Speaker Independent Recognition of Isolated Words Using Clustering Techniques

L. R. Rabiner

S. E. Levinson

A. E. Rosenberg

J. G. Wilpon

Acoustics Research Department
Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

A speaker independent, isolated word recognition system is proposed 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 data base consisting of 100 replications of each word (i.e. once by each of 100 talkers). The recognition system, which uses telephone recordings, 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 to lower the probability of error. Results are presented on two test sets of data which show error rates that are comparable to, or better than, those obtained with speaker trained, isolated word recognition systems.

I. Introduction

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. An earlier investigation [1] stimulated much of the research reported here.

II. Word Recognition System

Figure 1 shows a block diagram of the word recognition system [2]. 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. 1 is endpoint detection to determine points in time at which the unknown word begins and ends. The major causes of errors 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.

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

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

where a value of $a = 0.95$ was used in our simulations. Extensive experimental evidence has shown that pre-emphasis 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.

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 msec) using a Hamming window on the data. A total of 67 frames per second (i.e. every 15 msec) 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.

2.1 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) [3-5].

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 [2].

For recognition systems in which multiple reference templates are used for each word, the decision rule can be made more sophisticated than a simple nearest neighbor rule. 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 k^{th} nearest neighbor of the j^{th} word to the unknown sample x as $D[x,x_{[k]}^{(j)}]$, then for the KNN rule we compute the quantity r_j defined as

$$r_j = \frac{1}{K} \sum_{k=1}^{K} D[x,x_{[k]}^{(j)}] \tag{2}$$

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

$$r_{j^*} \leqslant r_j \quad j = 1, 2, \ldots, J \tag{3}$$

The quantity r_j of Eq. (2) is monotonically related to an estimate of the local probability density function of the j^{th} word.

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.

III. Clustering Algorithms

A series of four procedures was used to cluster a vocabulary of isolated word data [6]. These were:

1. The chainmap [7]
2. The shared nearest neighbor procedure [7]
3. The k-means iteration [8]
4. Isodata [9]

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 [6]. 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. A schematic representation of the use of the clustering procedures is shown in Fig. 2.

CH1379-7/79/0000-0574$00.75

To test the recognition and clustering algorithm a 39 word vocabulary 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].

IV. Recognition Results

Two distinct test sets of data were generated. We denote the individual test sets as TS1 and TS2. 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 data base 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.

4.1 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 (i.e. the standard dynamic time warping algorithm) 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 1 and Figure 3. 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 (I), and the number of ordered candidates that were considered (C). (The use of a list of recognition candidates is important when the recognition system is imbedded in a particular task in which other information is available.) The word templates were chosen in descending order based on the size of the cluster i.e. the $I = 1$ template was the cluster center of the largest cluster, the $I = 2$ 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. 3a. It is seen that the recognition accuracy is about 61% for $K = 1$ and $I = 2$, and about 51% for $K = 2$ and $I = 2$. As I increases, the $K = 2$ and $K = 3$ nearest neighbor rules yield higher recognition accuracies than the $K = 1$ or $K = 4$ rules. For $I = 12$ templates per word (the most used in our tests) the final recognition accuracy (for $C = 1$) was 79% for the $K = 2$ rule, and from 3 to 5% lower for the other rules.

Similar behavior of the curves of recognition accuracy versus I (for different K values) is seen for the $C = 2$ top candidates (Fig. 3b), and for the $C = 5$ top candidates (Fig. 3c). For the best 2 candidates, the recognition accuracy goes from about 75% (for the $K = 1$ rule) to 89% (for the $K = 2$ rule) as the number of templates per word goes from 2 to 12. For the top 5 candidates, the highest accuracy goes from 88% to 98.5% for a similar range of I.

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

4.2 Digit Recognition Results

Since the digits (zero to nine) were a subset of the 39 word vocabulary, it was a simple matter 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 utterances which 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% ($K = 1$), 98.2% ($K = 2$), 98.1% ($K = 3$), and 97.9% ($K = 4$). For the top 2 candidates ($C = 2$), the accuracies were within 0.1% of 99.6% for all 4 values of K. For the 10 talkers not in the original training set, the accuracy was 97% for $C = 1$, $K = 1$, 100% for $C = 1$, $K = 2$ and $K = 3$, and 98% for $C = 1$, $K = 4$.

