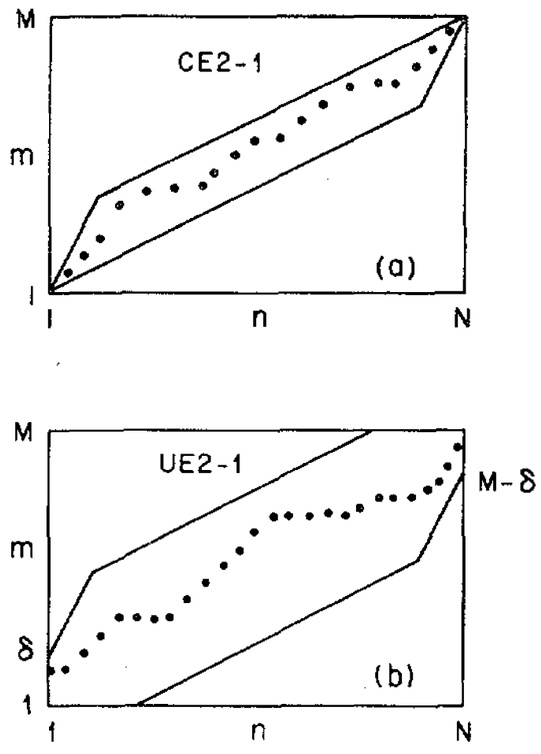

Fig. 3. Typical warping paths for the three dynamic time-warping techniques used in the recognition system.

3) UELM–Unconstrained Endpoints, Local Minimum: For this version, both the endpoint constraints are relaxed, and the allowable region of dynamic paths is constrained to follow the locally optimum path to within a range of $\pm\epsilon$ frames. A value of ϵ of 4 frames (60 ms) was used in our implementation.

Fig. 3 provides a summary of the three dynamic warping algorithms described above. Typical warping functions and the boundaries of the allowable regions of dynamic paths are shown in this figure.

C. Rejection Threshold

To speed up the distance calculation by eliminating unlikely reference patterns, an accumulated distance rejection threshold was used. If we denote the minimum accumulated distance at frame j as D_j, and the rejection threshold is denoted as T_j, then if

$$D_j > T_j \tag{5}$$

where

$$T_j = [T_{\min} + (j-1)T_{\text{slope}}] \cdot N \tag{6}$$

($T_{\min} = 0.3$, $T_{\text{slope}} = 0.7$ (typically), and N is the number of frames of the test sample), the scan is aborted at frame j and the vocabulary item corresponding to the reference pattern is rejected as a candidate for recognition. Generally, the minimum threshold $T_{\min}$ and the slope T_{slope}, are chosen to ensure that a sufficient number of candidates are not rejected. We discuss the effect of raising and lowering the rejection thresholds later in this paper.

Fig. 4 shows a plot of typical accumulated distances versus frame number for a recognition test. The rejection threshold is shown as the straight line at which the scans terminate. It

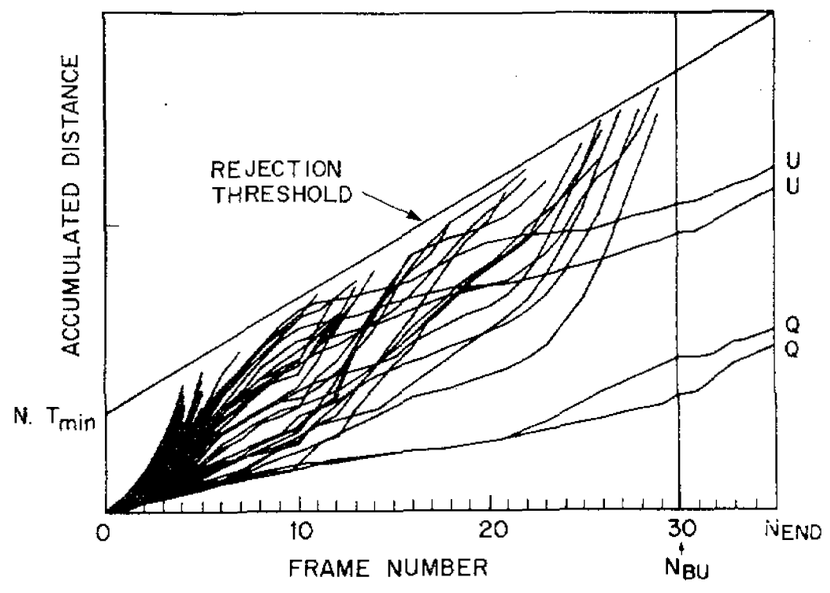

Fig. 4. Accumulated distance versus frame number for the test word Q.

is not unusual for the vast majority of incorrect words to be eliminated by the rejection threshold. Also shown in this figure is the back-up frame N_{BU}. For scans which fall below the rejection threshold until the backup frame, the total average distance for the entire scan is the quantity D, defined as

$$D = \min\left[\frac{D_{BU}}{N_{BU}}, \frac{D_{END}}{N_{END}}\right] \tag{7}$$

where N_{BU} is the frame number of the backup frame, N_{END} is the frame number of the ending frame, D_{BU} is the accumulated distance to the backup frame, and D_{END} is the accumulated distance to the ending frame (which may not be the last frame in the unknown word due to the boundary conditions of the warping algorithm). Thus, the backup frame serves as an alternate estimate of the endpoint of the unknown word, and if the accumulated distance rises rapidly after the backup frame, it is assumed that it is due to endpoint errors, and (7) uses the smaller accumulated average distance for the word.

D. Decision Rule for Recognition

For recognition systems in which a single (or perhaps 2) reference template(s) per word are used, the decision rule which is generally used is the nearest neighbor rule for which the vocabulary item whose average accumulated distance D is minimum is chosen. Thus, if we denote the candidate words by the index j, $j = 1, 2, \cdots, J$, then the nearest neighbor (NN) rule is

$$\text{Choose } i = i^* \ni D[x, x^{(i^*)}] \leqslant D[x, x^{(j)}] \qquad 1 \leqslant j \leqslant J \tag{8}$$

where $D[x, x^{(j)}]$ is the average distance between the unknown x and the reference template $x^{(j)}$.

For recognition systems in which multiple reference templates are used for each word, the decision rule can be made more sophisticated. For example, the K-nearest neighbor (KNN) rule can be used in which the vocabulary item whose average distance of the K-nearest neighbors to the unknown sample is minimum is chosen as the recognized word. If we denote the kth nearest neighbor of the jth word to the unknown sample x as $D[x, x^{(j)}_{[k]}]$, then for the KNN rule we compute the quantity r_j defined as

$$r_j = \frac{1}{K}\sum_{k=1}^{K} D[x, x^{(j)}_{[k]}] \tag{9}$$

and we recognize the unknown word as word j^* such that

$$r_{j^*} \leqslant r_j, \quad j = 1, 2, \cdots, J. \tag{10}$$

It is shown in the Appendix that the quantity r_j of (9) is monotonically related to an estimate of the local probability density function of the jth word. We shall see later that this estimator is well suited to our data.

It should be noted that for $K = 1$, the K-nearest neighbor decision rule becomes the nearest neighbor rule. In this paper we discuss results of recognition tests with values of K from 1 to 4 and for reference data with up to 12 templates per vocabulary word.

E. Other Considerations in the Recognition System

In general, the result of a recognition trial is a single vocabulary item. However, for some applications it is useful to give a *set* of candidate items rather than a single word [3]-[4]. Usually, the list of choices is ordered by the distance scores. Such a result is useful when the recognition output is itself subject to further processing. An example of such a system is the spoken spelled name recognizer for directory assistance proposed by Rosenberg and Schmidt [4]. In this system up to five word candidates were retained for each letter in the spelled name, and a directory search was used to resolve the correct candidate for each letter. For this system a median acoustic error rate of 20 percent led to a median string (name obtained from the directory search) error rate of 4 percent.

Another general comment about the recognition system of Fig. 2 is that the *only* feature that determines whether the recognition system is speaker-trained or speaker-independent is the reference template store. Thus, this recognition system is versatile enough to be used in a wide variety of applications.

III. Generation of Multiple Templates by Clustering

In order to implement a speaker-independent word recognizer, the variations among speakers in pronouncing the same word must be accounted for. One way of accomplishing this task is to select a gross set of features which are characteristic of the phonetic content of the word (e.g., nasality, fricative sound, vowel type, etc.) and rely on meausrements which are capable of predicting the presence or absence of such features [1]. Another possibility is to use an arbitrary set of features (e.g., formant frequencies, cepstral coefficients, LPC coefficients) and to form statistics on the variability of the features both in time and across talkers [16]. Each of the above solutions attempts to find a single characterization of each vocabulary word which is either independent of speakers, or can account for speaker variability in statistical terms.

Another possible solution, and the one which we have used, is to rely on statistical pattern clustering methods to obtain not just one, but a multiplicity of patterns which characterize the variability of the features (for a single word) across different talkers [11]. The basic assumption is that repetitions of a word by different speakers can be clustered into groups such that differences in the features within the group are small, but differences between groups within a word are relatively large. As such, each group or cluster can be represented by a single template, and the word can be represented by whatever number of templates are required to "span the space" of talkers. The local density of training words is a measure of the probability density function for the given word. (See the Appendix for further clarification of this point.) Thus, clusters with the largest number of tokens are those closest to a maximum of the probability density function for the word.

There are a number of important issues which are involved in implementing a set of clustering algorithms for isolated word data [11]. Although clustering techniques have become highly developed [17]-[19], it is still more of an art than a science to arbitrarily cluster data without making assumptions as to the form of its probability density function. Clustering is particularly a problem when the data are characterized by a set of distances (between pairs of words) rather than a set of features in a multidimensional space. In the latter case we can only indirectly estimate the probability density function of the data. Thus, one of the issues in clustering is to decide which types of algorithms are applicable to the given set of data.

