THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 61, No. 10, December 1982
Printed in U.S.A.

On the Use of Energy in LPC-Based Recognition of Isolated Words

By M. K. BROWN and L. R. RABINER

(Manuscript received May 17, 1982)

Recognition of isolated words by encoding speech into linear predictive coefficients (LPC) is well known and widely accepted as one of the better methods for speech recognition. One of the drawbacks in relying entirely on LPC measures for recognition, however, is that the energy information in the speech is removed during the LPC analysis. Consequently, attempts have been made to include energy pattern information along with the LPC pattern information to achieve greater recognition accuracy. This paper discusses problems involved in combining energy pattern information with the LPC pattern information and presents results of recognition experiments with one method. The energy information and LPC information are combined linearly in a (speech) frame-by-frame manner utilizing the dynamic time warping (DTW) method time alignment. The LPC log likelihood ratio distance function, which determines the spectral difference between two frames of speech, does not lend itself to direct statistical analysis in multiple dimensions. The method for obtaining the weighting for the linear combination involves an iterative minimization of a probability of error function. The combined energy and LPC distance function was tested using a 129-word "airline" vocabulary, which is designed for speaker-independent, isolated word recognition. The inclusion of energy information in the recognition feature space reduces recognition error rates by an average of about 25 percent as compared with LPC alone.

I. INTRODUCTION

In the last few years it has become common to use LPC coding techniques for speech recognition.[1-7] The speech to be represented is modeled by a linear digital filter with coefficients chosen so that the transfer function of the filter approximates the spectrum of the speech over some short interval of time. Typically, a speech recognition

system performs its task by comparing the unknown utterance or test with a number of previously stored reference patterns. Both the test and reference are characterized by a set of linear predictive coefficients. This is accomplished by digitizing the speech at some suitable rate and breaking the utterance into time windowed regions or "frames" upon which LPC analysis is performed. The frames of speech generally overlap and are typically spaced 10 to 20 ms apart in time. Thus, a typical 0.6-second utterance is represented by about 40 frames of linear predictive coefficients.

It is well known in the area of speech recognition that optimal time alignment of reference patterns to test patterns substantially reduces recognition errors for a vocabulary with polysyllabic words.[1] The most commonly used time alignment procedures, for the speech recognition problem, are the class of algorithms referred to as dynamic programming (DP) or dynamic time warping (DTW) methods.[2-5]

Let us assume that we are given a characterization of an isolated word that consists of a set of N vectors of LPC coefficients. The test pattern, T, is represented as:

$$\mathbf{T} = \{T(1), T(2), \cdots, T(N)\}, \tag{1}$$

where the vector $T(i)$ is a spectral (LPC) representation of the ith frame of the test word. In our system a set of nine autocorrelations constitutes the vector from which an 8th order LPC model is derived. The duration of the test utterance is N frames, where each frame represents 45 ms of speech, and adjacent frames are spaced 15 ms apart.

For a given vocabulary of V words, the reference R_v, is represented as:

$$\mathbf{R}_v = \{R_v(1), R_v(2), \cdots, R_v(M_v)\}, \tag{2}$$

where each vector, $R_v(i)$, is again a spectral representation of the corresponding frame within the reference pattern, and M_v is the number of frames in the vth reference.

To optimally align the time scale of the reference pattern (the dependent m index) to the test pattern (the independent n index), we must solve for a warping path function of the form:

$$m = w(n) \tag{3}$$

and thereby seek to minimize the total distance

$$D_v = \sum_{n=1}^{N} d\{T(n), R_v[w(n)]\} \tag{4}$$

over all possible paths, $w(n)$, where $d[T(n), R_v(m)]$ is the local distance between test frame n and reference frame $m = w(n)$. This operation

must be performed for each reference R_v in the vocabulary. The test pattern is classified as belonging to the class (i.e., the reference word) for which the smallest accumulated DTW distance, D_v, is obtained. In addition to the standard DTW algorithm, time normalization has been used on both the test and reference patterns, thereby allowing the widest range of time alignment paths to be considered. This procedure, called the normalize-and-warp method,[5] linearly normalizes the test and reference utterances to a fixed length (typically the average duration of all words in the vocabulary) before the DTW is performed. Experimental results have shown this method to be valid on several recognition vocabularies, including the one used in this study.[5]

Comparison of a test frame to a reference frame requires a measure of closeness (distance). Several distance measures have been investigated and used for utterance comparison purposes.[8] Virtually all of these distance measures are spectral in nature and generally do not explicitly consider the energy pattern of the speech. The LPC-based distance measure developed by Itakura[2] has been found to yield high recognition accuracy and cost relatively little in computation. This distance function, often referred to as the log likelihood ratio (LLR) yields numerical values that are indicative of the spectral energy difference between the two frames of speech. The form of the function is as follows:

$$d(T, R) = \log[(a_R V_T a'_R)/(a_T V_T a'_T)], \tag{5}$$

where T refers to a test frame and R refers to a reference frame, a is a vector consisting of the $(p + 1)$ LPC coefficients of a pth order LPC model of the speech, and V_T is the $(p + 1 \times p + 1)$ autocorrelation matrix of the test frame. [Itakura[2] has shown how the computation of eq. (5) can be performed with $(p + 1)$ multiplication and additions and one logarithm.] The LPC coefficients do not contain any energy information since they are derived from a normalized spectrum. The V matrix enters into both the numerator and denominator of the distance function (which is the ratio of two scalars), and thus contributes no energy information. Hence, $d(T, R)$ contains no energy information.