V. Discussion

The goals of this investigation included:

1. To investigate a supervised algorithm for clustering words using only accumulated LPC distances between words.
2. To study a novel decision rule which was linked to a multiple template (cluster) representation of the vocabulary words.

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:

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.
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.
4. High accuracies were obtained (98.2%) for speaker independent digit recognition.
5. 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.

5.1 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). An analysis of the confusions shows that the vast majority of confusions occur within classes of high acoustical and phonetic similarity. We have identified 6 such classes, namely:

1. The set of i sounds - b, c, d, e, g, p, t, v, z, 3
2. The set of el 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% of the errors occurred *within* each of the 6 classes, with the majority occurring within class 1. An error rate of less than 2.5% is obtained for the remaining 11 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.

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.

REFERENCES

1. L. R. Rabiner, "On Creating Reference Templates for Speaker Independent Recognition of Isolated Words," IEEE Trans. on Acoustics, Speech, and Signal Proc., Vol. ASSP-26, No. 3, pp. 34-42, Feb. 1978.
2. F. Itakura, "Minimum Prediction Residual Applied to Speech Recognition," IEEE Trans. Acoustics, Speech, and Signal Proc., Vol. ASSP-23, pp. 67-72, Feb. 1975.
3. H. Sakoe and S. Chiba, "A Dynamic Programming Approach to Continuous Speech Recognition," Proc. Int. Congress on Acoustics, Budapest, Hungary, paper 20 C-13, 1971.
4. H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," IEEE Trans on Acoustics, Speech, and Signal Proc., Vol. ASSP-26, No. 1, pp. 43-49, Feb. 1978.
5. L. R. Rabiner, A. E. Rosenberg, and S. E. Levinson, "Considerations in Dynamic Time Warping Algorithms For Discrete Word Recognition," IEEE Trans. on Acoustics, Speech, and Signal Proc., Vol. ASSP-26, No. 5, Oct. 1978 (to appear).
6. S. E. Levinson, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "Application of Clustering Techniques to Speaker Independent Word Recognition," submitted for publication.
7. E. A. Patrick, *Fundamentals of Pattern Recognition*, Prentice-hall, 1972.
8. J. Mac Queen, "Some Methods for Classification and Analysis of Multivariate Data," Proc 5th Berkeley Symposium on Probability and Statistics, Berkeley, CA, 1967.
9. G. H. Ball and D. J. Hall, "Isodata - An Iterative Methods of Multivariate Analysis and Pattern Classification," Proc. IFIPS, Congress, 1965.

C = 1 (Top Candidate)

	K = 1	*K = 2*	*K = 3*	*K = 4*
I	*CE2-1*	*CE2-1*	*CE2-1*	*CE2-1*
2	61.2	51.6		
4	69.2	69.4	60.9	58.3
6	68.4	71.2	68.1	61.2
8	72.3	75.6	75.1	73.0
10	73.3	77.6	75.3	74.0
12	74.6	79.2	75.8	73.3

C = 2 (2 Top Candidates)

	K = 1	*K = 2*	*K = 3*	*K = 4*
I	*CE2-1*	*CE2-1*	*CE2-1*	*CE2-1*
2	75.3	66.3		
4	82.5	82.0	74.5	69.4
6	84.1	83.3	80.7	74.0
8	85.4	87.9	85.9	83.3
10	87.4	87.9	86.6	86.1
12	87.4	89.0	87.4	85.9

C = 5 (5 Top Candidates)

	K = 1	*K = 2*	*K = 3*	*K = 4*
I	*CE2-1*	*CE2-1*	*CE2-1*	*CE2-1*
2	88.2	83.3		
4	93.6	92.8	88.4	83.3
6	95.6	92.6	91.3	88.4
8	96.7	96.1	95.1	91.5
10	97.4	97.7	97.4	95.6
12	97.9	98.5	96.7	94.9

