Application of an LPC Distance Measure to the Voiced–Unvoiced–Silence Detection Problem

LAWRENCE R. RABINER, FELLOW, IEEE, AND MARVIN R. SAMBUR

Abstract—One of the most difficult problems in speech analysis is reliable discrimination among silence, unvoiced speech, and voiced speech which has been transmitted over a telephone line. Although several methods have been proposed for making this three-level decision, these schemes have met with only modest success. In this paper, a novel approach to the voiced-unvoiced-silence detection problem is proposed in which a spectral characterization of each of the three classes of signal is obtained during a training session, and an LPC distance measure and an energy distance are nonlinearly combined to make the final discrimination. This algorithm has been tested over conventional switched telephone lines, across a variety of speakers, and has been found to have an error rate of about 5 percent, with the majority of the errors (about $\frac{2}{3}$) occurring at the boundaries between signal classes. The algorithm is currently being used in a speaker-independent word recognition system.

I. Introduction

The problem of reliably discriminating among voiced speech, unvoiced speech, and silence is one of the most difficult problems in speech analysis. There are several reasons why this is so. One problem is the large dynamic range of the speech signal itself in which a 20–40 dB variation of signal level is not uncommon within the speech of a single talker. Compounded with this is a 20–40 dB variation in level among talkers. Another problem is that sometimes the acoustic waveform does not provide accurate information about the signal classification [1], e.g., the vocal cords are vibrating (i.e., the signal is voiced speech) but no periodicity is seen in the acoustic waveform. Finally, all these problems are compounded by the degradations of telephone lines which include band-limiting, nonlinear phase distortion, center clipping, and noise addition.

Classically, the method for discriminating among these three signal classes is to use a level test to discriminate silence from speech, and then discriminate between voiced speech and unvoiced speech by a logical decision based on the values of certain measured features of the signal, e.g., energy, zero-crossings, etc., [2]. When used in conjunction with pitch detection, features of the pitch detector are often used to supplement the voiced-unvoiced decision [3]-[6]. Recently, Atal and Rabiner [7] proposed a statistical decision approach to voiced-unvoiced-silence classification in which a set of measured features were combined using a non-Euclidean distance metric to give a reliable decision. This method was optimized for telephone line inputs by Rabiner *et al.* [8]. Their results showed that reliable discrimination between voiced and nonvoiced speech could be obtained over telephone lines using the statistical approach; however, the overall error rate for the three-class decision was fairly high (11.7 percent) over telephone lines.

Based on the results of [8], it was felt that an alternative approach was required to lower the error rate for telephone line inputs. The problem with combining a set of features is that they can only partially represent the information present in the signal. To obtain a complete representation of the signal properties requires a classification procedure based on the signal waveform, or its spectrum. A novel approach was recently suggested by McAuley [9] in which a matched digital Wiener filter was designed for each of the signal classes, and the signal was processed by each of these filters. Based on the signal output from each of the filters, a distance was computed representing how closely the input signal was matched to the filter, and the minimum distance was used to make the final classification. Although this approach shows promise, it requires a large amount of signal processing and has not as yet been extensively tested.

An alternative procedure is suggested in this paper in which an average signal spectrum is measured (from a training set of data) for each of the three signal classes, and an LPC distance is used to measure similarity between the test signal and each of the reference patterns. Additionally, an energy distance is calculated, and the LPC and energy distances are nonlinearly combined to make the final class decision. The advantages of this technique are that all the spectral information in the signal is used in the classification algorithm, and that the LPC distance computation nonuniformly weights the spectrum in measuring overall similarity. In this way, a fairly reliable discrimination is obtained. Another advantage is that, like some other methods, the voiced-unvoiced-silence decision is made without the need for pitch detection. The main disadvantage of the method is the need for training the algorithm to obtain the average spectral representation for the three signal classes.

In the remainder of this paper we describe the algorithm (Section II), and show the results of a series of tests using conventional switched telephone lines (Section III).

II. Description of the Algorithm

Fig. 1 shows a block diagram of the signal processing used in the algorithm. The input signal $s(n)$ is sampled at a 6.67 kHz rate (to accommodate the 3.2 kHz cutoff of the telephone line) using a 16 bit A/D converter, and high-pass filtered at approximately 200 Hz to remove any dc, low-frequency hum, or noise components which might be present in the signal. An 8-pole LPC analysis is performed on each contiguous 15 ms (100 samples) section of signal using the covariance method of analysis [10]. A total of 67 analyses/s are performed. In

Manuscript received December 27, 1976; revised March 8, 1977.
The authors are with Bell Laboratories, Murray Hill, NJ 07974.