A few further observations about the LLR distance are in order. The distance function, $d(T, R)$, does not satisfy commutative or triangle inequality rules, (i.e., the function is not symmetric). The log likelihood ratio distance is related to the coefficient sensitivities of the LPC filter model of the test utterance. If the test filter model is very similar to the reference filter model, then it is reasonable to estimate the difference between the two filter models based on the test filter coefficient sensitivities. However, if the two models differ greatly, then the coefficient sensitivities of the reference model will be much different from

those of the test model and comparison of the two models yields inconsistent results. Thus, $d(T, R)$ is not monotonic when it exceeds certain values (usually about 0.6 per frame); however, it is quite useful in measuring spectral closeness (as opposed to spectral separation), and in this application serves very well for recognition purposes.

Since the log likelihood ratio distance measure normalizes energy out of the measurement, it is desirable to consider including this additional information in the distance calculation. In many pattern recognition disciplines, the addition of dimensions to a feature space is all that is normally required to add information to the distance measurement. Ordinarily, if the individual components of the distance measurement are available, an optimal weighting of features (LPC and energy) can be obtained by an analysis of feature covariance.[9,10] However, in speech recognition, there are complications that make simple addition of a dimension to the feature space difficult. One of these complications arises from the nature of the log likelihood ratio computation, which does not allow separation of the individual components of distance, i.e., $d(T, R)$ is a ratio of scalars and becomes meaningless if only one dimension of the LPC space is considered. DTW methods further complicate the addition of energy information to the feature space since the DTW path will be altered by the distances calculated during the DTW optimization. Thus, the addition of another dimension (frame energy) to the feature space is not trivial.

In the next section of this paper we will describe a new discriminant function that contains both spectral and energy information. A method for determining the weighting of the two components based on probability of recognition error will be derived. In Section III we discuss experimental results using this procedure on a vocabulary of 129 words of the "airlines" vocabulary described in Ref. 6.

II. ADDITION OF FEATURES TO THE LPC SPACE

To add the speech energy pattern information to the LPC feature space, the LPC part must be handled as a single dimension because of the log likelihood ratio distance function. Thus, if the total feature vector is otherwise treated as a linear combination of vectors, the total feature vector is of the form:

$$F = [\text{LPC}, E]', \tag{6}$$

where the feature vector, F, consists of a vector of autocorrelation coefficients (treated as scalar), and a value for peak normalized log energy.

The distance function chosen for comparing the test frame energy pattern with the reference frame energy pattern, referred to as peak normalized log energy ratio, is of the form

$$e(T, R) = |\log[E(T)E_{\max}(R)/E(R)E_{\max}(T)]|, \tag{7}$$

where $E(T)$ is the energy of a test frame and $E(R)$ is the energy of a reference frame, $E_{\max}(T)$ is the peak energy of the test utterance over frames $i = 1, 2, \cdots, N$, and $E_{\max}(R)$ is the peak energy of the reference utterance over frames $j = 1, 2, \cdots, M$. This is equivalent to a peak normalized log energy difference, i.e.,

$$e(T, R) = |NE(T) - NE(R)|, \tag{8}$$

where $NE(T)$ is normalized energy and

$$NE(T) = \log[E(T)] - \log[E_{\max}(T)]. \tag{9}$$

Then for the optimal linear classifier, distance is given by

$$d(T, R) = D(T, R)'\mathbf{W}\, D(T, R), \tag{10}$$

where (lower case indicates a scalar quantity)

$$D(T, R) = F(T) - F(R) \tag{11}$$

and

$$\mathbf{W} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}, \tag{12}$$

where a, b, and c are chosen to minimize the probability of recognition error, $P(E)$. Applying ordinary Bayesian techniques would result in the well-known Mahalonobis distance where the matrix $\mathbf{W}$ is the inverse covariance matrix of the feature space.[9,10] Because of the nature of the log likelihood ratio distance function, the LPC distance to the origin of the LPC space is generally too large to be within the 0.6 value required for monotonicity. Furthermore, no other point in the LPC space can easily be found that will allow this requirement to be satisfied for all possible samples in the space. Hence, the mean and, in turn, variance (and, hence, covariance) of the LPC component cannot be determined directly.

An alternate method of combining the energy measure with LPC has been developed which, although not able to determine the cross-product coefficients, will determine a weighting of the two measures based on probability of error. Let the distance function assume the form:

$$d(T, R) = [\mathrm{LLR}(T, R)] + \alpha[\mathrm{LER}(T, R)], \tag{13}$$

where $\mathrm{LLR}(T, R)$ is the log likelihood distance between a test frame T and a reference frame R, and $\mathrm{LER}(T, R)$ is the peak normalized log energy difference between test frame T and reference frame R, and α is a weighting coefficient. Equation (13) must be employed in (4) to determine the DTW function (3) and choose the closest reference to

a given test utterance. Consequently, the LLR distance is a function of the LER distance and vice versa, since the DTW path is a function of both distances.

