On Creating Reference Templates for Speaker Independent Recognition of Isolated Words

LAWRENCE R. RABINER, FELLOW, IEEE

Abstract—The three aspects of a statistical approach to a pattern recognition problem are the selection of features, choice of a measure of similarity, and a method for creating the reference templates (patterns) used in the statistical tests. This paper discusses a philosophy for creating reference templates for a speaker independent, isolated word recognition system. Although there remain many unanswered questions both about how to select appropriate features for recognition, and how to measure similarity between sets of features, such issues are not discussed here. Instead we concentrate on methods for creating the reference templates. In particular, a method of combining word patterns from a number of speakers is proposed in which a clustering type of analysis is used to determine which patterns are merged to create a word template. The creation of multiple templates, based on this method, is discussed and is shown to be of substantial value for as few as eight speakers in the training set. To test the ideas proposed here, a 54 word vocabulary word recognition system was implemented. All input words were recorded off a standard telephone line. The features used were the LPC coefficients of an 8-pole analysis, and the simple Itakura distance measure was used to measure similarity between patterns. With word templates obtained as described above, recognition accuracies of 85 percent were obtained in a forced choice recognition test on the 54 word vocabulary using eight new speakers. The correct word was within the top five choices 98 percent of the time. Using a strategy in which all the training words were used to create the templates, the recognition accuracy fell to 77 percent, and the correct word was within the top five choices only 89 percent of the time.

INTRODUCTION

ALTHOUGH a fairly large number of systems have been proposed for recognizing isolated words, [1]-[11] there remains a substantial number of unanswered questions about many aspects of such systems. This paper deals with one of these questions—namely the problem of how to create reference templates (patterns) from a large number of feature sets, for a speaker independent word recognition system.

To put this problem in its proper perspective requires some discussion of the framework of a general pattern recognition system. As shown in Fig. 1, there are three considerations in implementing a pattern recognition system—namely feature selection and measurement, selection of a pattern similarity measure, and creation of reference templates for recognition. The problem of feature selection has been investigated in a number of studies [1]-[11]. A wide variety of features have been used including time domain measurements such as energy, zero crossings, bandpass filter outputs; frequency domain measurements such as spectral coefficients, cepstral coefficients, spectral derivative; and most recently, LPC parameters (or suitable transformations of them). The rationale for choosing a feature set is related both to the information

Manuscript received July 6, 1977; revised October 3, 1977.
The author is with Bell Laboratories, Murray Hill, NJ 07974.

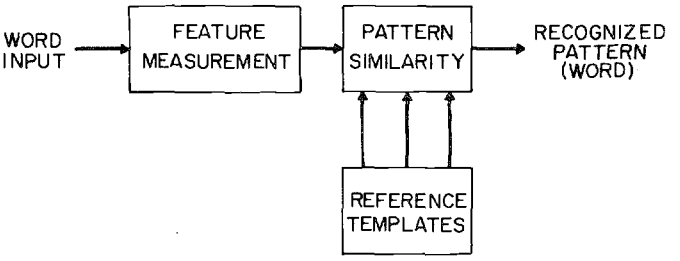

Fig. 1. Block diagram of overall word recognition scheme.

(about the patterns to be recognized) contained in the features, and the efficiency with which the features can represent this information. For the work to be discussed in this paper, the features selected were the LPC parameters for each frame of the input speech. This set of features (representation of speech) was chosen because of the fidelity with which the LPC parameters are capable of representing the speech waveform (as demonstrated in a number of LPC vocoders [12]-[15]), and because of the numerous theoretical interpretations of the LPC parameters in terms of spectral matching [16], vocal tract area functions [17], etc. The "ideal" set of features for speech recognition would, of course, be articulatory features describing the positions of the tongue, lips, velum, jaw, etc., as a function of time. From such "ideal" features the problems of recognizing words are substantially reduced due to the high degree of physical interpretation that can be applied to the feature measurements. Some preliminary attempts at the measurement of articulary features for recognition were made by Hafer and Coker [18]; however, much basic research remains to be done before any practical consideration is given to the use of articulatory features for recognition.

Once the set of features is chosen, the next important step is the choice of a pattern similarity measure which quantitatively tells how close a reference template is to the unknown spoken word. The choice of a similarity measure is intimately related to the chosen set of features. For the LPC coefficients, Itakura [6] has proposed a very powerful measure of similarity based on assumed statistics for the LPC parameter sets. For LPC sets a, and $\hat{a}$, the similarity measure proposed by Itakura is of the form

$$D(a, \hat{a}) = \log \left[\frac{aRa^t}{\hat{a}R\hat{a}^t} \right] \tag{1}$$

where R is the matrix of autocorrelations of the speech frame with LPC set $\hat{a}$. A computationally efficient form of (1) was also proposed by Itakura of the form

$$D(a, \hat{a}) = \log (a \cdot a) + \log [(br)/(\hat{a}r)] \tag{2}$$

0096-3518/78/0200-0034$00.75

where $(x \cdot y)$ is the inner product of the vectors x and y, r is the normalized correlation, and b is related to the autocorrelation coefficients associated with the inverse filter of the all-pole model, i.e.,

$$b(i) = 2 \sum_{j=0}^{p-i} a(j)a(j+i)/(a \cdot a). \tag{3}$$

A closely related distance measure to that of (1) is

$$D(a, \hat{a}) = (a - \hat{a}) \left[\frac{NR}{\hat{a}R\hat{a}^t} \right] (a - \hat{a})^t \tag{4}$$

where N is the effective frame size used for obtaining the LPC set of $\hat{a}$. As shown by Sambur and Rabiner [19], a somewhat more powerful distance measure can be obtained by postulating a statistical characterization of the reference vector a, leading to the distance

$$D(\hat{a}, a) = (\hat{a} - m_a)\, \Lambda^{-1} (\hat{a} - m_a)^t \tag{5}$$

where m_a is the mean value of vector a, and

$$\Lambda = S_a + R^{-1} \frac{(\hat{a}R\hat{a}^t)}{N} \tag{6}$$

where S_a is the covariance matrix of the a's. Computationally, (5) is two orders of magnitude more complicated than (2) for measuring pattern similarity. However, it has been shown to be a more powerful measure of similarity in cases where S_a, the covariance matrix of the a's is not negligible [19], e.g., connected digit recognition with large amounts of coarticulation between adjacent digits.