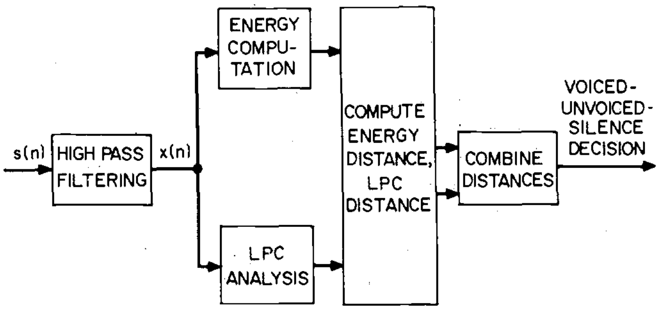

Fig. 1. Block diagram of signal classification method.

addition to the LPC parameters, the log energy of the 15 ms section is computed. For notational purposes, we refer to the LPC set for the ith frame as

$$a_i = (a(0), a_i(1), a_i(2), \cdots, a_i(8)) \tag{1}$$

where $a\,(0) = 1$, and the log energy for the ith frame as

$$E_i = 10 \log_{10} \left[\sum_{n=n_0}^{n_0+149} x^2(n) \right] \tag{2}$$

where $x(n)$ is the high-pass filtered signal and n_0 is the index of the initial sample in the ith frame.

The next step in the method is the computation of distances to the stored patterns for each of three signal classes. Both an energy and an LPC distance are computed. The energy distance is simply a normalized Euclidean distance of the form

$$D_E(j) = \left| \frac{E_i - \bar{E}(j)}{\sigma_E(j)} \right| \tag{3}$$

where $j = 1$, 2, and 3 represent silence, unvoiced speech, and voiced speech, respectively, and $\bar{E}(j)$ is the average log energy (as obtained from a training set of data) of the jth signal class, and $\sigma_E(j)$ is the standard deviation of the log energy for the jth signal class.

The LPC distance is based on the measure proposed by Itakura [11], and is of the form

$$D_a(j) = \frac{(a - m_j)(\phi)(a - m_j)^t}{(a\phi a^t)} \tag{4}$$

where m_j is a mean vector (with $m_0 = 1$) of LPC coefficients (again obtained from a training set of data) for the jth signal class, and ϕ is the matrix of correlations for the current frame. The denominator term in (4) is simply the residual error of the LPC analysis. The LPC distance measure of (4) is essentially a covariance weighting of the LPC coefficients, and is of the form $(E_j - E_{\min})/E_{\min}$ where E_j is the residual error from the use of the prediction given by m_j with the current frame of speech, and $E_{\min}$ is the minimum residual error for the current frame. The resulting distance is therefore a dimensionless, nonnegative quantity. This distance measure has been shown to provide a sensitive measure of similarity between frames with different sets of LPC coefficients [12]-[13]—hence its suitability for voiced-unvoiced-silence classification.

Based on the two sets of distances, $D_E(j)$ and $D_a(j)$, $j = 1, 2, 3$ and a small amount of logic, the final signal classification is made. Fig. 2 shows a flow diagram of the classification

