also denote I_m and I_n by I; the context making the choice obvious. Consider

$$\{(I - \delta A)\} \otimes I\} \{K \otimes Q - \mathcal{A} (K \otimes Q) \mathcal{A}^T - \mathcal{B}\mathcal{B}^T\} \cdot \{(I - \delta A^T) \otimes I\}$$
$$= [(I - \delta A)K(I - \delta A^T)] \otimes [Q - \alpha Q \alpha^T]$$
$$- [(I - \delta A)KA^T] \otimes [\alpha Q \gamma^T \beta^T]$$
$$- [AK(I - \delta A)^T] \otimes [\beta\gamma Q \alpha^T]$$
$$- (AKA^T) \otimes [\beta\gamma Q \gamma^T \beta^T]$$
$$- (bb^T) \otimes [\beta\beta^T].$$

Using (34)-(38), one finds that each of the five expressions on the right-hand sides of the above Kronecker products are scalar multiples of $\beta\beta^T$. Gathering terms leaves

$$[K - AKA^T - bb^T] \otimes [\beta\beta^T] = 0.$$

Since $1 - \delta^2 = \beta^T Q^{-1} \beta > 0$, we have $\delta^2 < 1$. Since the eigenvalues of A are all in the disk $|z| < 1$, so must the eigenvalues of δA. Therefore, $I - \delta A$ is invertible. It follows that

$$(K \otimes Q) = \mathcal{A}(K \otimes Q)\mathcal{A}^T + \mathcal{B}\mathcal{B}^T$$

which is to say that $\mathcal{K} = K \otimes Q$ satisfies (32). (Note: The equation $X = \mathcal{A} X \mathcal{A}^T + \mathcal{B}\mathcal{B}^T$ has a unique solution if the matrix $\mathcal{A}$ has no eigenvalues in the set $|z| \geqslant 1$. In view of our assumptions, the matrices α, A, $\mathcal{A}$ each have all of their eigenvalues in the open disk $|z| < 1$.)

References

[1] A. H. Gray, Jr., and J. D. Markel, "Digital lattice and ladder filter synthesis," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 491-500, Dec. 1973.

[2] ——, "A normalized digital filter structure," *IEEE Trans. Acoust., Speech, Signal Processing* (*Special Issue on 1974 Arden House Workshop on Digital Signal Processing*), vol. ASSP-23, pp. 268-277, June 1975.

[3] J. D. Markel and A. H. Gray, Jr., "Roundoff noise characteristics of a class of orthogonal polynomial structures," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 473-486, Oct. 1975.

[4] A. G. Constantinedes, "Spectral transformations for digital filters," *Proc. Inst. Elec. Eng.* (London), vol. 117, pp. 1585-1590, Aug. 1970.

[5] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing.* Englewood Cliffs, NJ: Prentice-Hall, 1975, pp. 226-230.

[6] W. Schüssler and W. Winkelnkemper, "Variable digital filters," *Arch. Elek. Übertragung*, vol. 24, pp. 524-525, 1970.

[7] C. T. Mullis and R. A. Roberts, "Synthesis of minimum roundoff noise fixed-point digital filters," *IEEE Trans. Circuits and Systems*, vol. CAS-23, pp. 551-512, Sept. 1976.

[8] J. F. Kaiser, "Some Practical Considerations in the Realization of Linear Digital Filters," in *Proc. 3rd Allerton Conf. Circuit and System Theory*, 1965, pp. 621-633.

[9] L. B. Jackson, "Roundoff-noise analysis for fixed-point digital filters realized in cascade or parallel form," *IEEE Trans. Audio Electroacoust.* (*Special Issue on Digital Filtering*), vol. AU-18, pp. 107-122, June 1970.

[10] W. S. Lee, "Optimization of digital filters for low roundoff noise," in *Proc. 1973 Int. Symp. Circuit Theory*, Toronto, Ont., Canada, 1973, pp. 381-383.

[11] T. R. Lapp, "Some useful algorithms for the analysis of multiplier roundoff noise in cascade digital filters," M.S. thesis, University of Colorado, Boulder, 1974.

[12] S. Y. Hwang, "On optimization of cascade fixed-point digital filters," *IEEE Trans. Circuits and Systems* (Lett.), vol. CAS-21, pp. 163-166, Jan. 1974.

[13] B. Liu and A. Peled, "Heuristic optimization of the cascade realization of fixed-point digital filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 464-473, Oct. 1975.

[14] S. Y. Hwang, "Minimum unit noise in the state space digital filtering," in *Proc. Conf. Acoustics, Speech, and Signal Processing*, Apr. 1976.

[15] R. Bellman, *Introduction to Matrix Analysis.* New York: McGraw-Hill, 1960.

[16] R. E. Kalman, M. A. Arbib, and P. L. Falb, *Topics in Mathematical System Theory.* New York: McGraw-Hill, 1969, p. 43.

A Statistical Decision Approach to the Recognition of Connected Digits

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

Abstract—A statistical decision approach to the recognition of connected digits is described in this paper. The method can be either speaker dependent (i.e., each new speaker must first train the system on representative digit strings before he can successfully use the system) or speaker independent. Multiple repetitions of each digit (spoken in connected strings) are used in the training sequence. Repetitions of the same digit are combined by linearly warping the individual reference patterns to the speakers' average length for the digit. Statistics of the mean and covariance of the recognition parameters between repetitions of the same digit are computed and are used in the recognition phase of the system.

Once a spoken digit string has been segmented, the recognition of each digit within the string is achieved using a distance measure based on an expanded form of the principle of minimum residual error. In

Manuscript received March 24, 1976; revised August 3, 1976. This work represents an equal contribution of the two authors.