There have been leveled many criticisms of the distance measures of (2) and (5), notably that they do not satisfy the necessary properties of a true metric. That is,

$$D(a, \hat{a}) \neq D(\hat{a}, a). \tag{7}$$

Furthermore, the derivation of the distance measure is based on a log likelihood measure assuming that the residual $\hat{a}R\hat{a}^t$ is the true minimum residual for the test frame. However, de Souza [20] has shown that if the residual $\hat{a}R\hat{a}^t$ is an *estimate* of the true residual (as it must be), then the distance D is not chi-square distributed with p degrees of freedom (p is the order of the LPC system), and thus, D is unsound as a test statistic.

In spite of these objections, the distance measures of (2) and (5) have been used in a wide variety of applications involving LPC parameters [21]-[25] with fairly good success. Alternative distance measures based on using LPC parameters are the cosh and cepstral distance metrics proposed by Gray and Markel [26]. All these LPC distance measures have been shown to correlate well with log spectral differences between the frames, and as such are good measures of spectral similarity between speech frames.

The third and final aspect of a pattern recognition system for isolated words is the creation of the reference templates for use in the recognition phase [27]. For systems which are trained to an individual speaker, this aspect of the system is fairly trivial in that repetitions of the vocabulary words can be combined in any number of simple ways and the resulting reference template will generally be quite satisfactory for most cases. This is because the variance between repetitions of the same word by the same speaker is relatively small, and thus simple averaging of the time normalized patterns of each word is a reasonable way of obtaining reference templates.

For systems which are intended to be speaker independent, the creation of reference templates is a far more critical problem. The issue here is how to intelligently combine reference tokens by different speakers to form template(s) for recognition when the variance between different speakers is often quite large. Fig. 2 illustrates the inherent problems in a fairly simple way. Each data point (denoted by an x) represents a reference token in a multidimensional feature space.[1] Due to the variability between speakers, the reference tokens span a large area in the space. Clearly, simple averaging of "time-warped" reference tokens, at least in this simple case, will lead to a meaningless average reference template which will be a poor representation of the word from which it came. Instead the data of Fig. 2 suggest that a clustering analysis be used in which only those reference tokens which form a cluster be averaged to form a reference template. For this simple example three clusters are obtained, denoted by $C1$, $C2$, and $C3$. This implies that at *least* three reference templates are required to adequately represent the data of Fig. 2. In addition, there are two "outliers" tokens (denoted by A and B) which do not cluster with any of the three clusters. In theory, two additional reference templates are required to represent these tokens properly in the final reference set.

Based on the above discussion, it is seen that there are several key issues in the creation of an adequate set of reference templates for speaker independent recognition of words. These are:

1) how to recognize clusters,
2) how to average tokens within a cluster to create a reference template,
3) what parameter set should be averaged in creating the reference template,
4) how many templates should be created for each word in the input set, and
5) how to handle outliers.

In this paper, we present some results on possible ways of handling the problems discussed above, and give numerical results on word recognition accuracy which illustrates typical performance achieved with one representative system.

II. Creation of Reference Templates

Assume we are given a training set of words. For each of the L words in the vocabulary, there is 1 replication of the word by J different speakers. For each word we are given the LPC parameter set $a(i, l, j)$ where

i = frame number, $\quad i = 1, 2, \cdots, I(l, j)$
l = word number, $\quad l = 1, 2, \cdots, L$
j = replication number, $\quad j = 1, 2, \cdots, J$

with $I(l, j)$ the number of frames of the jth speaker (replica-

[1] Time has been eliminated in this figure by using dynamic time-warping procedures and assigning a set of dimensions to each time aligned frame. Later we will discuss a practical procedure for time aligning the references.

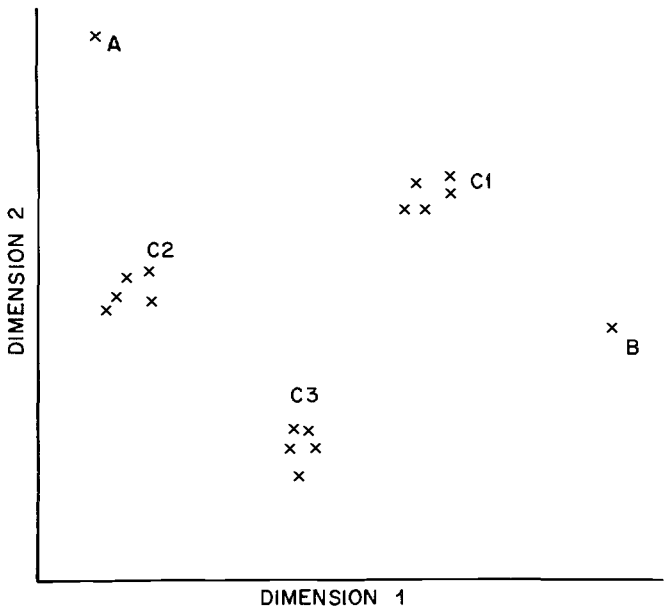

Fig. 2. Example showing clustering of reference tokens into three clusters ($C1$, $C2$, $C3$) with outliers A and B.

tion) of the lth word. In order to combine the j versions of the lth word, each word must be time-warped[2] to a standard duration which we denote as $I'(l)$. Thus, we denote the warped LPC parameter sets as $a'(i', l, j)$ where i' goes from 1 to $I'(l)$.

The problem in creating reference templates is to find a suitable way of combining the J versions of each word to give one or more templates whose properties reflect those of the distribution of LPC parameters for that word. If we denote the reference template for the lth word as $\hat{a}(i', l)$, then the problem becomes one of choosing $\hat{a}(i', l)$ to satisfy some optimality criterion based on the given training set of data.