Table 1

Recognition Accuracies (%) for Clusters from the CE2-1 Algorithm

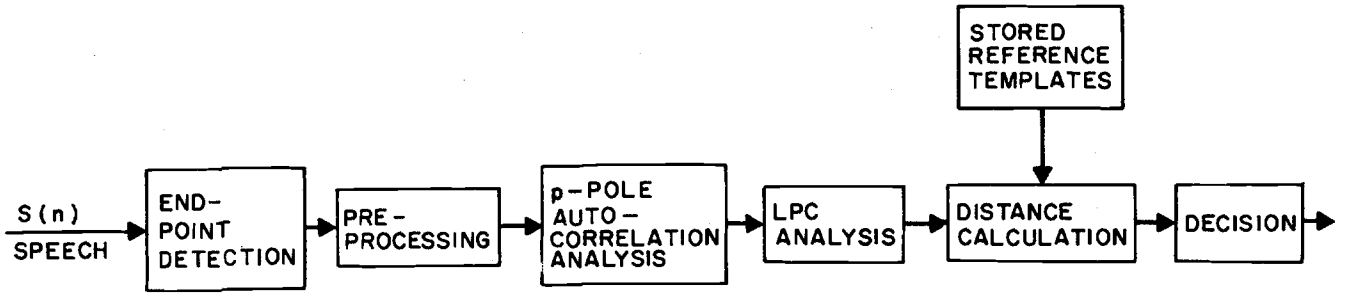

Fig. 1. Block diagram of the word recognition system.

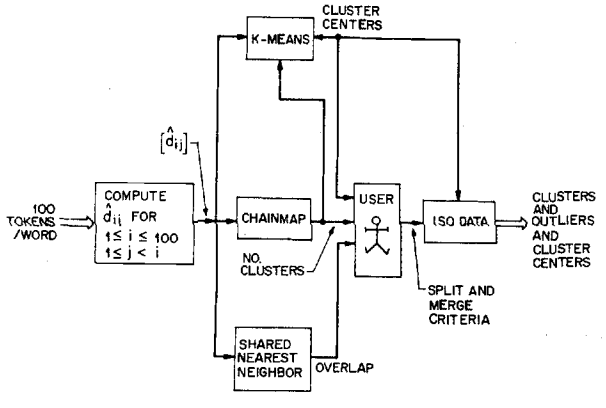

Fig. 2. The clustering procedures.

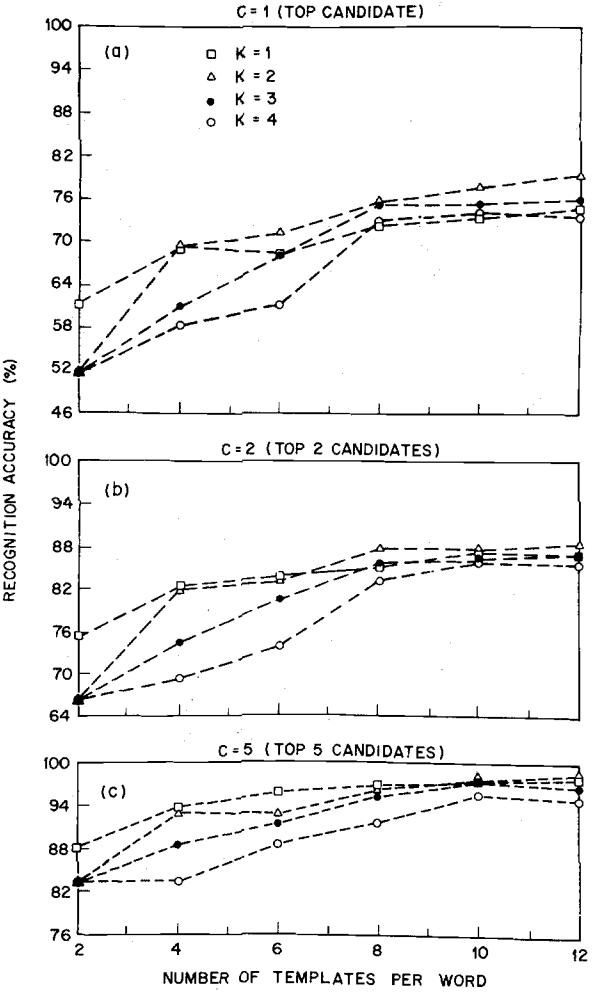

Fig. 3. Recognition accuracy (%) as a function of the number of templates per word for the CE2-1 clusters with $K = 1, 2, 3, 4$, and $C = 1, 2$ and 5.