ACKNOWLEDGMENT

The authors wish to thank C. H. Coker for his close collaboration in devising the interface transformation between our synthesizer control input and his text-synthesis program.

REFERENCES

[1] J. L. Flanagan, "Source-system interaction in the vocal tract," *Ann. New York Acad. Sci.*, vol. 155, art. 1, pp. 9-17, Nov. 20, 1968.

[2] K. Ishizaka and M. Matsudaira, "What makes the vocal cords vibrate," *Proc. 6th Int. Congr. Acoust.*, Aug. 1968, pt. II, pp. B, 9-12.

[3] J. L. Flanagan and L. L. Landgraf, "Self-oscillating source for vocal-tract synthesizers," *IEEE Trans. Audio Electroacoust. (Special Issue on Speech Communication and Processing–Part I)*, vol. AU-16, pp. 57-64, Mar. 1968.

[4] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *Bell Syst. Tech. J.*, vol. 50, pp. 1233-1268, July-Aug. 1972.

[5] J. L. Flanagan, *Speech Analysis, Synthesis and Perception*, second ed. New York: Springer-Verlag, 1972.

[6] K. Ishizaka, J. C. French, and J. L. Flanagan, "Direct determination of vocal tract wall impedance," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 370-373, Aug. 1975.

[7] S. N. Rshevkin, *A Course of Lectures on the Theory of Sound.* New York: Macmillan, 1963, pp. 400-405.

[8] W. Meyer-Eppler, "Die Reliefdarstellung von Zeit-Frequenz-Spektren durch photographische Differentiation," *Akust. Beih.*, vol. Ab-1-3, no. 1, 1951.

[9] J. L. Flanagan and L. Cherry, "Excitation of vocal tract synthesizers," *J. Acoust. Soc. Amer.*, vol. 45, pp. 764-769, 1969.

[10] G. Fant, *Acoustic Theory of Speech Production.* 's-Gravenhage, The Netherlands: Mouton, 1960.

[11] C. H. Coker, N. Umeda, and C. P. Browman, "Automatic synthesis from ordinary English text," *IEEE Trans. Audio Electroacoust. (Special Issue on 1972 Conference on Speech Communication and Processing)*, vol. AU-21, pp. 293-298, June 1973.

[12] Cooperative studies with Dr. T. Shipp, Veterans Hospital, San Francisco, CA.

[13] J. L. Flanagan, K. Ishizaka, and K. L. Shipley, "Synthesis of speech from a dynamic model of the vocal cords and vocal tract," *Bell Syst. Tech. J.*, vol. 54, pp. 485-505, Mar. 1975.

Some Preliminary Experiments in the Recognition of Connected Digits

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

Abstract—This paper describes an implementation of a speaker independent system which can recognize connected digits. The overall recognition system consists of two separate but interrelated parts. The function of the first part of the system is to segment the digit string into the individual digits which comprise the string; the second part of the system then recognizes the individual digits based on the results of the segmentation. The segmentation of the digits is based on a voiced-unvoiced analysis of the digit string, as well as information about the location and amplitude of minima in the energy contour of the utterance. The digit recognition strategy is similar to the algorithm used by Sambur and Rabiner [1] for isolated digits, but with several important modifications due to the impreciseness with which the exact digit boundaries can be located. To evaluate the accuracy of the system in segmenting and recognizing digit strings a series of experiments was conducted. Using high-quality recordings from a soundproof booth the segmentation accuracy was found to be about 99 percent, and the recognition accuracy was about 91 percent across ten speakers (five male, five female). With recordings made in a noisy computer room the segmentation accuracy remained close to 99 percent, and the recognition accuracy was about 87 percent across another group of ten speakers (five male, five female).

I. INTRODUCTION

ONE of the most interesting areas of speech recognition is the problem of digit recognition. Although the applications to telephony alone would justify serious consideration of digit recognition methods, there are a wide variety of additional applications where voice input to a machine is extremely useful [2]. Among these applications are banking by voice, data input to a computer, and inventory and stock record keeping by machine.

Traditionally the approach to research in digit recognition has been to study the issues involved in recognizing isolated digits, spoken by a designated or one of a set of designated speakers [1], [3]-[7]. Although Martin has amply demonstrated the wide range of applications for such systems it is clear that for many applications of digit recognition (e.g., in telephony), these restrictions are highly undesirable.

Earlier work by Sambur and Rabiner [1] showed that a digit recognition accuracy on the order of 97 percent could be achieved with isolated digits across a broad range of male and female speakers. It was felt that these high scores justified experimentation with speaker independent systems for recognizing continuous digits.[1] It is the purpose of this paper to discuss such a system and to present results of experiments in recognizing strings of connected digits.

Fig. 1 shows a block diagram of the digit recognition scheme. The recorded digit string is first subjected to an endpoint analysis to determine where in the given recording interval the speech data occurs. The endpoint analysis is based on self-normalized measures of the energy and zero crossings of the

Manuscript received September 22, 1975; revised December 3, 1975. This work represents an equal contribution of the two authors.

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

[1] The authors have recently become aware of similar work in the area of digit recognition at Perception Technology, Inc., Winchester, MA, Dialog Systems, Inc., Cambridge, MA, and Texas Instruments, Inc., Dallas, TX.

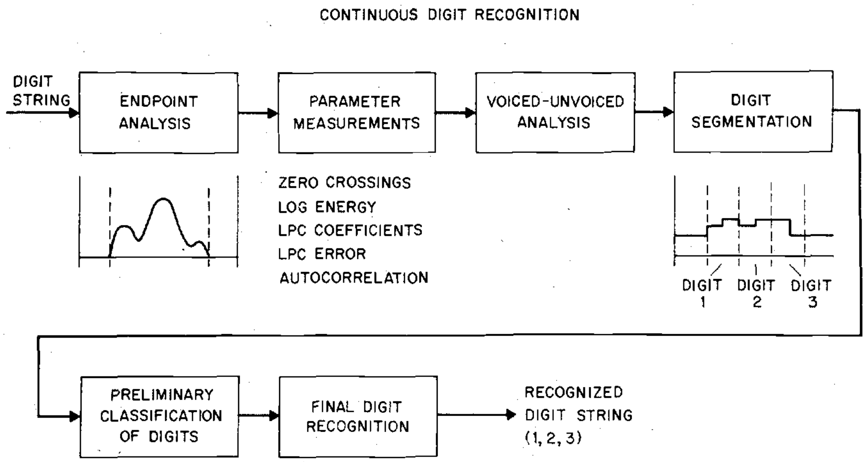

Fig. 1. Block diagram of the overall digit recognition system.

speech waveform, as described by Rabiner and Sambur [8]. Following endpoint alignment, the speech signal is analyzed 100 times/s, giving the following parameters:

1) Zero crossings
2) Log energy
3) Linear predictive coding (LPC) coefficients
4) LPC error
5) Autocorrelation coefficients

The measured parameters are then used in a statistical pattern recognition approach to give a voiced-unvoiced-silence contour of the utterance [9]. The voiced-unvoiced-silence contour is used along with some statistical information about the contour, and the speech energy measurements to segment the connected digit string into the individual digits. For the work to be described here it was assumed that all inputs were strings of three connected digits. This restriction on the number of digits could easily be removed with affecting any of the results to be presented. However, it is required that the number of digits in the string be specified.

Once the digits have been segmented, preliminary tests are made to check if the boundaries chosen by the segmentation algorithm might be grossly incorrect. In particular, classification tests on possibilities for any of the digits being 6, or 1, or 9 are made. Depending on the results of these tests the boundary locations of the individual digits are adjusted accordingly.

The final stage in the method is the digit recognition algorithm. This algorithm is similar in philosophy, but greatly different in details of implementation from the digit recognition algorithm described in [1]. The differences are due primarily to the coarticulation effects which are present with connected digits, but which are absent for isolated digits. In addition, there is a great deal less accuracy in the location of the digit boundaries for connected digits than there was for isolated digits. Thus a fairly large amount of processing is often required to make the final digit recognition.