Using the similarity measure of (2) to define distance between two frames of speech with LPC coefficients a and $\hat{a}$, an overall distance between the jth version and the reference for the lth word is given as

$$D_T(a', \hat{a}, l, j) = \sum_{i'=1}^{I'(l)} D(a'(i', l, j), \hat{a}(i', l)). \tag{8}$$

Different criteria can now be used to determine $\hat{a}$. For example, by using all the available versions of the lth word, one might set up an optimization problem of the form

$$\text{choose } \hat{a}(i', l) \text{ such that } \max_j [D_T(a', \hat{a}, l, j)] \text{ is minimized}$$

i.e., the reference template is chosen so that the maximum distance to any of the J reference tokens is minimized. (Clearly the same criterion can be used on a subset of the J reference tokens for the lth word.) This problem then becomes a minimax type of optimization.

A second possibility is to use an optimization criterion based on minimizing the probability of error using all the reference tokens in the training set. The ideas here are illustrated in Fig. 3. We assume that all reference tokens which represent the lth word are called set A, and all reference tokens which do not represent the lth word are called set B. For each token in each set we measure the quantity D_T of (8) to determine a probability density function for both sets of data. Typical examples of such probability density functions are shown in Fig. 3. Each of the density functions is a function of the reference template $\hat{a}$, and the criterion for choosing $\hat{a}$ would be to minimize the probability of error, which is computed simply as

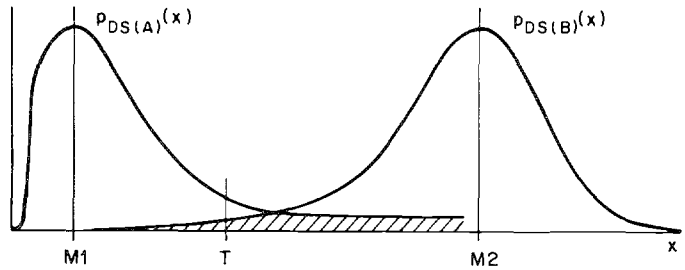

Fig. 3. Hypothetical probability density functions for the overall distance function for both the correct references and for all other words.

$$P(E) = \int_0^\infty \left[\int_T^\infty P_{DS(A)}(x)\,dx\right]\left[\int_{-\infty}^T P_{DS(B)}(x)\,dx\right] dT \tag{9}$$

where it is assumed that $P_{DS(A)}$ and $P_{DS(B)}$ are independent. In (9), T represents a tradeoff threshold between the two types of errors, namely, not recognizing the correct word, and incorrectly recognizing the wrong word.

Fig. 3 also suggests a simple way of reducing the probability of error even further from the value of (7) by including the possibility of not making a decision on a test word if the measured distance to the reference template is greater than some value $T(l)$ which depends on the statistics of the individual reference templates. Thus, by setting a value of $T(l)$ sufficiently low, the probability of a false alarm (i.e., not recognizing the correct word) can be made arbitrarily small; however, at the same time the probability of a miss (i.e., an incorrect word having a distance below the threshold) increases.

Based on the above discussion it would seem that fairly straightforward techniques can be used to solve the reference template problem. However, this is generally not the case in practice for several reasons. The major one is that the J versions of each word do not generally form a single stable cluster whose statistics are well behaved. Instead, their behavior is as described in Fig. 2, i.e., multiple clusters are formed with the possibility of one or more outliers which do not fit well into any cluster. As such, the major problem in creating reference templates is not the criterion for choosing the LPC reference from the reference tokens, but instead deciding on the appropriate subset of the J versions to be used in creating a template. The next problem is deciding on the number of templates to be used for each word in the vocabulary. Finally comes the problem of choosing an appropriate criterion for merging the individual tokens to form the reference template.

Rather than continuing on the theoretical discussion of how

[2]Dynamic programming is used for time warping throughout this paper.

to create reference templates, we now present the specific algorithm used in this paper. A flowchart of the algorithm is given in Fig. 4.[3] The procedure is an iterative one in which initially all replications of the lth word are used, and then those which fall outside the cluster [as determined by a distance measurement similar to (8)] are eliminated from the current cluster. An initial estimate of the template $\hat{a}_{(0)}$ is made by choosing the replication which is closest to the average length of the J versions. An updated estimate $\hat{a}_{(n)}$ is obtained by averaging the LPC sets (or suitable transformations of them such as the PARCOR coefficients) of the warped replications, i.e.,

$$\hat{a}(n) = \sum_{j} a'(i', l, j) \tag{10}$$

and the convergence criterion is when

$$D(\hat{a}_{(n)}, \hat{a}_{(n-1)}) < \delta \tag{11}$$

where δ is some small positive quantity. (To avoid cases in which no convergence was obtained, a maximum iteration count of 10 was used.) Following convergence, the distance between the final reference and each of the replications used to create the reference is computed. If the distance to any replication falls outside the cluster threshold,[4] this replication is removed from the cluster and the entire procedure is repeated to obtain a new reference template.

Once a stable reference template is obtained, all replications not used to create the template are used as the replication pool for determining another cluster from which another reference template is created. This procedure could be extended until all J versions are included in some cluster. (In the work described here, a maximum of two reference templates, per word, was allowed. All replications not falling within the two main clusters were effectively discarded.)

As just described, the technique for determining the reference template $\hat{a}$ from the tokens in a cluster is a fairly unsophisticated one—namely simple averaging of the time-warped LPC sets (or an appropriate transformation) from all the tokens in the cluster. This procedure was used, rather than one of the more sophisticated ones mentioned above, for several reasons. The first is that since we do not know which tokens will eventually cluster, use of a minimax or minimum error probability criterion can, and often will, lead to extremely poor reference templates due to multiple clusters and outliers which dominate the computation. One possible solution to this problem would be to segment the J versions of each word into clusters (some possibly with a single token) *before* attempting to create reference templates. Although several possibilities exist for determining these clusters, a fairly simple and intuitive one would be to let each of the J versions be the assumed reference and compute the total distance from it to each of the remaining $J - 1$ replications. Based on a predetermined cluster distance, the first cluster would be the set for which the maximum number of tokens fell within the predetermined cluster distance. After eliminating these tokens from the training set, additional clusters could be obtained in a similar manner.