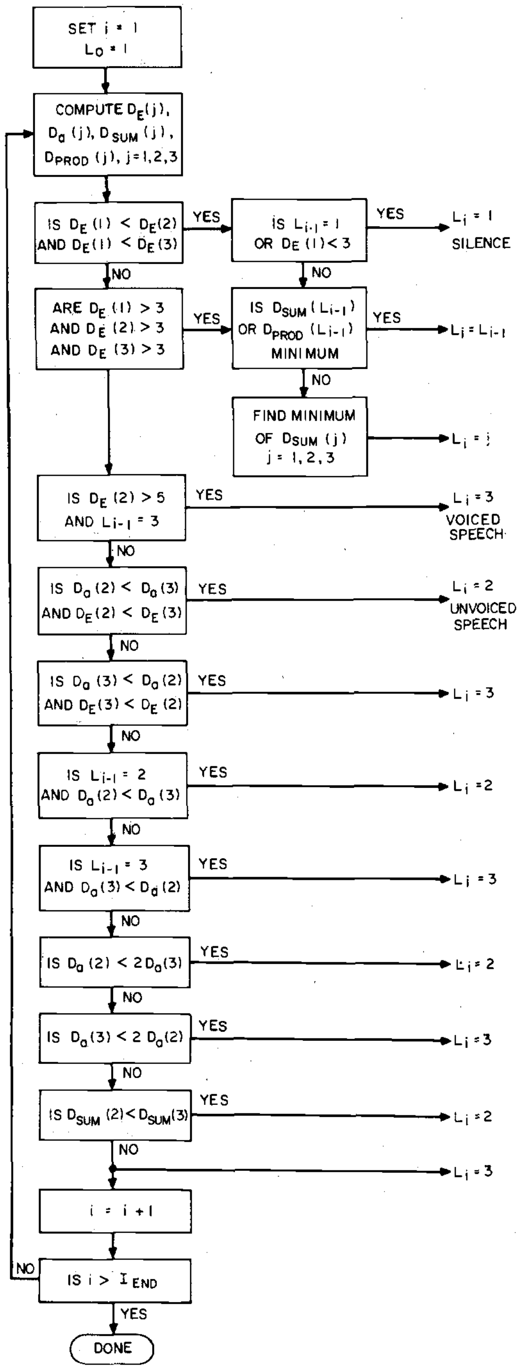
SET i = 1
L0 = 1
COMPUTE DE(j), Da(j), DSUM(j), DPROD(j), j=1,2,3
IS DE(1) < DE(2) AND DE(1) < DE(3)
YES
IS Li-1 = 1 OR DE(1) < 3
YES
Li = 1 SILENCE
NO
ARE DE(1) > 3 AND DE(2) > 3 AND DE(3) > 3
YES
IS DSUM(Li-1) OR DPROD(Li-1) MINIMUM
YES
Li = Li-1
NO
FIND MINIMUM OF DSUM(j) j = 1,2,3
Li = j
IS DE(2) > 5 AND Li-1 = 3
YES
Li = 3 VOICED SPEECH
NO
IS Da(2) < Da(3) AND DE(2) < DE(3)
YES
Li = 2 UNVOICED SPEECH
NO
IS Da(3) < Da(2) AND DE(3) < DE(2)
YES
Li = 3
NO
IS Li-1 = 2 AND Da(2) < Da(3)
YES
Li = 2
NO
IS Li-1 = 3 AND Da(3) < Da(2)
YES
Li = 3
NO
IS Da(2) < 2Da(3)
YES
Li = 2
NO
IS Da(3) < 2 Da(2)
YES
Li = 3
NO
IS DSUM(2) < DSUM(3)
YES
Li = 2
NO
Li = 3
i = i + 1
NO
IS i > IEND
YES
DONE

Fig. 2. Flow diagram of algorithm for combining energy distance and LPC distance to make signal classification.

algorithm. The most variable of the three signal classes is silence, so the algorithm first makes a decision as to whether the signal is silence based on the energy distances and one frame of memory. Thus, the first step is to classify the signal as silence if $D_E(1)$ is smaller than both $D_E(2)$ and $D_E(3)$ and if the value of $D_E(1)$ is less than 3 (standard deviations from the mean), or if the previous frame was classified as silence. This first step is based on the observation that energy is a much more reliable feature for classifying a signal as silence than the LPC distance.

If the minimum energy distance is not that of silence, a check is made to see if $D_E(j) \geqslant 3$ for all j, in which case either

TABLE I
TRAINING DATA USED IN THE ALGORITHM

	Signal Class		
	Silence (L=1)	Unvoiced Speech (L=2)	Voiced Speech (L=3)
$\bar{E}$	31.5 (db)	54.9 (db)	69.5 (db)
σ_E	3.3 (db)	2.6 (db)	6.2 (db)
m(1)	-0.427	-0.321	-1.025
m(2)	0.250	0.115	0.757
m(3)	-0.422	-0.111	-0.627
m(4)	0.221	0.243	0.620
m(5)	-0.198	-0.047	-0.346
m(6)	0.199	0.174	0.431
m(7)	-0.088	-0.027	-0.124
m(8)	0.140	0.067	0.148
Number of Frames	181	58	203

the one frame of memory is used to guide the decision, or the minimum combined distance is chosen. There are at least two ways of combining the two distances. One simple way is to sum the distances to give