Another important issue is the question of how to characterize the data within a cluster. One simple and effective way is to choose an element of the cluster which best (in some sense) characterizes the cluster. Another possibility is to combine the tokens within the cluster using some averaging technique. Intuitively, it is appealing to use a real token as the reference template (rather than some artificially created average). It is not clear how averaging affects the characteristic properties of the templates. However, for clusters with a small number of tokens, averaging may have some advantages.

Other important issues in clustering include deciding how many clusters to use in representing the data, setting thresholds for providing natural separations between clusters, modifying feature-dependent algorithms to use distance data, and the question of whether using distance data from other words will enhance the process of clustering a given word.

To handle a large speech data base of isolated words, a sophisticated clustering system was implemented [11]. In Section III-A we list the procedures used to cluster the data base. Then we discuss the actual data base used and present some statistics on the clustering output.

A. Clustering Algorithms

A series of four procedures was used to cluster the isolated word data. These were

1) the chainmap [18],
2) the shared nearest neighbor procedure [18],
3) the k-means iteration [20], and
4) Isodata [21].

Each of these procedures was used interactively on a matrix of distances between pairs of repetitions of a given word to produce a stable set of clusters for which σ the ratio of average intercluster distance to average intracluster distance, was maximized [11]. The total number of clusters per word is a variable which is determined interactively by examining the outputs of each of the above procedures and deciding whether to increase or decrease the total number of clusters in order to increase σ.

TABLE I
WORDS IN THE VOCABULARY

1. A	21. U
2. B	22. V
3. C	23. W
4. D	24. X
5. E	25. Y
6. F	26. Z
7. G	27. STOP
8. H	28. ERROR
9. I	29. REPEAT
10. J	30. ZERO
11. K	31. ONE
12. L	32. TWO
13. M	33. THREE
14. N	34. FOUR
15. O	35. FIVE
16. P	36. SIX
17. Q	37. SEVEN
18. R	38. EIGHT
19. S	39. NINE
20. T	

B. Speech Database

To test the recognition and clustering algorithm, the 39 word vocabulary of Table I was used. Included in the vocabulary were the letters of the alphabet, the digits, and the control words STOP, ERROR and REPEAT. This vocabulary is one which is suitable for a wide range of applications [4].

A group of 100 speakers (50 male, 50 female) recorded the complete vocabulary a total of four times each on different days. Recordings were made in a soundproof booth over a standard telephone line and recorded on analog tape. Each of the 15 600 words was manually edited to eliminate artifacts at the beginning and end of the utterances, and the autocorrelation frames for each word were stored in a file.

The first replication of each talker was used for training the system. Thus, for each vocabulary word, a total of 100 repetitions were used for clustering. If we denote the ith word for the jth speaker as $x^{(ij)}$, and the distance (dynamic time-warped) between the ith word for the lth speaker, and the ith word for the jth speaker as

$$d_{lj}^{(i)} = \frac{d[x^{(il)}, x^{(ij)}] + d[x^{(ij)}, x^{(il)}]}{2}, \tag{11}$$

then the input data to the clustering package for the ith word was the set of paired distances $d_{lj}^{(i)}$, $1 \leqslant l \leqslant 100$, $1 \leqslant j \leqslant 100$. The averaging of distances in (11) is done because, in general, the distance between tokens is not symmetric. Because of the imposed symmetry of (11), we get

$$d_{lj}^{(i)} = d_{jl}^{(i)} \tag{12}$$

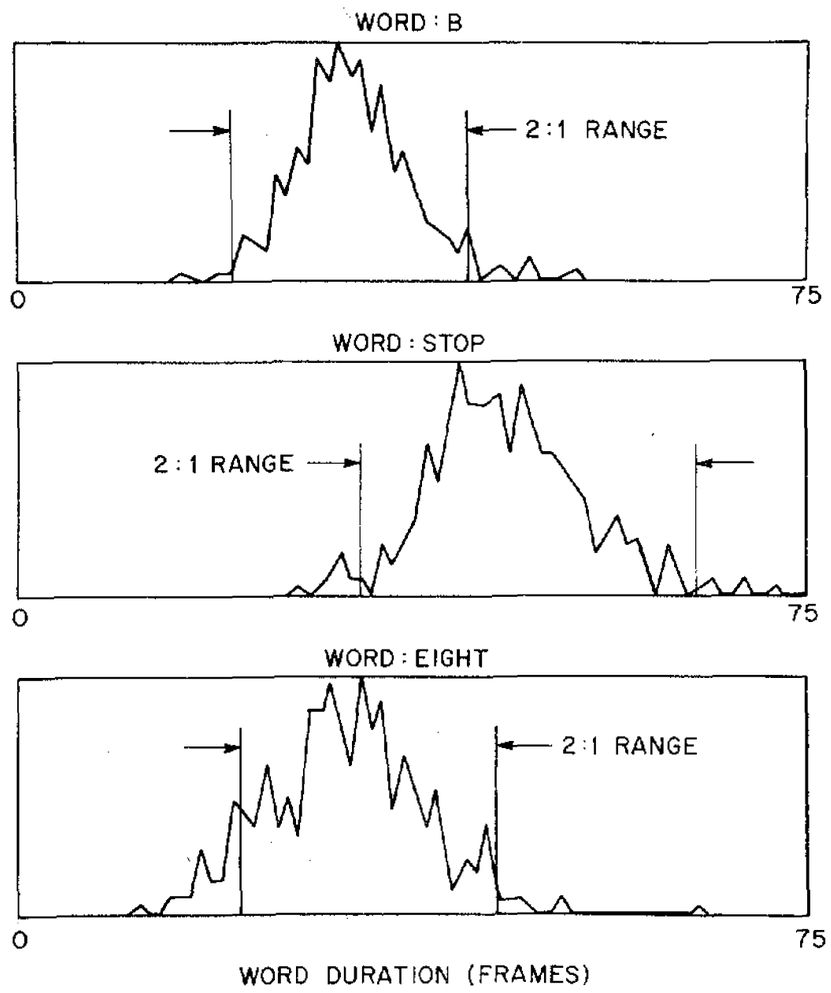

Fig. 5. Histogram plots of the durations of the words B, STOP, and EIGHT.

and thus only the lower triangular matrix $d_{lj}^{(i)}$, $1 \leqslant l \leqslant 100$, $1 \leqslant j \leqslant l$ is required.

Since the quantity $d[x^{(i)}, x^{(j)}]$ is determined by dynamically time-warping token $x^{(j)}$ to token $x^{(i)}$, it is possible that the distance is not defined for some pairs of tokens due to the slope constraints in some of the time-warping algorithms [1], [11]. Thus, prior to the distance calculation [(11)], a histogram of the durations of each of the 100 repetitions of each word was made, and a sliding window was used to find the range of the word durations such that the maximum number of repetitions satisfied the constraints and could be used in the clustering. Those repetitions falling outside the window were eliminated from the clustering.

Fig. 5 shows the plots of the histograms of word durations (for all four replications) for three of the words in the vocabulary (B, STOP, EIGHT). Included in the plots is the 2-to-1 range in which the maximum number of repetitions of the word occurred. (For the UELM time-warping algorithm, no slope constraint existed, and thus all 100 repetitions of each word were clustered.) For the entire 39 word vocabulary, the number of repetitions that were excluded (in the first replication) because of the slope constraint varied from 1 (for the word B) to 12 (for the word EIGHT), and the total number of words excluded was 152 out of 3900 words. A total of 4 talkers of the 100 spoke, 82 of the 152 excluded words; thus these talkers were effectively excluded from the training set about half the time.

C. Results of the Clustering of Word Data

The clustering algorithms of Section III-A were used interactively on the speech database to give an ordered set of clusters for each word in the vocabulary. The paired distance data were obtained from each of the three dynamic time-warping algorithms discussed in Section II-B. Table II presents a summary of the main statistics associated with each of the sets of clusters. In this table are shown, for each time-warping

TABLE II
STATISTICS OF THE WORD CLUSTERS FOR THE THREE DYNAMIC TIME WARPING ALGORITHMS

	CE2-1			*UE2-1*			*UELM*		
	AVG	*MIN*	*MAX*	*AVG*	*MIN*	*MAX*	*AVG*	*MIN*	*MAX*
Number of Clusters Per Word	13	8	19	13	9	18	14	9	19
Number of Outliers Per Word	8	3	16	13	5	19	9	4	16
Quality Ratio (σ)	2.95	2.41	3.88	3.38	2.67	4.68	3.20	2.54	4.47
Size of Largest Cluster	24	12	36	20	12	34	21	10	36

algorithm, the average and the minimum and maximum values of:

1) the number of clusters per word, where a cluster is a set with 2 or more tokens of the (approximately) 100 tokens being clustered;

2) the number of outliers per word, where an outlier is a token that falls outside all the word clusters for that word;

3) the σ or quality ratio for the word; and

4) the size (numbers of tokens) of the largest cluster.

From Table II we see that the number of clusters per word ranged from 6 to 19 and the average was about 13 for all 3 algorithms. The average number of outliers, however, for the UE2-1 case was 50 percent larger than for the other 2 cases. At the same time, however, the σ or quality ratio of the UE2-1 clusters was significantly larger than for the other 2 cases. This result is subject to some interpretation. The σ measure has the property that as the number of clusters increase, the σ ratio can become unbounded (i.e., as the number of clusters becomes equal to the number of data points, the average intracluster distance becomes 0 and σ becomes infinite). Thus, increases in σ are meaningful if the number of data points falling within the nonoutlier clusters stays the same, or increases.