Let $D(i, j)$ indicate the accumulated DTW distance between a test utterance (corresponding to word i) and reference pattern (corresponding to word j). A classification error will occur when $D(i, i) > D(i, j)$ for *any* j not equal to i. That is, if the distance between a test class (i) and a reference of the same class (i) is greater than the distance from class (i) to a different class (j), then a recognition error will occur. An alternate form of (13) and (14) is the discriminant function:

$$Q(i, j) = D(i, j) - D(i, i). \tag{14}$$

For this form a recognition error occurs only if $Q(i, j)$ is less than zero for any j not equal to i.

For notational convenience, let

$$L(i, j) = \mathrm{LLR}(i, j) - \mathrm{LLR}(i, i) \tag{15}$$

and

$$E(i, j) = \mathrm{LER}(i, j) - \mathrm{LER}(i, i), \tag{16}$$

where $\mathrm{LLR}(i, j)$ and $\mathrm{LER}(i, j)$ are the accumulated log likelihood ratio and log energy ratio on the DTW path, respectively. Then

$$Q(i, j) = L(i, j) + \alpha E(i, j). \tag{17}$$

There are four kinds of classification errors for which $Q(i, j)$ is less than zero and a test word will be misclassified, namely:

- (A) an LLR error for which $L(i, j) < 0$ for any $j \neq i$ and $E(i, j) > 0$ for all $j \neq i$
- (AB) an LLR error or an LER error but not both, i.e.,
  - (a) $L(i, j) < 0$ and $E(i, j) > 0$ for any $j \neq i$
  - (b) $L(i, k) > 0$ and $E(i, k) < 0$ for any $k \neq i, j$
- (B) an LER error for which $L(i, j) > 0$ for all $i \neq j$ and $E(i, j) < 0$ for any $j \neq i$
- (C) both errors for which $L(i, j) < 0$ and $E(i, j) < 0$ for any $j \neq i$.

The test word is correctly recognized when condition (A) exists if $Q(i, j) > 0$ for all $j \neq i$, which implies:

$$\alpha > |L(i, j)/E(i, j)| \quad \text{for all} \quad j \neq i. \tag{18}$$

Likewise, the test word is correctly recognized when condition (B) exists if:

$$\alpha < |L(i, j)/E(i, j)| \quad \text{for all} \quad j \neq i. \tag{19}$$

Condition (C) is not recoverable [i.e., $Q(i, j) < 0$] for any value of α.

In this case the test word will always be misrecognized since both the LLR and LER distances have made an error.

Condition (AB) is a special case where an LPC-type error $[L(i, j) < 0]$ occurs for a comparison of the test (i) with one reference (j) and an LER-type error $[E(i, k) < 0]$ occurs for a comparison with a different reference (k). In order for this test word to be properly recognized the following relation must be satisfied for all $j, k \neq i$:

$$|L(i, j)/E(i, j)| < \alpha < |L(i, k)/E(i, k)|. \tag{20}$$

Equation (20) may or may not be satisfiable for a given test word (i).

The probability of recognition error can be written with the further definition of the random variable:

$$x = |L(i, j)/E(i, j)|. \tag{21}$$

Define the probability density function of x conditioned on error type (A) as $p(x|A)$ and the probability density function of x conditioned on error type (B) as $p(x|B)$. Then the probability of error, $P(E, \alpha)$, is given by:

$$P(E, \alpha) = P(A) \int_{\alpha}^{\infty} p(x|A)dx + P(B) \int_{-\infty}^{\alpha} p(x|B)dx + P(C)$$

$$+ P(AB) \int_{\alpha}^{\infty} p(x|AB, a)dx + P(AB) \int_{-\infty}^{\alpha} p(x|AB, b)dx, \tag{22}$$

where $p(x|AB, a)$ is the probability density function conditioned on the existence of error type (AB) of the first kind $[L(i, j) < 0$ and $E(i, j) > 0]$ and $p(x|AB, b)$ is the density function conditioned on error type (AB) of the second kind $(L(i, k) > 0$ and $E(i, k) < 0)$. Optimizing (22) with respect to α yields the relation:

$$P(A)p(x = \alpha^*|A) + P(AB)p(x = \alpha^*|AB, a)$$

$$= P(B)p(x = \alpha^*|B) + P(AB)p(x = \alpha^*|AB, b), \tag{23}$$

where α^* is the optimal value of α. Inspection of (22), (23), and the error conditions (A), (AB, a), (B), and (AB, b) reveals that they are mutually exclusive events and that they can be combined. Thus, by relaxing the requirements "for all $j \neq i$" of condition (A) and (B) to "for any $j \neq i$" eq. (23) may be rewritten

$$P(A)p(x|A) = P(B)p(x|B), \tag{24}$$

keeping in mind the new properties of (A) and (B):

$$(A)\ L(i, j) < 0 \quad \text{and} \quad E(i, j) > 0 \quad \text{for any} \quad j \neq i$$

$$(B)\ L(i, j) > 0 \quad \text{and} \quad E(i, j) < 0 \quad \text{for any} \quad j \neq i.$$