The digit recognition system of Fig. 1 has been tested across a variety of speakers and transmission conditions. Experimental results have yielded about a 91 percent correct recognition score for ten subjects with recordings made in a soundproof booth using a high-quality microphone. For recordings made in a noisy computer room, an accuracy of 87 percent correct digit recognition was achieved across another group of ten speakers. Informal experiments showed the system to be capable of working on telephone line data if some modifications were made in the detailed recognition rules, and in the training data for the voiced-unvoiced-silence decision algorithm.

The organization for this paper is as follows. In Section II we discuss the method used to segment the connected digit string into the individual digits. Included in the section is a discussion of the various parameters which were used in the voiced-unvoiced-silence analysis, and for segmentation. In Section III the algorithm used to recognize the individual digits is described. Also described in this section is the method by which the digits are preliminarily classified to eliminate gross boundary errors, and the method of adjusting the boundaries based on the classification results. In Section IV an experimental evaluation of the overall recognition algorithm is described. Confusion matrices are included to show the types and distribution of recognition errors across the digits, and across different speakers. Finally, in Section V a general discussion of the problems of connected digit recognition is given.

II. Connected Digit Segmentation

In order to recognize the digits in the input string, it is necessary to accurately segment the input into the individual digits. Segmentation of speech is inherently an extremely difficult problem. However, for connected digit strings this problem is much more readily solved than for continuous speech in general. This is because by restricting the speech to be a string of digits known properties of the 10 possible digits can be used to accurately locate the digit boundaries. For example, it is easily shown that an interval of unvoiced speech or silence within the digit string denotes the beginning or end of a digit, i.e., there are no internal unvoiced or silence regions within the 10 digits (0-9).[2] Another observation about the digits is that, with a couple of exceptions, the energy contours of the digits have no internal local minima. Thus, excluding the noted exceptions, local minima of the energy

[2] Digits 6 and 8 can have internal silence or unvoiced regions. Such cases are treated directly using preliminary digit classification rules as discussed in Section III-B.

contour also are strong indications of digit boundaries within the digit string.

A. Voiced-Unvoiced Analysis

Based on the observations given above about the digits, a segmentation algorithm was studied which relied on an accurate voiced-unvoiced-silence analysis of the digit string, and the energy contour of the utterance. The voiced-unvoiced-silence analysis was made using a pattern recognition approach described by Atal and Rabiner [9]. The way in which this algorithm works is as follows. Each 10 ms interval of the utterance is classified as voiced speech, unvoiced speech, or silence, based entirely on measurements made on the signal during the prescribed interval. The classification of the interval into one of the three possibilities (voiced-unvoiced-silence) is based on a classical hypothesis-testing procedure in which a non-Euclidean distance is computed from the given set of measurements, for each of the hypotheses, and the minimum distance determines which class is selected. The classification algorithm had to be trained, a priori, to obtain the necessary statistics of the measurements for each of the three classifications.

The measurements which were used in the digit recognition experiment were the zero crossing count (number of zero crossings per 10 ms interval), the log energy of the interval, the first autocorrelation coefficient, the first LPC coefficient in a two-pole LPC analysis, and the LPC error of the two-pole analysis. These measurements differ from the ones described in [9] in that a two-pole LPC analysis was used rather than a 12-pole LPC analysis. The two-pole analysis was used because it was required in the final digit recognition phase of the system, and rather than perform two distinct LPC analyses, the method was trained on the two-pole analysis data. Additionally, the two-pole LPC analysis reduced the computation time considerably over an equivalent twelve-pole LPC analysis.

As mentioned above, the voiced-unvoiced classification algorithm requires a priori training to obtain the statistics for the measurements used in the analysis. Although the training is speaker independent (as long as a reasonable sampling of different speakers is used in the training set), the training is not independent of the transmission medium. Since three different transmission mediums were used in formal and informal evaluations of the recognition system, three sets of training data were used. The three conditions were the following.

1) Recordings made in a double-walled, soundproof booth using a high-quality microphone. This condition is called the high-quality speech.

2) Recordings made in a noisy computer room using a high-quality microphone. This condition is called the room-quality speech.

3) Recordings made over a dialed up telephone line using a telephone handset. These recordings went through the local Bell Laboratories PBX. This condition is called the telephone-quality speech.

Tables I-III show the training statistics for each of the five measurements, for each of the three classes, and for each of the three speech conditions. Table I is for high-quality speech; Table II for room-quality speech; and Table III is for telephone-quality speech. Each of these tables shows the measured means, standard deviations, and covariance matrices for the five measurements, for each of the classes, i.e., silence ($L = 1$), unvoiced speech ($L = 2$), and voiced speech ($L = 3$). Several observations can be made about the behavior of the measurements for these three conditions. It can be seen that the distributions of the five measurements for the room- and high-quality conditions are essentially the same for voiced and unvoiced speech; however significant changes have occurred in the distributions for silence. This result is what one would anticipate since the transmission system (the microphone) was the same in both cases; only the background silence characteristics have varied. This result is most readily seen in comparing the difference in means of the log energy between voiced speech and silence for high- and room-quality speech. For high-quality speech this difference is about 39.6 dB whereas for room-quality speech this difference is about 33.4 dB. Also the silence correlation coefficient mean is about 0.73 for high-quality speech, whereas it is about 0.91 for room-quality speech—a quite significant difference.

In comparing the distributions of the measurements for telephone-quality speech (Table III) with high-quality speech (Table I), many more differences are apparent. Although the statistical distributions of the measurements for voiced speech are quite similar, the strong band-limiting effects of a telephone line and a telephone handset are quite apparent for unvoiced speech and silence. For example, the average number of zero crossings for unvoiced speech is 49.5 for high-quality speech, whereas for telephone speech it is about 20.0. Thus the effectiveness of the zero crossing measurement in differentiating between voiced and unvoiced speech is greatly reduced for telephone speech. The reason this measurement is retained rather than replaced by a more effective measurement is for uniformity in the use of the digit recognition system. Since the same measurements are used for all three recording conditions, the only change required in the recognition system is the substitution of one training set of data for another set.

For purposes of illustration, Figs. 2 and 3 show plots of the distribution of two of the measurements for silence, unvoiced, and voiced speech for the room-quality speech. Also included in these plots are Gaussian curves that have the same mean and standard deviation as the measured data. Fig. 2 shows plots of the distributions of zero crossings; whereas Fig. 3 shows plots of the distributions of the first LPC coefficient in the two-pole LPC analysis. The differences in mean and standard deviations of the measurements, for the different classes, are readily obvious from these plots.

B. Digit Segmentation Rules

The digit segmentation algorithm is based on the results of the voiced-unvoiced-silence analysis discussed in the previous section, and also uses the energy measurements to aid in locating the digit boundaries. Fig. 4 shows a flow diagram of the digit segmentation algorithm. Based on N_D, the number of digits in the input string ($N_D = 3$ throughout this paper), the algorithm searches for two external boundaries

TABLE I
MEANS, STANDARD DEVIATIONS, AND COVARIANCE MATRICES FOR HIGH-QUALITY SPEECH

	Zero Crossings	Log Energy (Decibels)	First Auto-correlation	First LPC	LPC Log Error (Decibels)
1) Silence					
Mean	23.270	9.448	0.728	-0.955	6.672
Standard Deviation	7.672	3.424	0.120	0.205	3.403
	1.000	-0.557	-0.900	0.669	-0.749
Covariance Matrix	-0.557	1.000	0.529	-0.805	0.946
	-0.900	0.529	1.000	-0.671	0.741
	0.669	-0.805	-0.671	1.000	-0.859
	-0.749	0.946	0.741	-0.859	1.000
2) Unvoiced					
Mean	49.532	25.365	0.024	-0.046	11.436
Standard Deviation	8.891	5.747	0.260	0.400	2.892
	1.000	0.068	-0.921	0.895	-0.092
Covariance Matrix	0.068	1.000	-0.032	-0.007	0.943
	-0.921	-0.032	1.000	-0.978	0.143
	0.895	-0.007	-0.978	1.000	-0.200
	-0.092	0.943	0.143	-0.200	1.000
3) Voiced					
Mean	12.589	49.679	0.880	-1.469	26.793
Standard Deviation	5.367	6.095	0.085	0.289	4.241
	1.000	0.371	-0.821	0.450	-0.271
Covariance Matrix	0.371	1.000	-0.309	-0.041	0.525
	-0.821	-0.309	1.000	-0.711	0.447
	0.450	-0.041	-0.711	1.000	-0.780
	-0.271	0.525	0.447	-0.780	1.000