Another reason that the sophisticated objective criteria for determining the reference templates were not used is that although they are both physically and mathematically appealing, there is no simple way of implementing the computation so as to guarantee obtaining the optimum solution. Preliminary attempts at implementing the minimax solution led to limit cycle behavior in which the reference template cycled between two states corresponding to the extreme distances in the cluster set.

In summary, we have implemented a procedure for creating multiple reference templates for speaker independent recognition of isolated words in which clusters of reference tokens are obtained as a byproduct of the procedure, and for which a new template is obtained for each cluster. In the next section, we present experimental results which show both the statistics of the referencing algorithm and the resulting performance of the overall recognition system for both single and double templates for each word in the recognition set.

III. Isolated Word Recognition Results

Fig. 5 shows a block diagram of the overall word recognition system. The input words were recorded off a standard telephone line, filtered between 200 and 3000 Hz, and sampled at a 6.67 kHz rate. The recording interval for each word was 2 s and the endpoint analysis made preliminary estimates of the beginning and end of the words [31]. A voiced-unvoiced-silence (VUS) analysis was then performed to obtain refined estimates of the endpoints of the utterance [32]. The purpose of the analysis was to extend the initial boundary to include weak fricatives (such as *f*, *th*) at the beginning of the utterance, and to identify stop bursts at the end of the utterance, so as to eliminate them from consideration. Stop bursts were eliminated both because of their high variability, and because they were not always present. Thus, whenever a speaker released a final plosive, the most expedient thing to do was to adjust the endpoint not to include such bursts. (In the training set used to create the reference templates, the bursts were eliminated manually.)

Following VUS analysis, an LPC analysis was performed on the word. The speech was first preemphasized using a simple first order network of the form

$$H(z) = 1 - 0.95z^{-1}$$

and then an 8-pole LPC analysis was performed using the autocorrelation method with a Hamming window. A 30 ms frame size was used and the analysis was performed 67 times per second, i.e., 15 ms overlap between adjacent LPC frames.

The next block in Fig. 5 is the recognition computation in which the test word was dynamically time warped to the duration of each of the reference templates, and a distance of the form of (8) was computed between each reference and the test word. The recognized word was then selected as the

[3] This algorithm is similar in several respects to ones used by Lummis [28], Rosenberg and Sambur [29], and Rosenberg [30].

[4] The cluster threshold was determined empirically from the reference data. Its value was not critical in that tokens outside most clusters had distances significantly greater than this threshold.

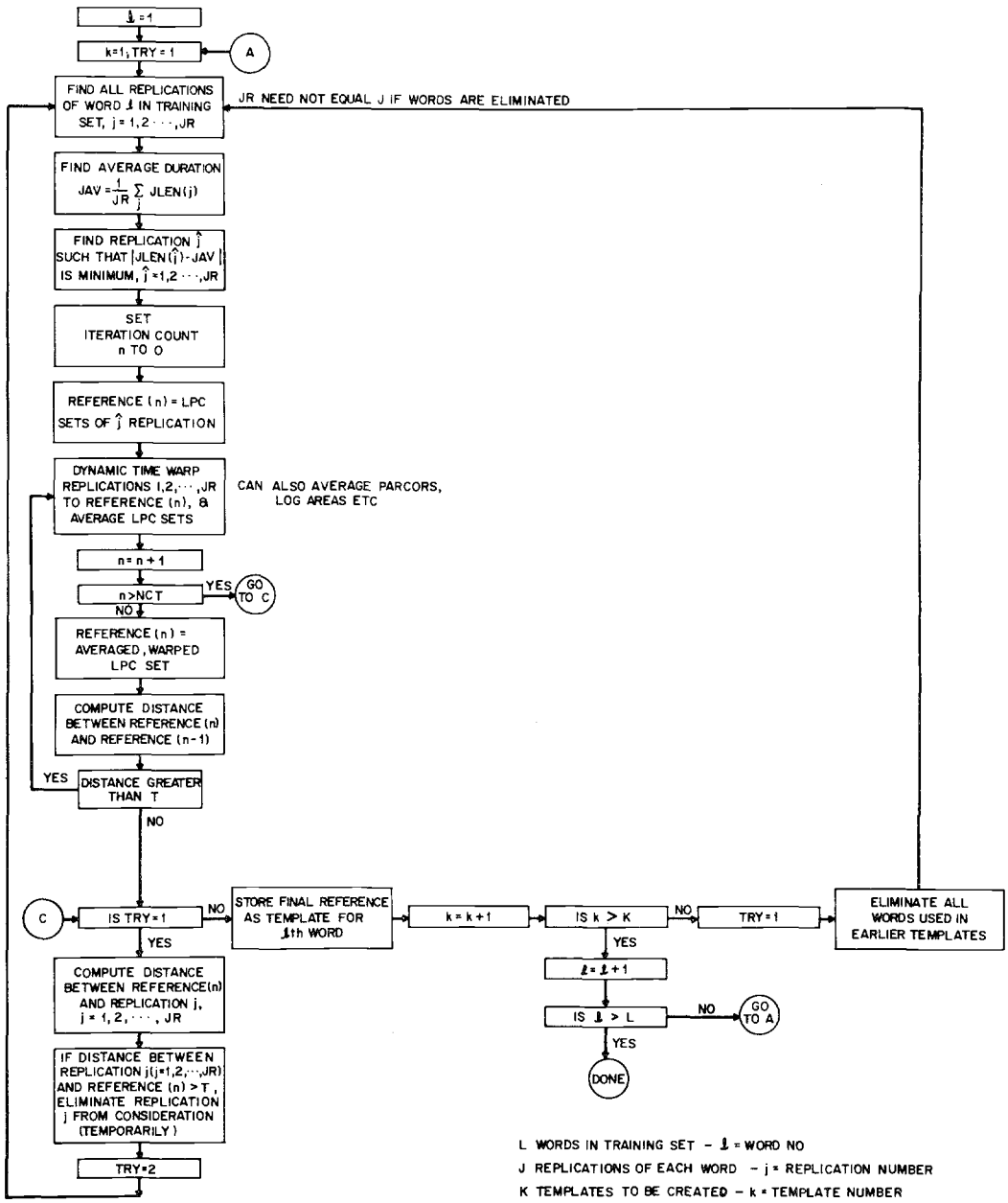

Fig. 4. Flow diagram of algorithm used to combine K replications of a reference word into clusters and to form one reference template per cluster.

reference which was closest (minimum distance) to the test word.