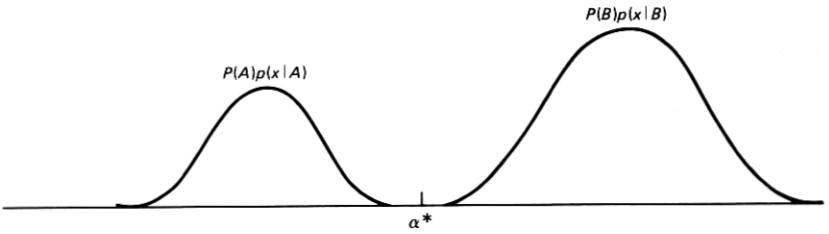

Fig. 1—Conditions for minimum-recognition error rate.

Several observations concerning (18) through (24) are worth noting. First, (22) indicates that the best performance that can be expected for any choice of α is

$$P(E, \alpha^*) = P(C), \tag{25}$$

where α^* is the value of α for which $P(E, \alpha)$ is minimized. This condition occurs only if α can be chosen so that the distribution $P(A)p(x|A)$ lies entirely below α and $P(B)p(x|B)$ lies entirely above α, as shown in Fig. 1. If a recognizer is implemented with only the LLR distance function, the probability of error will be

$$P(E|\text{LLR only}) = P(A) + P(C) \tag{26}$$

since error conditions (A) and (C) are mutually exclusive LLR-type errors. Hence, if $P(A) > 0$, then

$$P(C) < P(E|\text{LLR only}), \tag{27}$$

indicating with (25) that

$$P(E, \alpha^*) < P(E|\text{LLR only}) \quad \text{if} \quad P(A) > 0 \tag{28}$$

and therefore recognition performance better than that obtained with LPC alone can be achieved by using distance function (13). Note also by (21) that random variable $x \geq 0$. The lower limits of the second and fourth integrals in (22) can be made 0 without effect on $P(E, \alpha)$. Then, under the worst-case condition that $P(A) = 0$, the optimal value for α is $\alpha^* = 0$. Inspection of (13) indicates that, for $\alpha^* = 0$, $D(i, j)$ is identical to the LPC component of the distance function, $\text{LLR}(i, j)$. Thus, the worst performance that can be expected is the LPC error rate.

Under good conditions, $|L(i, j)|$ will be small for condition (A) and $|E(i, j)|$ will be small for condition (B). Thus, the mean of the distribution $P(A)p(x|A)$ will be low, while the mean of the distribution $P(B)p(x|B)$ will be large, as shown in Fig. 2. The probability of recognition error, which consists of $P(C)$ plus the shaded area shown in Fig. 2, is minimized by the value of α^* shown. If we assume that $P(A)p(x|A)$ and $P(B)p(x|B)$ are normally distributed as in Fig. 2, then the shaded area will be less than the area of $P(A)p(x|A)$, i.e.,

$$P(E, \alpha^*) < P(A) + P(C), \tag{29}$$

which by (26) yields

$$P(E, \alpha^*) < P(E|\text{LLR only}). \tag{30}$$

Thus, not only can we guarantee that performance will be no worse than LPC recognition performance, but, if the conditions of Fig. 2 can be established, we can guarantee better performance by properly choosing α.

Differing DTW paths owing to the interaction of the LLR and LER components of $d(T, R)$ will cause the distributions $P(A)p(x|A)$ and $P(B)p(x|B)$ to vary with different values for coefficient α. Consequently, the determination of coefficient α may require several iterations of selecting α and redefining (13) until a stable value for α is obtained. Under adverse conditions $P(A)p(x|A)$ and $P(B)p(x|B)$ may overlap considerably and the number of recoverable errors may be small.

III. EXPERIMENTAL RESULTS

Experiments were conducted on speech collected from several talkers speaking an "airlines" vocabulary.[6] This vocabulary consisted of 129 words that would commonly be used to obtain information from airline scheduling service. The range of words is broad enough to be considered a good cross section of English speech. The reference words were generated by clustering the speech of several male and female talkers to form speaker-independent templates.[7] Six clustered templates for each of the 129 words were generated resulting in a reference file of 774 templates. The average duration of all words was 42 frames and this was the duration used for word length normalization.

Test sets were studied for two male and two female talkers (not of the reference set) in order to choose an approximate value for the weighting coefficient α. The LLR and LER distance functions were not linked together for these experiments. Thus, two different DTW paths were generated, one for LLR distances and one for LER dis-

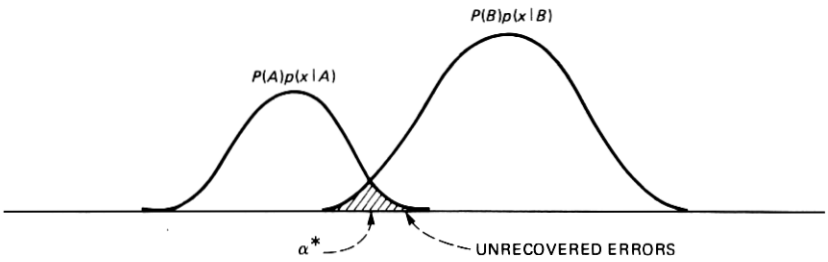

Fig. 2—Conditions for low-recognition error rate.

