A Pattern Recognition Approach to Voiced–Unvoiced–Silence Classification with Applications to Speech Recognition

BISHNU S. ATAL, MEMBER, IEEE, AND LAWRENCE R. RABINER, FELLOW, IEEE

Abstract—**In speech analysis, the voiced-unvoiced decision is usually performed in conjunction with pitch analysis. The linking of voiced-unvoiced (V-UV) decision to pitch analysis not only results in unnecessary complexity, but makes it difficult to classify short speech segments which are less than a few pitch periods in duration. In this paper, we describe a pattern recognition approach for deciding whether a given segment of a speech signal should be classified as voiced speech, unvoiced speech, or silence, based on measurements made on the signal. In this method, five different measurements are made on the speech segment to be classified. The measured parameters are the zero-crossing rate, the speech energy, the correlation between adjacent speech samples, the first predictor coefficient from a 12-pole linear predictive coding (LPC) analysis, and the energy in the prediction error. The speech segment is assigned to a particular class based on a minimum-distance rule obtained under the assumption that the measured parameters are distributed according to the multidimensional Gaussian probability density function. The means and covariances for the Gaussian distribution are determined from manually classified speech data included in a training set. The method has been found to provide reliable classification with speech segments as short as 10 ms and has been used for both speech analysis-synthesis and recognition applications. A simple nonlinear smoothing algorithm is described to provide a smooth 3-level contour of an utterance for use in speech recognition applications. Quantitative results and several examples illustrating the performance of the method are included in the paper.**

I. Introduction

THE NEED for deciding whether a given segment of a speech waveform should be classified as voiced speech, unvoiced speech, or silence (absence of speech) arises in many speech analysis systems. A variety of approaches have been described in the speech literature for making this decision [1]-[6]. Methods for voiced-unvoiced (V-UV) decision usually work in conjunction with pitch analysis. For example, in the well-known cepstral pitch detector [2], the V-UV decision is made on the basis of the amplitude of the largest peak in the cepstrum. There are two disadvantages in this approach to V-UV decision. First, the decision is based on a single feature—the degree of voice periodicity. Voiced speech is only approximately periodic; sudden changes in articulation and the idiosyncracies of vocal cord vibrations can produce speech waveforms which are not periodic. In such cases, a feature such as the amplitude of the largest cepstral peak will fail to distinguish voiced speech from unvoiced. In practice, additional features, such as the rate of zero crossings of the speech waveform, the ratio of low to high-frequency energy, etc. must be included in the decision procedure. Second, the V-UV decision is tied to the pitch detection which may be acceptable for speech synthesis applications. But, for other applications, such as speech segmentation or speech recognition, the linking of V-UV decision to pitch detection can result in unnecessary complexity as well as in poorer performance, particularly at the boundaries between voiced and unvoiced speech. For pitch detection, a large speech segment, 30-40 ms long, is necessary, which can result in unwarranted mixing of voiced and unvoiced speech. By separating the V-UV decision from pitch detection, it is possible to perform the V-UV decision on a much shorter speech segment, thereby enabling one to track fast changes in speech from one class to another.

In this paper, we describe a method which uses a pattern recognition approach for classifying a given speech segment into three classes: voiced speech, unvoiced speech, and silence. The pattern recognition approach provides an effective method of combining the contributions of a number of speech measurements—which individually may not be sufficient to discriminate between the classes—into a single measure capable of providing reliable separation between the three classes. The method presented here is essentially a classical hypothesis-testing procedure based on the statistical decision theory. In this method, for each of the three classes, a non-Euclidean distance measure is computed from a set of measurements made on the speech segment to be classified and the segment is assigned to the class with the minimum distance. The distance function is chosen so as to provide minimum classification error for normally distributed measurements. The normal distribution has important advantages due to its computational simplicity. The decision rule in this case is completely determined by the mean vector and the covariance matrix of the probability density function for each class. Results based on the computed one-dimensional distributions of the chosen measurements suggest that the assumption of normal distribution is a reasonable one.

The success of a hypothesis-testing procedure depends, to a considerable extent, upon the measurements or features which are used in the decision criterion. The basic problem is of selecting features which are simple to derive from speech and, yet, are highly effective in differentiating between the three classes: voiced speech, unvoiced speech, and silence. The following five measurements have been used in the implementation described in this paper.

1) Energy of the signal.
2) Zero-crossing rate of the signal.

Manuscript received June 10, 1975; revised January 9, 1976.
The authors are with Bell Laboratories, Murray Hill, NJ 07974.

3) Autocorrelation coefficient at unit sample delay.
4) First predictor coefficient.
5) Energy of the prediction error.

The choice of these particular parameters is based partly on the experimental evidence that the parameters vary consistently from one class to another and partly on our knowledge of the method in which voiced and unvoiced speech sounds are generated in the human vocal tract. We will discuss these reasons in greater detail later in this paper when we describe the nature of the dependence of these parameters on different classes of sounds. Of course, the above set of measurements is not the only possibility and, quite likely, a better choice could be realized by careful evaluation of different parameter sets. The above set of parameters does, however, represent a good compromise between the complexity of their measurement procedures and their ability to discriminate between the three classes reliably across a wide variety of speakers.

The organization of the remainder of this paper is as follows. In Section II, we describe the basic algorithm for making the voiced-unvoiced-silence (VUS) decision. We also discuss the methods for computing the means and covariances of the probability distributions for the three classes. In Section III, we provide some quantitative results about the performance of the algorithm, and discuss its application in an experimental digit-recognition system. An error detection and correction method is also described to correct isolated errors in the VUS contour. Finally, we present several examples of VUS contours obtained during the course of the digit recognition experiments.

II. Speech Measurements and Decision Algorithm

A block diagram of the analysis and decision algorithm is shown in Fig. 1. The speech signal is low-pass filtered to 4 kHz, sampled at 10 kHz, and each sample is quantized with an accuracy of 12 bits. Prior to analysis, the speech signal is high-pass filtered at approximately 200 Hz to remove any dc, low-frequency hum, or noise components which might be present in the speech signal. The high-pass filter has two poles and two zeros and its transfer function is given by

$$H(z) = \frac{1 - 2z^{-1} + z^2}{1 - 2e^{-aT}\cos(bT)z^{-1} + e^{-2aT}z^{-2}}, \tag{1}$$

where

$$a = 130 \cdot 2\pi,$$

$$b = 200 \cdot 2\pi,$$

$$T = 10^{-4}.$$

Following high-pass filtering, the speech is formatted into blocks of 100 samples (an interval of 10 ms at 10 kHz sampling frequency), with each block spaced 100 samples apart. For each block, we define $s(n)$, $n = 1, 2, \cdots, N$, to be the nth sample in the block. The samples N, $N-1$, $N-2$, etc. of the previous block are numbered 0, -1, -2, etc. Thus $s(0)$ is the last sample of the previous block.

A. Measurements

1) Zero-crossing count, N_z, the number of zero crossings in the block.