It is of interest to examine the distribution of tokens among the clusters. An important question is whether the tokens are distributed uniformly among the clusters, or do a relatively small number of clusters account for most of the tokens. To answer this question, the function $g_n(l)$, which represents the accumulated number of tokens in the l largest clusters for the nth word, was computed. The range on l was 1 to 10 (representing the 10 largest clusters for each word), and the range of n was 1 to 39. From $g_n(l)$, the average across n was computed for each value of l and is plotted in Fig. 6 for each of the three time-warping algorithms. Also included in the plots are individual curves for the word with the fewest number of tokens in the 10 largest clusters, and the word with the largest number of tokens in the 10 largest clusters. These curves serve to approximately delineate the range of $g_n(l)$ for each warping algorithm. The plots show a highly nonuniform distribution of tokens within the clusters. They also show that, on average, about 80 percent of the tokens are included in the 10 largest clusters.

For each of the five largest clusters for each word, histograms were made of the estimated probability density functions of the distance of the tokens within the cluster and of the durations of the words. No unusual distributions of the duration of tokens within the cluster were found for any of the clusters. The only correlation between clusters and physical quantities that was observed was the tendency of the largest clusters to consist almost entirely of tokens by either male speakers or female speakers, but not both together. The histograms of distances were essentially Gaussian with mean distances of about 0.2 to 0.4.

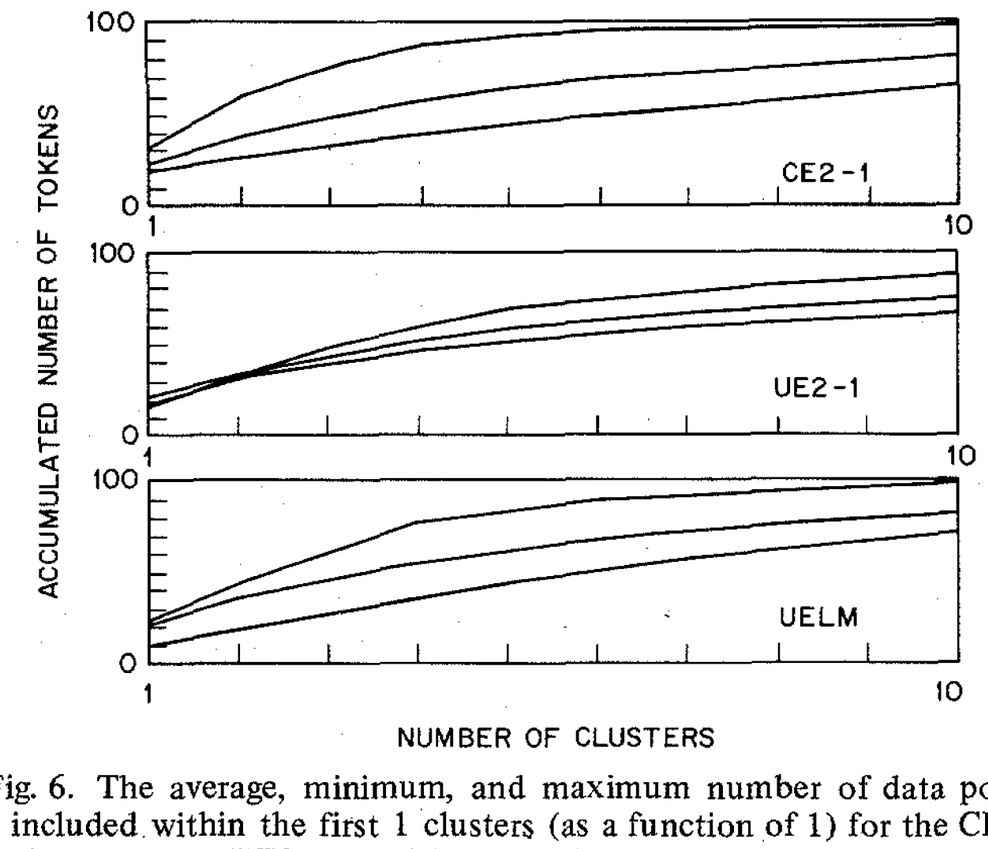

Fig. 6. The average, minimum, and maximum number of data points included within the first 1 clusters (as a function of 1) for the CE2-1 algorithm, the UE2-1 algorithm, and the UELM algorithm.

IV. RECOGNITION RESULTS

To test the clustering analysis, several recognition tests were performed. In this section we describe the different data test sets which were generated, and then describe the individual experiments which were done.

A. Recognition Test Sets

Four distinct test sets of data were generated to test both the recognition system of Section II, and the clustering analysis of Section III. We denote the individual test sets as TS1 to TS4. The test material was as follows.

TS1—Each of 10 talkers (5 male, 5 female) spoke the 39 word vocabulary once over a dialed-up telephone line. The 10 talkers were all subjects who were *not* part of the original 100 talker database used for the clustering analysis. A new dialed connection was used for each talker. On-line editing (manual) of the endpoints was done on this data set to correct gross errors made in recording, e.g., erroneous clicks, pops, etc. which were not part of the recording process. A total of 390 words were in TS1.

TS2—Each of 8 new talkers (4 male, 4 female) spoke the 39 word vocabulary once over dialed-up telephone lines. Again the 8 talkers were not in the original training set. A high speed array processor (CSP MAP-200) performed the autocorrelation analysis of the input speech in real time and thus no manual editing of the endpoints was performed. A total of 312 words were in TS2.

TS3—This test set consisted of a random selection of talkers and words from the 100 talker database. The random selection was made from the three replications of the original recordings which were not used in the training set. For each of the 39 words, a random selection of 10 of the 300 tokens

TABLE III
RECOGNITION ACCURACIES (%) FOR CLUSTERS FROM THE CE2-1 ALGORITHM AND FOR THE RANDOMLY CHOSEN CLUSTERS

C = 1 (Top Candidate)

	K = 1		K = 2		K = 3		K = 4	
l	CE2-1	Random	CE2-1	Random	CE2-1	Random	CE2-1	Random
2	61.2	49.9	51.6	47.1				
4	69.2	55.8	69.4	58.6	60.9	55.0	58.3	49.6
6	68.4	64.8	71.2	68.1	68.1	64.0	61.2	64.3
8	72.3	67.4	75.6	67.6	75.1	67.9	73.0	66.3
10	73.3	66.6	77.6	70.4	75.3	69.1	74.0	67.9
12	74.6	69.7	79.2	69.9	75.8	72.0	73.3	71.5

C = 2 (2 Top Candidates)

	K = 1		K = 2		K = 3		K = 4	
l	CE2-1	Random	CE2-1	Random	CE2-1	Random	CE2-1	Random
2	75.3	59.1	66.3	55.0				
4	82.5	70.7	82.0	70.7	74.5	67.4	69.4	63.3
6	84.1	80.7	83.3	80.7	80.7	79.7	74.0	74.0
8	85.4	81.2	87.9	81.2	85.9	79.4	83.3	75.8
10	87.4	80.7	87.9	83.3	86.6	83.0	86.1	81.5
12	87.4	82.5	89.0	82.3	87.4	84.1	85.9	83.8

C = 5 (5 Top Candidates)

	K = 1		K = 2		K = 3		K = 4	
l	CE2-1	Random	CE2-1	Random	CE2-1	Random	CE2-1	Random
2	88.2	77.4	83.3	76.9				
4	93.6	89.5	92.8	88.7	88.4	84.8	83.3	80.7
6	95.6	91.5	92.6	92.3	91.3	91.5	88.4	87.7
8	96.7	93.3	96.1	92.3	95.1	91.0	91.5	88.4
10	97.4	94.6	97.7	94.1	97.4	94.6	95.6	92.3
12	97.9	94.4	98.5	95.4	96.7	94.6	94.9	93.8

(100 talkers times 3 replications) was made. A total of 390 words were in TS3.

TS4—This test set consisted of all the tokens from the training set which were out of range for the constrained dynamic warping algorithms, i.e., tokens which were unusually long, or short as compared to the average duration for the word. This set represents an extremely difficult test set because of the extremes of the duration of the words. The number of words in the test set was 162, and the words were nonuniformly distributed across the vocabulary, as mentioned earlier.