Table I—Statistics on LLR and LER distance measures for the four data sets

Data Set (Talker Sex)		1 (M)	2 (M)	3 (F)	4 (F)
LLR Recognition Error	First Choice	9.3%	12.4%	22.5%	21.7%
	First Two Choices	3.9%	5.4%	10.8%	12.4%
LER Recognition Error	First Choice	77.5%	72.1%	75.2%	73.6%
	First Two Choices	66.7%	60.5%	63.6%	58.9%
Correlation LLR vs. LER	Correct Words	0.093	0.183	0.205	0.227
	Incorrect Words	0.205	0.175	0.143	0.144

tances. The accumulated LER distance was also determined along the DTW path generated by the LLR distances to evaluate the effects of the warp path on the accumulated LER distance. (It was found, early in the investigation, that the LLR distance would be the dominant force in directing the DTW path.) Statistics were gathered on the number of recognition errors for both distance measures, the distributions for LLR distances and LER distances were calculated, and the distributions $P(A)p(x|A)$ and $P(B)p(x|B)$ were determined. In addition, correlation matrices were calculated for LLR, LER, and LER along the LLR path. The results are shown in Table I and Fig. 3. The

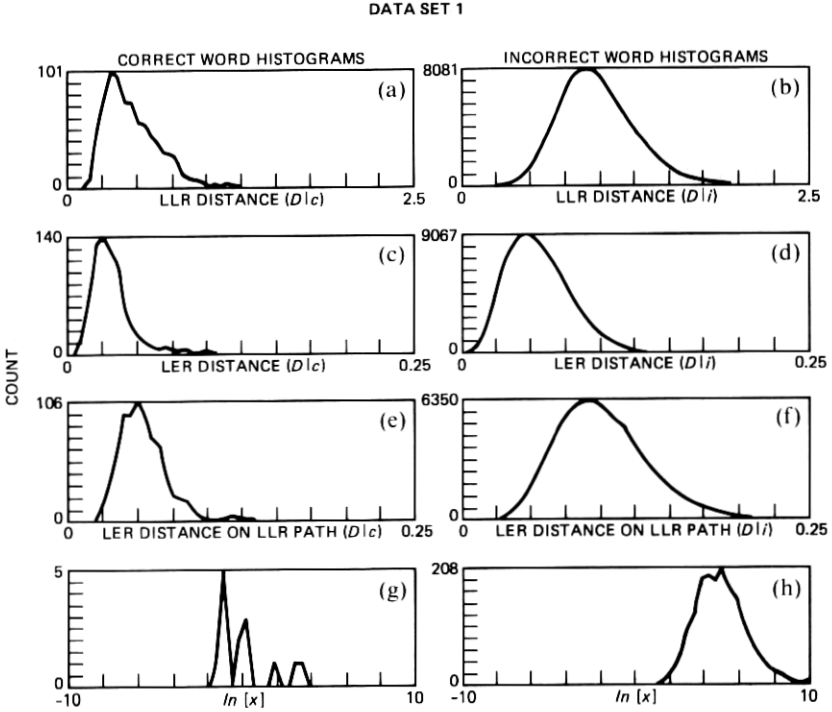

Fig. 3—DTW statistics for unlinked LLR and LER distances.

captions in Fig. 3 are read as follows. $D|c$ is distance to correct reference, $D|i$ is distance to incorrect reference, and LER(LLR) is the LER distance along the LLR-controlled DTW path.

The plots of $P(A)p(x|A)$ and $P(B)p(x|B)$ of Fig. 3 indicate that an initial value of α should be between 1.0 and 5.0. The correlations of LLR and LER (Table I) indicate that there is little redundancy and, hence, useful information may be obtainable by combining the LLR and LER distance measures. The plots of $P(A)p(x|A)$, as shown in the figure, are quite jagged due to the low number of samples obtained; however, the optimal value of α was actually determined by inspection of the samples of the two distributions. Although this is a somewhat tedious process, it ensures that the best value for α is obtained from the available information. The graphical display of $P(A)p(x|A)$ and $P(B)p(x|B)$ gives good indication that the requirements for (29) and (30) are being met.

Table I indicates that the recognition error rate for LER distance measures alone is very high. Even among the top two candidates, the correct word is found less than half of the time. Obviously, LER alone is not a good discriminating feature. The average LER distance [eq. (7)] from the test to a correct reference (i.e., from the same class as the test) is low in comparison to the LER distance to an incorrect class (about 1:1.75). This would normally indicate a good discriminating feature; however, the LER distances to both correct and incorrect references are very widely distributed and the amount of overlap of the two distributions is considerable (see Fig. 3). The LER distances along the DTW path of the LLR measure are less widely distributed but the overlap is still large. For comparison, the ratio of test to correct reference LLR distance with respect to incorrect reference LLR distance averages about 1:1.97, and the amount of overlap of the distributions is relatively small.