The authors are with Bell Laboratories, Murray Hill, NJ 07974.

cases where a great deal of coarticulation can be anticipated between adjacent digits (i.e., between digits bounded by voiced regions) a second distance metric is employed. This metric includes both the effects of the analysis estimation error and the effects of coarticulation. The analysis parameters used in this system are the linear prediction coefficients (LPC's) of a 10-pole LPC analysis. For stability purposes, the linear predictive coding (LPC) coefficients are converted to parcor or reflection coefficients prior to the linear warping, and then the warped parcor coefficients are converted back to LPC coefficients for recognition purposes. The recognition system was tested on six speakers in the speaker-dependent mode with recognition accuracies of from 97 to 100 percent. It was also tested with 10 new speakers in the speaker-independent mode, with a digit recognition accuracy of 95 percent.

I. Introduction

PREVIOUS research in the area of digit recognition has generally concerned itself with the recognition of isolated digits [1]-[3]. Recently Rabiner and Sambur investigated a speaker-independent, connected digit recognition system [4]. One of the main features of this system was a highly reliable method of digit segmentation which isolated the individual digits in the string. However, the recognition algorithm used in this study (a parallel processing decision tree logic structure) was unable to achieve recognition accuracies higher than about 87-91 percent. One reason for the low scores was the high variability of the recognition features across both speakers and repetitions of each digit. Additionally, a significant degree of digit coarticulation existed for certain digits, e.g., 1, 8, and 9 when preceded by a digit ending in a voiced sound.

In this paper we present a new approach to the recognition of connected digits. The method used is a statistical pattern recognition approach which can be applied in either a speaker-dependent or speaker-independent mode. This system builds on earlier work in that it uses both the endpoint location method of Rabiner and Sambur [5], as well as the digit segmentation algorithm discussed in [4]. However, the recognition strategy is based on a distance measure which specifies the distance between the test digit (the one to be recognized) and stored reference patterns for each of the 10 digits. The manner in which the stored reference patterns are obtained, as well as the way in which different distance measures are used are among the issues to be discussed in this paper.

II. Overview of the Recognition System

Fig. 1 shows a block diagram of the overall recognition system. The utterance (consisting of a string of three digits) is first analyzed to find the endpoints, and to give a voice-unvoiced-silence contour of the utterance [6]. The digit string is then segmented into individual digits by using the voiced-unvoiced contour and some additional energy information. For each segmented digit, the voicing region is analyzed using a 10-pole LPC formulation during each nonoverlapping 10-ms frame. The parameters used for recognition are the 10 linear predictive coding (LPC) coefficients in each frame and the duration of the voiced region of the segmented digit.

In order to use the recognition system, the digit reference files must first be created. Fig. 2 shows a flow diagram of how the training is done. The user has to specify both the reference digit and its voiced region. (The automatic voiced-unvoiced-silence algorithm provides this information directly to the user.) For each frame in the voiced region of the digit, the LPC coefficients are converted to parcor or reflection coefficients[1] which are then linearly warped[2] to a precomputed average digit length. The precomputed average digit length is either the average duration of the digit for an individual speaker (speaker-dependent case) or the average duration of the digit across speakers (speaker-independent case). The linearly warped parcor coefficients are then converted back to LPC coefficients.

The digit reference files consist of a statistical description of the behavior of the LPC coefficients for each frame and for each digit. Information about both the mean and variation of the LPC coefficients across repetitions and speakers is contained in these files. A description of the statistical analysis used is given in Section III, along with a discussion of the selection of digit strings for the training phase of the system.

Fig. 3 shows a flow diagram of the recognition phase of the system. The test digit is linearly warped to the duration of each of the reference digits.[3] For each reference digit a distance based on the average of the prediction residuals across the voiced region of the digit is computed. The distance measure used was originally proposed by Itakura [8] for the recognition of isolated polysyllabic words. The reference digit whose average distance to the test digit is smallest is chosen as the correct digit. In cases where the initial boundary of the digit occurs at a voiced to voiced boundary, e.g., the boundary between 1 and 9 in the string 196, a second distance measure is computed.[4] This second distance measure is based on both the prediction residual [8], as well as the interreplication variation of the LPC parameters for the digit. The digit with the minimum distance is again chosen as the spoken digit. Additional discussion of the recognition system is given in Section IV.

The system has been tested on six speakers in the speaker-dependent mode, and ten additional speakers in the speaker-independent mode. The overall results and an analysis of the system performance are given in Section V.

In the next section we present the mathematics behind the selection and use of the two distance measures used in the recognition system.

III. Distance Measures for LPC Analysis Parameters

Let $s(n)$, $n = 1, 2, \cdots, N$ be a frame of speech, digitized at a 10-kHz rate. Linear prediction methods model the speech

[1] Parcor coefficients [7] are used rather than LPC coefficients because they can be linearly warped (interpolated) and still transform back to a stable system.

[2] For the digit 7 (the only polysyllabic digit), a piecewise-linear warping characteristic is used. There are two pieces in the warping. The extra warping point is obtained as the local internal energy minimum of the digit. This point corresponds to the /v/ in seven.

[3] If the ratio between the duration of the reference digit and the test digit either exceeds 2 to 1, or is less than 1 to 2, the reference digit is omitted from consideration.

[4] For a speaker independent system, the rule for applying the second distance measure is slightly different (see Section V).