TABLE II
MEANS, STANDARD DEVIATIONS, AND COVARIANCE MATRICES FOR ROOM-QUALITY SPEECH

	Zero Crossings	Log Energy (Decibels)	First Auto-correlation	First LPC	LPC Log Error (Decibels)
1) Silence					
Mean	12.167	18.138	0.911	-1.388	13.061
Standard Deviation	3.739	3.928	0.053	0.180	2.372
	1.000	-0.239	-0.591	0.524	-0.499
Covariance Matrix	-0.239	1.000	0.124	-0.281	0.863
	-0.591	0.124	1.000	-0.802	0.515
	0.524	-0.281	-0.802	1.000	-0.682
	-0.499	0.863	0.515	-0.682	1.000
2) Unvoiced					
Mean	49.860	30.568	0.007	-0.041	13.086
Standard Deviation	16.093	4.860	0.441	0.482	2.161
	1.000	0.361	-0.971	0.948	0.331
Covariance Matrix	0.361	1.000	-0.370	0.366	0.954
	-0.971	-0.370	1.000	-0.977	-0.341
	0.948	0.366	-0.977	1.000	0.319
	0.331	0.954	-0.341	0.319	1.000
3) Voiced					
Mean	12.858	51.522	0.878	-1.490	28.373
Standard Deviation	5.491	5.745	0.109	0.379	4.715
	1.000	0.208	-0.778	0.635	-0.471
Covariance Matrix	0.208	1.000	-0.055	-0.158	0.487
	-0.778	-0.055	1.000	-0.838	0.627
	0.635	-0.158	-0.838	1.000	-0.847
	-0.471	0.487	0.627	-0.847	1.000

TABLE III
MEANS, STANDARD DEVIATIONS, AND COVARIANCE MATRICES FOR TELEPHONE-QUALITY SPEECH

	Zero Crossings	Log Energy (Decibels)	First Auto-correlation	First LPC	LPC Log Error (Decibels)
1) Silence					
Mean	13.804	14.711	0.891	-1.374	11.495
Standard Deviation	4.554	4.991	0.068	0.189	3.463
	1.000	-0.242	-0.771	0.499	-0.437
Covariance Matrix	-0.242	1.000	0.158	-0.665	0.939
	-0.771	0.158	1.000	-0.597	0.417
	0.499	-0.665	-0.597	1.000	-0.830
	-0.437	0.939	0.417	-0.830	1.000
2) Unvoiced					
Mean	19.987	29.819	0.788	-1.305	16.332
Standard Deviation	5.707	8.224	0.096	0.173	3.702
	1.000	0.095	-0.857	0.746	-0.159
Covariance Matrix	0.095	1.000	-0.054	-0.236	0.942
	-0.857	-0.054	1.000	-0.883	0.231
	0.746	-0.236	-0.883	1.000	-0.531
	-0.159	0.942	0.231	-0.531	1.000
3) Voiced					
Mean	12.858	51.319	0.896	-1.651	29.289
Standard Deviation	5.213	6.171	0.069	0.173	3.726
	1.000	0.361	-0.886	0.676	-0.256
Covariance Matrix	0.361	1.000	-0.316	0.094	0.598
	-0.886	-0.316	1.000	-0.837	0.372
	0.676	0.094	-0.837	1.000	-0.670
	-0.256	0.598	0.372	-0.670	1.000

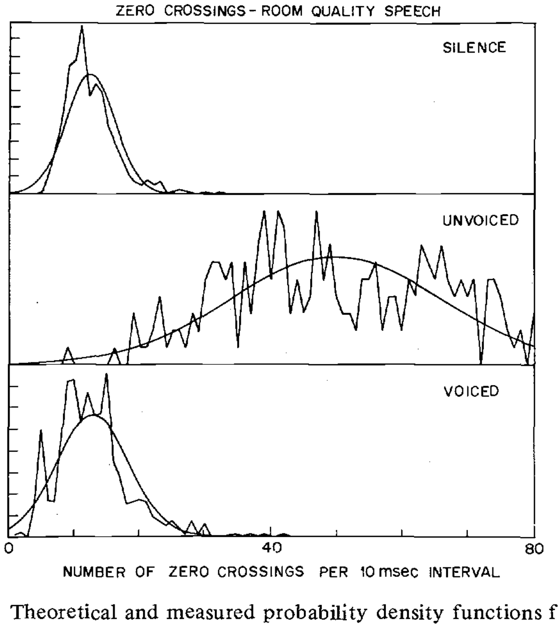

Fig. 2. Theoretical and measured probability density functions for the zero crossing measurement for room-quality speech.

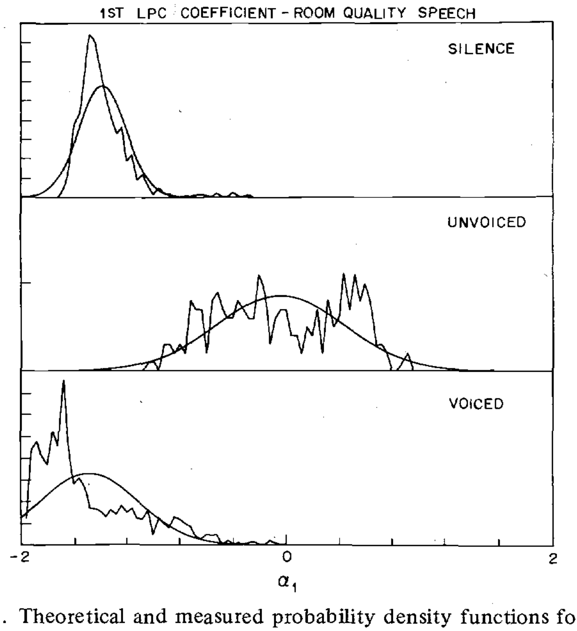

Fig. 3. Theoretical and measured probability density functions for the first LPC coefficient in a two-pole analysis for room-quality speech.

and (N_D - 1) internal boundaries. The external boundaries are generally the endpoints of the digit string as determined in the endpoint analysis. However the speech interval used for analysis is 100 ms longer both at the beginning and at the end of the utterance to account for possible endpoint location errors which can be corrected by the more sophisticated voiced-unvoiced-silence analysis. Thus the external boundaries are almost always located within 100 ms of the beginning and end of the analysis interval—as depicted in Fig. 1.

The internal boundaries are located using the procedure outlined in Fig. 4. First, all strong local minima of the energy contour of the utterance which occur in voiced regions are located and tagged. These minima are only possible candidates because the energy of the utterance can exhibit a strong dip both because of a word boundary, and because of a voiced fricative, e.g., /v/ in seven.