Before describing the experimental results obtained with this system, two comments should be made. Although the recognition computation is made solely on the basis of the computed LPC distances, some additional information about the correct word is available from the VUS analysis. Thus, if the distance computation gives a result which is inconsistent with the VUS analysis, there is feedback that either (or both) the results are in error. In any case, one can and should use such information in an ultimate word recognition system. Such information was not used here for a variety of reasons related to the purpose of the current work (i.e., to test template creation methods).

The second comment concerns the LPC distance. Sambur and Rabiner [7] have shown that a more powerful LPC distance measure that exploits the statistics of the LPC parameters themselves (i.e., covariances and means) can yield better scores then the simple LPC distance of (1) and (2). However, these more powerful distances are computationally inefficient and require considerably more training data than what was available here. As such, they were not used in these experiments.

To evaluate the various methods of creating speaker independent word templates, a series of experiments was carried out using the system of Fig. 5. The vocabulary, shown in Table I, was the 54 word computer vocabulary proposed originally by Gold [11]. For training the system, both a four speaker and an eight speaker training set were used. For testing the system, a new set of eight speakers was used. (For both the training and testing sets, an equal number of male and female speakers were used.)

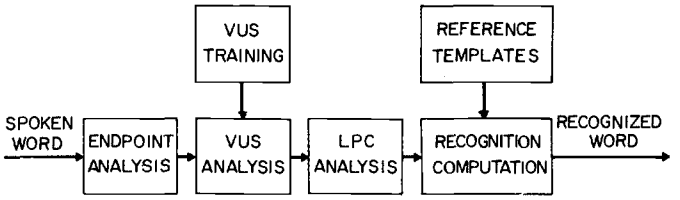

Fig. 5. Block diagram of the signal processing in the recognition algorithm.

For the four speaker training set, four distinct sets of reference templates were created. The four sets were the following:

Set 1–4 individual templates per word, corresponding to the 4 speakers in the training set.

Set 2–1 template per word using the algorithm of Fig. 4 with a distance threshold of infinity.

Set 3–1 or 2 templates per word (depending on how many words clustered in the first groups) based on averaging LPC coefficients using the algorithm of Fig. 4.

Set 4–1 or 2 templates per word based on averaging PARCOR coefficients and then transforming back to LPC coefficients.

For the eight speaker training set, only three distinct sets of reference templates were created (corresponding to Sets 2–4 above) since it was not feasible to store eight complete sets of reference templates. In creating reference Sets 3 and 4, a value for the cluster distance threshold (which determined which words went into a cluster) was obtained experimentally.

A complete breakdown of the results for each of the above conditions is given in Table II. Each of the eight test speakers spoke the entire word vocabulary once using a normal telephone handset. The spoken words were recorded at the receiver. The transmission system was that of a locally switched line through a local PBX. A different dialed-up line was used for each group of five words. Thus, a total of 11 dialed-up lines were used in testing each speaker. In Table II, the columns labeled E, C, $C(E)$, $T2$, and $T5$ correspond to:

E = number of errors made in one repetition of the 54 word vocabulary,

C = number of words in which the ratio between the next-to-minimum distance and the minimum distance fell below a threshold of 1.10,

$C(E)$ = the number of words in which the condition on C was met, and for which the chosen word was in error,

$T2$ = the number of times the correct word was not within the top two candidates based on the measured distances,

$T5$ = the number of times the correct word was not within the top five candidates.

The quantity E is an absolute measure of the accuracy of the recognition system for each of the different training sets. The quantities C and $C(E)$ are measures of separation between the candidates with the lowest and next lowest distances. As seen in Table II, about half the time that the condition on C was met (i.e., the first two choices were close in distance) an error was made in recognition. Thus, by using a threshold on the distances, a number of errors could be eliminated. At the same time a category of "no decision" would be the result a fairly large amount of the time. Finally, the quantities $T2$ and $T5$ are the error rates for the top two and five choices and these reflect the ultimate capabilities of the training and recognition system, since one would hope that in almost all cases, the correct word was within the top five choices.

TABLE I
WORDS IN THE VOCABULARY

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

By examining the data of Table II several interesting observations can be made. For the four speaker training set, the minimum error rate (11.3 percent) occurred for the training in which four individual templates were stored for each word. The error rate for the four template system was 4.8 percent after two tries, and 1.6 percent after five tries. In addition, the ambiguity rate (when the C condition was met) was 13.2 percent for this training set.

For the three other sets of training (for the four speaker case), the results in *all* the error categories were quite comparable, although somewhat larger than for the four-template method. The differences in scores between single and double templates were marginal for the E and $T2$ categories; however, the $T5$ scores for the double template approach with averaged PARCOR coefficients were about half as large as for the single template method. When the LPC coefficients were averaged in the double template method, no significant differences in scores were obtained for the $T5$ category.

For the eight speaker training set, a clearer picture emerges as to the importance of creating multiple templates. As seen at the bottom of Table II, the scores in *every* category for the double template reference were significantly lower than for the single template reference. This result shows that as more speakers are added to the training set, the distribution of LPC parameters tends to have larger and larger variance; as such any attempt at combining *all* the reference data into a single composite template will have great difficulty.

In comparing the recognition scores obtained from the eight speaker training set to those of the four speaker training set (using a double template for both cases) it is seen that no significant differences existed in any of the error categories.