Tests were next conducted on the same speech data with the LLR and LER distances linked for DTW path determination. An initial value of $\alpha = 1.0$ was used as a starting point for each test set. Statistics were gathered as previously, except that the combined LLR and LER distance, LLR component, and LER component were items of interest. The distributions of $P(A)p(x|A)$ and $P(B)p(x|B)$ were again calculated.

Several iterations of testing and evaluating were performed to obtain a good estimate for the coefficient α^*. For speech data set 1 the value obtained for α^* was 1.8. The statistical results are shown in Fig. 4 and Table II. A plot of the number of recognition errors predicted by the distributions of $P(A)p(x|A)$ and $P(B)p(x|B)$ as a function of x is shown in Fig. 5a for $\alpha = 3.0$. The same test was run on the second set of speech data (at $\alpha = 3.0$), and the plot shown in Figure 5b was

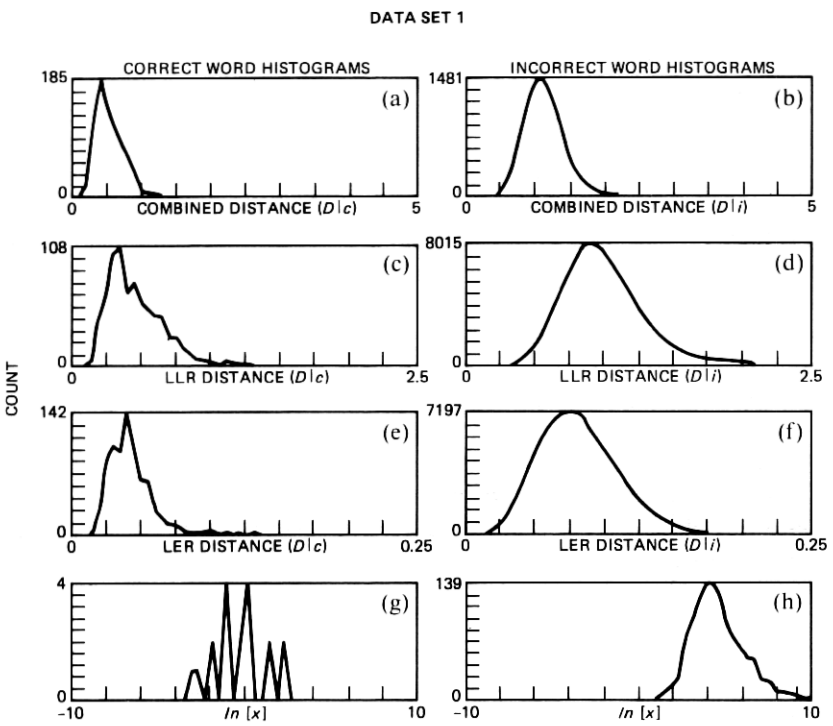

Fig. 4—DTW statistics for linked LLR and LER distances.

obtained. Iterative application of the evaluation procedure yielded a final value of $\alpha = 5.0$ for test set 2. Obviously, the selection of α is very sensitive to the speech data. In all cases, a value for α^* could be found to improve the recognition accuracy over that for LPC-based recognition alone (indeed, for $\alpha = 3.0$ the recognition error rate was lower for both data sets 1 and 2) but the optimal values predicted by error rate plots like those of Fig. 5 were considerably different.

Further testing indicated that a single value for α could be chosen so that an improvement in recognition performance would be obtained for all of the testing sets. The values of $\alpha = 2.0$ and $\alpha = 3.0$, as indicated in Table II, show significant improvement for all four test sets. The average reduction in error rate is 23.9 percent for $\alpha = 2.0$ and 29.8 percent for $\alpha = 3.0$. Thus, a speaker-independent recognizer can make use of the improvement in performance available by the inclusion of energy information using this technique.

3.1 Analysis of the recognition errors

It is interesting to examine the specific errors that were connected by using the combined energy plus LPC feature set. A list of all such

Table II—Word error rates as a function of α for the four data sets

Data Set (Talker Sex)		1 (M) (%)	2 (M) (%)	3 (F) (%)	4 (F) (%)
$\alpha = 0.0$*	First Choice	9.3	12.4	22.5	21.7
	First Two	3.9	5.4	10.8	12.4
$\alpha = 1.0$	First Choice	6.2	10.8	17.8	20.9
	First Two	3.1	5.4	8.5	8.5
$\alpha = 1.8$	First Choice	3.1	10.1	...	...
	First Two	2.3	7.0	...	...
$\alpha = 2.0$	First Choice	3.9	10.8	20.2	18.6
	First Two	2.3	7.0	9.3	9.3
$\alpha = 2.2$	First Choice	...	...	...	18.6
	First Two	...	...	...	9.3
$\alpha = 2.8$	First Choice	...	...	18.6	...
	First Two	...	...	7.7	...
$\alpha = 3.0$	First Choice	3.9	10.1	17.8	17.0
	First Two	2.3	7.0	7.7	10.8
$\alpha = 3.4$	First Choice	...	...	17.0	...
	First Two	...	...	9.3	...
$\alpha = 3.7$	First Choice	...	...	...	18.6
	First Two	...	...	...	10.8
$\alpha = 5.0$	First Choice	...	9.3	...	...
	First Two	...	5.4	...	...
$\alpha = 5.5$	First Choice	...	10.1	...	...
	First Two	...	5.4	...	...