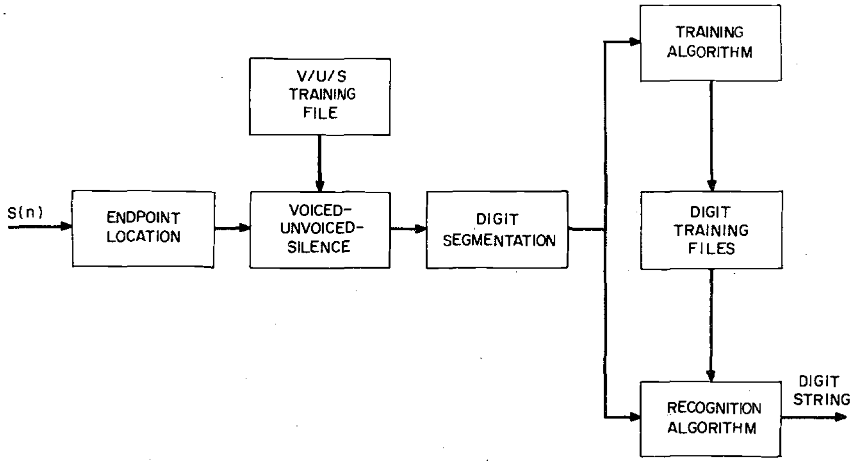

Fig. 1. Block diagram of overall recognition system.

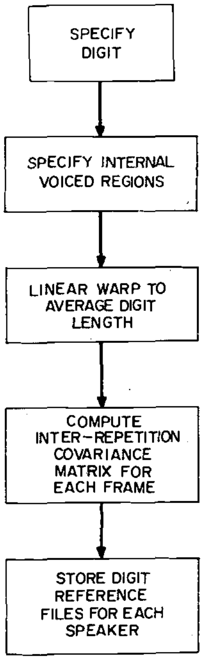
SPECIFY DIGIT
SPECIFY INTERNAL VOICED REGIONS
LINEAR WARP TO AVERAGE DIGIT LENGTH
COMPUTE INTER-REPETITION COVARIANCE MATRIX FOR EACH FRAME
STORE DIGIT REFERENCE FILES FOR EACH SPEAKER

Fig. 2. Flow chart of the training algorithm.

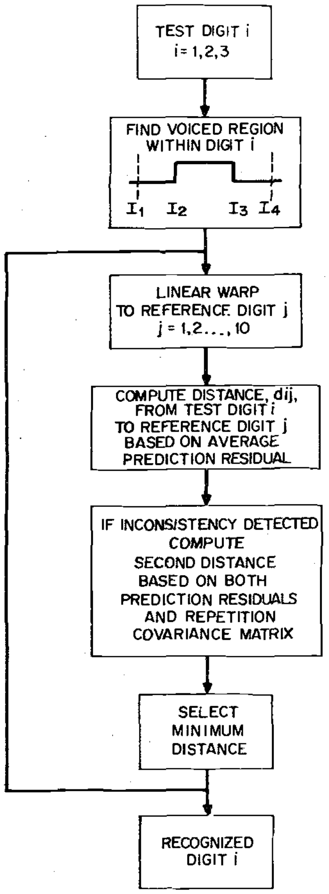
TEST DIGIT i
i = 1,2,3
FIND VOICED REGION WITHIN DIGIT i
I1 I2 I3 I4
LINEAR WARP TO REFERENCE DIGIT j
j = 1,2..., 10
COMPUTE DISTANCE, dij, FROM TEST DIGIT i TO REFERENCE DIGIT j BASED ON AVERAGE PREDICTION RESIDUAL
IF INCONSISTENCY DETECTED COMPUTE SECOND DISTANCE BASED ON BOTH PREDICTION RESIDUALS AND REPETITION COVARIANCE MATRIX
SELECT MINIMUM DISTANCE
RECOGNIZED DIGIT i

Fig. 3. Flowchart of the recognition algorithm.

waveform as the output of an all-pole filter, $H(z)$, with transfer function

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{1}$$

where G is the gain of the filter, p is the number of poles used in the model, and the a_k's are the linear prediction coefficients (LPC coefficients). The LPC coefficients model the combined effects of the vocal tract, glottal source, and radiation load, and as such are a natural choice for speech recognition feature parameters. However, to use the LPC coefficients as parameters for speech recognition, a suitable measure for assessing distance in the feature space spanned by the LPC parameters is required. Such a distance measure was recently proposed by Itakura in a paper on the recognition of isolated words [8].

Although Itakura derived the distance measure in a somewhat different manner, this distance measure can be obtained by the following reasoning. It can be argued that because of noise as well as the inadequacies of the linear prediction model, it is not possible to measure the true LPC coefficients associated with a segment of speech. It is only possible to estimate (i.e., measure) the underlying LPC coefficients for the speech segment [9]. Assume that we are given a segment of speech with estimated LPC coefficients $\hat{a}$ where $\hat{a}$ is a row vector $(\hat{a}_1, \hat{a}_2, \cdots, \hat{a}_p)$. The problem is to determine the probability that $\hat{a}$ is from a speech segment with true LPC a $(a_1, a_2, \cdots, a_p)$. Once this probability is determined, an effective measure for assessing dissimilarity can be obtained. The distance measure should have the feature that the greater the probability that $\hat{a}$ is a member of the ensemble of speech segments with true LPC a, the smaller the dissimilarity between $\hat{a}$ and a.