$$D_{\mathrm{SUM}}(j) = D_E(j) + D_a(j). \tag{5}$$

The rationale behind summing distances is that each distance is a dimensionless quantity, and if the features were independent, the distances associated with the features would add. The problem with summing distances is that the features of energy and LPC coefficients are not independent; hence, some measure of their correlation should be used. The second way of combining the distances is multiplicatively to give

$$D_{\mathrm{PROD}}(j) = D_E(j) \cdot D_a(j). \tag{6}$$

The rationale here is that the distances are related to probabilities of occurrence of the three classes, and an overall distance based on the product of distances gives an overall measure of probability. Although a good theoretical justification of using a product distance is lacking, it is easily seen that in cases when $D_E(j)$ is large for all j (as occurs when $D_{\mathrm{PROD}}(j)$ is used by the algorithm), then summing distances is meaningless since $D_a(j)$ will generally be much less than $D_E(j)$. In such cases, $D_{\mathrm{PROD}}(j)$ provides a more reasonable method of combining distances. Theoretically, the proper way of combining $D_a(j)$ and $D_E(j)$ would be to account for the correlation between E and a. For reasons related to the specific implementation, such a combined distance was not used.

Based on the combined distance, the signal class is chosen based on either memory (if either combined distance is a minimum) or strictly on the $D_{\mathrm{SUM}}(j)$ distance, as shown in Fig. 2.

At this point in the algorithm, we have completely eliminated silence as the signal classification in that the signal has either been classified as silence, or it has not in which case only the unvoiced or voiced speech classes remain. The remainder of the algorithm is a series of steps which use $D_a(j)$, $D_E(j)$, and $D_{\mathrm{SUM}}(j)$ to decide whether to classify the signal as unvoiced or voiced speech. For cases in which $D_a(j)$ and $D_E(j)$ are both minimum for the same value of j, the signal class is chosen as that value. Otherwise, the final decision is based on the exact values and relationships between $D_a(2)$, $D_a(3)$, $D_E(2)$, and $D_E(3)$, as shown in Fig. 2. The different thresholds, and the sequence of logical decisions shown in Fig. 2, are based on experimentation with the method.

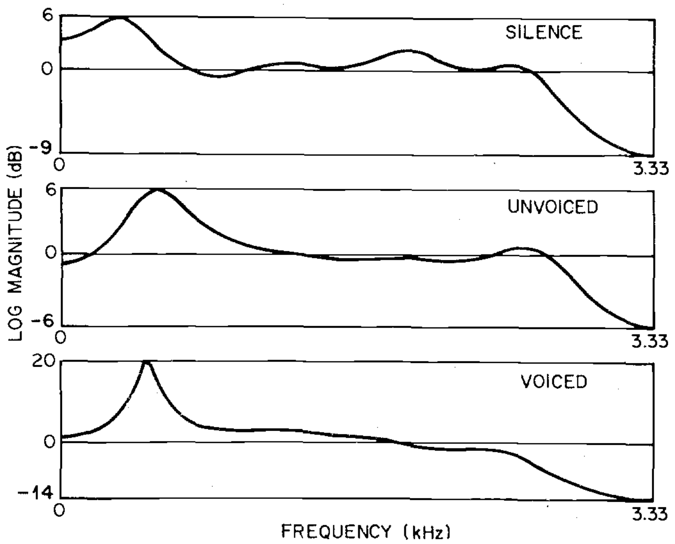

Fig. 3. Average spectra for three signal classes from the training data.

Before giving examples to show how the algorithm performed on some test data, we first discuss some issues involved in obtaining the reference data for each of the signal classes.

A. Training the Algorithm

In order to compute the energy and LPC distances for each signal class, a set of reference frames must be used to train the algorithm, i.e., to provide values for $\bar{E}(j)$, $\sigma_E(j)$, and m_j in (3) and (4).

The way in which these quantities are computed is as follows. Consider a set of frames with each frame manually classified as silence, unvoiced, or voiced speech. Thus, for each frame we have E_i, a_i, and k_i where $k_i = 1, 2,$ or 3, depending on the manual classification. The quantities $\bar{E}(j)$, $\sigma_E(j)$, and m_j are obtained as

$$\bar{E}(j) = \frac{1}{N_j} \sum_{\substack{i=1 \\ (k_i=j)}}^{I} E_i \tag{7}$$

$$\sigma_E(j) = \left[\frac{1}{N_j} \sum_{\substack{i=1 \\ (k_i=j)}}^{I} E_i^2 - (\bar{E}(j))^2 \right]^{1/2} \tag{8}$$

and

$$m_j = \frac{1}{N_j} \sum_{\substack{i=1 \\ (k_i=j)}}^{I} a_i \tag{9}$$

where N_j is the number of frames in the training set for which $k_i = j$ and I is the total number of frames in the training set.