2) Log energy E_s—defined as

$$E_s = 10 * \log_{10}\left(\epsilon + \frac{1}{N}\sum_{n=1}^{N} s^2(n)\right) \tag{2}$$

where ϵ is a small positive constant added to prevent the computing of log of zero. Obviously, $\epsilon \ll$ mean-squared value of the speech samples. In our implementation, speech samples ranged between ± 2048, and ϵ was set to 10^{-5}.

3) Normalized autocorrelation coefficient at unit sample delay, C_1, which is defined as

$$C_1 = \frac{\sum_{n=1}^{N} s(n)s(n-1)}{\sqrt{\left(\sum_{n=1}^{N} s^2(n)\right)\left(\sum_{n=0}^{N-1} s^2(n)\right)}}. \tag{3}$$

4) First predictor coefficient, α_1, of a p-pole ($p = 12$ typically) linear predictive coding (LPC) analysis using the covariance method.

5) Normalized prediction error, E_p, expressed in decibels, which is defined as

$$E_p = E_s - 10 * \log_{10}\left(10^{-6} + \left|\sum_{k=1}^{p} \alpha_k \phi(0, k) + \phi(0, 0)\right|\right) \tag{4}$$

where E_s is defined above, and

$$\phi(i, k) = \frac{1}{N}\sum_{n=1}^{N} s(n-i)s(n-k) \tag{5}$$

is the (i, k) term of the covariance matrix of the speech samples, and the α_k's are the predictor coefficients. The predictor coefficients are obtained by minimizing the mean-squared prediction error E defined as

$$E = \frac{1}{N}\sum_{n=1}^{N}\left[s(n) + \sum_{k=1}^{p} \alpha_k s(n-k)\right]^2. \tag{6}$$

Note that, p samples of the previous block are required for computing the covariance term $\phi(i, k)$ in (5). In (4), the quantity inside the absolute sign is positive by definition; however, roundoff errors in the computation may yield a small negative value which can result in the computing of the log of a negative number. Once again 10^{-6} is added for reasons identical to ones mentioned for (2).

Before proceeding to a detailed discussion of the decision algorithm, it is worthwhile discussing the expected nature of variation of each of the above parameters for the three classes. Measured distributions of these parameters for the different classes are shown in Figs. 2-6. These results will be discussed later in Section II-C.

The zero-crossing parameter, N_z, is an indicator of the frequency at which the energy is concentrated in the signal spectrum. Voiced speech is produced as a result of excitation of

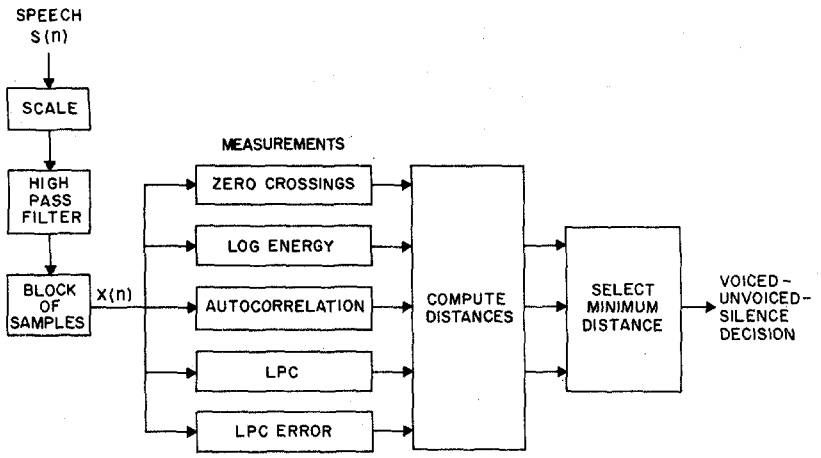

Fig. 1. Block diagram of the analysis system.

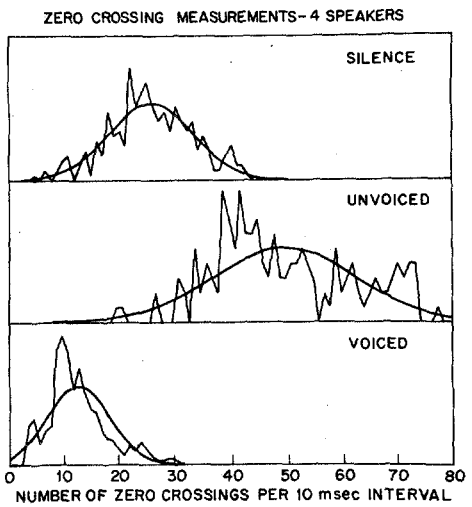

Fig. 2. Theoretical and measured probability density functions for the zero-crossing measurement.

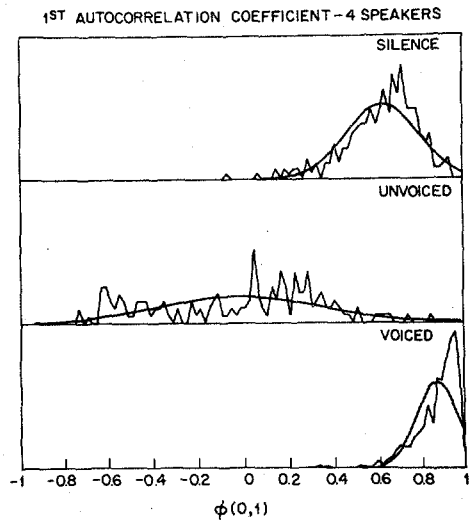

Fig. 4. Theoretical and measured probability density functions for the first autocorrelation coefficient measurement.

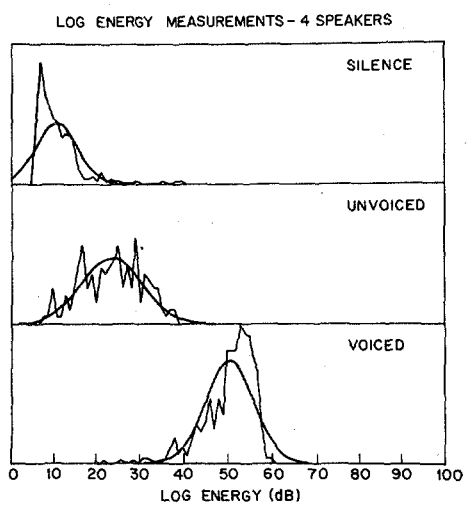

Fig. 3. Theoretical and measured probability density functions for the energy measurement.

