Voiced-Unvoiced-Silence Detection Using the Itakura LPC Distance Measure

L. R. Rabiner
M. R. Sambur

Bell Laboratories
Murray Hill, New Jersey 07974

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 3-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 3 classes of signal is obtained during a training session, and an LPC distance metric 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%, with the majority of the errors (about 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 bandlimiting, 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 3-class decision was fairly high (11.7%) over telephone lines.

Based on the results of reference 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 3 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 similaity. In this way a fairly robust, reliable discrimination is obtained.

II. Description of the Algorithm

Figure 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 accomodate the 3.2 kHz cutoff of the telephone line) 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 contingous 15 msec (100 samples) section of signal using the covariance method of analysis. A total of 67 analyses per second are performed. In addition to the LPC parameters, the log energy of the 15 msec section is computed. For notational purposes we refer to the LPC set for the i^{th} frame as

$$a_i = (a_i(1), a_i(2), \ldots, a_i(8)) \tag{1}$$

and the log energy for the i^{th} frame as

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

where x(n) is the highpass filtered signal and n_o is the index of the initial sample in the i^{th} frame.

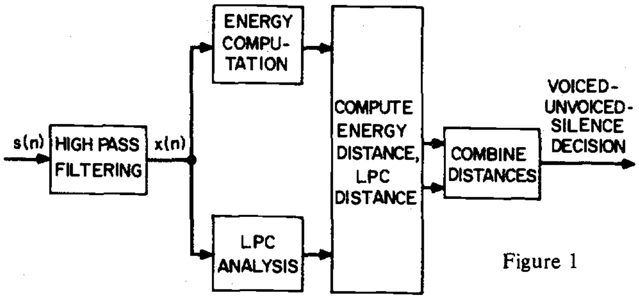

Figure 1

Block diagram of signal classification method.

The next step in the method is the computation of distances to the stored patterns for each of the 3 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 j^{th} signal class, and $\sigma_E(j)$ is the standard deviation of the log energy for the j^{th} signal class.

The LPC distance is based on the measure proposed by Itakura [10], and is of the form

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

where m_j is a mean vector of LPC coefficients, (again obtained from a training set of data) for the j^{th} signal class, and ϕ is the matrix of correlations for the current frame. The denominator term in Eq. (4) is simply the residual error of the LPC analysis. The LPC distance measure of Eq. (4) is essentially a covariance weighting of the LPC coefficients, and has been shown to provide a sensitive measure of similarity between frames with different sets of LPC coefficients [11-12] - 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. Figure 2 shows a flow diagram of the 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 1 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 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

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

The second way of combining the distances is multiplicatively to give

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

Theoretically the proper way of combining $D_a(j)$ and $D_E(j)$ would be to account for the correlation between E and a. Because of the covariance weighting of the a's 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 $D_{SUM}(j)$ distance as shown in Figure 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 hasn't 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 $D_{SUM}(j)$ 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 Figure 2.

2.1. 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 Eqs. (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

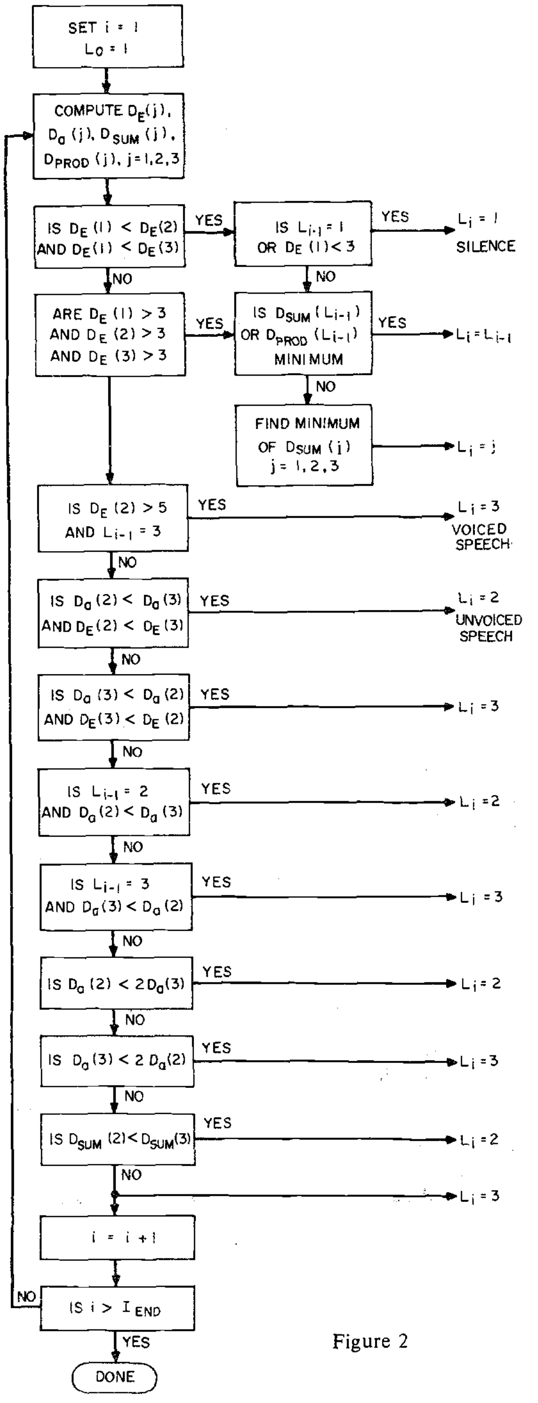
SET i = 1
L₀ = 1
COMPUTE D_E(j), D_a(j), D_SUM(j), D_PROD(j), j=1,2,3
IS D_E(1) < D_E(2) AND D_E(1) < D_E(3)
YES
IS L_{i-1} = 1 OR D_E(1) < 3
YES
L_i = 1 SILENCE
NO
NO
ARE D_E(1) > 3 AND D_E(2) > 3 AND D_E(3) > 3
YES
IS D_SUM(L_{i-1}) OR D_PROD(L_{i-1}) MINIMUM
YES
L_i = L_{i-1}
NO
FIND MINIMUM OF D_SUM(j) j = 1,2,3
L_i = j
IS D_E(2) > 5 AND L_{i-1} = 3
YES
L_i = 3 VOICED SPEECH
NO
IS D_a(2) < D_a(3) AND D_E(2) < D_E(3)
YES
L_i = 2 UNVOICED SPEECH
NO
IS D_a(3) < D_a(2) AND D_E(3) < D_E(2)
YES
L_i = 3
NO
IS L_{i-1} = 2 AND D_a(2) < D_a(3)
YES
L_i = 2
NO
IS L_{i-1} = 3 AND D_a(3) < D_a(2)
YES
L_i = 3
NO
IS D_a(2) < 2D_a(3)
YES
L_i = 2
NO
IS D_a(3) < 2 D_a(2)
YES
L_i = 3
NO
IS D_SUM(2) < D_SUM(3)
YES
L_i = 2
NO
L_i = 3
i = i + 1
IS i > I_END
NO
YES
DONE