It has been shown by Mann and Wald [9] that the probability distribution governing the estimates of a can be reasonably modeled as a multidimensional Gaussian distribution with mean a, and estimated covariance matrix Λ, defined as

$$\Lambda = R^{-1}\left(\frac{\hat{a}R\hat{a}^t}{N}\right) \tag{2}$$

where R is the correlation matrix of the speech segment, with elements

$$r_{ij} = r(|i-j|) = r(l) = \frac{1}{N}\sum_{n=1}^{N-l} s(n)s(n+l) \tag{3}$$

and the t denotes the transpose of the row vector, i.e., a column vector. Thus the probability of obtaining the estimate $\hat{a}$ when the underlying LPC is a is

$$P(\hat{a}/a) = [(2\pi)^{p/2}\,|\Lambda|^{1/2}]^{-1}\exp\left[-\tfrac{1}{2}(\hat{a}-a)\Lambda^{-1}(\hat{a}-a)^t\right] \tag{4}$$

where $|\Lambda|$ is the determinant of the matrix Λ. An appropriate distance measure is obtained by taking the logarithm of (4), and neglecting the bias term due to $|\Lambda|$. The resulting distance measure is

$$d(\hat{a}, a) = (\hat{a}-a)\left[\frac{NR}{\hat{a}R\hat{a}^t}\right](\hat{a}-a)^t. \tag{5}$$

It is readily seen that the greater the probability that $\hat{a}$ came from the distribution with underlying LPC a, the smaller the distance computed using the metric of (5). It should be noted that for computational considerations, Itakura proposed the closely related distance measure

$$d'(\hat{a}, a) = \log(aRa^t/\hat{a}R\hat{a}^t). \tag{6}$$

The key assumption in the above analysis is that the ensemble of all possible speech segments derived from the same speech sound are similar in that the underlying LPC coefficients a are identical. The differences in the measured LPC coefficients for these speech segments are attributed primarily to the effects of statistical sampling. For the recognition of isolated words spoken by a designated individual, this assumption is quite reasonable. However, for connected digits, the underlying or true LPC coefficients are very much influenced by the surrounding digits. Thus it cannot be safely assumed that the true LPC coefficients associated with a given speech segment are constant.

For a complete characterization of a particular segment derived from a given digit, it is necessary to determine the probability that the true LPC coefficient set is a when it is known that the segment is from the ith digit. (Call this probability $P(a/i\text{th digit})$.) Thus the overall probability of measuring $\hat{a}$ for a segment from the ith digit is

$$P(\hat{a}/i\text{th digit}) = \int P(\hat{a}/a)P(a/i\text{th digit})\,da \tag{7}$$

where the integration is performed over the vector a [10]. If we assume that the distribution of a is Gaussian with mean m and covariance matrix S, then the integral in (7) can be solved to give $P(\hat{a}/i\text{th digit})$ as a Gaussian distribution with mean m, and covariance matrix C of the form

$$C = S + R^{-1}\left(\frac{\hat{a}R\hat{a}^t}{N}\right). \tag{8}$$

Thus an appropriate distance measure which incorporates the effects of both coarticulation and estimation error is

$$D_2 = D(\hat{a}/i\text{th digit}) = (\hat{a}-m)C^{-1}(\hat{a}-m)^t. \tag{9}$$

To use the distance measure of (9), the quantities m and S must be determined for each digit segment from a training set. In the next section we discuss the manner in which the quantities m and S are obtained from the training set.

IV. Training the System

The training set consists of J repetitions ($J = 15$ for this work) of each digit by each of K speakers. The digits used in the training set were spoken in connected strings. A balanced list of 50 strings of three digits each (see Table I) was used. Each digit appeared equally often in each position in the string.

Let us define $\hat{a}(n)_{ijkl}$ as the measured LPC coefficient set where

n	LPC coefficient number,	$n = 1, 2, \cdots, 10$
i	frame number,	$i = 1, 2, \cdots, I(j, k, l)$
j	repetition number,	$j = 1, 2, \cdots, J(J = 15)$
k	speaker number,	$k = 1, 2, \cdots, K(K = 6)$
l	digit number,	$l = 1, 2, \cdots, 10.$

From the training set of data, the mean vector m and the covariance matrix S have to be estimated. The main difficulty in obtaining m and S directly from the set of $\hat{a}(n)_{ijkl}$ is the lack of time alignment between repetitions of the same digit. Thus the first step in estimating m and S is to time align the repetitions of the same digit by each speaker, so that the *same* speech event occurs at the *same* time for all J repetitions of the ith digit by the kth speaker. Once the speech events are time aligned, m and S are obtained by conventional statistical techniques as discussed below.

For the recognition of monosyllabic digits, White [11] has shown that a linear warping of the time axis is sufficient for time alignment of different repetitions of the same digit. Based on using a linear warping to achieve time alignment, the procedure for estimating $m = m_{ikl}$, and $S = S_{ikl}$ for a speaker-dependent digit reference is as follows.

1) Find the average length of the lth digit for speaker k, as

$$\bar{I}_{kl} = \frac{1}{J}\sum_{j=1}^{J} I(j, k, l). \tag{10}$$

2) Linearly stretch or contract the J repetitions of the lth digit by speaker k to a standard length of $\bar{I}_{kl}$ frames. For

TABLE I

Training Sequence (from Martin)

525	990	631	005	033
759	583	349	140	477
101	171	565	819	680
626	098	113	974	306
202	232	460	357	915
727	670	892	212	782
366	854	964	551	248
044	386	076	161	887
843	795	228	453	939
418	429	737	508	694

notational purposes we denote the linearly warped LPC coefficients as $a'(n)_{i'jkl}$ where the index i' extends from 1 to $\bar{I}_{kl}$.