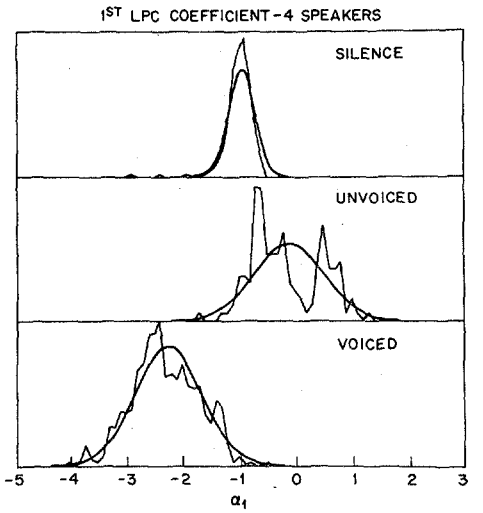

Fig. 5. Theoretical and measured probability density functions for the first LPC coefficient measurement.

approximately 12 dB/octave thereby producing a concentration of energy at low frequencies in the speech signal. Voiced speech usually shows a low zero-crossing count—typically in the range 0 to 30. Unvoiced speech is produced due to excitation of the vocal tract by a noise-like source at a point of constriction in the interior of the vocal tract. While the spectrum of the noise source is flat, the vocal-tract response usually increases with frequency. Thus, the unvoiced speech has a concentration of energy at high frequencies and shows a high zero-crossing count—typically in the range 10 to 100.

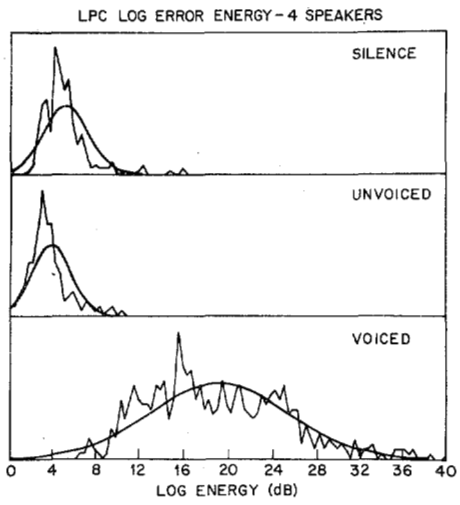

Fig. 6. Theoretical and measured probability density functions for the LPC error measurement.

The zero-crossing count for silence can vary considerably from one speaking environment to another reflecting the variable characteristics of the room noise. Quite often, the spectrum of room noise is concentrated at low and middle frequencies. In such cases, the zero-crossing count for silence would be expected to be lower than for unvoiced speech, but quite comparable to that for voiced speech.

The energy parameter, E_s, depends upon many factors such as the sensitivity of the microphone and the amplifiers, the quantizing characteristics of the analog/digital (A/D) converter, etc. Generally speaking, the energy of voiced sounds is much higher than the energy of silence. The energy of unvoiced sounds is usually lower than for voiced sounds, but often higher than for silence.

The parameter C_1 is the correlation between adjacent speech samples, and, by definition, varies between -1 and +1. Due to the concentration of low-frequency energy in voiced sounds, adjacent samples of voiced speech waveform are highly correlated and the parameter C_1 is close to unity. On the other hand, the correlation is close to zero for unvoiced speech.

The first LPC coefficient is usually thought as a number coming out of LPC analysis. It can be shown that it is identical (with a negative sign) to the cepstrum of the signal at unit sample delay [7], [8]. Thus α_1 is the negative of the Fourier component of the log spectrum at unit sample delay. Since spectra of the three classes—voiced, unvoiced, and silence—differ considerably, so does the first LPC coefficient. It can vary from a value of about -5 for voiced speech to a value of about 1 for unvoiced speech.

The normalized prediction error is a measure of the nonuniformity of the spectrum. More precisely, it is the ratio of the geometric to the arithmetic mean of the spectrum [9]. The more nonuniform the spectrum is, the smaller is the prediction error. The spectrum of voiced speech has a well-defined formant structure which results in smaller prediction error as compared to unvoiced speech or silence. The parameter E_p defined in (4) can vary from 0 to about 40 dB. It may be pointed out that it is the only measurement among the five which does not use the low-frequency predominance of voiced speech and the high-frequency predominance of unvoiced speech to differentiate between the classes. The prediction error depends solely upon the variation of the spectrum from one frequency to another and not the location of these frequency components.

The five parameters discussed above are correlated with each other. These correlations vary between the parameters and between the classes. The decision algorithm discussed in the next section makes use of these correlations to optimally combine their contributions in differentiating between the classes.

B. Decision Algorithm

As shown in Fig. 1, the five measurements are used to classify the block of signal as either silence, unvoiced, or voiced speech. To make this decision, a classical minimum probability-of-error decision rule is used in which it is assumed that the joint probability density function of the possible values of the measurements for the ith class is a multidimensional Gaussian distribution with known mean $\boldsymbol{m}_i$ and covariance matrix W_i, where $i = 1, 2, 3$ corresponds to class 1 (silence), class 2 (unvoiced), and class 3 (voiced), respectively. The assumption of normal distribution for the measurements can be justified from several considerations: first, for the decision rule to be correct, it is not necessary that the distribution be exactly normal. In the case of unimodal distributions, it is sufficient that the distribution be normal in the center of its range—a property often found to be true for physical measurements. Moreover, as mentioned earlier, the decision rule is optimum for a class of probability densities which are related to the Gaussian density through arbitrary monotonic functional relationships. Finally, the decision rule based on the normal distribution requires information only about the first two moments of the distribution. Accurate estimation of higher order moments is usually difficult in practical situations.

Let $\boldsymbol{x}$ be an L-dimensional column vector (in our case, $L = 5$) representing the measurements, that is, the kth component of $\boldsymbol{x}$ is the kth measurement. The L-dimensional Gaussian density function for $\boldsymbol{x}$ with mean vector $\boldsymbol{m}_i$ and covariance matrix W_i is given by

$$g_i(\boldsymbol{x}) = (2\pi)^{-L/2} |W_i|^{-1/2} \exp\left[-\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{m}_i)^t W_i^{-1} (\boldsymbol{x} - \boldsymbol{m}_i)\right], \tag{7}$$

where W_i^{-1} is the inverse of the matrix W_i,[1] $|W_i|$ is the determinant of W_i, and the superscript t denotes the transpose of a vector. The decision rule which minimizes the probability of erroneous classifications states that the measurement vector $\mathbf{x}$ should be assigned to class i if

$$p_i g_i(\mathbf{x}) \geq p_j g_j(\mathbf{x}) \quad \text{for all} \quad i \neq j, \tag{8}$$

where p_i is the *a priori* probability that $\boldsymbol{x}$ belongs to the ith class [10]. Since $\ln y$ is a monotonically increasing function of its argument y, the decision rule of (8) can be considerably simplified and rewritten as follows.

[1] It is assumed that W_i is a nonsingular matrix. A singular W_i implies that one of more measurements are linear combinations of the remaining measurements. Such redundant measurements can be identified by determining the eigenvectors of W_i with zero eigenvalue.