B. Recognition Experiments and Results

1) Recognition as a Function of the Number of Templates per Word: The purpose of the first recognition experiment was to measure recognition accuracy as a function of the number of templates per word in the training set. For this purpose the reference templates were chosen from the CE2-1 clustering results. The test set was TS1. For all the recognition experiments to be described in this paper, results were obtained for values of K (in the K-nearest neighbor rule) from $K = 1$ to $K = 4$. The results of this first test are given in Table III (the columns labeled random will be described later) and Fig. 7. The results are given as the mean accuracy (averaged over talkers) for each nearest neighbor rule (K) as a function of the number of templates per word (l), and the number of ordered candidates that were considered (C). The word templates were chosen in descending order based on the size of the cluster, i.e., the $l = 1$ template was the cluster center of the largest cluster, the $l = 2$ template was the cluster center of the next largest cluster, etc. For $C = 1$, only the top candidate was considered. The results for this case are shown in Fig. 7(a). It is seen that the recognition accuracy is about 61

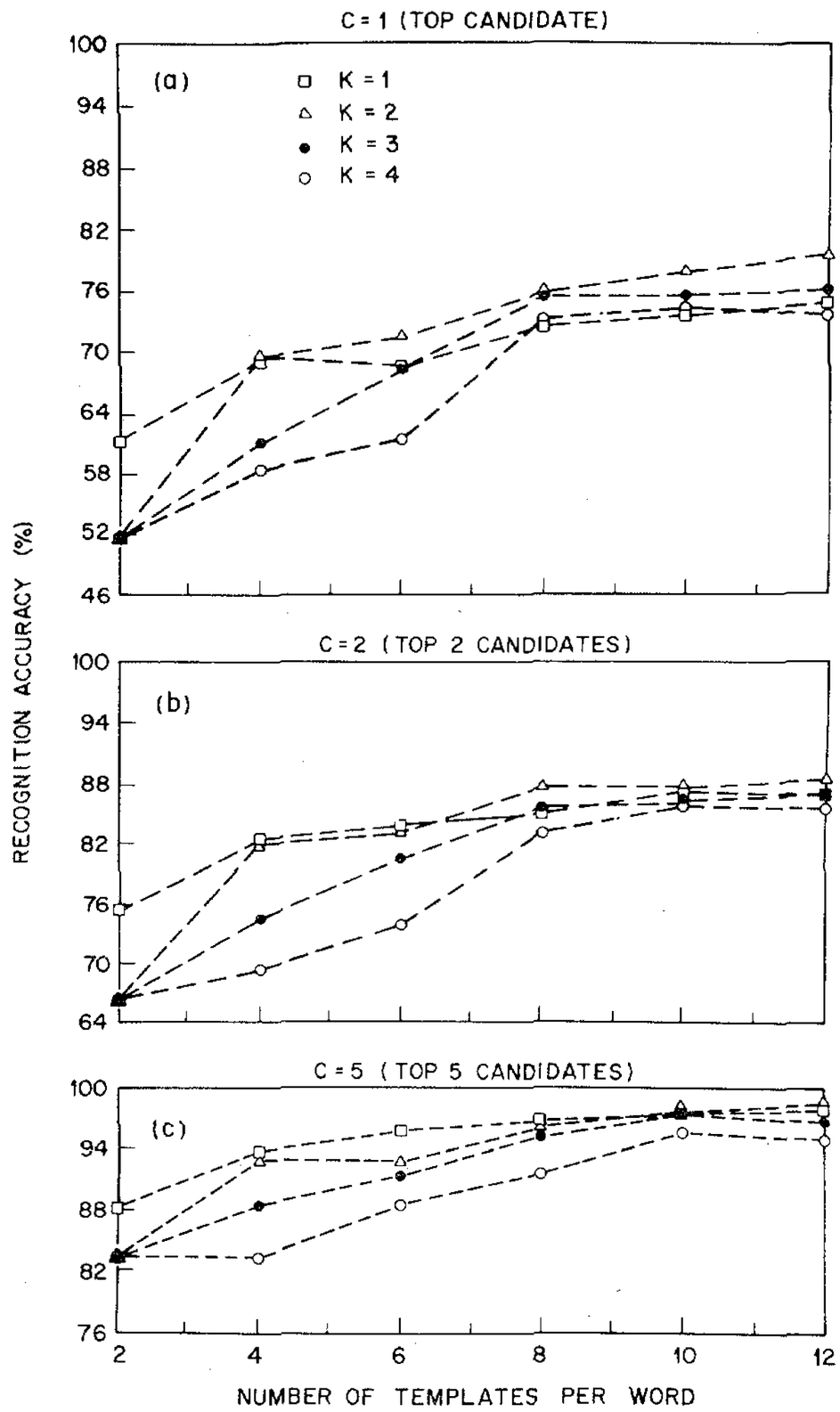

Fig. 7. Recognition accuracy (percent) as a function of the number of templates per word for the CE2-1 clusters with $K = 1$, 2, 3, 4, and $C = 1$, 2, and 5.

percent for $K = 1$ and $l = 2$, and about 51 percent for $K = 2$ and $l = 2$. As l increases, the $K = 2$ and $K = 3$ nearest neighbor rules yield higher recognition accuracies than the $K = 1$ or $K = 4$ rules. For $l = 12$ templates per word (the most used in our tests), the final recognition accuracy (for $C = 1$) was 79 percent for the $K = 2$ rule, and from 3 to 5 percent lower for the other rules.

Similar behavior of the curves of recognition accuracy versus l (for different K values) is seen for the $C = 2$ top candidates [Fig. 7(b)], and for the $C = 5$ top candidates [Fig. 7(c)]. For the best two candidates, the recognition accuracy goes from about 75 percent (for the $K = 1$ rule) to 89 percent (for the $K = 2$ rule) as the number of templates per word goes from 2 to 12. For the top five candidates, the highest accuracy goes from 88 to 98.5 percent for a similar range of l.

The overall shape of the curve of recognition accuracy versus l (for all the values of K and C) shows a sharp rise near $l = 2$ and a gradual steadying off near $l = 10$ to 12. Thus, increases in the number of templates per word beyond 12 would produce marginal (if any) increases in recognition accuracy.

2) Comparisons of the Three Dynamic Warping Algorithms: A series of tests was performed to compare the recognition rates using the CE2-1, UE2-1 and UELM dynamic warping

TABLE IV
RECOGNITION ACCURACIES (%) FOR THE FOUR TEST SETS

	$C=1$				$C=2$				$C=5$			
Clusters	$K=1$	$K=2$	$K=3$	$K=4$	$K=1$	$K=2$	$K=3$	$K=4$	$K=1$	$K=2$	$K=3$	$K=4$
CE2-1	75	79	76	73	87	89	87	86	98	99	97	95
UE2-1	69	74	74	74	84	84	86	85	97	96	96	96
UELM	68	73	73	72	84	87	87	85	95	96	96	95

(a)

	$C=1$				$C=2$				$C=5$			
Clusters	$K=1$	$K=2$	$K=3$	$K=4$	$K=1$	$K=2$	$K=3$	$K=4$	$K=1$	$K=2$	$K=3$	$K=4$
CE2-1	74	76	72	71	86	88	89	88	96	98	97	97
UE2-1	72	73	74	71	85	86	85	83	96	95	96	95
UELM	65	68	68	66	82	82	80	80	92	92	92	91

(b)

	$C=1$				$C=2$				$C=5$			
Clusters	$K=1$	$K=2$	$K=3$	$K=4$	$K=1$	$K=2$	$K=3$	$K=4$	$K=1$	$K=2$	$K=3$	$K=4$
CE2-1	79	82	82	82	90	91	91	91	99	98	99	99
UE2-1	75	80	80	81	90	92	91	91	98	99	99	98
UELM	71	76	75	76	86	87	90	89	95	97	97	97

(c)

	$C=1$				$C=2$				$C=5$			
Clusters	$K=1$	$K=2$	$K=3$	$K=4$	$K=1$	$K=2$	$K=3$	$K=4$	$K=1$	$K=2$	$K=3$	$K=4$
CE2-1	60	65	69	62	75	81	80	80	90	93	91	91
UE2-1	59	67	70	70	78	82	83	80	91	92	94	93
UELM	62	70	72	70	81	82	83	80	91	95	97	97

(d)

algorithms. All four test sets (TS1-TS4) were used in these tests. A total of $I = 12$ templates per word were used in each test. Table IV shows the average recognition accuracy for each set of data for each dynamic warping algorithm, as a function of K (nearest neighbor rule) and C (number of candidates).

Table IV(a) shows that the CE2-1 algorithm consistently performed as well or better than the other two warping algorithms for the data of TS1. For $K = 1$ and 2, the CE2-1 algorithm gave recognition accuracies from 1 to 6 percent higher than the next best warping method. For $K = 3$, the differences in error rate were small, but remained consistent. For $K = 4$, the differences between all three methods were small.

Tables IV(b) and IV(c) shows that for the data of TS2 and TS3, the CE2-1 again performed consistently as well as or better than the other two algorithms. Higher recognition accuracies of from 1 to 5 percent were obtained for different values of K and C.

The data of Table IV(d), however, show that for the data of TS4 (the out-of-range candidates) the recognition accuracy of all three dynamic warping methods fell considerably for $C = 1$ and $C = 2$. The data show that the UELM performed consistently the best and achieved recognition accuracies of 97 percent for $K = 3$ and $K = 4$ for $C = 5$ top candidates, whereas the CE2-1 and UE2-1 algorithms had from 3 to 6 percent lower accuracies for these cases. For $C = 1$ and $C = 2$, the peak recognition accuracies of 72 percent and 83 percent were considerably lower than for the earlier test sets of data.

3) Effects of Changes of the Rejection Threshold on Recognition Accuracy: All the preceding recognition tests were run using a fixed linear rejection threshold on the dynamic warping accumulated distance. The threshold was of the form

$$R(n) = R_{\min}(N) + (R_{\max} - R_{\min}) \frac{(n-1)}{(N-1)} (N)$$

$$n = 1, 2, \cdots, N \quad (13)$$

where $R_{\min}$ was chosen to be 0.3 and $R_{\max}$ was chosen to be 1.0, and N was the number of frames of the test (unknown) utterance. The quantities $R_{\min}$ and $R_{\max}$ can be shown to be related to the anticipated range of LPC distances for a given frame based on the distribution of LPC distances [1], [22]. One short experiment was run to show the effects of adjusting $R_{\min}$ and $R_{\max}$ on the recognition accuracy. For this experiment, the test set was TS1 and the templates from the CE2-1 training set were used. Again a total of 12 templates per word were used. The results are given in Table V. This experiment showed that when $R_{\min}$ and $R_{\max}$ were raised by 50 percent (allowing more templates to go to termination in the dynamic warping), *no change* occurred in the recognition accuracy. However, when $R_{\min}$ and $R_{\max}$ were lowered by 50 percent (rejecting a greater percentage of templates), an increase in error rate of from 2 to 8 percent occurred for different values of K and C. These tests conclusively showed the validity of the chosen values of $R_{\min}$ and $R_{\max}$ for this system.