Two key questions arise from an analysis of the results given in Table II. The first relates to the required number of templates to adequately characterize a given size population of speakers. The main point is how to handle speakers who do not fall into the main clusters which are used to create the reference templates. Since such speakers are inherently a

TABLE II
OVERALL RECOGNITION RESULTS FOR FOUR- AND EIGHT-SPEAKER TRAINING SETS

	Training Set 1					Training Set 2				
Test Speaker	E	C	C(E)	T2	T5	E	C	C(E)	T2	T5
1	3	6	1	1	0	7	8	3	1	0
2	5	2	0	2	0	5	5	1	1	0
3	6	11	5	2	1	5	9	5	4	1
4	3	5	2	0	0	6	14	5	3	1
5	14	16	8	7	4	6	10	5	1	0
6	8	9	3	4	2	19	15	8	12	6
7	5	5	4	3	0	4	9	4	1	0
8	3	3	1	2	0	6	5	3	3	1
Totals	49	57	24	21	7	58	75	34	26	9
Per Cent	11.3	13.2	5.6	4.8	1.6	13.4	17.4	7.9	6.0	2.1

	Training Set 3					Training Set 4				
Test Speaker	E	C	C(E)	T2	T5	E	C	C(E)	T2	T5
1	9	8	4	5	1	6	7	2	3	0
2	8	6	3	1	0	7	8	4	1	0
3	5	6	4	4	0	3	9	3	2	0
4	7	12	4	4	1	9	10	4	6	0
5	7	10	5	2	0	6	14	4	2	0
6	19	24	11	13	10	20	18	14	13	5
7	3	6	2	1	0	6	8	3	1	0
8	6	5	3	2	0	5	4	2	2	0
Totals	64	77	36	32	12	62	78	36	30	5
Per Cent	14.8	17.8	8.3	7.4	2.8	14.4	18.1	8.3	6.9	1.2

(a) Results for 4-Speaker Training Set

	Training Set 2					Training Set 3				
Test Speaker	E	C	C(E)	T2	T5	E	C	C(E)	T2	T5
1	9	10	5	7	5	4	4	1	1	0
2	9	9	4	8	5	4	5	2	1	0
3	12	8	5	11	8	6	10	4	2	1
4	14	15	11	9	5	7	11	6	4	1
5	11	12	7	6	6	8	7	4	5	1
6	18	15	13	13	8	15	12	8	8	5
7	14	12	9	8	6	10	10	7	4	1
8	12	10	8	6	5	9	7	5	4	0
Totals	99	91	62	68	48	63	66	37	29	9
Per Cent	22.9	21.1	14.4	15.7	11.1	14.6	15.3	8.6	6.7	2.1

	Training Set 4				
Test Speaker	E	C	C(E)	T2	T5
1	5	7	3	2	2
2	2	7	1	0	0
3	8	10	6	2	1
4	13	14	8	10	3
5	9	7	4	5	1
6	15	11	8	10	8
7	12	7	5	8	3
8	5	8	3	2	0
Totals	69	71	38	39	18.
Per Cent	15.9	16.4	8.8	9.0	4.2

(b) Results for 8-Speaker Training Set

TABLE III
TRAINING STATISTICS USING EIGHT-SPEAKER TRAINING SET USING AVERAGED LPC COEFFICIENTS (SET 3)

	Training Speaker No.									
Word No.	1	2	3	4	5	6	7	8	N1	N2
1	1	2	1	1	1	1	1		6	1
2	2		1		1	1			3	1
3	1	1	1		1	1	1	2	6	1
4	1	1	1		1	2	1		5	1
5	2				1				1	1
6			1	2	1	2	1		3	2
7	2		2	1	1	1	1		4	2
8	1	1	1	1	1	1	1	1	8	0
9	1	2	1	1	1	1	1	1	7	1
10	1		1	1	1	1	1	2	6	1
11			1		2	1	1		3	1
12	2	2	1	1	1	1	1	1	6	2
13	1	2			1		1		3	1
14	1	1	1	1	1	1	1	1	8	0
15	1	2	1	1	1	1	1	2	6	2
16	1	2	1	1	1	1	1		6	1
17	1		1	1	1	1	1	2	6	1
18	1	2	1		1	1	1		5	1
19	1	1	1	1	1	1	1	2	7	1
20	1	2	1	1	1	1	1	1	7	1
21	1	1	1	1	1	1	1	1	8	0
22	1	1	1	1	1	1	1	2	7	1
23	2		1		1	1	1		4	1
24	2		1	1	2	1	1	2	4	3
25	1	1	1	1	1	1	1	1	8	0
26	1	2	1	1	1	1	1	1	7	1
27	1	2	1	1	1	1	1	2	6	2
28	1		1	1	2	1	2		4	2
29	1	2	1	1	1	1	1		6	1
30	2		1	1	1	1	1	2	5	2
31	1	1	1	1	1	1	1	1	8	0
32	1	2	1	2	1	1	1		5	2
33	1	1	1	2	1	1	1		6	1
34	1	1	1	1	1	1	1	2	7	1
35	1	1	1	1	1	1	1	1	8	0
36	2	1	1	1		1	1	2	5	2
37	2	1	1	1	1	1	1	2	6	2
38	1	1	1	2	1	1	1	2	6	2
39	1	1	1	1	1	1	1	1	8	0
40	2					1			1	1
41			1	2	1	1	1		4	1
42	2	1	1	1	1	1	1	1	7	1
43		2	1		1	1	1		4	1
44	1	1	1	1	1	1	1	2	7	1
45			1	1	1	1	1	2	5	1
46	2	1	1		1		1		4	1
47	2		1	1	1	1	1		5	1
48		2	1	1	1	1	1	2	5	2
49	2	1	1	1	1	1	1	2	6	2
50	2	1		1	1		1		4	1
51	1	1	1	1	1	1	1	2	7	1
52		2	1		1	1	1		4	1
53	1	1	1	1	1	1	1	2	7	1
54	2	1	1	1	2	1	1	1	6	2
S1	31	24	49	37	48	48	50	13		
S2	16	15	1	5	4	2	1	19		
SN	7	15	4	12	2	4	3	22		

large distance from the reference templates, the probability of correct identification in a test environment is not generally high. The second question concerns the correlation between the words in which errors were most often made and the statistics of the clustering in the training set for those words. The point here is whether or not a high error rate on a word is related to the way in which the training tokens of that word clustered together.