Decide class i if

$$d_i(x) = \tfrac{1}{2}(x - m_i)^t W_i^{-1}(x - m_i) + \tfrac{1}{2}\ln|W_i| - \ln P_i \leqslant d_j(x) \quad \text{for all } i \neq j. \tag{9}$$

The last two terms on the right side of (9) do not depend on the measurement vector x, and can be thought of as a constant term representing a certain bias towards the ith class. As a practical matter, we have found that such a bias term—even if it is carefully selected for the measurements—does not provide any significant advantage over a decision rule based exclusively on the first term on the right side of (9). Thus, instead of the decision rule of (9), the quantity or distance measure $\hat{d}_i$, defined as

$$\hat{d}_i = (x - m_i)^t W_i^{-1}(x - m_i) \tag{10}$$

is computed and the index i is chosen such that $\hat{d}_i$ is minimized. Finally, based on the individual distances $\hat{d}_1$, $\hat{d}_2$, and $\hat{d}_3$, a probability measure P_i, $i = 1, 2, 3$ is created by the rule

$$P_1 = \frac{\hat{d}_2\hat{d}_3}{\hat{d}_1\hat{d}_2 + \hat{d}_2\hat{d}_3 + \hat{d}_1\hat{d}_3}, \tag{11a}$$

$$P_2 = \frac{\hat{d}_1\hat{d}_3}{\hat{d}_1\hat{d}_2 + \hat{d}_2\hat{d}_3 + \hat{d}_1\hat{d}_3}, \tag{11b}$$

and

$$P_3 = \frac{\hat{d}_1\hat{d}_2}{\hat{d}_1\hat{d}_2 + \hat{d}_2\hat{d}_3 + \hat{d}_1\hat{d}_3}. \tag{11c}$$

The higher the probability measure P_i for the class having the minimum distance $\hat{d}_i$, the more assurance the algorithm has that it has chosen the correct class. These probabilities are useful for correcting isolated classification errors. For example, they have been used to aid a smoothing algorithm which has been applied to the analysis data in the speech recognition experiments.

C. Estimation of the Means and the Covariances

In order to use the above decision algorithm, a training set of data is required to obtain the mean vector and the covariance matrix for each class. This training set is created by manually segmenting natural speech into regions of silence, unvoiced speech, and voiced speech. The speech segments are subdivided into 10-ms blocks and the set of measurements defined in Section II-A is made on each block of data. These measurements, along with the manual classification of the interval, are saved in a training set file. If we let $x_i(n)$ denote the measurement vector for the nth block of the class ($i = 1$, 2, or 3) and N_i denote the number of 10-ms blocks manually classified as class i in the training set, then, the mean vector m_i and the covariance matrix W_i of class i are given by

$$m_i = \frac{1}{N_i}\sum_{n=1}^{N_i} x_i(n), \tag{12}$$

and

$$W_i = \frac{1}{N_i}\sum_{n=1}^{N_i} x_i(n)x_i^t(n) - m_i m_i^t. \tag{13}$$

The minimum value of N_i necessary to ensure nonsingularity of the covariance matrix W_i is obviously L. In practice, however, a much larger value of N_i is desirable for several reasons. First, for a training set to be representative of the entire population, it must include a variety of speech utterances spoken by a number of speakers. Second, manual segmentation of speech into the three classes is not always perfect. The influence of segmentation errors can be minimized by using a large training set. It is also necessary for the satisfactory operation of the algorithm that the recording conditions—frequency response of the recording system, background noise conditions, etc.—remain reasonably stable. A new set of training data is usually needed if there are drastic changes in the recording conditions. In practice, a good training set for voiced or unvoiced speech can be obtained without difficulty. The greatest difficulty is encountered in the characterization of the "silence" class which is strongly dependent on the speaking environment.

Table I shows the means, standard deviations, and the normalized covariance matrices[2] for the three classes for a typical set of training data. The index i refers to the class; $i = 1$ is silence, $i = 2$ is unvoiced speech, and $i = 3$ is voiced speech. These recordings were made using a dynamic microphone in a double-walled soundproof booth and stored on magnetic tape using a high-quality analog tape recorder. Approximately 6 s of speech for each of 4 speakers (2 male, 2 female) were used in the training set. The columns in Table I correspond to the five measurements discussed earlier. The off-diagonal terms of the covariance matrices are a measure of the correlation between the different parameters. If the measurements were all independent and uncorrelated, then all off-diagonal elements would be 0. It can be seen that the magnitudes of the off-diagonal elements vary from 0.00 to 0.96, indicating varying degree of correlations between the different parameters.

The means and standard deviations of Table I characterize completely the assumed one-dimensional normal probability density functions for each of the measurements, for each class. These normal distributions are compared with the actual computed one-dimensional distributions for each of the five measurement variables in Figs. 2-6, respectively.

Fig. 2 shows the results for the zero-crossing measurements. It can be seen that a simple Gaussian fit to the data is quite good for all three classes. As mentioned earlier, high zero crossings tend to distinguish unvoiced sounds from silence and voiced sounds. However, the distributions overlap indicating that the zero crossings alone—as in the case for the other four parameters too—cannot separate the three classes.

Fig. 3 shows the results for the log energy measurement. In this case the Gaussian fit for both silence and voiced speech is not quite as good. The actual distributions are decidedly skewed and therefore a symmetrical distribution is not completely appropriate. As seen from Table I, the difference in mean between the log energy for voiced speech (50.6 dB) and for silence (10.8 dB) provides an underbound to the estimate

[2]The term in the ith row and jth column of the normalized matrix is obtained by dividing the corresponding term of the matrix W by the square-root of the product of the ith and jth diagonal terms.

TABLE I
TYPICAL MEANS, STANDARD DEVIATIONS, AND COVARIANCE MATRICES FOR THE THREE CLASSES (FOUR SPEAKERS USED IN TRAINING SET)

	Zero Crossings	Log Energy	First Auto-correlation	First LPC	LPC Log Error
1) Silence ($i = 1$)					
Mean	25.663	10.781	0.649	−0.935	4.976
Standard Deviation	7.534	4.715	0.158	0.234	1.994
Covariance Matrix	1.000	−0.032	−0.842	0.386	−0.629
	−0.032	1.000	−0.098	−0.558	0.580
	−0.842	−0.098	1.000	−0.442	0.596
	0.386	−0.558	−0.442	1.000	−0.710
	−0.629	0.580	0.596	−0.710	1.000
2) Unvoiced ($i = 2$)					
Mean	49.914	23.439	0.007	−0.107	3.661
Standard Deviation	12.680	6.985	0.365	0.618	1.763
Covariance Matrix	1.000	0.471	−0.959	0.909	−0.019
	0.471	1.000	−0.454	0.437	0.447
	−0.959	−0.454	1.000	−0.947	0.028
	0.909	0.437	−0.947	1.000	−0.044
	−0.019	0.447	0.028	−0.044	1.000
3) Voiced ($i = 3$)					
Mean	12.775	50.608	0.881	−2.256	18.944
Standard Deviation	5.546	5.530	0.090	0.582	6.151
Covariance Matrix	1.000	0.250	−0.882	0.276	−0.626
	0.250	1.000	−0.200	−0.130	−0.051
	−0.882	−0.200	1.000	−0.380	0.728
	0.276	−0.130	−0.380	1.000	−0.603
	−0.626	−0.051	0.728	−0.603	1.000

of the signal-to-background noise ratio of the recording environment. It can also be seen from Fig. 3 that high values of log energy tend to separate voiced speech from silence and unvoiced speech.