Table I gives values of $\bar{E}(j)$, $\sigma_E(j)$, and m_j for the three signal classes discussed in this paper, and Fig. 3 shows LPC spectra derived from the m_j's of (9). The spectra were obtained from the relation

TABLE II
COMPARISONS OF VALUES OF m FOR VOICED SPEECH AS A FUNCTION OF THE PARAMETER SET BEING AVERAGED

	Parameter Set Being Averaged			
	LPC Coefficients	Log Area Ratios	PARCOR Coefficients	Autocorrelation Coefficients
m(1)	-1.025	-1.067	-0.983	-0.973
m(2)	0.757	0.951	0.840	0.667
m(3)	-0.627	-0.770	-0.688	-0.599
m(4)	0.620	0.733	0.664	0.678
m(5)	-0.346	-0.491	-0.431	-0.483
m(6)	0.431	0.536	0.506	0.508
m(7)	-0.124	-0.132	-0.122	-0.273
m(8)	0.148	0.151	0.148	0.219

$$M(e^{j\omega}) = 20 \log_{10} \left[\frac{1}{1 - \sum_{k=1}^{8} m_j(k) e^{-j\omega k}} \right]. \qquad (10)$$

As seen in Table I, the average log energies for silence, unvoiced, and voiced speech were 31.5, 54.9, and 69.3 dB, respectively; however, the standard deviations were 3.3, 2.6, and 6.2 dB, respectively. Thus, considerable overlap between the individual distributions existed.

Examination of the "average" spectra for the three signal classes (Fig. 3) shows a strong similarity between the spectra for silence and unvoiced speech, and some fairly prominent differences for voiced speech sounds. For voiced sounds, a large spectral range (34 dB) is obtained, with a noticeable trend in the spectral shape due to a prominent first formant, and quite broad second and third formants.

One side issue in the computation of the training data is the way in which m_j is calculated in (9). Theoretically, one cannot average a set of LPC coefficients (the a's), and ensure that the resulting average value represents a stable system. Thus, theoretically, some transformation of the a's should be used which will give a parameter set that can be averaged, and for which stability is guaranteed. There are many such parameter sets which can be used including the autocorrelation coefficients of the impulse response of the linear predictor, the PARCOR coefficients, and the log area ratio coefficients. As a check on the validity of averaging the a's, the training data were transformed to each of the three alternate parameter sets (i.e., log area ratios, PARCOR's, and autocorrelation coefficients); these parameters were averaged, and the resulting average values were transformed back to LPC coefficients. Thus, m_j of (9) was computed as

$$m_j = G^{-1} \left[\frac{1}{N_j} \sum_{\substack{i=1 \\ (k_i=j)}}^{I} G(a_i) \right] \qquad (11)$$

where G and G^{-1} represented the transformations from a's to each of the parameter sets used. The results of this experiment are given in Table II and Fig. 4. Table II shows the values of m_j for each of the three transformations, and Fig. 4 shows the resulting average spectrum for voiced speech. Comparisons of the results show only slight differences in the average spectrum as a function of the transformation used for averaging the a's. Thus, heuristically it can be argued that for these data, averaging a's (untransformed) is valid since it gives a set of values which have the minimum variance, and at the same time gives a stable result.