The next step in the algorithm is to locate one boundary in each nonvoiced region within the utterance, e.g., one digit

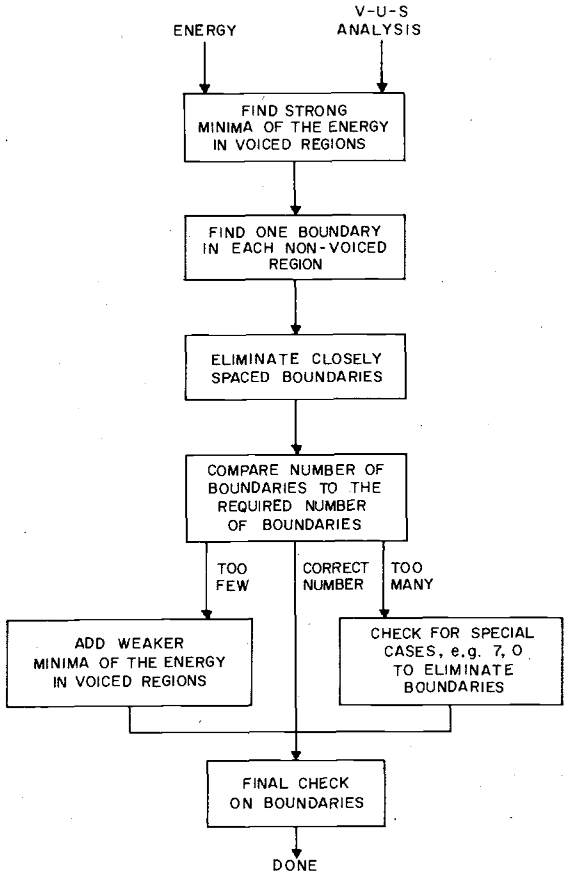
ENERGY
V-U-S ANALYSIS
FIND STRONG MINIMA OF THE ENERGY IN VOICED REGIONS
FIND ONE BOUNDARY IN EACH NON-VOICED REGION
ELIMINATE CLOSELY SPACED BOUNDARIES
COMPARE NUMBER OF BOUNDARIES TO THE REQUIRED NUMBER OF BOUNDARIES
TOO FEW
CORRECT NUMBER
TOO MANY
ADD WEAKER MINIMA OF THE ENERGY IN VOICED REGIONS
CHECK FOR SPECIAL CASES, e.g. 7, 0 TO ELIMINATE BOUNDARIES
FINAL CHECK ON BOUNDARIES
DONE

Fig. 4. Flow diagram for the digit segmentation algorithm.

boundary must occur with a silence, or unvoiced interval. The boundary locations are placed at the first nonvoiced interval within the region. The only exceptions to this rule (i.e., that the digit boundary occurs at the beginning of the nonvoiced region) are six, which ends in an unvoiced region, and eight, which ends in an interval of silence, and is sometimes followed by a stop burst at the release of the /t/. Methods of correcting the boundary locations for these cases are described in Section III where preliminary classification rules are used to adjust the boundaries within nonvoiced regions.

The first two steps in the flowchart of Fig. 4 showed ways of choosing candidates for the digit boundaries. The next step is to eliminate one boundary whenever two boundary candidates are too closely spaced (i.e., less than 150 ms apart). Appropriate logic decides which boundary candidate to eliminate in such cases.

The next step in the segmentation strategy is to compare the number of boundary candidates to the required number of internal boundaries. If there are too few choices, weaker minima of the energy contour are located and added as boundary candidates. If there are too many boundary candidates, checks are made for digits like seven, etc., for which spurious boundary candidates often occur and such boundary candidates are eliminated.

When the correct number of boundary candidates has been obtained, a final check is made on the boundaries for internal consistency, e.g., checks are made to insure each digit consists of only a single voiced region, etc.

To illustrate the operation of the segmentation rules, Figs. 5-9 show several examples of digit strings which were segmented using the rules discussed above. In each of these figures parts (a), (b), (c), and (d) show plots of the zero crossing contour, the log energy contour, a statistical measure of the certainty of the voiced-unvoiced-silence analysis, and the voiced-unvoiced-silence contour of the utterance, respectively. The statistical parameter shown in (c) is a measure of the probability that the decision made by the voiced-unvoiced-silence analysis is correct, and it varies from 0 to 1.0. The voiced-unvoiced-silence contour of (d) is a 3-level contour where level 1 is silence, level 2 is unvoiced speech, and level 3 is voiced speech.

Fig. 5 shows the segmentation boundaries for the digit string /721/. The initial boundary was placed at the beginning of the first unvoiced region, i.e., the /s/ is seven. The second boundary was placed at the initial interval of the second unvoiced region—corresponding to the /t/ in two. The third boundary was placed in the region of a local minimum of the log energy contour within the second voiced region. The exact boundary location is not at the absolute minimum of the log energy, but instead occurs somewhere within the region of the minimum. The exact location is determined by a series of complex decisions which will not be discussed. Although the correct location of the third boundary is not readily determined, it has been found that precise location of the boundaries within voiced regions is not required for reliable digit recognition. The final digit boundary is located at the beginning of the last silence region.

It should be pointed out that another possible candidate for a boundary location in Fig. 5 is at the strong local minimum in the log energy contour at the /v/ in seven. However, the segmentation rules were able to eliminate this case quite readily and instead choose the minimum in the second voiced region.

Fig. 6 shows a somewhat more complicated digit string—the string /191/. This input is all voiced—thus there are no convenient boundaries in unvoiced regions. In addition the local minima in the log energy plot are not very strong ones (e.g., the energy dips are not large ones), and the widths of these minima are quite large. Thus the choice of boundary locations was made at the next to last step in the segmentation algorithm. Listening tests showed that the location of these boundaries within the all-voiced regions was not critical due to the high degree of coarticulation in the speaking of such all-voiced digit strings.

Fig. 7 shows results for the segmentation of the digit string /650/. For this case there were three unvoiced regions and one boundary occurred at the beginning of each region. However the second boundary should actually be within the unvoiced region—not at the beginning of it. The dashed line in the figure shows where the boundary was moved by the preliminary classification algorithm which classified the initial digit as a six, and classified the following digit as one which might begin with an unvoiced region.

Finally, Figs. 8 and 9 show the three-digit sequence with which the segmentation rules had the most difficulty. It was the digit string /387/. Fig. 8 shows the segmentation

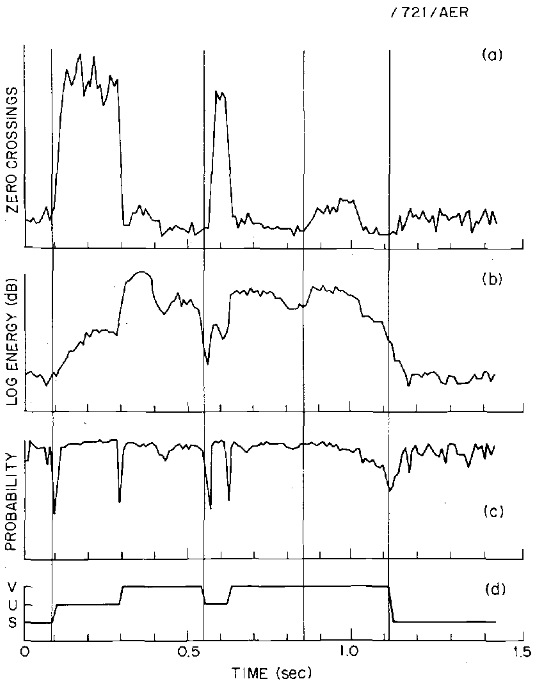

Fig. 5. Typical measurement contours and the resulting boundary locations for the utterance /721/.

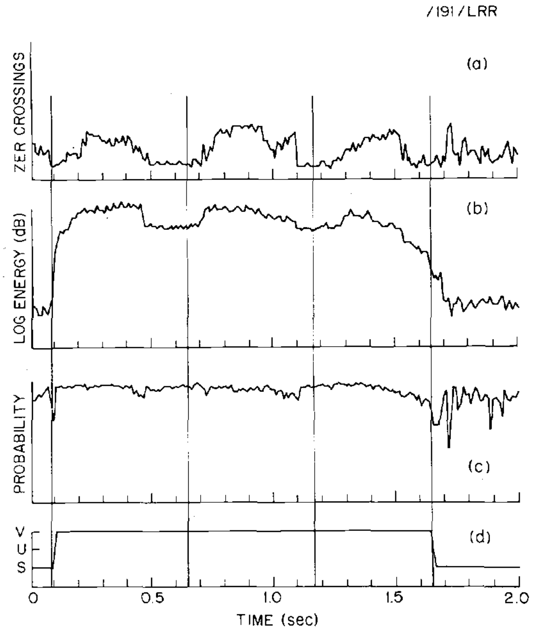

Fig. 6. Typical measurement contours and the resulting boundary locations for the utterance /191/.