4) Effect of the Backup Frame: One brief experiment was run again using the test data of TS1, and the reference data of the CE2-1 templates (12 per word) in which the backup frame was eliminated (i.e., the backup frame was chosen as the last frame in the utterance). The results of this experiment are also given in Table V. It can be seen that without the backup frame, a small but consistent increase in error rate occurs for different values of K and C. Increases of up to 3 percent ($K = 2$, $C = 1$) of the error rate can be seen in the table. These results, however, should not be considered con-

TABLE V
RECOGNITION ACCURACY AS A FUNCTION OF REJECTION THRESHOLD AND BACKUP FRAME

			$C=1$				$C=2$				$C=5$			
R_{min}	R_{max}	BU	$K=1$	$K=2$	$K=3$	$K=4$	$K=1$	$K=2$	$K=3$	$K=4$	$K=1$	$K=2$	$K=3$	$K=4$
0.3	1.0	Yes	75	79	76	73	87	89	87	86	98	99	97	95
0.15	0.6	Yes	71	74	72	71	82	83	82	81	90	91	90	90
0.5	1.5	Yes	75	79	76	73	87	90	87	86	98	99	97	95
0.3	1.0	No	75	76	75	74	87	89	88	86	98	97	97	95

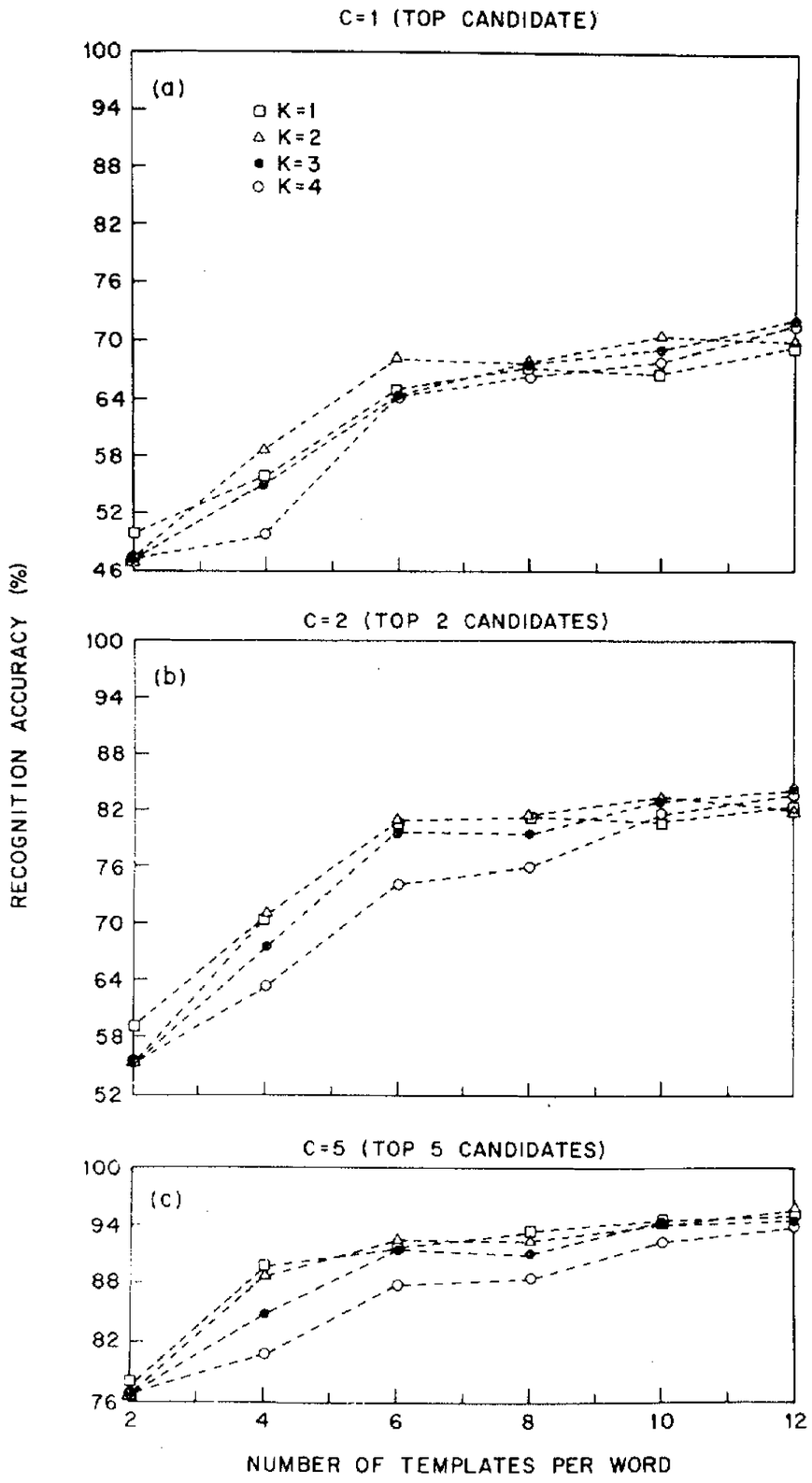

Fig. 8. Recognition accuracy (percent) as a function of the number of templates per word for random templates with $K = 1, 2, 3, 4$, and $C = 1, 2$, and 5.

clusive since the endpoints of the words of TS1 were manually corrected to eliminate clicks, pops, etc. A more definitive test needs to be done to check the utility of the backup frame.

5) Results Obtained Using Random Templates: To verify that the clustering analysis was providing any benefit, a set of templates was selected by choosing at random l out of the 100 templates for each word. Using the data of TS1, the recognition accuracy was measured as a function of l and C for $K = 1$ to 4 using the random templates. The results are given in Table III (columns labeled random) and Fig. 8. It can be seen (by comparing the random and the clustered results of Table III) that decreases in recognition accuracy of from 1 to 16 percent are the result of using randomly selected templates. For $K = 2$, $l = 12$ the differences in recognition accuracy are

9.3 percent for $C = 1$ (top candidate)
6.7 percent for $C = 2$ (top 2 candidates)
3.1 percent for $C = 5$ (top 5 candidates).

Although as l and C get large and the differences in accuracy decrease, there are substantial differences in recognition accuracy for all values of K, l, and C. Thus, this analysis shows the effectiveness of the clustering algorithms.

Similar comparisons were made for the data of TS2 and TS3 for a value of $l = 12$ templates per word. The results are entirely consistent with those of TS1; namely, significantly increased error rates for the randomly chosen templates.

6) Digit Recognition Results: Since the digits (zero to nine) were a subset of the 39 word vocabulary, it was a simple mat-

ter to perform an experiment to see how well digits spoken in isolation could be recognized using the clustered digit data. Thus, a test set was created with 2100 digits from 110 talkers, 100 of which were in the training set (using replications 3 and 4 of their data, which, of course, were *not* used in the training), and 10 who were not in the training set. The reference templates were obtained from the CE2-1 clusters for the digits. A total of 12 clusters per digit were used. The overall accuracies for the top candidate ($C = 1$) was 97.5 percent ($K = 1$), 98.2 percent ($K = 2$), 98.1 percent ($K = 3$), and 97.9 percent ($K = 4$). For the top 2 candidates ($C = 2$), the accuracies were within 0.1 percent of 99.6 percent for all 4 values of K. For the 10 talkers not in the original training set, the accuracy was 97 percent for $C = 1, K = 1$, 100 percent for $C = 1, K = 2$, and $K = 3$, and 98 percent for $C = 1, K = 4$.

7) Ratio Test Threshold: Based on the interpretation of the K-nearest neighbor distance $d_i(x)$ as being an estimate of the probability density function of the ith word at the point x (where x represents an unknown sample), the ordered distance data provide an interesting and useful way of setting a threshold to give a possible "no decision" as the outcome of each recognition trial. The assumption is made that if the average distances to two classes are essentially the same, the distributions have significant overlap in the region of x, and it is impossible to perform reliable recognition. The rule we have studied is a simple one. If we denote the ordered K-nearest neighbor distances as $d_{i_1}(x), d_{i_2}(x), \cdots, d_{i_q}(x)$ with

$$d_{i_1}(x) \leqslant d_{i_2}(x) \leqslant \cdots \leqslant d_{i_q}(x), \tag{14}$$

then a rejection occurs if

$$D_R = \frac{d_{i_1}(x)}{d_{i_2}(x)} \leqslant \frac{1}{T} \tag{15}$$

and a classification of the unknown as word i_1 occurs if the inequality of (15) is reversed.

To illustrate this rule, Fig. 9, shows a series of plots of recognition accuracy, rejection rate, and error rate as a function of the parameter $1/T$ for values of $1/T$ from 1.0 to 2.0. The plot of Fig. 9(a) is for the digit experiment with $K = 2$ ($C = 1$, of course). It can be seen that the error rate for the digits can be kept below 0.5 percent with a 3.9 percent rejection rate for a value of $1/T = 1.1$.

Fig. 9(b) shows results for $K = 2$ for the data of TS1 using 12 templates per word obtained from the CE2-1 algorithm. Fig. 9(c) shows results for TS2 and Fig. 9(d) shows results for TS3 with the same set of templates as for the data of Fig. 9(b). It can readily be seen that the error rate decreases substantially as $1/T$ goes from 1.0 to 1.1, and gradually beyond $1/T = 1.1$. For words in which D_R of (15) was in the range 1.0 to 1.1 [for the full vocabulary tests of Fig. 9(b) to 9(d)] about half the time the recognition result was in error. Since about 25 percent of the words fell into this range, the decrease of about 12.5 percent in error rate (for $1/T = 1.1$) generally brought the error rate down to around 6–8 percent, with a reject rate of about 25 percent. Further increases in $1/T$ brought about significantly larger increases in rejection rate with only small decreases in error rate.