Fig. 4 shows results for the first autocorrelation coefficient. For voiced speech, the distribution is particularly skewed. These results suggest that a suitable nonlinear function of the autocorrelation coefficient, such as the inverse hyperbolic tangent, would be more appropriate here. The distributions, for silence and unvoiced speech, are well approximated by the Gaussian. For unvoiced speech, the distribution has a particularly large variance.

Figs. 5 and 6 show the results for the first LPC coefficient for a 12-pole analysis, and the log energy of the resulting prediction error. In almost all cases, the Gaussian fit to the distribution is a reasonable one.

Before proceeding to show some examples of how the algorithm worked in some typical cases, it is worthwhile discussing the limitations imposed by the necessity for training the algorithm. Strictly speaking, the training data is particular to one set of recording conditions. Thus, whenever, the transmission system varies or the background noise level varies, a new set of training data is required. If the recording conditions differ considerably from one occasion to another, it may be possible to adapt the algorithm by continuously updating the training data based on some measure of the relative distances to each of the classes. Whether or not the algorithm can successfully adapt is as yet unclear; however, it is worth investigation. It is also our experience that the training data obtained from one speaker can be used for another—even from male speakers to female speakers. A second limitation in the training is the necessity for manually locating the regions of silence, unvoiced, and voiced speech in an utterance—a task which could be both tedious, and time consuming. It is also difficult to locate the exact time at which the speech becomes voiced or unvoiced. This does not however present a major problem because the times at which speech changes from voiced to unvoiced, or vice versa, need not be accurately pinpointed. In fact, using only intervals where the character of the signal is completely clear in the training set tends to enhance the capability of the method, rather than detract from it. This is so because the distance measure essentially provides a smooth transition between voiced and unvoiced and silence. Thus, one need not include these ambiguous cases in the training set.

III. RESULTS

A. Numerical Evaluation

The algorithm has been tested on a wide variety of speech material for both speech synthesis and segmentation applications. The V-UV decision has performed satisfactorily as a part of a speech analysis-synthesis system based on linear prediction [4]. We present in this section some numerical results regarding the performance of the algorithm for classifying speech into the three classes. We will also present results

about the comparative performance of each of the five individual speech measurements.

For the results presented in this section, the experimental speech data was divided into a training set and a testing set. The training set was used to compute the means and the covariance matrices for the three classes. The speech data in the training set consisted of two utterances "Should we chase those young outlaw cowboys?" and "Few thieves are never sent to the jug" spoken by a male speaker. The testing set was used to evaluate the performance of the algorithm. The speech data in the testing set consisted of two utterances: a sentence "It's time we rounded up that herd of asian cattle" spoken by a male speaker different from the one who spoke the utterance in the training set and another sentence "High altitudes jets whiz past screaming" spoken by a female speaker. The speech material was recorded in an anechoic chamber with a condenser microphone. The average signal-to-noise ratio for voiced speech was 34 dB and for unvoiced speech was 14 dB.

Fig. 7 shows the waveform of a portion of the speech data used in the training set. Manually classified regions corresponding to the three classes are also marked on the figure with the symbol "V" indicating voiced speech, the symbol "U" indicating unvoiced speech, and the symbol "S" indicating silence. The mean and covariance matrix data for the three classes is shown in Table II. The algorithm was first run on the training set itself to see how well it performs with the data on which it was trained. The results are presented in Table III(a) in the form of a matrix of incorrect identifications (confusion matrix). The total number of 10-ms long voiced, unvoiced, and silent segments in the training set were 313, 57, and 76, respectively.

The algorithm was next used on the speech data in the test set. The total number of 10-ms long voiced, unvoiced, and silent segments in the test set were 375, 82, and 94, respectively. The confusion matrix for this case is presented in Table III(b). Most of the errors occurred at the boundaries between the different classes. Since the classification was made on the basis of consecutive 10-ms long speech segments, a segment at the boundary often included data from two classes. The voiced speech generally has a higher energy than either the unvoiced speech or the silence and, therefore, the classification decision shows a bias towards voiced speech for boundary segments. This accounts for the relatively fewer number of errors in the case of voiced speech as compared to unvoiced speech. An example of the speech waveform showing the various voiced, unvoiced, and silence regions as determined by the algorithm is shown in Fig. 8.

The algorithm was used on the speech data in the test set using only one of the five measurement parameters at a time. The total number of errors found for each class for each of the five parameters is shown in Table IV. For comparison, the corresponding errors where all of the five parameters are used is also shown on the last row of Table IV. It can be seen that none of the parameters by itself is capable of identifying a class with sufficiently high accuracy. The performance of the five parameters when used in combination is quite good in view of the fact that most of the errors occur at the boundaries between the different classes.

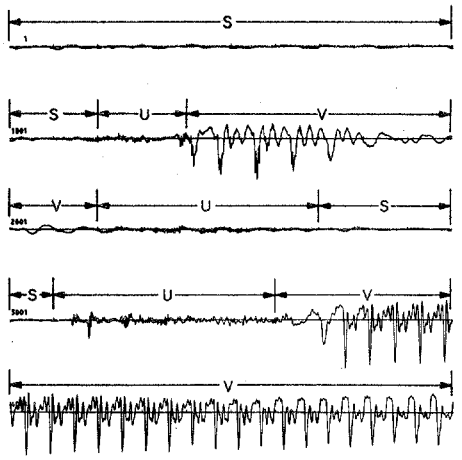

Fig. 7. Example of speech waveform, showing manually classified regions, used in the training data. The symbol "V" indicates voiced speech, the symbol "U" indicates unvoiced speech, and the symbol "S" indicates silence.

The performance of any of the parameters is not equally good for discriminating between all of the three classes. It is of interest to know the distribution of errors for pairwise discrimination between the different classes. These errors are shown in Table V for the five parameters for three kinds of confusions: V-UV, UV-S, and V-S. The ranking of the parameters is different for the three cases. The autocorrelation coefficient C_1 and the prediction error E_p are among the best candidates for V-UV decision. The zero-crossing parameter is not far behind for this decision. On the other hand, signal energy E_s comes out as the leading parameter for discrimination between speech and silence. These results are not very surprising and would have been expected from the nature of voiced and unvoiced sounds.