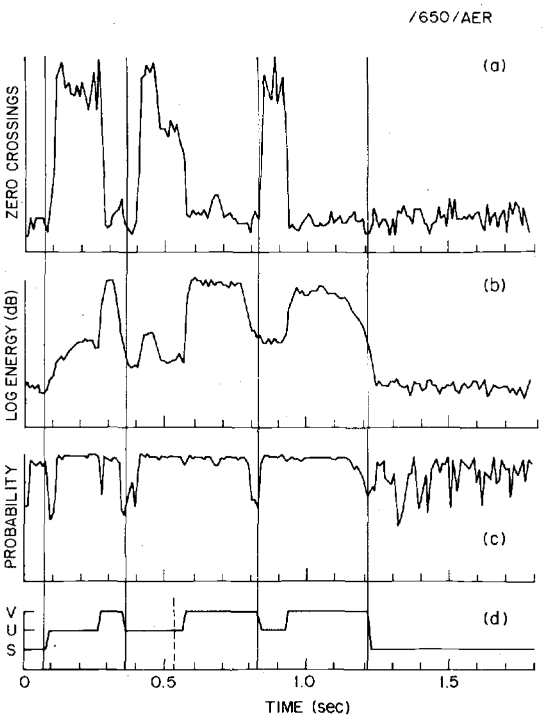

Fig. 7. Typical measurement contours and the resulting boundary locations for the utterance /650/.

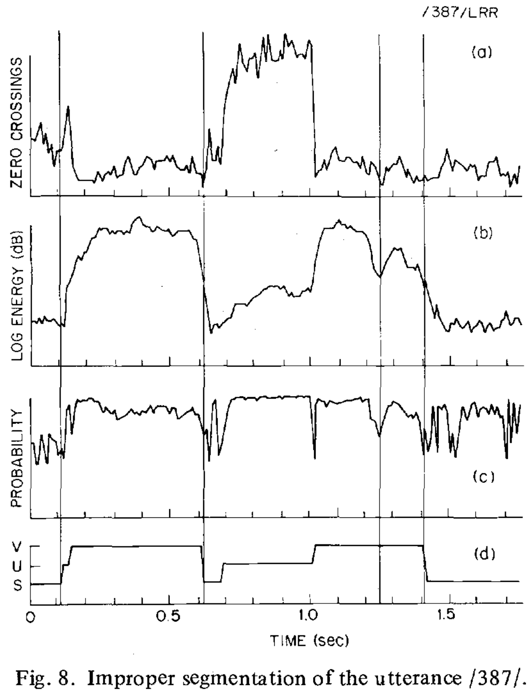

Fig. 8. Improper segmentation of the utterance /387/.

when the digit string was spoken fast, and Fig. 9 shows the segmentation when the digit string was spoken slowly and with emphasis on the /38/ rather than on the /7/.

The difference in emphasis is reflected in the relative durations of the frication in the initial /3/, and in the duration of the voiced regions of the /38/. The segmentation rules were not correct in placing a boundary at the /v/ in seven for the data of Fig. 8, whereas there was little difficulty segmenting the data of Fig. 9. It is also seen that the log energy during the voiced region of the /38/ of both Fig. 8 and Fig. 9 show no strong (or even weak) minima, indicating the presence of a boundary. However, for the data of Fig. 9, the relative durations strongly implied the presence of a boundary in the initial voiced region; even though no good location for the boundary could be found. For the data of Fig. 8, the relative durations implied the presence of a boundary in the second voiced region; hence the segmentation error.

An evaluation of the segmentation accuracy was made using 200 sequences of three digits recorded by ten different speakers in the computer room. Out of the 600 digits, only four were improperly segmented. An improper segmentation

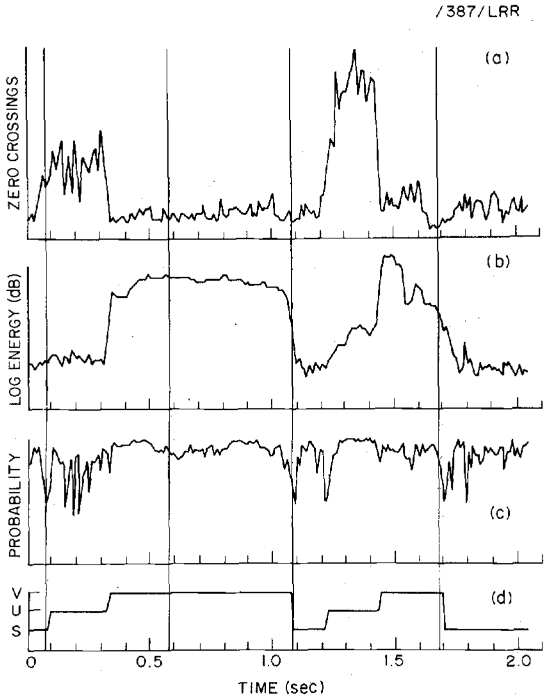

Fig. 9. Proper segmentation of the utterance /387/.

was defined as one in which a distinctly audible part of the preceding digit was included within the boundaries of the current digit being segmented. All of the errors occurred in cases similar to the one illustrated in Figs. 8 and 9. Thus a segmentation accuracy of about 99 percent was obtained on these three-digit strings.

III. Digit Recognition Algorithm

The next stage in the implementation of the recognition system is the identification of the digits within the boundaries determined by the segmentation algorithm. The digit recognition strategy that was used is similar to the one discussed by Sambur and Rabiner [1] for recognizing isolated digits but with several important modifications. The modifications are necessary to account for certain limitations in the segmentation algorithm. These limitations include:

1) impreciseness of the exact digit boundaries,
2) ambiguities in the assignment of boundaries for digits ending in unvoiced sounds, and
3) incorrect unvoiced-voiced decisions.

In addition, the effects of coarticulation must be incorporated in the recognition of connected digits. Before discussing the new aspects of the digit recognition algorithm, we shall briefly review the structure of the isolated digit scheme since this forms the framework of the present system.

A. Philosophy of Recognition

In the system for recognizing isolated digits, the approach was to describe the digits in terms of the six broad speech categories listed in Table IV. A set of robust measurements were then used to classify the sounds into the above categories. By robust measurements, we mean acoustic parameters that give a general indication of the gross nature of each phoneme without being overly dependent on the individual speaker's voice characteristics. The robust measurements selected were:

TABLE IV[a]
Sound Classes Characteristic of the Digits

Digit	Sequence of Sound Classes
0	VNLC → FV → VLC → BV
1	VLC → MV → VLC
2	UVNLC → FV → BV
3	UVNLC → VLC → FV
4	UVNLC → BV → MV
5	UVNLC → MV → FV → VNLC
6	UVNLC → FV → UVNLC
7	UVNLC → FV → VNLC → FV → VLC
8	FV → UVNLC
9	VLC → MV → FV → VLC

VNLC	Voiced, noise-like consonant.
UVNLC	Unvoiced, noise-like consonant.
VLC	Vowel-like consonant.
FV	Front vowel.
MV	Middle vowel.
BV	Back vowel.

[a] After Martin [3].

1) Zero-crossing rate (ZCR), which is defined as the number of zero crossings in a fixed-frame length (on the order of 10 ms).
2) Energy, which is defined as the sum of the squared values of the speech waveform in a given frame.
3) Normalized error obtained from a two-pole LPC analysis of a given speech frame.
4) Pole frequency (or frequencies) obtained from a two-pole LPC analysis of a given speech frame.