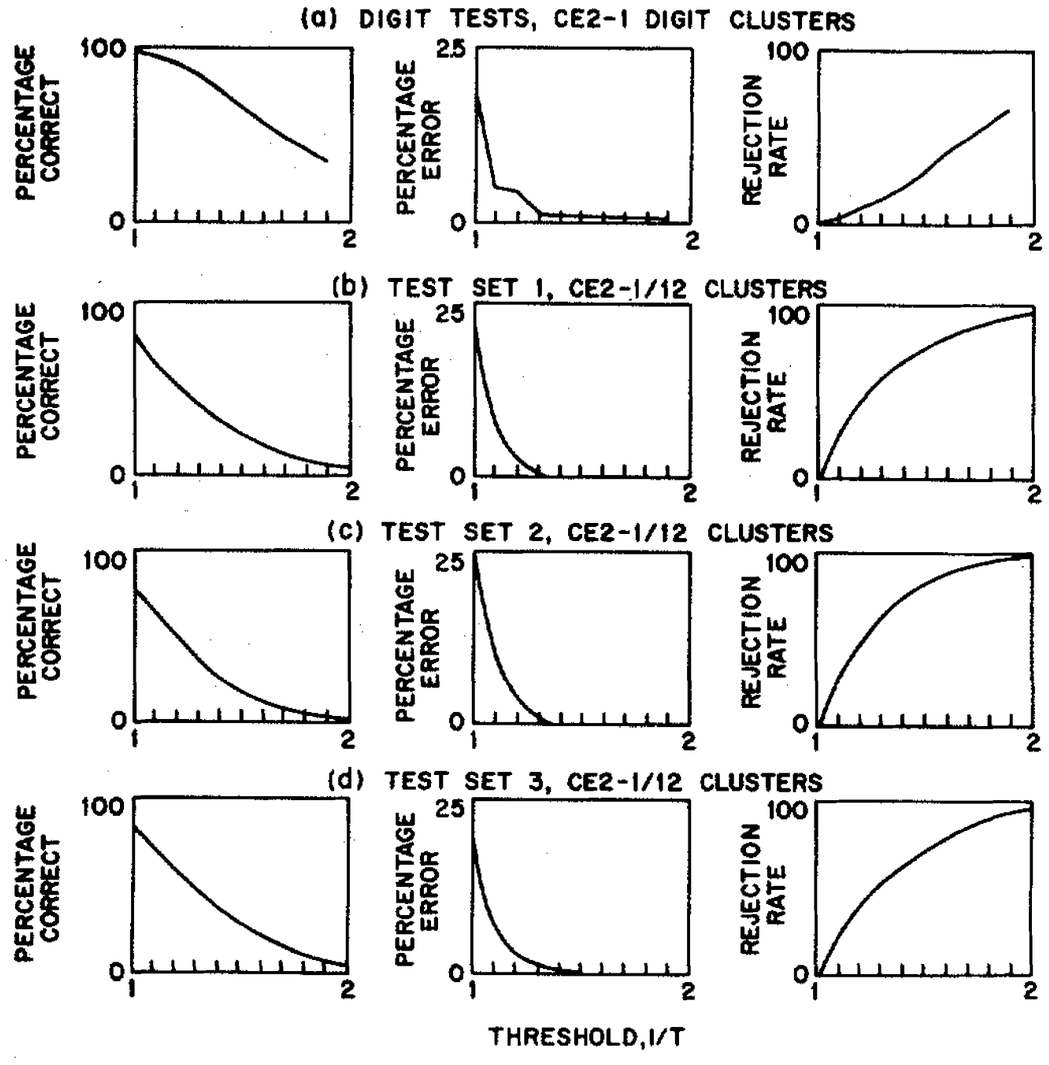

Fig. 9. Curves of recognition rate ($P(C)$) and error rate ($P(E)$) as a function of the rejection threshold T for: a) digit tests, b) test set 1, c) test set 2, and d) test set 3.

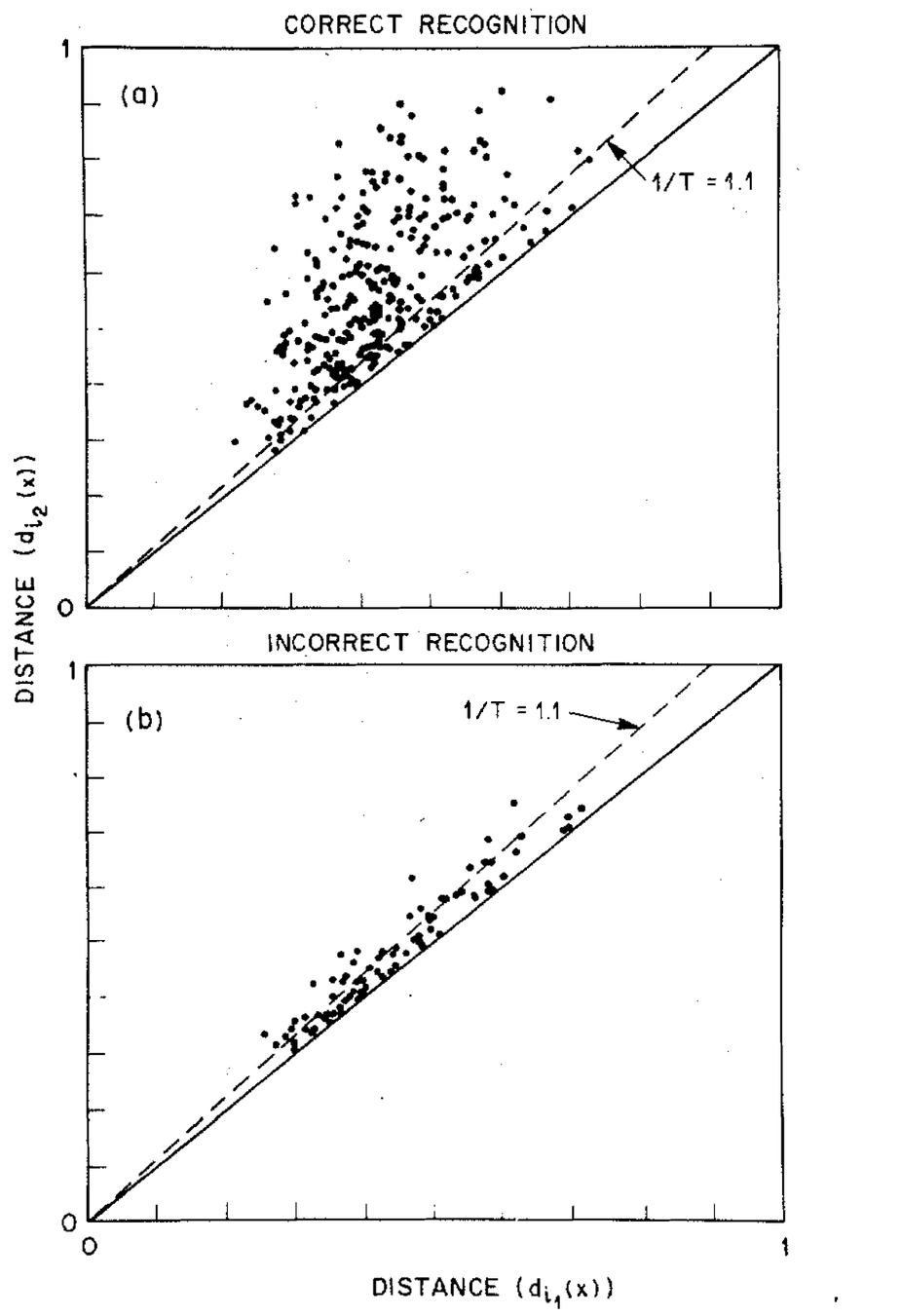

Fig. 10. Plots of the distance of the second candidate versus the distance of the first candidate for: a) words recognized correctly, and b) words incorrectly recognized.

To illustrate that a threshold of the type given in (15) (i.e., a ratio test) is more suitable than an absolute distance threshold, Fig. 10 shows plots of the quantity $d_{i_1}(x)$ versus $d_{i_2}(x)$

for all the words in TS1 using the set of 12 templates per word from the CE2-1 algorithm. The dashed line in the figure shows the set of points where $d_{i_2}(x) = 1.1\,d_{i_1}(x)$. Fig. 10(a) shows data for words which were correctly recognized as the first candidate (using the $K = 2$ rule), and Fig. 10(b) shows data for words which were incorrectly recognized. It is readily seen that in both cases, values of $d_{i_1}(x)$, the minimum distance, range from about 0.2 to 0.8. However, for the misrecognized words the quantity $d_{i_2}(x)$ was generally close to $d_{i_1}(x)$ as indicated by the points around the diagonal. It is this observation that led to the decision rule of (15).

We reiterate the result that, in cases in which absolute word identification is required (as opposed to giving an ordered list of candidates), the use of a decision rejection threshold can substantially reduce error rate at the expense of a finite rejection rate. Since rejections are data-dependent [(15)], cases in which an inherently high *a posteriori* error rate exist are rejected.

V. Discussion

In this paper we had several goals. These were:

1) to investigate a supervised algorithm for clustering words using only accumulated LPC distances between words,

2) to study the effects of variations on the dynamic time-warping algorithm on both clustering and the resulting recognition accuracy.