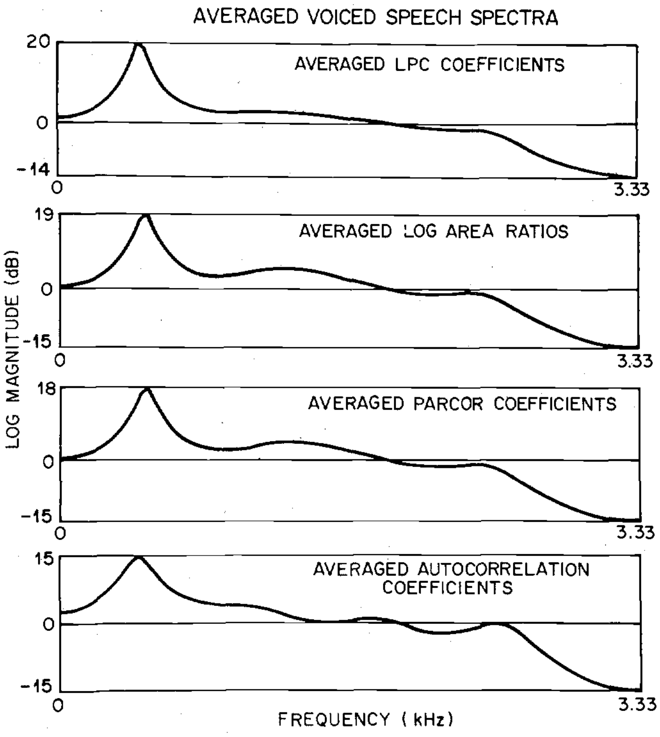

Fig. 4. Comparisons of spectra derived from averaging four different parameter sets.

Aside from the computational issues involved in training the method, another important issue is the selection of a reasonable set of data which is representative of each of the three signal classes to be discriminated. For the class of silence, no major difficulties exist. It is important to obtain a sampling of telephone lines to get a good distribution of telephone silence. For unvoiced sounds, the training set excluded extremely weak fricatives since these were not effectively transmitted over telephone lines. All other unvoiced sounds (bursts, fricatives, etc.) were included in the training set. For voiced sounds, efforts were made to include representative examples of each of the classes of voiced sounds, e.g., vowels, voiced fricatives, nasals, etc. In particular, the training set used in the results to be presented in Section III consisted of 218 frames (15 ms each) of silence, 108 frames of unvoiced speech, and 279 frames of voiced speech. The training data were obtained from a single speaker. Alternative training sets obtained from multiple speakers showed no significant differences in the training statistics.

III. EVALUATION TESTS

To evaluate the method, a total of six speakers (three male, three female) each spoke two utterances over dialed-up telephone lines. None of the six speakers was in the training set, and each individual utterance was made over a different telephone line. A manual classification was made for each 15 ms frame based on both the acoustic waveform and a phonetic transcription of the utterance. Two independent manual classifications were made and each 15 ms frame was given one of the following classifications.

TABLE III
Breakdown of Testing Data into Signal Class and Degree of Certainty

Manual Classification	*Silence*	*Unvoiced*	*Voiced*	*Total*
Certain	285	246	787	1318
Uncertain-Single Classification	14	54	43	111
Uncertain-Double Classification	-	-	-	120

TABLE IV
Analysis of the Signal Classifications

	SS	*SU*	*SV*	*US*	*UU*	*UV*	*VS*	*VU*	*VV*	*S*	*U*	*V*	*Overall*
TS1	93.0	6.0	1.0	4.5	90.7	4.9	0.3	2.0	97.7	7.0	9.4	2.3	4.6
TS2	92.6	6.4	1.0	6.7	84.6	8.7	0.2	2.4	97.4	7.4	15.4	2.6	6.3
TS3	-	-	-	-	-	-	-	-	-	-	-	-	6.4

(a) Percentage Classification Rates for Test Utterances

	SS	*SU*	*SV*	*US*	*UU*	*UV*	*VS*	*VU*	*VV*	*S*	*U*	*V*	*Overall*
TS1	265	17	3	11	233	12	2	16	769	285	246	787	1318
TS2	277	19	3	20	254	26	2	20	808	299	300	830	1429

(b) Breakdown of Number of Occurrences of Each of the Signal Classifications

1) *Certain*—Both manual classifications were in agreement.

2) *Uncertain*—The manual classifications did not agree with each other, or the individual classifications were in doubt. The uncertain intervals were given either a single or a double classification based on the individual results.

A total of 1549 frames were used in the test set. Table III gives a breakdown of the test set into both signal classes and degree of certainty. Table IV gives an analysis of the results obtained on these data sets. For notational purposes, we refer to TS1 as the set of data containing only frames for which the classification was certain, TS2 as the set of data containing all single class decisions, and TS3 as the total set of data (i.e., including the frames for which a double classification was made). The notation SU, etc., in Table IV refers to the case when the signal was silence and was classified as unvoiced. Thus SS, UU, and VV denote correct decisions. For TS1, an overall error rate of 4.6 percent was obtained, for which about 75 percent of the errors occurred at a boundary frame, i.e., one in which a transition occurred between signal classes. Such frames are prone to error since they invariably contain a mix of the signals which occur on both sides of the boundary.