3) Compute the mean vector $\boldsymbol{m}_{i'kl}$ for the (i')th frame for speaker k and digit l as

$$\boldsymbol{m}_{i'kl} = (m(1)_{i'kl}, \quad m(2)_{i'kl}, \cdots, m(10)_{i'kl}) \tag{11}$$

where

$$m(n)_{i'kl} = \frac{1}{J} \sum_{j=1}^{J} \hat{a}'(n)_{i'jkl}. \tag{12}$$

4) Compute the covariance matrix $S_{i'kl}$ for the (i')th frame for speaker k for digit l, with matrix entry $s(n,p)_{i'kl}$ as[5]

$$s(n,p)_{i'kl} = \frac{1}{J} \sum_{j=1}^{J} \hat{a}'(n)_{i'jkl} \hat{a}'(p)_{i'jkl} - m(n)_{i'kl} m(p)_{i'kl}. \tag{13}$$

It should be restated that the above procedure is applied to only the LPC coefficients obtained within the *voiced regions* of each digit.[6] Also the linear warping is not applied directly to the LPC coefficients, but instead to the parcor coefficients which are obtained directly from the LPC coefficients. The parcor coefficients have the desirable property that they can be linearly interpolated and still ensure that the system derived from the resulting LPC coefficients will be stable.

Although the linear warping is appropriate for most digits, it is not strictly appropriate for the bisyllabic digit seven. Thus for seven a piecewise-linear warping procedure was used, consisting of two segments. The first segment was defined from the initiation of voicing to the pronounced dip in energy in the middle of the voiced region. This pronounced dip in energy is due to the /v/ in seven and occurs for all speakers. The second segment was defined from the energy dip until the end of the utterance. Since linear warping was used on each of the pieces of the warp, the above algorithm was used to determine $\boldsymbol{m}$ and S for each segment of the warping.

For a speaker-independent system, all of the above training procedures are modified by averaging over the index k. Thus one obtains an average duration for the lth digit, $\bar{I}_l$, and the set of means $\boldsymbol{m}_{i'l}$, and covariance matrices $S_{i'l}$. No other modifications to the procedure are required.

V. The Recognition System

Once the reference information is obtained, the distance measure of (9) can be used to find the digit which is closest in distance to the test digit. In cases when the effects of coarticulation are small (i.e., a digit which begins with an unvoiced region such as 2, 3, 4, 5, 6, 7, and 0), a considerable savings in computation can be obtained (with only a small loss in recognition accuracy) by using the modified distance measure

$$D_1 = (\hat{\boldsymbol{a}} - \boldsymbol{m}) \Lambda^{-1} (\hat{\boldsymbol{a}} - \boldsymbol{m})^t. \tag{14}$$

The gain in speed is due to the fact that Λ^{-1} can be efficiently computed for each analysis frame directly from (2) and (3) and is independent of the digit being tested; whereas the computation of C^{-1} of (9) requires the computation of two 10×10 matrix inverses for each analysis frame, for each digit being tested. Thus the gain in speed is on the order of 1 to 2 orders of magnitude. Additionally, the distance measure of (14) is also appropriate for the recognition of isolated words as the use of $\boldsymbol{m}$ allows the use of multiple training.

Based on the above discussion, the use of the system for recognition is quite straightforward. For the speaker-dependent case, the speaker inputs his identity and then speaks a given digit string. Using either the distance measure of (14) or (9), the system sequentially computes the average distance to each reference digit and chooses the digit for which the average distance is smallest. In cases where the segmentation boundary occurred within a voiced region the distance measure of (9) is used because of the anticipated high degree of coarticulation between digits. In cases where the segmentation boundary occurred at a voiced-unvoiced, or unvoiced-voiced transition, the modified (simplified) distance measure of (14) is used.

Figs. 4-7 illustrate some typical examples of the frame by frame distances obtained using the distance measure of both (14) and (9). Fig. 4 shows the recognition candidates for the initial 1 in the digit string /123/. The average distance for the digit 1 was 0.20 and the absolute distance for each frame was fairly uniform across the digit. The next candidate was 9 whose average distance was 0.44, a distance of more than 2 times that of 1. It can also be seen that the absolute distance for each frame was uniformly much higher for the 9 than the 1. The other candidates (whose average distance was less than the arbitrary plotting threshold of 1.25) were the digits 5, 4, and 3. It can be seen in this figure that the absolute distance for some frames for each of these digits was small; however, the average distance across the digit was generally quite large.

Fig. 5 shows the results of the distance calculations for the

[5] Equation (13) is a biased estimate of the covariance matrix S because the effects of prediction residual have not been removed.

[6] The information in the unvoiced regions was found to be unreliable (highly variable) and thus was not used in the decision algorithm.

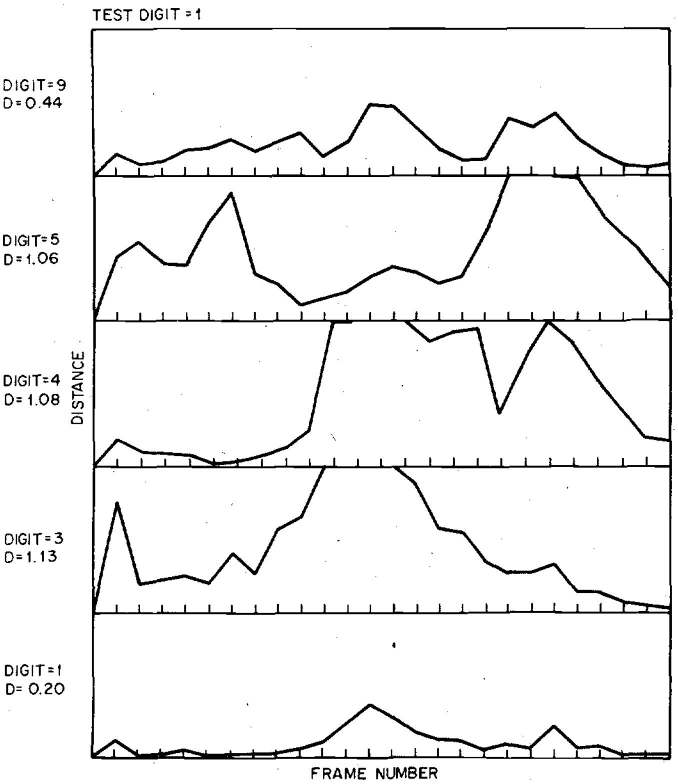

Fig. 4. Frame-by-frame distances for the test digit 1.