Figure 2

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

classification. The quantities $\bar{E}(j)$, $\sigma_E(j)$ and m_j are obtained as

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

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

and

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

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.

Figure 3 shows LPC spectra derived from the m_j's of Eq. (9). The spectra were obtained from the relation

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

Examination of the "average" spectra for the 3 signal classes (Figure 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.

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 3 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 msec each) of silence, 108 frames of unvoiced speech, and 279 frames of voiced speech.

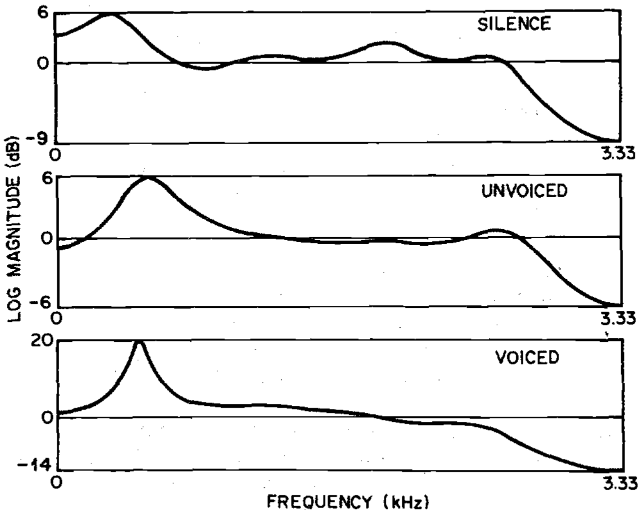

Figure 3

Average spectra for 3 signal classes from the training data.

III. Evaluation Tests

To evaluate the method a total of 6 speakers (3 male, 3 female) each spoke 2 utterances over dialed-up telephone lines. None of the 6 speakers were in the training set, and each individual utterance was made over a different telephone line. A manual classification was made for each 15 msec frame based on both the acoustic waveform, and a phonetic transcription of the utterance. Two independent manual classifications were made and each 15 msec frame was given one of the following 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 1 gives an analysis of the results obtained on three 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 1 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% was obtained, for which about 75% 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%, of which about 64% of the errors occurred at signal boundaries. Finally the overall error rate for TS3 was 6.3%.

If the categories of silence and unvoiced speech are merged to give the category nonvoiced speech (NV) then an overall error of 2.5% is obtained for TS1 and an error rate of 3.6% is obtained for TS2. Table 2 shows a breakdown of the errors for this 2-class decision.