The discrimination errors are a consequence of the fact that the distributions of the parameters overlap as shown in Figs. 2–6. The effectiveness of each parameter can also be determined by obtaining a measure of separability of the classes for a given data set. One example of such a measure of separability is the "divergence" originally defined by Kullback as a measure of information [11], [12]. For Gaussian probability density, the divergence is determined by the mean vectors and the covariance matrix for each class. The divergence H between the classes i and j is given by [12]

$$H = \tfrac{1}{2}\,\text{trace}\,(W_i - W_j)(W_j^{-1} - W_i^{-1}) + \tfrac{1}{2}\,\text{trace}\,(W_i^{-1} + W_j^{-1})(m_i - m_j)(m_i - m_j)^t, \quad (14)$$

where m_i is the mean vector and W_i is the covariance matrix for the ith class. We will not discuss the properties of divergence in detail here but will refer to the two references cited above for more information. The divergence was computed for each parameter for the three binary discrimination tasks and is shown also in Table V. A high value of divergence is indicative of a high degree of separability between the classes. The rank-order of each parameter based on the divergence measure is shown on the last column of Table V. It is interesting to know that the two rank orders—one based on the discrimination errors and the other based on the divergence—are almost identical. Since the divergence can be computed

TABLE II
MEANS, STANDARD DEVIATIONS, AND COVARIANCE MATRICES FOR THE THREE CLASSES FOR THE TRAINING DATA DESCRIBED IN SECTION III-A (1 SPEAKER—2 UTTERANCES)

	Zero Crossings	Log Energy	First Auto-correlation	First LPC	LPC log Error
1) Silence					
Mean	23.743	11.156	0.629	−0.485	5.770
Standard deviation	14.564	4.626	0.358	0.354	3.336
Covariance matrix	1.000	0.241	−0.967	0.736	−0.272
	0.241	1.000	−0.278	−0.162	0.737
	−0.967	−0.278	1.000	−0.773	0.244
	0.736	−0.162	−0.773	1.000	−0.417
	−0.272	0.737	0.244	−0.417	1.000
2) Unvoiced					
Mean	48.614	25.201	0.032	−0.402	6.206
Standard deviation	14.911	7.367	0.411	0.543	2.796
Covariance matrix	1.000	0.290	−0.981	0.888	0.033
	0.290	1.000	−0.271	0.021	0.790
	−0.981	−0.271	1.000	−0.915	−0.008
	0.888	0.021	−0.915	1.000	−0.215
	0.033	0.790	−0.008	−0.215	1.000
3) Voiced					
Mean	11.224	45.324	0.912	−2.263	20.495
Standard deviation	4.838	7.483	0.073	0.562	3.440
Covariance matrix	1.000	0.361	−0.819	−0.036	−0.526
	0.361	1.000	−0.248	−0.649	−0.035
	0.819	−0.248	1.000	−0.159	0.576
	−0.036	−0.649	−0.159	1.000	−0.197
	−0.526	−0.035	0.576	−0.197	1.000

TABLE III

(a) Actual class →	Silence	Unvoiced	Voiced
Identified as			
Silence	65	0	0
Unvoiced	9	56	3
Voiced	2	1	310
Total	76	57	313

(b) Actual class →	Silence	Unvoiced	Voiced
Identified as			
Silence	91	7	3
Unvoiced	2	70	1
Voiced	1	5	371
Total	94	82	375

(a) Matrix of incorrect identifications for the three classes for the speech data in the training set.
(b) Matrix of incorrect identifications for the three classes for the speech data in the testing set.

directly from the means and covariance matrices, it offers a convenient method of comparing the performance of different parameters for a given classification task.

B. Examples of the Use of the Algorithm in a Speech Recognition Application

One of the practical problems in using the classification algorithm just described is the proper characterization of the silence. This is because, in many applications, the silence distribution is nonstationary. Thus, the ability of the algorithm to correctly classify silence is extremely dependent on how closely the properties of the actual silence matches the trained distribution. Another problem is that different speakers use widely varying talking levels. To make the voicing decision as accurate as possible, the signal level is scaled on a long-term basis such that the maximum level is about 2048. However, this means that the background silence level is also scaled so that, for a weak talker, the measured

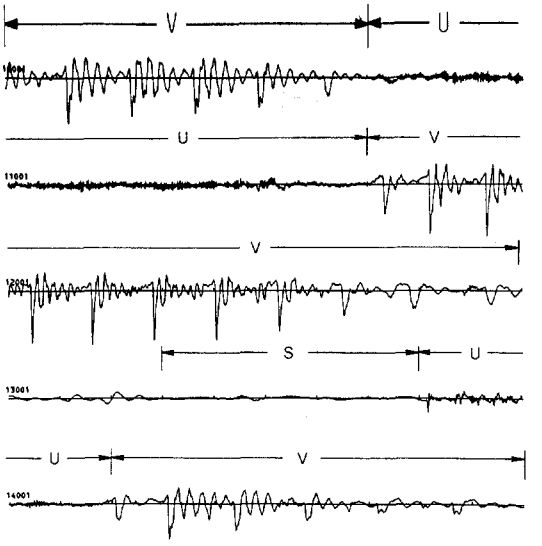

Fig. 8. Example of speech waveform showing the various voiced, unvoiced, and silence regions as determined by the algorithm.

TABLE IV

Number	Parameter Used	Errors: Silence	Errors: Unvoiced	Errors: Voiced
1	N_z	34	52	35
2	E_s	3	23	23
3	C_1	24	51	22
4	α_1	39	60	10
5	E_p	19	48	4
6	all five parameters	3	12	4

Total number of identification errors for the different classes with different sets of parameters. The total number of segments were 94 for the silence, 82 for the unvoiced, and 375 for the voiced class.

TABLE V
COMPARATIVE PERFORMANCE OF DIFFERENT PARAMETERS FOR THE THREE BINARY DISCRIMINATION TASKS

Binary Task	Parameter	Total Number of Errors	Rank Order Based on Errors	Divergence	Rank Order Based on Divergence
Voiced-Unvoiced	N_z	10	3	37	2
	E_s	38	5	7	5
	C_1	4	1	90	1
	α_1	32	4	11	4
	E_p	7	2	22	3
Unvoiced-Silence	N_z	49	2	3	2
	E_s	11	1	7	1
	C_1	53	3	3	3
	α_1	66	5	0	4
	E_p	60	4	0	5
Voiced-Silence	N_z	62	5	7	5
	E_s	0	1	38	1
	C_1	40	4	19	3
	α_1	11	3	18	4
	E_p	4	2	19	2

energy of the silence is often much larger than the mean energy for the silence distribution. This strongly biases the decision rule away from silence—often to a voiced classification.

To compensate for these problems, a "smoothing" or correction algorithm is applied to the VUS contour to make the results appropriate for experiments in continuous digit recognition [13]. (It should be noted that for more stable background silence conditions, and some form of voice gain adjustment (VOGAD), the required smoothing and error correction algorithm would be considerably simpler.) For the continu-

ous digit recognition experiment [13], a subject spoke a three digit sequence during a specified time interval. Thus, it is assumed *a priori* in the smoothing algorithm that each contour begins and ends with silence. After the speech begins, it is generally sufficient to locate only the intervals of voicing and unvoicing—even if there are internal silences (e.g., the stop gaps in two, eight, and six, or pauses between words). In this particular speech recognition application [13] the subsequent processing is essentially insensitive to whether an interval is classified as unvoiced or silence. Of course, this distinction may be important for other applications.