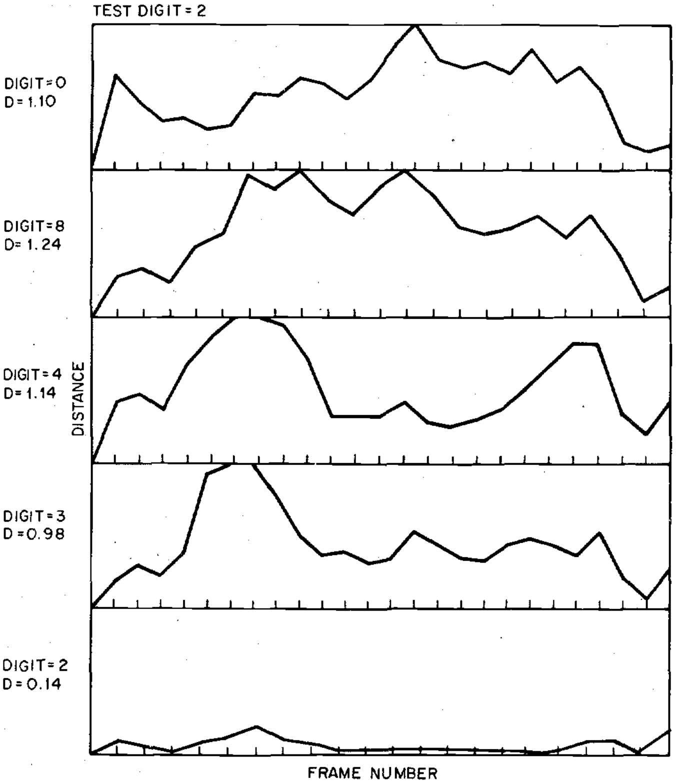

Fig. 5. Frame-by-frame distances for the test digit 2.

digit 2 in the string /123/. The absolute distance for each frame for the digit 2 was uniformly quite small, whereas for the other 4 candidates (3, 4, 8, and 0) the absolute distance for each frame was generally quite large.

Fig. 6 shows an example of the distance calculations for the digit 5. The similarity between the diphthongs in both 5 and 9 is seen by the small values of distance for the middle frames of 9. However, the initial and final regions of 9 and 5 were considerably different—giving an average distance for 9 of about 2.5 times the average distance for 5.

Finally, Fig. 7 shows the results of the distance calculations for the digit 8 in the string /486/. Using the distance measure of (14), the digit is recognized improperly as 3 because of the high degree of coarticulation between the final /r/ in four, and the initial $/e^{I}/$ in eight. Since this was a case in which the distance measure of (9) was required, the distance calculations were repeated giving the second set of curves shown in Fig. 7. The reduction in distance at the beginning of the digit eight is noteworthy; no other such reduction in distance is seen for the other candidates. Using the distance measure of (9), the digit is properly recognized as /8/ by a substantial margin over the nearest candidate.

For the speaker-independent mode the recognition procedure is essentially the same as the dependent mode. The only modification in the procedure is that the distance measure of (9) is utilized in cases where the initial boundary of the digit occurs at a voiced region, and whenever the ratio between the distances of the first two recognition candidates is less than 1.1. In addition, the measure of (9) is also used whenever the results of the recognition procedure using (14) yield an answer that is in conflict with the unvoiced-voiced information. For example, if the digit begins with an unvoiced region, and the smallest average distance is measured for either the digits 1, 9, or 8, the procedure is repeated using the measure of (9). If a conflict in unvoiced-voiced information is found and the preceding digit is not 5, 6, 8 (these digits sometimes end in

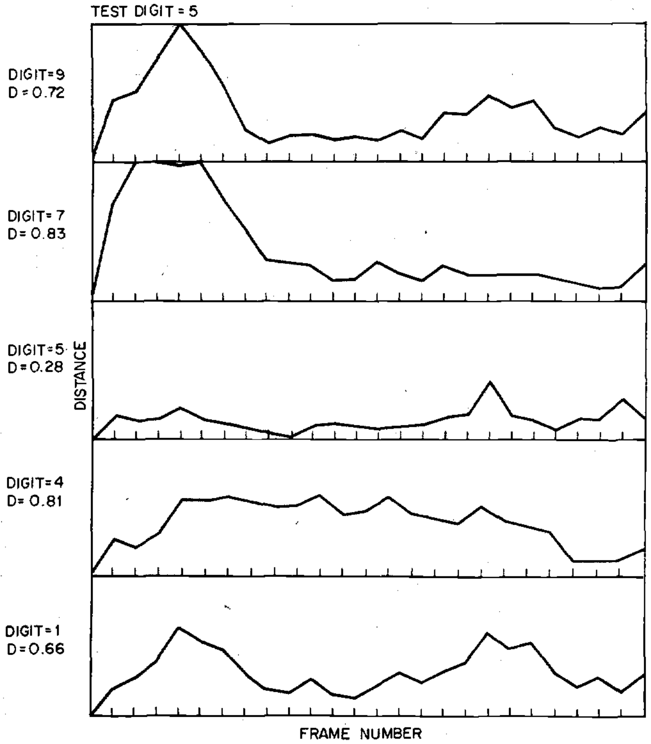

Fig. 6. Frame-by-frame distances for the test digit 5.