Finally, Figure 4 shows a typical example illustrating the operation of the method. Part a of this figure shows the raw analysis contour; part b shows the results of nonlinearly smoothing the analysis contour using a median smoother [13]; parts c, d, and e show plots of the probability of correct classification 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

$$P(s) = \frac{D_u D_v}{D_s D_u + D_s D_v + D_u D_v} \qquad (11a)$$

	SS	*SU*	*SV*	*US*	*UU*	*UV*	*VS*	*VU*	*VV*
TS1	93.0	6.0	1.0	4.5	90.7	4.9	0.3	2.0	97.7
TS2	92.6	6.4	1.0	6.7	84.6	8.7	0.2	2.4	97.4

(a) Percentage Error Rates for Test Utterances

	SS	*SU*	*SV*	*US*	*UU*	*UV*	*VS*	*VU*	*VV*
TS1	265	17	3	11	233	12	2	16	769
TS2	277	19	3	20	254	26	2	20	808

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

Table 1

Analysis of the Signal Classification Errors

$$P(u) = \frac{D_s D_v}{D_s D_u + D_s D_v + D_u D_v} \tag{11b}$$

$$P(v) = \frac{D_s D_u}{D_s D_u + D_s D_v + D_u D_v} \tag{11c}$$

where D_s, D_u, and D_v denote the distances for silence, unvoiced, and voiced speech respectively. Figure 4 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. The results of nonlinear smoothing corrected a couple of the boundary errors, and converted a short unvoiced interval between two silence interval into silence. Otherwise the contour was unchanged.

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 robust 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% 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. [14].

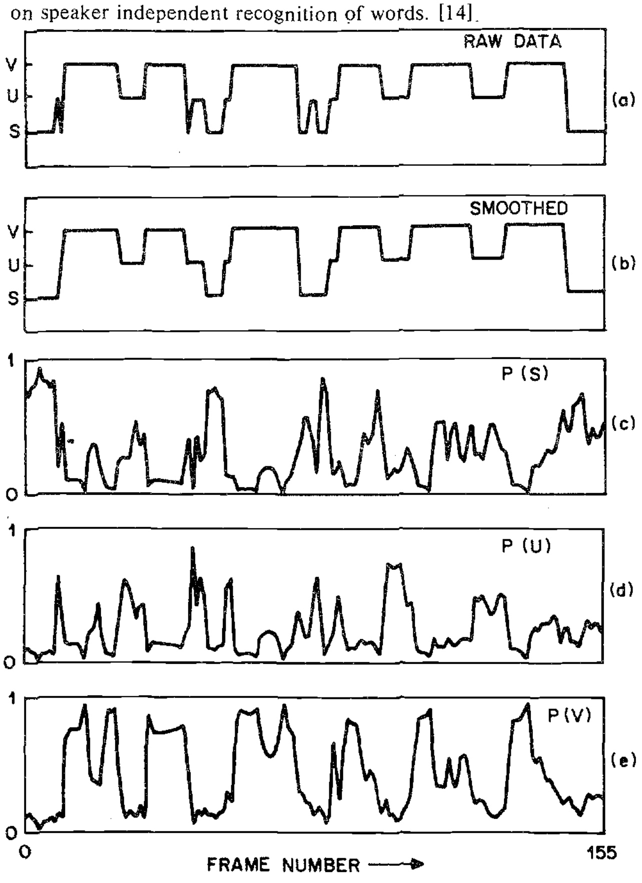

Figure 4

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

	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 Error 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

Table 2

Analysis of the Signal Classification
Errors for a 2-Class Decision

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. Am., Vol. 60, No. 2, pp. 446-455, August 1976.
2. B. Gold, "Note on Buzz-Hiss Detection," J. Acoust. Soc. Am., Vol. 36, pp. 1659-1661, 1964.
3. A. M. Noll, "Cepstrum Pitch Determination," J. Acoust. Soc. Am., Vol. 41, pp. 293-309, February 1967.
4. J. D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation," IEEE Trans. on Audio and Elect., Vol. AU-20, No. 5, 367-377, December 1972.
5. R. W. Schafer and L. R. Rabiner, "System for Automatic Formant Analysis of Voiced Speech," J. Acoust. Soc. Am., Vol. 47, No. 2, pp. 634-648, February 1970.
6. J. J. Dubnowksi, R. W. Schafer, and L. R. Rabiner, "Real-Time Digital Hardware Pitch Detector," IEEE Trans. on Acoustics, Speech, and Signal Proc., Vol. ASSP-24, No. 1, pp. 2-8, February 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. on Acoustics, Speech, and Signal Processing, Vol. ASSP-24, No. 3, pp. 201-121, 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 System Tech. J., March 1977.
9. R. J. McAulay, "Optimum Classification of Voiced Speech, Unvoiced Speech and Silence in the Presence of Noise and Interference," Lincoln Laboratory Technical Note 1976-7, June 1976.
10. F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-23, No. 1, pp. 67-72, February 1975.
11. M. R. Sambur and N. S. Jayant, "Speech Encryption by Manipulations of LPC Parameters," Bell Syst. Tech. J., Vol. 55, No. 9, pp. 1373-1388, November 1976.
12. J. R. Makhoul, L. Viswanathan, L. Cosel, and W. Russel, "Natural Communication with Computers: Speech Compression Research at BBN," BBN Report No. 2976, December 1974.
13. L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, "Applications of a Nonlinear Smoothing Algorithm to Speech Processing," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-23, No. 6, pp. 552-557, December 1975.
14. L. R. Rabiner and M. R. Sambur, "Systems for Speaker Independent Recognition of Words," Proc. 9th ICA, Madrid Spain, 1977.