* Equivalent to LLR only.

errors is given in Table III. This table gives the correct word, the word recognized using LPC alone ($\alpha^* = 0$), the test set in which the error occurred, and a classification as to the type of error initially made. The classification code describes the phonetic nature of the correct and misrecognized words as one of the following:

(*i*) MS—Simple monosyllabic word

(*ii*) MS + A—monosyllabic word plus affix (final stop or fricative consonant)

(*iii*) PS—polysyllabic word.

An examination of the 34 words in Table III shows that 11 of the corrections involved a polysyllabic word (as either the correct word or the word recognized using LPC alone), nine of the corrections involved monosyllabic words with affixes, and the remaining 14 corrections involved monosyllabic words.

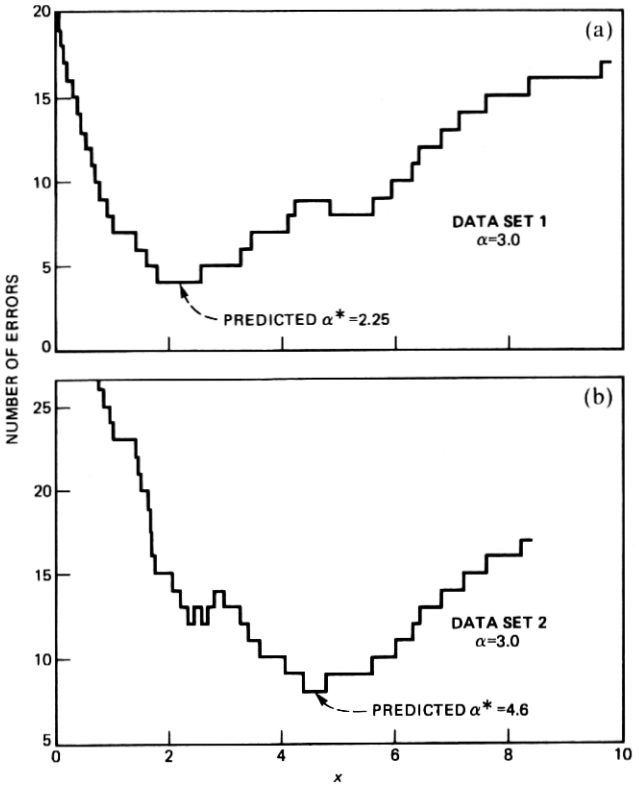

Fig. 5—Prediction of α from statistics on x.

A similar list of the words that were incorrectly recognized using the combined distance metric is given in Table IV. The format for this table is similar to that of Table III except that the classification code refers to the correct word and the word originally recognized using the LPC distance alone. It can be seen that eight new errors are introduced by the combined metric that were not present using LPC alone. Hence, a total net improvement of 26 words was obtained using the combined distance metric.

Of the 53 errors given in Table IV, 16 involve a polysyllabic word (either the correct or the LPC misrecognized word), 23 involve only monosyllabic words, and 13 involve polysyllabic words with affixes.

The data of Tables III and IV indicate that the inclusion of energy into the distance metric leads to a fairly uniform improvement in accuracy across all three types of word classifications. The results also indicate that, for the most part, the remaining errors involve acoustically similar words, whereas the corrections generally come from errors involving acoustically different sounding words.

Table III—List of words misrecognized using LPC alone but correctly recognized using energy combined with LPC

Correct Word	Word Recognized From LPC Alone	Test Set	Classification Code
Boston	Washington	3	(PS, PS)
Card	I	3	(MS+A, MS)
Depart	July	4	(PS, PS)
Detroit	Information	1	(PS, PS)
Does	Five	3	(MS, MS)
Eight	Seat	1	(MS+A, MS+A)
First	Express	3	(MS+A, PS)
First	Express	4	(MS+A, PS)
I	Lockheed	3	(MS, PS)
Like	Flight	4	(MS+A, MS+A)
Many	May	3	(MS, MS)
May	Change	2	(MS, MS)
Morning	Miami	4	(PS, PS)
My	By	2	(MS, MS)
Number	November	4	(PS, PS)
Oh	Of	4	(MS, MS)
On	Arrival	4	(MS, PS)
One	What	3	(MS, MS+A)
Pay	A	2	(MS, MS)
Phone	Five	2	(MS, MS)
Please	Seat	1	(MS, MS+A)
Seats	Seat	1	(MS+A, MS+A)
Seats	Seat	3	(MS+A, MS+A)
Some	From	3	(MS, MS)
Some	From	4	(MS, MS)
Ten	Afternoon	4	(MS, PS)
There	Fare	3	(MS, MS)
Three	Detroit	2	(MS, PS)
Time	Card	3	(MS, MS+A)
Times	Five	1	(MS+A, MS)
To	Do	3	(MS, MS)
Twelve	Five	4	(MS, MS)
Uh	How	3	(MS, MS)
When	Would	1	(MS, MS)

IV. CONCLUSIONS

The addition of energy information to the LPC distance improves recognition performance significantly. It is likely that the energy information would significantly improve the error rate on certain kinds of anomalies in the speech, such as partially voiced words and lip pops. These anomalies sometimes cause a test word to match reference words that have similar spectral patterns but significantly different energy patterns. Unfortunately, data bases of speech containing these kinds of erroneous sounds are not yet readily available and testing of this hypothesis must be deferred until such data are generated.