To enable the system to perform without having to be trained every time a different speaker wishes to use the system, the technique of self-normalization was introduced. Self-normalization implies that the system does not use fixed threshold levels in the decision process, but instead calculates the appropriate decision thresholds from the speech input itself. Thus, for example, in the case of setting thresholds on ZCR to determine whether a sound is noise-like or nasal, a statistical description of the ZCR was made for each isolated digit. The statistical description consisted of measuring the mean of the ZCR and its standard deviation over the region of strong energy (i.e., the region where the energy exceeded 10 percent of the maximum energy of the utterance). Based on ZCR measurements, one criterion for classifying a segment as noise-like was if its ZCR exceeded a level one standard deviation above the mean during the segment. Fig. 10 shows the ZCR measurements for the word "seven." Indicated in this figure are the average ZCR and a range of one standard deviation around this average. During the initial /s/, the ZCR is significantly above the threshold, as anticipated.

The technique of "self-normalization" also implies that the decision process should avoid the use of absolute decisions and instead classify sounds according to the transitional nature of the various measurements. It was shown that for speech sounds, the relative magnitude of the normalized error generally increases from sonorants to vowels and then to frica-

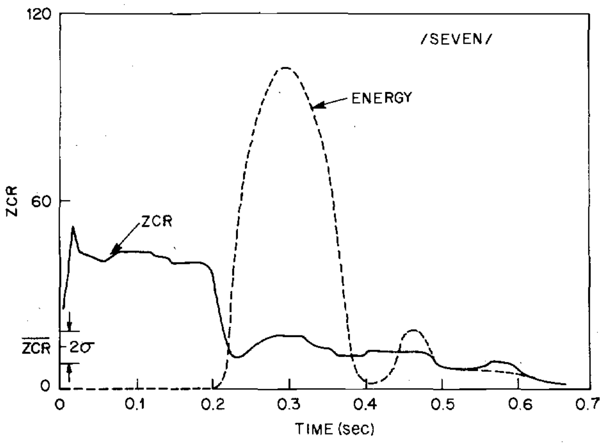

Fig. 10. Energy and ZCR for the digit 7.

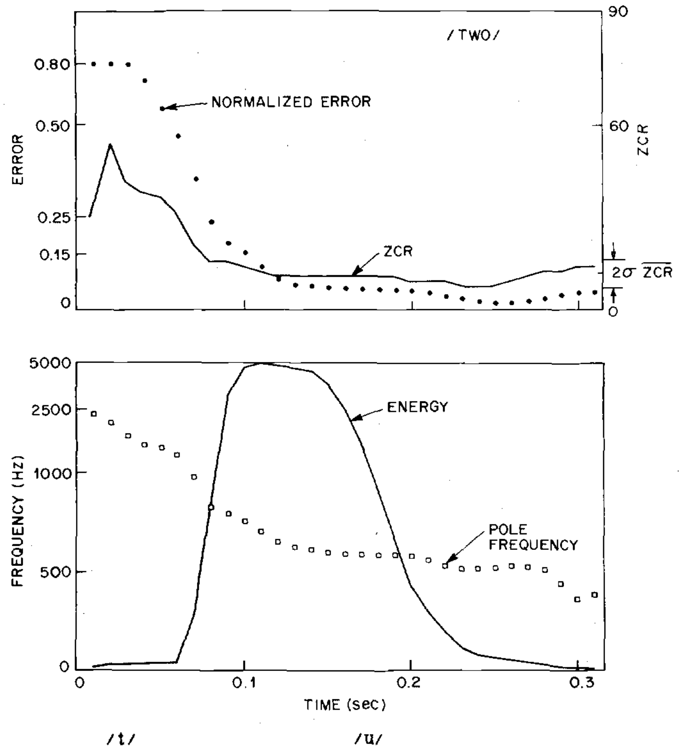

Fig. 11. Complete set of measurements for the digit 2.

tives. Within the three vowel types, the back vowels have the lowest relative normalized error and the front vowels have the highest. By observing the relative changes in the pole frequency and normalized error, important information about the structure of the voiced region of the word can be obtained. As an example, Fig. 11 shows the normalized error and pole frequency throughout the word "two." After the frication region, which is marked by high normalized error and low energy, the normalized error uniformly decreases. Thus the constituent structure of the voiced section is changing from a front vowel to a back vowel.

The final classification of the isolated digits was achieved by a "hypothesize-verification" scheme. In this method, the identity of the digit is first proposed and then the acoustic parameters are checked to see if they are consistent with this hypothesis. The sequence of consistency check was designed to verify the most obvious characteristic and then to proceed to the more difficult decisions. The essential strategy of the connected digit recognition is the same as the isolated scheme, but as noted above certain modifications were necessary. These modifications are described in the following subsections.

B. Preliminary Decisions

In Section II-B it was noted that the segmentation algorithm assigns a boundary location at the first nonvoiced interval within an unvoiced region. This rule is clearly not adequate for both the digit 6, which ends in an unvoiced region, and for cases when the final /t/ in 8 is released.

To partially overcome the limitation of this rule, a preliminary decision scheme was incorporated in the recognition system. One function of the preliminary digit classifier was to check for occurrences of the digit 6. The decision process first selects possible candidates for the digit 6. A "6 candidate" is defined as a digit with a voiced interval of less than 210 ms in duration that is surrounded by nonvoiced intervals and does not contain an internal boundary. Whenever a digit is preliminarily identified as a 6 using the above tests, it is automatically checked for the existence of a stop gap following the voiced interval. The determination of the presence of a stop gap is made by making measurements of the energy and zero crossings during the following unvoiced interval. A significant local minimum in the energy and zero crossing contours at the initial portion of the unvoiced interval is a strong indication of a stop gap. If the decision is that a stop gap exists, the voiced region is checked for front vowel like characteristics [1] and a positive indication confirms the digit as 6.

Assuming that a digit is classified as 6, there is still some ambiguity in locating the proper initial boundary for the following digit (of course if 6 is the final word, no problem exists). If the digit 6 is followed by 1, 9, or 8, then the boundary location should be placed at the first voiced interval following the 6; otherwise the location is within the preceding unvoiced interval. To resolve this ambiguity, a preliminary decision box has been incorporated to detect the digits 1 or 9.

Basically the 1, 9 decision involves checking the voiced interval following the digit classified as 6 to determine the existence of nasal like segments. A self-normalized ZCR threshold procedure was used as one indication of a nasal-like sound. The statistics of the ZCR were calculated in the voiced interval under investigation over those segments in which the energy exceeded 10 percent of the maximum energy in the interval and for which the two-pole frequency exceeded 375 Hz. The first criterion for a nasal-like sound to meet is that the ZCR rate should fall below a level one standard deviation below the average ZCR for three consecutive intervals. In addition, the nasal-like segment should exhibit a "relatively" low normalized error and the two-pole frequency should exhibit abrupt changes at the nasal-like boundaries. Fig. 12 shows the measured parameters for a typical pronounciation of the word "nine." The nasal segments of the word are clearly consistent with the characteristics verified by the decision rule.

To arrive at the final decision for the existence of the digits

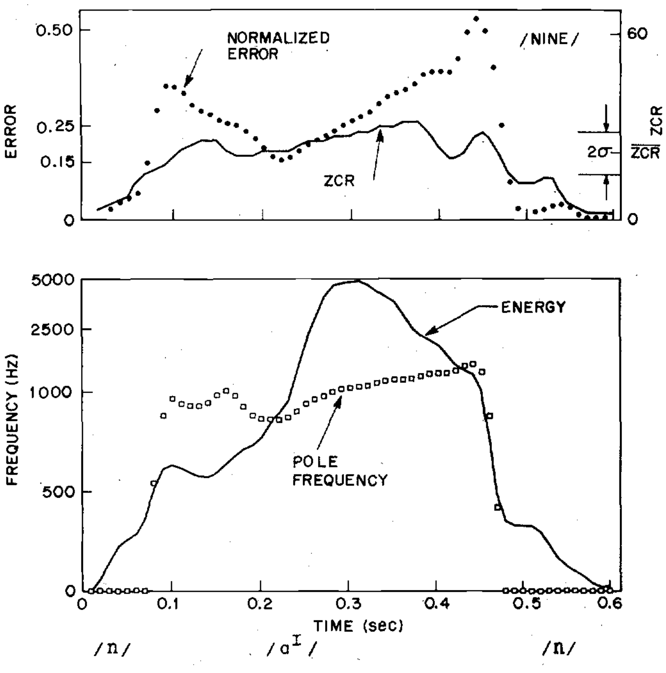

Fig. 12. Complete set of measurements for the digit 9.