Fig. 9 gives a flowchart of the smoothing algorithm. The first step in the process is to examine the overall VUS contour and to reclassify any low energy voiced intervals as silence. An interval is called a low energy interval if it falls by a specified amount below the maximum energy during the utterance. For example, if the maximum normalized log energy of the utterance is 60 dB and the anticipated signal-to-noise ratio of the system is 30 dB, any interval whose log energy falls below 30 dB is reclassified as silence. This step in the smoothing is used primarily for weak talkers to overcome the inherent difficulty discussed above.

The next step in the algorithm is to find regions of the VUS contour with high probability of correct classification, based on the probability measure discussed earlier. For each of these highly reliable intervals, the endpoints of these regions are then found. Next, long unvoiced regions are added to the list of tentatively good regions, even if the probability score does not exceed the threshold. A tentative check of the list of regions is made and any overly short voiced intervals (i.e., less than 30 ms in duration) are eliminated.

Next, the beginning and end silence regions are located. These regions are chosen as the interval from the beginning (end) of the contour until the first (last) interval which is reliably classified as voiced or unvoiced.

The only remaining step is to smooth across the boundaries between the regions in the list. The interpolation rule is fairly simple. If the regions on both sides of the boundary are voiced or unvoiced, and the duration of the boundary is less than 50 ms, the boundary region is classified as voiced or unvoiced as appropriate. If the boundary duration exceeds 50 ms, the boundary values are classified as silence.

The final step in the algorithm is to apply a median smoother to the VUS contour. For most cases, the median smoother is all that is required to smooth the contour—i.e., the preceding steps in the flowchart leave the contour unchanged. However, such sophistication and checking is required for the unusual cases and thus is included for all cases.

Figs. 10-13 illustrate typical examples of VUS contours obtained in the digit recognition experiment. In each of these figures, there are five curves. The first curve is a plot of the probability of correct classification of the VUS contour. The fourth and fifth curves are the unsmoothed and smoothed VUS contours. These contours can assume one of three levels where level 1 is silence, level 2 is unvoiced, and level 3 is voiced.

Fig. 10 shows contours for the digit sequence 852 by a female speaker (SAW). The recordings were made in a relatively noisy computer room using a dynamic omnidirectional microphone. It can be seen that the analysis made essentially one error in classifying an unvoiced interval as voiced at the beginning of the digit two. The smoothing algorithm corrected this error and essentially left the rest of the VUS contour alone.

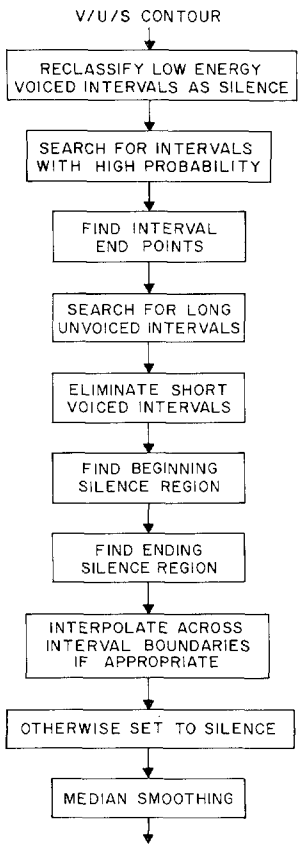

Fig. 9. Flowchart of the smoothing algorithm for the VUS contour.

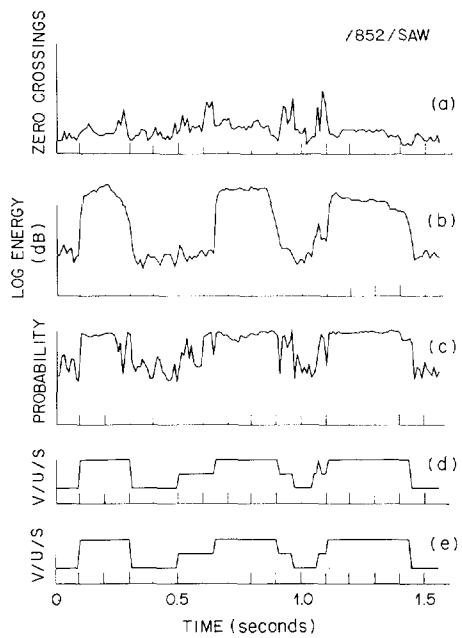

Fig. 10. (a)-(e) Typical measurement contours for the utterance /852/ spoken by a female speaker.

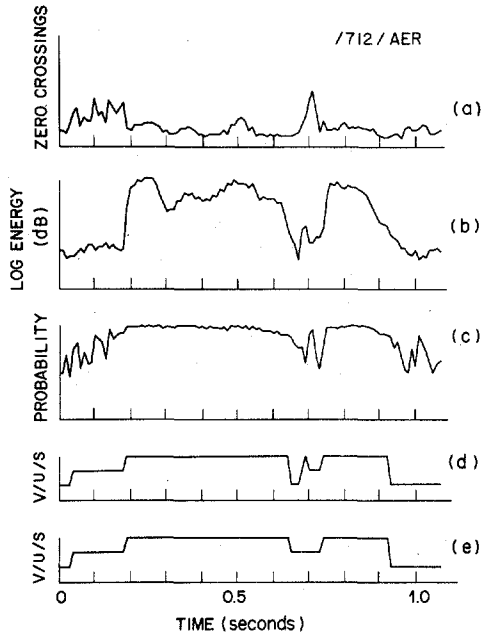

Fig. 11. (a)–(e) Typical measurement contours for the utterance /712/ spoken by a male speaker.

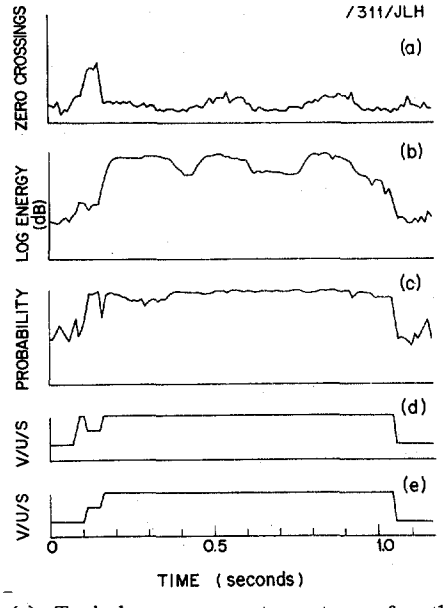

Fig. 12. (a)–(e) Typical measurement contours for the utterance /311/ spoken by a male speaker.

Fig. 11 shows contours for the digit sequence 712 spoken by a male speaker (AER). Again, the analysis algorithm made only one error in classification where an unvoiced interval was classified as voiced; however, the smoothing algorithm smoothed the silence interval preceding the error and reclassified it as unvoiced since the duration of the silence was very short (30 ms). Otherwise, the remainder of the contour was left unchanged by the smoothing algorithm.