Some partial answers to these questions are provided in the data in Tables III and IV. Table III shows the statistics of the clustering for each word in the vocabulary for the eight speaker training set using averaged LPC coefficients. For a reference speaker, the number 1 corresponds to inclusion in the first (largest) cluster and the number 2 corresponds to inclusion in the second cluster. The columns labeled $N1$ and $N2$ indicate the number of tokens (out of a maximum of eight) in the first and second clusters, respectively. The rows (at the bottom) labeled $S1$, $S2$, and SN indicate the total number of times the speaker was included in the first ($S1$), second ($S2$), or no cluster (SN).

From Table III it can be seen that for about half the words (24 of 54), all eight tokens of a test word were included in the two reference templates. The statistics for the remaining 30 words show that:

1) for 12 words, 1 token was not included in either template,
2) for 6 words, 2 tokens were not included in either template,
3) for 7 words, 3 tokens were not included in either template,
4) for 3 words, 4 tokens were not included in either template, and
5) for 2 words, 6 tokens were not included in either template.

(The two words for which no clustering was possible were "read" and "exchange".) The overall statistics of Table III

TABLE IV
TESTING STATISTICS USING EIGHT-SPEAKER TRAINING SET WITH AVERAGED LPC COEFFICIENTS (SET 3)

	Test Speaker No.											
Word No.	1	2	3	4	5	6	7	8	TX	TC	TXC	T
1				C						1		1
2				X		X	XC		2		1	3
3							XC	X	1		1	2
4			X						1			1
5			C					X	1	1		2
6		C				XC		X	1	1	1	3
7			C			X	XC		1	1	1	3
8												0
9												0
10								C		1		1
11		C								1		1
12								C		1		1
13						XC					1	1
14												0
15												0
16						X			1			1
17			XC					XC			2	2
18						X			1			1
19				XC			X		1		1	2
20	X			X	X		X	XC	4		1	5
21												0
22	C			C		X	C		1	3		4
23								X	1			1
24	XC					X			1		1	2
25												0
26			C		X	C			1	2		3
27			C				XC			1	1	2
28		C			C	XC				2	1	3
29							X		1			1
30												0
31			C	XC	X	C			1	2	1	4
32					X		C		1	1		2
33				C						1		1
34			XC				X	X	2		1	3
35												0
36	C				X	X			2	1		3
37					C					1		1
38												0
39				XC				XC			2	2
40	C	XC	XC	C		C				3	2	5
41				X	XC				1		1	2
42						X			1			1
43					C					1		1
44			XC								1	1
45							C			1		1
46												0
47		XC				X		X	2		1	3
48	X	X	X						3			3
49				X	X	XC	XC		2		2	4
50	X	X			XC	X	X		4		1	5
51												0
52			C	C		C				3		3
53												0
54												0

show that seven of the eight speakers were in reference templates for 39 or more of the 54 words, and that the eighth speaker was represented in 32 of the 54 words.

Table IV gives a complete breakdown of the errors from the eight speaker, double template training set obtained from averaging LPC coefficients. An *X* in a column indicates that an error was made in recognizing the word; a *C* indicates that, although no error was made, the ratio of the distances between the second and first candidates exceeded the threshold of 1.1; and an *XC* indicates that an error in recognition was made *and* that the ratio of distances exceeded the threshold. The columns labeled *TX*, *TC*, *TXC*, and *T* indicate the total number of row entries with an *X*, a *C*, an *XC*, and the sum total of all entries, respectively.

The statistics of the word errors (based on the total (*T*) column) show that six of the 54 words had four or five total error or close entries. Of these six words, four of the six had all eight tokens included in one of the two reference templates for that word. Only one of the words (exchange) was among the two words for which no clustering occurred. Of the eight words with three total error or close entries, three had all eight tokens included in the reference templates. Thus, the preliminary indication is that little or no correlation exists between word errors and the number of tokens which cluster together from a given size training set. However a somewhat larger training set is required before much confidence can be given to this result. In addition, some of the alternative ways of pooling the tokens which cluster together, such as those discussed in Section II, should be investigated more thoroughly.

IV. DISCUSSION

Although the main purpose of this paper was to propose and study several alternative methods of creating reference templates for a speaker independent word recognition system, it is worthwhile comparing the recognition results obtained here with those of two other major studies. Gold [11], using the same word vocabulary, achieved a recognition accuracy of 86 percent using high quality input speech, and an accuracy of about 95 percent when using the top two choices. The system Gold used was a feature measurement scheme which used a 16-channel spectrum analyzer, a pitch extractor, and a voicing detector. The decision process was a probabilistic scoring algorithm based on the presence or absence of key features within the word. The recognition accuracies reported here are approximately the same as those given by Gold; however, here the input speech was severely bandlimited by the telephone line and therefore the recognition task was much harder.

The second basis for comparison is the word recognition system of Levinson, Rosenberg, and Flanagan [33]. In this work, the vocabulary was 127 isolated words and the recognition algorithm was similar to the one used here. Although the system was intended to be used as a speaker *dependent* recognition system, results were obtained for speaker independent recognition as well. Their input words were also recorded off a standard telephone line. For their vocabulary, the median error rate was 11.7 percent for a designated speaker (speaker dependent mode) and 34.9 percent for a composite of four male speakers in which both female and male speakers were in the test set.[5] Using the top five choices, the error rates were 1.8 percent and 20 percent for the speaker dependent and speaker independent training sets. Since the vocabularies were different, it is difficult to make exact comparisons between results; however, the similarities in the analysis and recognition algorithms suggest that the clustering analysis and reference template creation is a good way of significantly reducing the recognition error rate for the speaker independent mode. Furthermore, it is seen that the speaker independent error rate for our scheme is not significantly larger than the speaker dependent error rate for the Levinson *et al.* recognition system.

[5]The speaker independent error rates were considerably higher for the female speakers than for the male speakers. However, the error rates for both males and females were higher than those reported herein.