When the single classification uncertain frames were included in the test set (TS2), the error rate was 6.3 percent, of which about 64 percent of the errors occurred at signal boundaries. Finally, the overall error rate for TS3 was 6.4 percent.

A detailed breakdown of the errors is included in Table IV. For all test sets, the errors distributed themselves almost uniformly across the possible error categories, with the exception of silence-voiced, voiced-silence, and voiced-unvoiced errors for which low error rates were attained.

If the categories of silence and unvoiced speech are merged to give the category nonvoiced speech (NV), then an overall error of 2.5 percent is obtained for TS1 and an error rate of 3.6 percent is obtained for TS2. Table V shows a breakdown of the errors for this two-class decision.

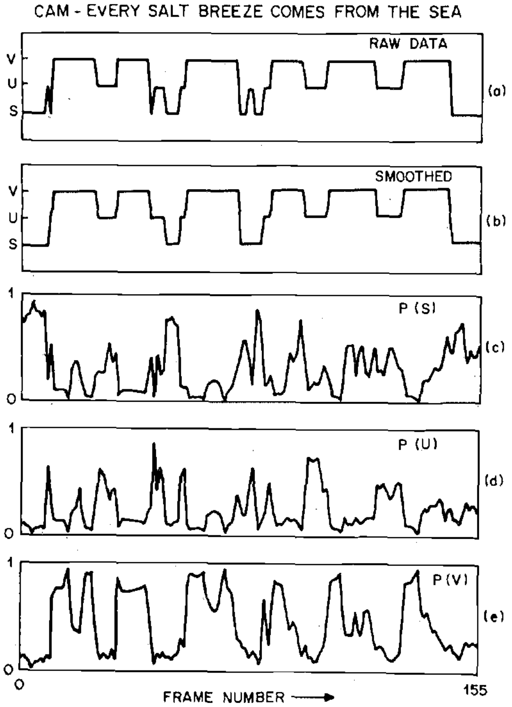

Fig. 5. Analysis results for utterance, "Every salt breeze comes from the sea" by a female talker.

Finally, Figs. 5 and 6 show typical examples illustrating the operation of the method. Part (a) of each figure shows the raw analysis contour as obtained using the algorithm of Fig. 2; part (b) shows the results of nonlinearly smoothing the analysis contour using a median smoother [14]; parts (c), (d), and (e) show plots of a probability measure based on the particular distance used for each signal class, i.e., for silence, the energy distance is generally used, whereas for unvoiced or voiced speech, the LPC distance is generally used. The probability measure is obtained as

$$P(S) = \frac{D_U D_V}{D_S D_U + D_S D_V + D_U D_V} \tag{12a}$$

$$P(U) = \frac{D_S D_V}{D_S D_U + D_S D_V + D_U D_V} \tag{12b}$$

$$P(V) = \frac{D_S D_U}{D_S D_U + D_S D_V + D_U D_V} \tag{12c}$$

where D_S, D_U, and D_V denote the distances for silence, unvoiced, and voiced speech, respectively. This probability measure provides an indication of the probability of correct classification to the extent that if the distance for a particular classification is small relative to the other classification distances, then the probability of the signal being properly classified is high. Fig. 5 shows the results for a female speaker for the utterance, "Every salt breeze comes from the sea." For this utterance, a total of 11 errors were recorded. All these 'errors occurred at boundaries between signal classes.

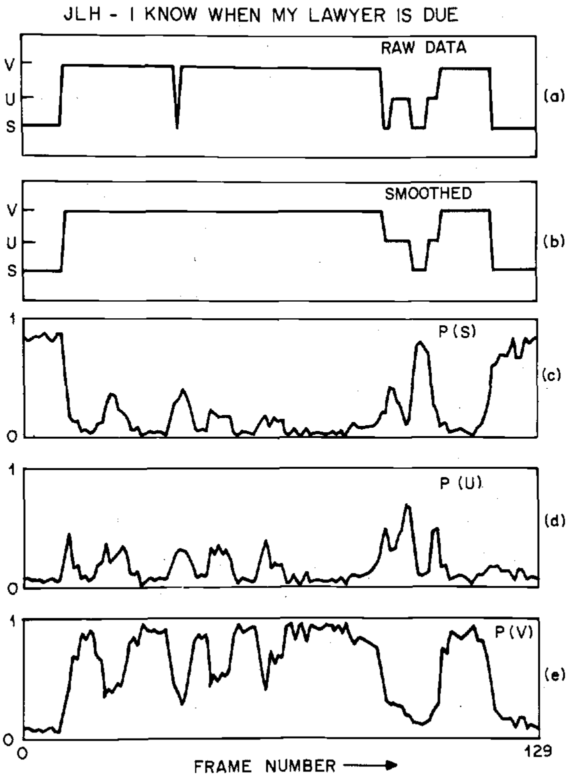

Fig. 6. Analysis results for utterance, "I know when my lawyer is due" by a male talker.