The selection of the weighting coefficient, α, is quite sensitive to the speech data. For that reason, the value of α^* will usually be different

Table IV—Word incorrectly recognized using the combined distance metric

Correct Word	Word Recognized From LPC Alone	Word Recognized From LPC + Energy	Test Set	Classification Code
A	May	Pay	2	(MS, MS)
A	*A*	Pay	3	(MS, MS)
Boston	*Boston*	What	2	(PS, MS+A)
By	I	I	2	(MS, MS)
By	I	I	3	(MS, MS)
By	Like	I	4	(MS, MS+A)
Change	*Change*	Stops	4	(MS, MS+A)
Code	Card	Card	3	(MS+A, MS+A)
Code	Card	Card	4	(MS+A, MS+A)
Code	Coach	Coach	2	(MS+A, MS)
DC	BAC	BAC	4	(PS, PS)
Do	Will	Code	4	(MS, MS)
Eight	*Eight*	Take	3	(MS+A, MS+A)
Flight	*Flight*	Flights	4	(MS+A, MS+A)
Flights	Flight	Flight	3	(MS+A, MS+A)
Flights	Flight	Flight	4	(MS+A, MS+A)
For	Five	Five	4	(MS, MS)
Four	Five	Five	4	(MS, MS)
From	*From*	Eleven	4	(MS, PS)
Go	Twelve	Club	3	(MS, MS)
Home	How	How	4	(MS, MS)
I	By	By	1	(MS, MS)
In	*In*	AM	3	(MS, PS)
Is	In	In	4	(MS, MS)
Leave	Please	Please	3	(MS, MS)
Many	Pay	Pay	1	(PS, MS)
March	Much	Much	4	(MS, MS)
Much	March	March	3	(MS, MS)
My	I	By	3	(MS, MS)
Nine	Washington	Morning	4	(MS, PS)
October	September	September	1	(PS, PS)
Oh	How	How	2	(MS, MS)
Oh	How	How	3	(MS, MS)
On	Five	Five	2	(MS, MS)
One	When	When	4	(MS, MS)
PM	Seattle	Seattle	4	(PS, PS)
Pay	Friday	Friday	3	(MS, PS)
Phone	From	From	4	(MS, MS)
Please	April	Thee	4	(MS, PS)
Prefer	There	Fare	3	(PS, MS)
Seat	Seats	Seats	3	(MS+A, MS+A)
Sunday	Saturday	Saturday	3	(PS, PS)
Ten	AM	PM	2	(MS, PS)
The	Five	Five	3	(MS, MS)
Thee	DC	DC	1	(MS, PS)
Thee	Make	Three	2	(MS, MS+A)
There	Fare	Fare	2	(MS, MS)
Times	Office	Five	3	(MS, PS)
Two	Do	Do	4	(MS, MS)
Want	What	What	4	(MS+A, MS+A)
What	Want	Want	2	(MS+A, MS+A)
When	*When*	Morning	4	(MS, PS)

for each talker in the testing set. Consequently, this method is likely to be more effective for speaker-trained recognition than for speaker-independent recognition systems. However, a value of $\alpha = 2$ or $\alpha = 3$ has been found to work well for the test sets used in this study and will probably work for most test sets.

REFERENCES

1. G. M. White and R. B. Neely, "Speech Recognition Experiments with Linear Prediction, Bandpass Filtering, and Dynamic Programming," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-24*, No. 2 (April 1976), pp. 183–8.
2. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-23*, No. 1 (February 1975), pp. 67–72.
3. H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-26*, No. 1 (February 1978), pp. 43–9.
4. L. R. Rabiner, A. E. Rosenberg, and S. E. Levinson, "Considerations in Dynamic Time Warping for Discrete Word Recognition," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-26*, No. 6 (December 1978), pp. 575–82.
5. C. S. Myers, L. R. Rabiner, and A. E. Rosenberg, "Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-28*, No. 6 (December 1980), pp. 622–33.
6. J. G. Wilpon, L. R. Rabiner, and A. Bergh, "Speaker Independent Isolated Word Recognition Using a 129 Word Airline Vocabulary," J. Acoustical Society of America, *72*, No. 2 (August 1982), pp. 390–6.
7. L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon, "Speaker Independent Recognition for Isolated Words Using Clustering Techniques," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-27*, No. 4 (August 1979), pp. 336–49.
8. A. H. Gray and J. D. Markel, "Distance Measures for Speech Processing," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-24*, No. 5 (October 1976), pp. 380–91.
9. G. S. Sebestyen, *Decision Making Processes in Pattern Recognition*, New York: Macmillan, 1962.
10. N. J. Nilsson, *Learning Machines: Foundations of Trainable Pattern Classifying Systems*, New York: McGraw-Hill, 1965.
11. J. M. Tribolet and L. R. Rabiner, "Statistical Properites of an LPC Distance Measure," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-27*, No. 5 (October 1979), pp. 550–8.