3) to study a novel decision rule which was linked to a multiple template (cluster) representation of the vocabulary words, and

4) to compare recognition accuracies obtained from multiple templates for each word and used in a speaker-independent manner to those obtained in systems which were trained to the individual speaker.

The discussion in Sections II and III, and the data of Section IV have provided partial answers to many of our original questions. The key results have been the following.

1) The pattern recognition clustering algorithms have provided an effective method of finding structure in the speech data. Evaluations of the resulting clusters in terms of both a quality measure of clustering and in terms of recognition accuracies have shown the data to fall naturally into a small number of clusters each of which could be adequately represented by a single point, the so-called cluster center. Recognition accuracies on test sets containing both new talkers and talkers from the test set were essentially identical across all conditions. This result shows that the clustered data provide, to a first approximation, a universal data set for the given vocabulary words.

2) The constrained endpoint (CE2-1) warping algorithm provided the highest recognition accuracies for almost all the data sets and recognition variables that were tested. For words which were anomolously long, or short, the UELM warping algorithm provided the best results.

3) The $K = 2$ and $K = 3$ nearest neighbor decision rules provided a significant improvement in recognition accuracy over the $K = 1$ (minimum distance) and $K = 4$ rules. This result was anticipated based on the interpretation of the KNN rule distance as an estimate of the local probability density function of the ith word. For a finite set of clusters which are used to "span the entire space" of the ith word, large values of K would be anticipated to give poor results due to sparse sampling of the space.

4) To obtain the best results in the recognition tests, conservative thresholds are required for the rejection threshold to guarantee that at least two or three valid candidates from each cluster set are used to give the KNN estimate of distance. The use of a backup frame to provide protection against spurious, nonspeech sounds at the end of a word provided a small but consistent increase in recognition accuracy.

5) The error rate of the system could be reduced at the expense of a finite rejection rate for applications in which a specified error rate had to be maintained. A simple, effective measurement was discussed which automatically identified those trials with which a high probability of error was associated.

6) High accuracies were obtained (98.2 percent) for speaker-independent digit recognition.

7) Experiments with randomly selected templates clearly showed the superiority of the clustering methods in giving an efficient representation of the structure for each word class.

In the remainder of this section we analyze the types of recognition errors that were made, compare the recognition accuracies that we obtained to those of other investigators, and discuss relevant issues that remain to be investigated.

A. Analysis of Recognition Errors

To analyze the performance of the overall recognition systems, the results of test sets 1, 2, and 3 were merged (using reference data obtained from the CE2-1 warping algorithm with 12 clusters per word), and a confusion matrix of errors was obtained for the $C = 1$ (first candidate) condition using $K = 2$ (nearest neighbor rule). A series of subsets of the resulting confusion matrix are given in Table VI. It is readily seen that the vast majority of confusions occur within classes of high acoustical and phonetic similarity. We have identified six such classes, namely:

1) the set of i sounds—b, c, d, e, g, p, t, v, z, 3,
2) the set of eI sounds—a, j, k, 8, h,
3) the set of e sounds—l, m, n,
4) the set of final fricatives with e or I—f, s, x, 6,
5) the set of aI sounds—i, y, 5,
6) the set of u sounds—q, u, 2.

A total of about 75 percent of the errors occurred *within* each of the six classes, with the majority occurring within Class 1. An error rate of less than 2.5 percent is obtained for the remaining eleven words of the vocabulary. Based on both previous experience and similar experiments with this vocabulary [4], it is felt that the overall error rate of this recognition system is fundamentally controlled by the acoustic similarities between words within each class (especially for band-limited) telephone speech), and not by the clustering results or any particular aspect of the recognition system.

B. Comparisons With Other Recognition Results

The full 39 word vocabulary has been used in two previous research projects by Itakura [1] and Rosenberg and Schmidt

TABLE VI
CONFUSIONS AMONG SUBCLASSES OF THE 39 WORD VOCABULARY

Recognized Word

	B	C	D	E	G	P	T	V	Z	3	Other
B	8		2	2		4	1	8	3		
C		23	1				1		2		1
D	3		7	3		4	4	6	1		
E	6		1	18				3			
G			2		19		3		2		2
P	2	1	2			14	7	1	1		
T	2	1	3	1	2	2	13	2	2		
V	1		3	1				16	5	1	1
Z		3					1	3	20		1
3		1	1			1		1		22	2

	A	J	K	H	8	Other
A	17	2	5		1	3
J		22	5			1
K	2	7	19			
H				25	3	3
8				1	24	

	L	M	N	Other
L	27			1
M	2	20	4	2
N		5	19	4

	F	S	X	6
F	24	2	1	1
S	2	25	1	
X		1	27	
6				28

	I	Y	5	Other
I	19	5	2	2
Y		27		1
5	2		24	2

	Q	U	2	Other
Q	22	3	3	
U	1	27		
2	4	1	21	2

[4]. Itakura tested the vocabulary on a single speaker for which the system was trained. Itakura reported a recognition accuracy of 88.6 percent on the top candidate. This score was about 9 percent higher than those obtained here. However, there were several speakers (of the 28 tested here) who had the same or higher recognition accuracy than Itakura; thus, it is difficult to assess our results based on Itakura's score.

Rosenberg and Schmidt trained their system to each of 10 talkers who recited 364 letters which spelled a group of 50 names with initials [4]. For this group of talkers, average recognition accuracies of

1) 79 percent for $C = 1$ (top candidate),
2) 88.5 percent for $C = 2$ (2 top candidates, and
3) 96 percent for $C = 5$ (5 top candidates),

were obtained.

The average recognition accuracies obtained across 28 talkers (merged test sets), none of whom individually trained the system, were

1) 79 percent for $C = 1$,
2) 89.3 percent for $C = 2$, and
3) 98.5 percent for $C = 5$.

Thus, our average recognition accuracies for speaker-independent recognition were comparable or somewhat higher than those of a speaker trained system. Since Rosenberg and Schmidt used only the letters, these comparisons are not strictly valid. However, they do serve to illustrate that carefully chosen word templates can yield essentially equivalent recognition scores for speaker-trained and speaker-independent systems.

For the digit vocabulary, we compare our results to those of Sambur and Rabiner [23] for speaker-independent recognition of isolated digits, Martin [5] for speaker-dependent recognition of isolated digits, and Rosenberg and Itakura [2] for speaker-dependent recognition of isolated digits. The results obtained by these investigators were

1) 95.6 percent accuracy by Sambur and Rabiner,
2) 99.5 percent accuracy by Martin, and
3) 96 percent accuracy by Rosenberg and Itakura.

Clearly the digit recognition accuracies obtained from multiple templates are essentially comparable to or better than those obtained in earlier investigations.

C. Issues for Further Investigation

Although we have attempted to explore a number of issues related to both clustering of word data for creating templates and recognition, there remain a number of interesting questions which require subsequent investigation. These include the following.

1) The applicability of the clustering approach to speaker-dependent recognition systems.
2) The effects of a totally unsupervised clustering of data on the quality of the clusters, and the resulting recognition accuracy.
3) The applicability of the ordered list of word candidates from the speaker-independent recognizer to a spelled spoken-speech system as described by Rosenberg and Schmidt [4].
4) The effects of averaging tokens within a cluster to give a cluster template, rather than choosing a cluster center based on minimax distance.

We hope to investigate these issues more fully in subsequent research.

VI. SUMMARY

In this paper we have discussed the suitability of using sophisticated pattern recognition techniques to provide multiple speaker-independent word templates for an isolated word recognition system. We have shown that such methods do indeed provide templates which give recognition accuracies that are comparable to equivalent recognition systems that are trained to an individual talker.

APPENDIX
THE GENERALIZED K-NEAREST NEIGHBOR DECISION RULE

Consider the set of points $\{X_1, X_2, \cdots, X_n\}$ and the point Z shown in Fig. 11. The set $\{X_i\}_{i=1}^n$ is a set of n observations drawn from a random process characterized by the probability density function $f(Z)$. The point Z is an arbitrary point in the observation space. As shown in Fig. 11, X_2 is the nearest neighbor to Z; X_1 is its second nearest neighbor; X_3 is its third, and so on. In general, we shall designate the Kth nearest neighbor to Z by $Z_{[K]}$ and we have

$$\|Z - Z_{[1]}\| \leqslant \|Z - Z_{[2]}\| \leqslant \cdots \leqslant \|Z - Z_{[K]}\| \cdot \leqslant \cdots \leqslant \|Z - Z_{[n]}\|. \tag{A1}$$

We define r_K as the average distance from Z to its K-nearest neighbors; thus

$$r_K = \frac{1}{K} \sum_{j=1}^{K} \|Z - Z_{[j]}\|. \tag{A2}$$

Following Fraser [24], we define the tolerance region T as the hypersphere of radius r_K centered at Z. We denote its volume as $\Phi(Z)$. For convenience we shall designate the complement of T in the observation space as T^*. Finally, let m be the number of observations in T. Clearly, $1 \leqslant m < K$.

Let $\hat{f}(Z)$ be an estimate of $f(Z)$ where

$$\hat{f}(Z) = \frac{K}{n} \cdot \frac{1}{\Phi(Z)}. \tag{A3}$$

It can be shown that $\hat{f}(Z)$ is a consistent estimator of $f(Z)$, i.e.,

$$\lim_{n \to \infty} \|\hat{f}(Z) - f(Z)\| = 0. \tag{A4}$$

The proof of (A4) is identical to that given by Loftsgaarden and Queensberry [25] with r_K substituted for $\|Z - Z_{[K]}\|$, the distance from Z to its Kth nearest neighbor. We shall not reproduce the proof here.