Aside from the possibility of combining the reference tokens within a cluster in a more optimal manner, there is one other obvious way of lowering the error rate without much additional computational effort. This is to make use of the VUS contour of the utterance to eliminate from consideration those words which cannot possibly be candidates. For example, a word with a clear fricative beginning would eliminate 33 words from consideration, thereby reducing significantly the amount of computation involved in the distance algorithm, as well as reducing the possibility of incorrect identifications. If too much reliance is placed on the VUS contour, then serious mistakes can occur causing recognition errors. More work needs to be done in this area to explore these possibilities.

V. Summary

In this paper we have given some possibilities for combining a set of reference words from several different speakers into a set of templates which characterize the properties of the word. Many issues are involved in the choice of an algorithm to determine which words cluster together, and how to effectively combine these words into a single composite template and we have provided some ideas on how to handle these problems. Using four and eight speaker training sets we showed that the creation of multiple templates can and does offer advantages over a single template when there are a sufficient number of references to be combined. Recognition accuracies of about 85 percent were obtained for speaker independent recognition of a 54 word vocabulary with eight speakers. The correct word was in the top five choices about 98 percent of the time.

Acknowledgment

The author gratefully acknowledges the contributions of Dr. M. Sambur, formerly at Bell Laboratories, currently at ITT, in the early stages of this work. Dr. Sambur contributed significantly towards the implementation of the recognition and training algorithms. The helpful comments and criticisms of Dr. J. Flanagan, Dr. A. Rosenberg, and Dr. S. Levinson were greatly appreciated.

References

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

[2] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," *J. Acoust. Soc. Amer.*, vol. 24, p. 637, 1952.

[3] H. Dudley and S. Balashek, "Automatic recognition of phonetic patterns in speech," *J. Acoust. Soc. Amer.*, vol. 30, pp. 721-732, 1958.

[4] P. B. Denes and M. V. Mathews, "Spoken digit recognition using time-frequency pattern matching," *J. Acoust. Soc. Amer.*, vol. 32, pp. 1450-1455, 1960.

[5] J. H. King and C. J. Tunis, "Some experiments in spoken word recognition," *IBM J. Res. Dev.*, vol. 10, no. 1, pp. 65-79, Jan. 1966.

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

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

[8] 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, 1967.

[9] T. G. Von Keller, "An on-line recognition system for spoken digits," *J. Acoust. Soc. Amer.*, vol. 49, pp. 1288-1296, Apr. 1971.

[10] C. F. Teacher, H. G. Kellet, and L. R. Focht, "Experimental limited-vocabulary speech recognizer," *IEEE Trans. Audio Electroacoust.*, vol. AU-15, pp. 127-130, 1967.

[11] B. Gold, "Word-recognition computer program," Mass. Inst. Technol., Cambridge, RLE Tech. Rep. 452, June 1966.

[12] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, no. 2, pp. 637-655, 1971.

[13] J. D. Markel and A. H. Gray, Jr., "A linear prediction vocoder simulation based upon the autocorrelation method," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 124-134, 1974.

[14] G. S. Kang and D. C. Coulter, "600-bit-per-second voice digitizer (linear predictive formant vocoder)," Naval Res. Lab., Washington, DC, Rep. 8043, Nov. 1976.

[15] B. S. Atal, M. R. Schroeder, and V. Stover, "Voice-excited predictive coding system for low bit-rate transmission of speech," *Proc. ICC*, pp. 30-37 to 30-40, 1975.

[16] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561-580, Apr. 1975.

[17] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 417-427, Oct. 1973.

[18] E. Hafer and C. H. Coker, "Determining tongue body motion from the acoustic speech wave," *J. Acoust. Soc. Amer.*, vol. 57, Sup. 1, p. 53 (A), 1975.

[19] 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.

[20] P. V. de Souza, "Statistical tests and distance measures for LPC coefficients," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 554-559, Dec. 1977.

[21] J. Makhoul, R. Viswanathan, L. Cosell, and W. Russell, "Natural communication with computers," BBN Rep. No. 2976, Dec. 1974.

[22] D. T. Magill, "Adaptive speech compression for packet communication systems," *Proc. NTC*, pp. 29D1-29D5, Nov. 1973.

[23] M. R. Sambur and N. S. Jayant, "LPC analysis/synthesis from speech inputs containing quantizing noise or additive white noise," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 488-494, Dec. 1976.

[24] R. E. Crochiere, D. J. Goodman, L. R. Rabiner, and M. R. Sambur, "Tandem connections of wideband and narrowband speech communications systems: Part 1, narrowband to wideband link," *Bell Syst. Tech. J.*, vol. 56, no. 9, Nov. 1977.

[25] L. R. Rabiner, M. R. Sambur, R. E. Crochiere, and D. J. Goodman, "Tandem connections of wideband and narrowband speech communications systems: Part 2, wideband to narrowband link," *Bell Syst. Tech. J.*, vol. 56, no. 9, Nov. 1977.

[26] A. H. Gray, Jr., and J. D. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 380-391, Oct. 1976.

[27] P. B. Scott, "VICI-A speaker independent word recognition system," *Conf. Rec., 1976 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Philadelphia, PA, pp. 210-213, Apr. 1976.

[28] R. C. Lummis, "Speaker verification by computer using speech intensity for temporal registration," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 80-89, Apr. 1973.

[29] A. E. Rosenberg and M. R. Sambur, "New techniques for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 169-176, 1975.

[30] A. E. Rosenberg, "Evaluation of an automatic speaker-verification system over telephone lines," *Bell Syst. Tech. J.*, vol. 55, no. 6, pp. 723-744, July-Aug., 1976.

[31] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech. J.*, vol. 54, pp. 297-315, Feb. 1975.

[32] ——, "Application of an LPC distance measure to the voiced-unvoiced-silence detection problem," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 338-343, Aug. 1977.

[33] S. E. Levinson, A. E. Rosenberg, and J. L. Flanagan, "Evaluation of a word recognition system using syntax analysis," *1977 Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 483-486, May 1977.