The results of nonlinear smoothing corrected a couple of the boundary errors, and converted a short unvoiced interval between two silence intervals into silence. Otherwise, the contour was unchanged.

Fig. 6 shows a second set of results for the utterance, "I know when my lawyer is due," spoken by a male speaker. For this utterance, a total of three errors were made, two of which occurred at boundaries. The voiced-to-silence error occurring at the beginning of the utterance is readily seen in Fig. 6(a). The nonlinear smoother corrected all three errors for this utterance, giving the contour shown in Fig. 6(b).

IV. Summary

We have presented a new approach to the problem of reliably discriminating among the signal classes of silence, unvoiced, and voiced speech over telephone lines. We have tried to combine some analytical measures of similarity (the LPC distance and the energy distance) with some logic for combining these measures in a meaningful way to give a reliable signal classification. A novel aspect of the analysis is that all the information in the signal is used in computing similarity—not just a small set of features.

The algorithm was tested using a number of different speakers, telephone lines, and utterances. Overall error rates of about 5 percent were obtained, based on manual classification of the frames. This result compares favorably to error scores obtained using statistical decision techniques on telephone line utterances [8]. Currently, the algorithm is being used as an analysis tool in research on speaker-independent recognition of words [15].

TABLE V
Analysis of the Signal Classifications for a Two-Class Decision

	NV-NV	*NV-V*	*V-NV*	*V-V*	*NV*	*V*	*Overall*
TS1	97.0	3.0	2.0	98.0	3.0	2.0	2.5
TS2	95.2	4.8	2.7	97.3	4.8	2.7	3.6

(a) Percentage Classification Rates for 2-Class Decision

	NV-NV	*NV-V*	*V-NV*	*V-V*	*NV*	*V*	*Overall*
TS1	516	15	18	769	531	787	1318
TS2	570	29	22	808	599	830	1429

(b) Number of Occurrences of Each of the Signal Classifications

References

[1] J. L. Flanagan, L. R. Rabiner, D. K. Christopher, and D. E. Bock, "Digital analysis of laryngeal control in speech production," *J. Acoust. Soc. Amer.*, vol. 60, pp. 446-455, Aug. 1976.

[2] B. Gold, "Note on buzz-hiss detection," *J. Acoust. Soc. Amer.*, vol. 36, pp. 1659-1661, 1964.

[3] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, pp. 293-309, Feb. 1967.

[4] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 367-377, Dec. 1972.

[5] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *J. Acoust. Soc. Amer.*, vol. 47, pp. 634-648, Feb. 1970.

[6] J. J. Dubnowski, R. W. Schafer, and L. R. Rabiner, "Real-time digital hardware pitch detector," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 2-8, Feb. 1976.

[7] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 201-212, June 1976.

[8] L. R. Rabiner, C. E. Schmidt, and B. S. Atal, "Evaluation of a statistical approach to voiced-unvoiced-silence analysis for telephone quality speech," *Bell Syst. Tech. J.*, Mar. 1977.

[9] R. J. McAulay, "Optimum classification of voiced speech, unvoiced speech and silence in the presence of noise and interference," M.I.T. Lincoln Lab., Lexington, MA, Tech. Note 1976-7, June 1976.

[10] 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.

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

[12] M. R. Sambur and N. S. Jayant, "Speech encryption by manipulations of LPC parameters," *Bell Syst. Tech. J.*, vol. 55, pp. 1373-1388, Nov. 1976.

[13] J. R. Makhoul, L. Viswanathan, L. Cosel, and W. Russel, "Natural communications with computers: Speech compression research at BBN," BBN Rep. 2976, Dec. 1974.

[14] L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, "Applications of a nonlinear smoothing algorithm to speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 552-557, Dec. 1975.

[15] L. R. Rabiner and M. R. Sambur, "Systems for speaker independent recognition of words," in *Proc. 9th. Int. Conf. Acoust.*, Madrid, Spain, 1977.