1 or 9, the initial and final regions of the voiced interval must exhibit nasal-like characteristics. To avoid any possible confusion with the digit 7, an additional check is made any significant dips in the energy contour. These dips are strongly characteristic of the sound /v/ in 7. If the voiced interval contains a boundary location, this boundary location is allowed to float or wobble to account for a certain impreciseness in assigning internal boundaries in voiced regions. The wobbling of the boundaries is restricted to a range of ±30 ms and each possible boundary is checked for adjacent nasal like segments.

The preliminary recognition scheme does not check for the existence of the digit 8, but it is implicitly recognized in the final decision algorithm that the digit 8 is possible after the detection of the digit 6. In addition, if the digit 8 is recognized in the final decision algorithm and an unvoiced interval follows the recognized digit, it is implicitly assumed that the unvoiced interval is due to the release of the /t/ and a 1 or 9 check is undertaken. We now discuss the final digit recognition algorithm.

C. Digit Recognition

Fig. 13 illustrates the decision process used in the digit recognition algorithm. As can be noted from this figure, the results of the unvoiced-voiced analysis are *not* assumed to be error free. Thus if the initial segment of the bounded interval is classified as voiced, the decision is not automatically restricted to the digits 1, 9, and 8. Instead these digits are postulated as the most likely candidates and the acoustic parameters are then checked to verify these possibilities. The likelihood that the digit is either 1, 9, or 8 is quantified in terms of the probability (as measured in the voiced-unvoiced analysis) that the initial section of the word is voiced and the position of the word in the digit string. If the word is the first digit then there is a good chance that the weak fricatives /f/ and /th/ may not be classified as unvoiced.

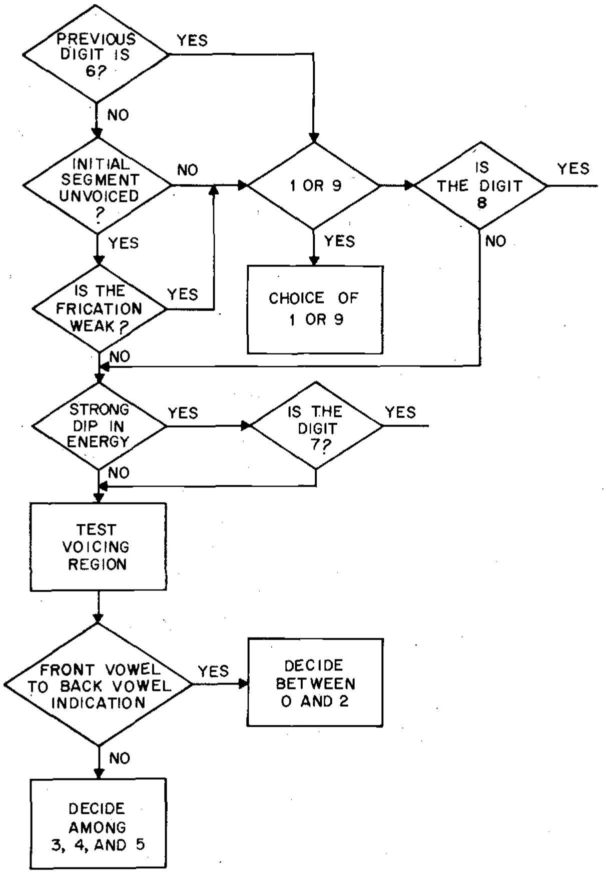

Fig. 13. Decision tree for the digit recognition algorithm.

The decision boxes shown in Fig. 13 are, for the most part, similar to the corresponding boxes used in the isolated digit recognition algorithm [1]. However, to account for the effects of coarticulation, the use of transitional changes in the second formant contour has been incorporated as a measurement feature. Used in parallel with the measured transitional changes in two-pole frequency and normalized error, an accurate decision can be obtained. The transitional nature of the second formant was derived by measuring F_2 at 3 locations in the segmented digit. The locations used in the measurement are the first point at which the probability of voicing exceeds 80 percent for three consecutive intervals, the point of maximum energy, and the first point after the maximum at which the energy dips below 15 percent of the maximum value. The formants were measured over a 20 ms interval by computing a linear prediction spectrum [10]. A peak-picking algorithm determined the second formant frequency. Fig. 14 illustrates the three LPC spectrums obtained for the digit 2. For this digit a fairly large drop in the second formant can be seen in comparing Fig. 14(a) and (c). Also, the third formant falls somewhat during the voiced region.

Since the complete details of the recognition algorithm are quite involved, we will only discuss the most prominent aspects of the decision rules for each box in Fig. 13. After a digit has been classified as beginning with a frication region, the energy contour is checked to determine if the energy dipped below 15 percent of the maximum value and then rose to a value at least 18 percent of the maximum. This

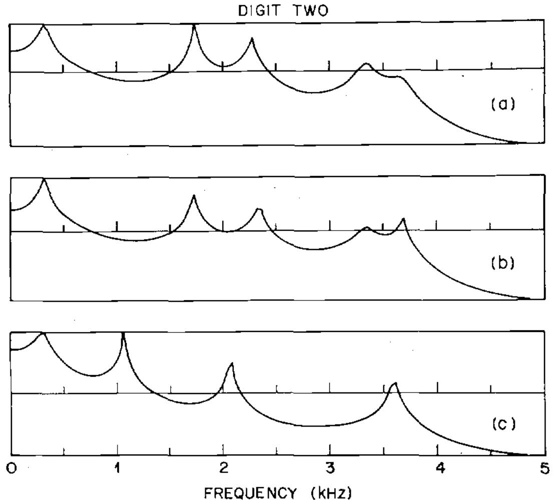

Fig. 14. Application of formants in the final recognition decision.

dip is an almost unfailing characteristic of the /v/ in the digit 7. Should such a dip occur, the ZCR of the digit is checked to see if it falls below the average during the end portion of the digit (after the second local maxima). If this condition is fulfilled, the digit is classified as 7.

If there is no strong dip in energy, the variations in the normalized error, the second formant and the two-pole frequency are calculated during the voicing region. A decrease in any two of these three measurements establishes the voiced region as having a front vowel to back vowel indication. The digits 0 and 2 are then the only candidates for recognition. The digit 0 differs from the digit 2 in the character of the energy contour and two-pole frequency. The digit 0 usually has a small but significant dip in energy during the consonant /r/ and a corresponding dip in two-pole frequency.

If the transitions in the parameters mentioned above are increasing rather than decreasing, the candidates for recognition are 3, 4, and 5. The digit 3 is characterized by it relatively high value of F_2 in the last measured formant location ($F_2 > 1700$ Hz). In addition, the two-pole frequency of the digit 3 is frequently zero at some point in the interval after the maximum energy, but before the energy dips below 20 percent of the maximum. This zero in the location of the two-pole frequency is attributed to the high second formant of the vowel /i/ in 3 [1], and is almost never observed for the other possible candidates (e.g., 4 and 5). The digit 5 can be separated from the digit 4 by its higher F_2 value in the end region and its higher values of two-pole frequency and normalized error. In addition, the two-pole frequency contour of the digit 4 is more uniformly increasing than the contour of 5.