Nonparametric density estimators similar to that of (A3) have been studied by Loftsgaarden and Queensberry [25], Cover and Hart [26], and Patrick and Fischer [27]. These estimators lead to the nearest neighbor, majority vote, and generalized nearest neighbor decision rules, respectively. However, these estimators have different properties from that of (A3). To see what the differences are we define the coverage C_T of a tolerance region to be the probability that an observation drawn from $f(Z)$ will fall in the region; thus

$$C_T = P_r\{X \epsilon T\} = \int_T f(\xi)\, d\xi. \tag{A5}$$

Since T is defined in terms of the observations which are random variables, C_T is a random variable with $0 \leqslant C_T \leqslant 1$ and having density function, say, $g(C_T)$.

In the case when C_T is independent of $f(Z)$, that is, if the coverage of the tolerance region is independent of the underlying statistics of the observations, T is said to be distribution-free. The estimates described in the references [25] - [27] cited above are all based upon distribution-free tolerance regions.

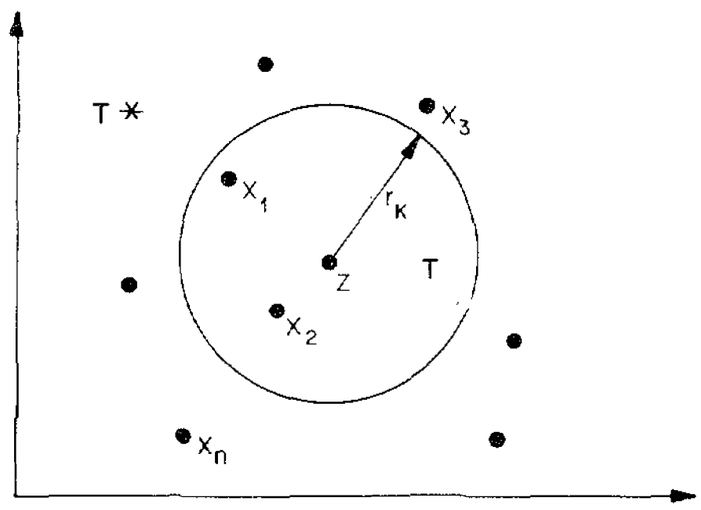

Fig. 11. Points in the observation space.

Using an argument similar to that given by Wilks [28], we see that the tolerance regions as defined above are not distribution-free. We first note that the probability that exactly m observations lie in T, and $n - m$ lie in T^* is given by the binomial distribution

$$P_r[|\{X_i \epsilon T\}| = m] = \frac{n!}{n!(n-m)!} C_T^m (1 - C_T)^{n-m} \tag{A6}$$

from which it follows that $g(C_T)$ has the beta distribution $\beta(m, n - m)$, independent of $f(Z)$.

However, the value of m varies between 1 and $K - 1$ depending on Z, even for a fixed set of observations, so that the density function for the coverage of T is the finite mixture

$$g(C_T) = \sum_{m=1}^{K-1} W_m \beta(m, n - m) \tag{A7}$$

in which the weights W_m depend on $f(Z)$. Thus, the tolerance regions as defined here are "nearly" distribution-free in the sense that their coverages are restricted to a parametric family of distributions. We shall see the significance of this later.

To describe the decision rule imagine that the n observations are taken from M classes, $\omega_1, \omega_2, \cdots \omega_M$, and ω_i has n_i observations and class-conditional density function $f(Z|\omega_i)$. The Bayes decision rule says assign the unknown observation Z to the class for which the conditional density function at Z is the largest, or

$$Z \epsilon \omega_i \text{ iff } f(Z|\omega_i) \geqslant f(Z|\omega_l) \qquad 1 \leqslant l \leqslant M. \tag{A8}$$

Now consider the decision rule

$$Z \epsilon \omega_i \text{ iff } \frac{1}{K_i} \sum_{j=1}^{K_i} \|Z - Z_{[j]}^{(i)}\| \leqslant \frac{1}{K_l} \sum_{j=1}^{K_1} \|Z - Z_{[j]}^{(l)}\| \ \forall l \tag{A9}$$

where $Z_{[j]}^{(i)}$ is the jth nearest observation to Z in ω_i and $K_l \leqslant \sqrt{n}$. The decision rule of (A8) is simply: assign an unknown observation to the class for which the average distance to its K_l-nearest neighbors is minimum.

By definition of r_K, (A9) becomes

$$Z \epsilon \omega_i \text{ iff } r_{K_i} \leqslant r_{K_l} \qquad 1 \leqslant l \leqslant M, \tag{A10}$$

and since $\Phi(Z)$ varies inversely with r_K according to

$$\Phi(Z) = \frac{N\Gamma\left(\frac{N}{2}\right)}{2r_K^N \pi^{(N/2)}} \tag{A11}$$

where N is the dimensionality of the observation space and $\Gamma(\cdot)$ is the gamma function. We have

$$Z \epsilon \omega_i \text{ iff } \frac{K_i}{N_i \Phi_i(Z)} \geqslant \frac{K_l}{N_l \Phi_l(Z)} \ \forall l \tag{A12}$$

where the subscripts are class indices. Finally, from (A8) we have

$$Z \epsilon \omega_i \text{ iff } \hat{f}(Z|\omega_i) \leqslant \hat{f}(Z|\omega_l) \quad 1 \leqslant l \leqslant M \tag{A13}$$

which is exactly Bayes' rule, (A8).

Patrick [29] has observed that while distribution-free tolerance regions provide satisfactory density estimates, better estimates may be obtained if the tolerance regions are constructed in a way that takes into account the special properties of the data.

The peculiarities of the data with which we are most concerned are the small sample size and occasional artifacts introduced by the time alignment procedure. The effect of the small sample size is to make the variance of the estimator of (A3) large which, in turn, introduces classification errors. Averaging the distance to the K-nearest neighbors reduces the variance by a factor of $1/K$. The nonlinear time registration procedure sometimes forces the distance from a sample to an incorrect template to be uncharacteristically small. Clearly, the averaging operation will mitigate the adverse effects of pathologically small distances.

We have compared the decision rule of (A9) to the nearest neighbor rule of Loftsgaarden and Queensberry [26] and that of Patrick and Fischer [27] in which a sample is classified according to the distances of its kth nearest neighbor in each class. On test set TS1, our rule showed improvements of 6.6 and 8.5 percent over the Patrick and Fischer scheme for $K = 2$ and $K = 3$, respectively, and 4.6 and 1.2 percent improvements over the nearest neighbor rule for $K = 2$ and $K = 3$, respectively.

Acknowledgment

The authors wish to acknowledge the programming support and assistance of C. Schmidt for data input and analysis for the clustering phase of the system.

References

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

[2] 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 (abstr.), Nov. 1976.

[3] 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.

[4] 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. 563 (abstr.), Dec. 1977.

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

[6] J. N. Shearme and P. F. Leach, "Some experiments with a simple word recognition system," *IEEE Trans Audio Electroacoust.*, vol. AU-16, pp. 256–261, June 1968.

[7] B. Gold, "Word recognition computer program," Res. Lab Electron., Massachusetts Inst. Tech., Cambridge, Tech. Rep. 452, June 1966.

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

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

[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, "Application of clustering techniques to speaker-independent word recognition," to be published.

[12] L. R. Rabiner, B. S. Atal, and M. R. Sambur, "LPC prediction error–Analysis of its variation with the position of the analysis frame," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 434–442, Oct. 1977.

[13] H. Sakoe and S. Chiba, "A dynamic programming approach to continuous speech recognition," in *Proc. Int. Congress on Acoustics*, Budapest, Hungary, Paper 20 C-13, 1971.

[14] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 43–49, Feb. 1978.

[15] L. R. Rabiner, A. E. Rosenberg, and S. E. Levinson, "Considerations in dynamic time-warping algorithms for discrete word recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, Oct. 1978.

[16] M. R. Sambur and L. R. Rabiner, "A statistical decision approach to the recognition of connected digits," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 550–558, Dec. 1976.

[17] J. T. Tou and R. C. Gonzalez, *Pattern Recognition Principles*. Reading, MA: Addison-Wesley, 1974.

[18] E. A. Patrick, *Fundamentals of Pattern Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1972.

[19] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[20] J. Mac Queen, "Some methods for classification and analysis of multivariate data," in *Proc. 5th Berkeley Symp. Probability and Statistics*, Berkeley, CA, 1967.

[21] G. H. Ball and D. J. Hall, "Isodata–An iterative method of multivariate analysis and pattern classification," in *Proc. IFIPS Congress*, 1965.

[22] J. M. Tribolet and L. R. Rabiner, "Statistical properties of the log likelihood ratio for LPC coefficients," to appear in *IEEE Trans. Acoust., Speech, Signal Processing*, 1979.

[23] M. R. Sambur and L. R. Rabiner, "A speaker-independent digit-recognition system," *Bell Syst. Tech. J.*, vol. 54, pp. 81–102, Jan. 1975.

[24] D. A. S. Fraser, *Nonparametric Methods in Statistics*. New York. Wiley, 1957.

[25] D. O. Loftsgaarden and C. P. Queensberry, "A nonparametric estimate of a multivariate density function," *Ann. Math. Statist.*, vol. 36, 1965.

[26] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. IT-13, p. 21–27, Jan. 1967.

[27] E. A. Patrick and F. P. Fischer, "A generalized K-nearest neighbor rule," *Inform. Contr.*, vol. 16, 1970.

[28] S. S. Wilks, "Determination of sample sizes for setting tolerance limits," *Ann. Math. Statist.*

[29] E. A. Patrick, *Fundamentals of Pattern Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1973, 12, 1941.