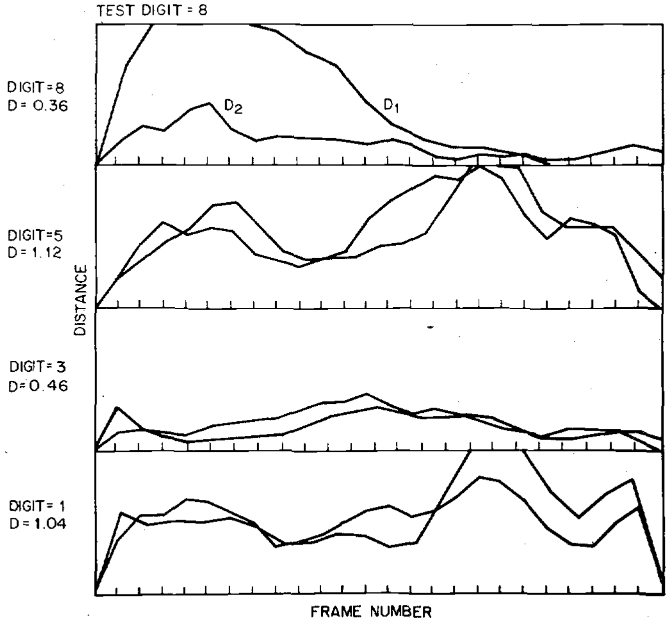

Fig. 7. Frame-by-frame distances for the test digit 8 using both the $D1$ and $D2$ distance measures.

TABLE II

Testing Sequence

027	310	912	181	279
435	813	132	291	103
908	886	629	864	424
292	035	318	206	761
806	712	161	943	045
628	786	938	786	451
533	327	370	554	075
975	480	147	563	697
542	259	660	677	894
550	354	798	409	049

unvoiced sounds), the digits 1, 9 and 8 are eliminated from considerations whenever the initial unvoiced interval exceeds 70 ms.

VI. System Evaluation

The recognition system was evaluated in both the speaker-dependent and the speaker-independent modes. The results of these experiments are shown in Tables III and IV.

To test the system each speaker read a balanced list of 50 three-digit strings (see Table II) over a dynamic microphone in a computer room environment. The test digit strings were in no way related to the digit strings used to train the system. For the speaker-dependent test six speakers were used—four male and two female. The recognition accuracies for these speakers are given in Table I. Four of the six speakers were tested using only the distance measure of (14). The remaining two speakers used the system when both distance measures were being evaluated.

As seen in Table III, the absolute recognition accuracy ranged from 97 to 100 percent across the six speakers. When a threshold was used to eliminate decisions when the ratio of the distance from the second candidate to the distance from the first candidate fell below 1.2 (this is called the 20 percent rule), several of the errors became cases where no decision was made. However, several of the correct decisions also were classified as cases where no decision was made. The number of no decisions and errors for each of the six speakers using the 20 percent rule is given in Table III.

Although the number of errors was small, there was a distinct pattern to the errors. Most of the errors occurred with the digits 1, 8, and 9. As discussed previously, these digits are most susceptible to errors due to the high degree of variation in their acoustic properties due to coarticulation when imbedded in digit strings.

Table IV shows the results for a speaker-independent experiment involving ten male speakers. The training data for this

TABLE III
RESULTS OF SPEAKER-DEPENDENT EVALUATION TESTS

Males	# Trials	# Recognition Errors	% Correct	# No Decision	# Errors	% Correct
LRR	150	4	97.3	4	2	96.0
MRS	150	2	98.7	0	2	98.7
AER*	150	1	99.3	0	1	99.3
REC*	150	1	99.3	2	0	98.7
TOTALS	600	8	98.7	6	5	98.2
Females						
KD	150	0	100.0	1	0	99.3
KS	150	3	98.0	6	2	94.7
TOTALS	300	3	99.0	7	2	97.0

* Used both distance measures. All other speakers used only the distance metric of Eq. (13).

TABLE IV
RECOGNITION ACCURACIES FOR SPEAKER-INDEPENDENT DIGIT RECOGNITION SYSTEM

Speaker	No. Presentations	No. Correct	No. Errors	Recognition Accuracy
JH	60	59	1	98.3%
JM	60	59	1	98.3
OJ	60	56	4	93.3
JF	60	57	3	95.0
CC	60	55	5	91.7
MS	60	54	6	90.0
ML	60	58	2	96.7
FP	60	60	0	100.0
MC	60	54	6	90.0
JO	60	59	1	98.3
TOTAL	600	572	28	95.3%

TABLE V
CONFUSION MATRIX FOR SPEAKER-INDEPENDENT DIGIT RECOGNITION SYSTEM

		Digit Recognized										
		0	1	2	3	4	5	6	7	8	9	Total
	0	53	1	3		2			1			60
	1		60									60
	2			59	1							60
Digit	3			2	56					2		60
	4					60						60
Spoken	5		1				54		4		1	60
	6			1				58		1		60
	7			3					57			60
	8									60		60
	9		5	1							54	60
Total		53	67	69	57	62	54	58	62	63	55	600

experiment were collected from the data of the four male speakers were not included in the speaker-independent test.) speakers were not included in the speaker-independent test). In this experiment each speaker sequentially read a list of 20 three-digit strings from Tables I and II. The total recognition accuracy for the experiment was 95.3 percent and the accuracy for any given speaker was no worse than 90 percent. As seen in Table V, the most common confusions in this experiment were between the digits 1 and 9.