Fig. 12 shows an example in which a 20 ms interval of silence (the speaker burbled slightly prior to articulating the digit sequence) was classified as voiced but the smoothing algorithm reclassified this interval as silence since it was too short, and it occurred at the beginning of the utterance. The

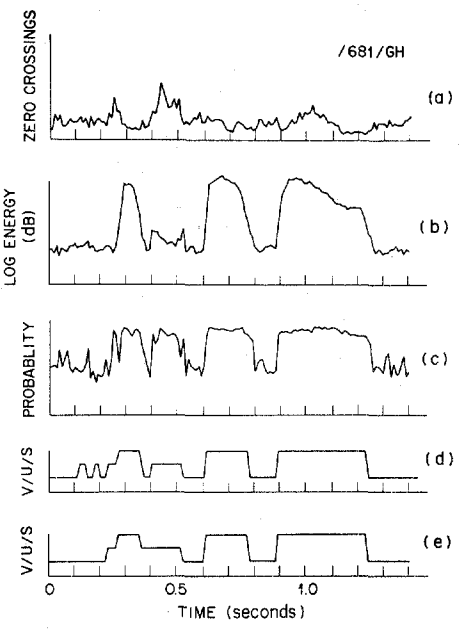

Fig. 13. (a)–(e) Typical measurement contours for the utterance /681/ spoken by a female speaker.

digit sequence here was 311 spoken by another male speaker (JLH). Finally, Fig. 13 shows an example in which the analysis oscillated between silence and unvoiced at the beginning of the digit sequence 681 spoken by another female speaker (GH). Because of the constraint that the initial part of the VUS contour be silence, and since the probability of correct classification during the short unvoiced intervals was low, both these short intervals were reclassified as silence. Also, the short silence interval following the initial voiced interval was reclassified as unvoiced by the smoothing algorithm since it was of sufficiently short duration.

These four examples illustrate how the training data obtained for a completely different set of utterances can be used over a wide variety of speakers. For these examples, *none* of the speakers were used in the training set, yet the algorithm still performed quite well. We have generally found this behavior to be the case—i.e., when the recording conditions and the background noise level remain relatively stable, the training data can be used across a large number of speakers with no apparent degradation.

IV. Summary

A fairly general framework based on a pattern recognition approach to VUS classification has been described in which a set of measurements are made on the interval being classified, and a minimum non-Euclidean distance measure is used to select the appropriate class. Almost any set of measurements can be used so long as there is some physical basis for assuming that the measurements are capable of reliably distinguishing between these three classes. Although a non-Euclidean distance measure was used, other distance measures may be equally appropriate. Finally, a smoothing algorithm was discussed which was appropriate for a digit recognition algorithm in which errors in the analysis were corrected, and unusually short intervals were eliminated. The classification algorithm

has been extensively tested for both speech analysis-synthesis and recognition applications over a wide range of recording conditions and has been found to provide satisfactory results.

The major limitation of the method is the necessity for training the algorithm on the specific set of measurements chosen, and for the particular recording conditions. For nonstationary speaking environments, it may be preferable to adapt the means and covariance matrices continuously.

References

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

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

[3] M. M. Sondhi, "New methods of pitch extraction," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 262-266, June 1968.

[4] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pt. 2, pp. 637-655, Aug. 1971.

[5] J. I. Makhoul and J. J. Wolf, "Linear prediction and the spectral analysis of speech," Bolt Beranek and Newman Inc., Cambridge, MA, BBN Rep. 2304, Aug. 1972.

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

[7] A. V. Oppenheim and R. W. Schafer, "Homomorphic analysis of speech," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 221-226, June 1968.

[8] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, pp. 1304-1312, June 1974.

[9] A. H. Gray, Jr., and J. D. Markel, "A spectral flatness measure for studying the autocorrelation method of linear prediction of speech analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 207-217, June 1974.

[10] E. A. Patrick, *Fundamentals of Pattern Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1972, pp. 193-194.

[11] S. Kullback, *Information Theory and Statistics*. New York: Dover, 1968, pp. 189-191.

[12] T. Marill and D. M. Green, "On the effectiveness of receptors in recognition systems," *IEEE Trans. Inform. Theory*, vol. IT-9, pp. 11-17, Jan. 1963.

[13] L. R. Rabiner and M. R. Sambur, "Some preliminary experiments in the recognition of connected digits," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 170-182, Apr. 1976.

Tone Detection for Automatic Control of Audio Tape Drives

JOHN J. DUBNOWSKI, JOSEPH C. FRENCH, AND LAWRENCE R. RABINER, FELLOW, IEEE

Abstract—**This paper describes digital hardware for automatically stopping a cassette recorder upon detection of a prerecorded tone. This hardware is used in conjunction with experiments on computer assisted voice wiring experiments being performed at Western Electric locations [1]. For these experiments a sequence of instructions is automatically recorded on a cassette tape by a computer voice response system. At the end of each instruction, a tone is recorded. The hardware detects this tone and stops the cassette recorder. The operator, after performing the prescribed wiring instruction, manually restarts the cassette recorder for the next instruction. The technique used to detect the tone is a simple digital method comparing the axis crossings of the signal to a fixed threshold. This threshold is determined based on knowledge of the tone frequency, duration, and amplitude. When the signal axis crossings exceeds this threshold during two consecutive 40 ms nonoverlapping intervals the tone is detected and the tape recorder is stopped. The method described is a robust one which is rather insensitive to normal tape recorder problems, e.g., wow and loss of signal level due to battery drainage. The tone detection hardware requires nominal power and is portable.**

Introduction

ONE very promising application of computer voice response systems is in computer assisted wiring of electronic circuitry by voice. In essence, a complex printed list of wiring instructions is replaced by a spoken sequence of instructions on cassette tapes. During playback, all the information required to make the wiring connections is defined by a verbal instruction. The advantage of this procedure is that a craftsman is able to maintain continual visual contact with the wiring assembly. This avoids any disorientation usually caused by referring to printed wire-lists; thus, the operator's performance should be improved through increased efficiency and a reduction in errors. Furthermore, a training period is avoided in which the operator must learn how to read complex wiring diagrams.

A sequence of spoken instructions for wiring of electronic assemblies appears to have many advantages over the more conventional methods of wiring. To study the practical aspects of the feasibility of this method, the multiline computer voice response system [2] of the Acoustics Research Department has been used to generate a variety of spoken wiring lists for experimental evaluation at several Western Electric locations.

Along with the spoken wiring instructions, a tone cue is recorded to indicate the end of each instruction. By using special hardware to automatically detect this tone and subsequently halt the cassette drive, the operator is free to execute the latest wiring instruction without the added distriction of halting the drive.

Manuscript received September 23, 1975; revised December 23, 1975.
The authors are with Bell Laboratories, Murray Hill, NJ 07974.