VII. CONCLUSIONS

A system for the recognition of connected digits was described in this paper. The system incorporated a distance measure that was based on an expanded form of the principle of minimum residual error. The expanded measure included the effects of coarticulation and multiple repetitions and can be used for both speaker-independent and speaker-dependent situations.

In an experimental evaluation the system achieved an accuracy of nearly 99 percent for the speaker-dependent mode and 95 percent for the speaker-independent mode. Although the problem of continuous digit recognition is significantly more difficult than isolated digit recognition, these results compare quite favorably to the scores reported for isolated digit recognition [1], [3]. The use of the expanded distance measure was a major factor in the success of the recognition system.

REFERENCES

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

[2] T. B. Martin, "Acoustic recognition of a limited vocabulary in continuous speech," Ph.D. dissertation, Univ. Pennsylvania, Philadelphia, 1970.

[3] ——, "Practical applications of voice input to machine," *Proc. IEEE (Special Issue on Man-Machine Communication by Voice)*, vol. 64, pp. 487-501, Apr. 1976.

[4] L. R. Rabiner and M. R. Sambur, "Some preliminary results on

the recognition of connected digits," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 170-182, Apr. 1976.

[5] ——, "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech. J.*, vol. 54, pp. 297-315, Feb. 1975.

[6] 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.

[7] J. D. Markel, A. H. Gray, Jr., and H. Wakita, "Linear prediction of speech-theory and practice," Speech Commun. Res. Lab., Santa Barbara, CA, Monograph 10, 1973.

[8] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing (Special Issue on IEEE Symposium on Speech Recognition)*, vol. ASSP-23, pp. 67-72, Feb. 1975.

[9] H. Mann and A. Wald, "On the statistical treatment of linear stochastic difference equations," *Econometrica*, vol. II, no. 3 and 4, July-Oct. 1943.

[10] H. Van Trees, *Detection, Estimation and Modulation Theory.* New York: Wiley, 1968.

[11] G. M. White, "Speech Recognition: A tutorial overview," *Comput.*, 1976.

The Design of Markov Chains for Waveform Generation

LEAH J. SIEGEL, STUDENT MEMBER, IEEE, KENNETH STEIGLITZ, MEMBER, IEEE, AND MARK ZUCKERMAN

Abstract—**A linear programming algorithm is described for designing circulant Markov chains which generate random waveforms whose spectral densities have specified poles. These chains can be implemented by random jumping in a fixed table, and have the advantages of speed and simplicity. Possible applications to speech synthesis are discussed.**

I. Introduction

In many digital signal processing applications, it is necessary to generate random signals with prescribed spectral densities. Computer generation of music and speech, and the generation of test signals for system simulation are two such applications. One straightforward way of generating such signals is to filter white noise with a suitable digital filter. We present here an alternate method which has three distinct advantages: 1) the method requires no multiplies, 2) the speed is independent of the number of poles in the spectral density, and 3) the maximum output amplitude and rms values are precisely controllable. The method is also naturally suited to hardware implementation.

In [1] it was shown that a circulant Markov chain (CMC) can be used to generate random processes whose spectral densities are rational and have a simple and convenient form. CMC's can be implemented by randomly jumping in a fixed table, and we restate the main result in the form of an algorithm.

Algorithm CMC:

1) Initialize the following variables:

a) the table size N;

b) the *table* $a(i), i = 0, \cdots, N-1$;

c) the *probabilities* $p(i), i = 0, \cdots, N-1$, where $p(i) \geq 0$, and $\Sigma_{i=0}^{N-1} p(i) = 1$.

2) Initialize the pointer $j = 1$, and time $t = 0$.

3) Using a random number generator, set the increment $k = i$ with probability $p(i), i = 0, \cdots, N-1$.

4) Set $j = (j + k) \bmod N$.

5) Set the output variable $y(t) = a(j)$.

6) Set $t = t + 1$ and go to 3).

The random variable $y(t)$ has the autocorrelation function

$$\phi_{yy}(n) = \sum_{k=1}^{N-1} |A(k)|^2 [P(k)]^{|n|} \tag{1}$$

where A is the IDFT of the table a:

$$A(k) = \frac{1}{N} \sum_{i=0}^{N-1} a(i) W^{ik}, \qquad W = e^{j2\pi/N} \tag{2}$$

and P is the conjugate DFT of the probabilities

$$P(k) = \sum_{i=0}^{N-1} p(i) W^{ik}. \tag{3}$$

The corresponding spectral density is therefore

$$\Phi_{yy}(z) = \sum_{k=1}^{N-1} |A(k)|^2 \cdot \left[\frac{1}{1 - P(k)z^{-1}} + \frac{1}{1 - P(k)z} - 1\right] \tag{4}$$

Manuscript received December 12, 1975; revised May 27, 1976 and August 2, 1976. This work was supported by the NSF under Grant GK-42048 and by the U.S. Army Research Office at Durham under Grant DAHCO4-75-G-0192. An earlier version of this paper was presented at the 1975 IEEE Electronics and Aerospace Systems Conference, Washington, DC, Oct. 1975.

The authors are with the Department of Electrical Engineering and Computer Science, Princeton University, Princeton, NJ 08540.