When the candidates for recognition are 1, 9, and 8, the first decision made is whether the digit is 1 or 9. This decision has been discussed in the previous subsection. The digit 9 is separated from the digit 1 by the impulsive nature of the two-pole frequency and normalized error after the initial nasal-like section. The digit is identified as 9 if a total jump in two-pole frequency exceeding 200 Hz is obtained within 20 ms of the initial nasal-like section and the two-pole frequency in the interval also exceeds 600 Hz. The identification of the digit 8 is heavily dependent on formant measurements. The criterion employed for recognition of the digit 8 is that F_2 must exceed 1650 Hz in the three measured intervals. Based upon the Peterson and Barney [11] data, this criterion should be easily realized for this particular vowel. Since the digit 8 frequently begins with an unvoiced sound (8 is often pronounced with an initial /h/), the condition $F_2 > 2000$ Hz for all three intervals is applied for the recognition of 8 when the digit begins with an unvoiced sound. To aid in the recognition of 8, a burst detection algorithm, based on energy considerations, is also used.

It can be seen from the above discussion that the individual recognition rules are not simply written down. The sequential nature of the recognition algorithm implies vastly different recognition criteria for the digits depending on the results of earlier tests along the decision tree. The purpose of the above discussion was to provide a rough idea of the major features used to recognize the digits, without getting too involved with the individual tests which are made in the parallel decision mode.

IV. Experimental Evaluation

The entire digit recognition scheme of Fig. 1 was experimentally evaluated in two separate experiments. In one experiment ten speakers recorded a sequence of 100 7-digit telephone numbers read from a randomly generated list of telephone numbers. These recordings were included in the high-quality speech condition discussed in Section II-A. From these data 20 telephone numbers were chosen at random as test data for each of the speakers. (The processing time per utterance was several minutes. This limitation precluded using a much larger data base for evaluating the system.) The results of this experiment are described in Section IV-A.

The second experiment consisted of an evaluation of the system using room-quality recordings. For this experiment ten speakers (four of them were not in the first experiment) recorded ten randomly selected groups of three digits each. (Again the size of the data base was severely limited by the processing time per utterance.) The results of this experiment are discussed in Section IV-B.

Finally, an informal evaluation of the system was made using telephone-quality speech Section IV-C provides a discussion of the additional problems created using telephone-quality speech.

A. Experiment 1

The results of the first experiment are shown in Tables V and VI. Table V gives the error scores for each of the ten speakers (five female, five male). The range of the individual results is from 81.7 percent correct digit recognition, to 93.3 percent. The accumulated recognition scores for female and male speakers were 91.3 percent and 90.7 percent correct digit recognition, respectively. Thus the overall recognition accuracy for the ten speakers was 91 percent.

Table VI shows an analysis of where the digit errors occurred. This table shows the overall confusion matrix for the experiment. The entries which are circled are the cases which

TABLE V
ERROR SCORES FOR EXPERIMENT 1

Women	Correct	Errors	Percent Correct
SAW	56	4	93.3
CAM	53	7	88.3
SP	55	5	91.7
KD	56	4	93.3
BJM	54	6	90.0
Total	274	26	91.3
Men	**Correct**	**Errors**	**Percent Correct**
JLH	56	4	93.3
RWS	55	5	91.7
MRS	56	4	93.3
AER	49	11	81.7
LRR	56	4	93.3
Total	272	28	90.7
Overall	546	54	91.0

TABLE VI
CONFUSION MATRIX FOR EXPERIMENT 1

Experiment 1

		Digit Recognized											
		0	1	2	3	4	5	6	7	8	9	Number of Tries	Number of Errors
	0	36	1							2	2	41	5
	1		31	1							⑦	39	⑧
	2	⑤		59	1	1						66	⑦
	3				68					1		69	1
Digit	4					81	2					83	2
Spoken	5		1				49					50	1
	6							60				60	0
	7					1		1	63		1	66	3
	8		1	1	⑪				1	57	1	72	⑮
	9	1	2	2	2		2			3	42	54	⑫
												600	54

had the highest number of confusions. It is seen from Table VI that the following confusions occurred most frequently.

1) The digit 8 was recognized as the digit 3 eleven times out of 72 tries.
2) The digit 1 was recognized as the digit 9 seven times out of 39 tries.
3) The digit 2 was recognized as the digit 0 five times out of 66 tries.
4) The digit 9 was recognized as either 0, 1, 2, 3, 5, or 8 twelve times out of 54 tries.

These results show that the greatest number of recognition errors occurred when the digits 8, 9, 1, and 2 were spoken. This result is anticipated by the discussion in Sections II and III since these digits (especially in combinations) were the most difficult to segment reliably. The uncertainty as to the exact digit boundary led to higher reliance on the decisions within the digit for accurate recognition results, thereby leading to the results shown in Table VI.

Another source of error in this experiment was due to a low-level type of breath noise that several speakers used preceding initial eights. Because of the high-quality recording method, the analysis classified this initial breath noise as a short interval of unvoiced speech preceding the digit string. When the digit string began with an initial 8 (as occurred a large number of times) the combination of breath noise and eight caused the incorrect recognition of the eight as a three. This accounted for more than 20 percent of the errors in this experiment, as seen in Table VI.

B. Experiment 2

The results of the second experiment in which room-quality recordings were used are shown in Tables VII and VIII. Table VII gives the error scores for each of the ten speakers (five female, five male) in this experiment. The individual recognition accuracy varied from 73.3 percent to 96.7 percent in this experiment. The overall recognition accuracy was 86.7 percent across the ten speakers. For the female speakers the accuracy was 85 percent, whereas for the male speakers it was 88.4 percent.

Table VIII shows the confusion matrix for this experiment. Although the amount of data is quite small, the following confusions seem to occur most frequently.

1) The digit 1 was recognized as the digit 9 five times in 34 tries.
2) The digit 1 was recognized as the digit 4 four times in 34 tries.
3) The digit 2 was recognized as the digit 3 four times in 25 tries.
4) The digit 9 was recognized as either 0, 1, 3, 4, or 5 eight times in 33 tries.

The confusions occurring with the digits 1 and 9 were similar in nature to those which occurred in Experiment 1. The eight confusions for initial 8's were essentially eliminated in this experiment since the background noise level was sufficiently high to classify any breath noise as silence. Thus initial 8's were correctly recognized in most cases.

Although the room noise had some beneficial masking effects (e.g., the 8-3 confusion mentioned above), it can be seen that the higher noise level lowered the recognition accuracy by about 4 percent.

C. Telephone Recognition of Digits

An informal test of the recognition system was made using 3-digit sequences recorded over switched telephone lines. The purpose of the experiment was to see how much the band-limiting effects of the telephone line degraded the system performance using the same system as was used for high-quality and room-quality conditions.

For this informal experiment the voiced-unvoiced analysis used the training data of Table III. Two speakers (JLH and SAW) were used in this test. In spite of the fact that the five analysis parameters were not equally effective (and in some cases not at all effective) in separating silence from unvoiced speech from voiced speech, the analysis gave accurate results for 19 of 20 digit strings. For these 19 digit strings the segmentation program accurately (to within the stated limitations) segmented all 19 strings. The recognition accuracy, however, fell to about 64 percent for these 19 strings. An analysis of the types of errors showed that the effects of the

TABLE VII
ERROR SCORES FOR EXPERIMENT 2

Women	Correct	Errors	Percent Correct
CES	22	8	73.3
CAM	27	3	90.0
IE	27	3	90.0
SAW	27	3	90.0
GH	22	5	81.5
Total	125	22	85.0
Men	**Correct**	**Errors**	**Percent Correct**
REC	21	6	77.8
LRR	28	2	93.3
MRS	27	3	90.0
AER	29	1	96.7
JLH	25	5	83.3
Total	130	17	88.4
Overall	255	39	86.7

TABLE VIII
CONFUSION MATRIX FOR EXPERIMENT 2

Experiment 2

		0	1	2	3	4	5	6	7	8	9	Number of Tries	Number of Errors
	0	25			1	1						27	2
	1		22			④	1			2	⑤	34	⑫
	2			21	④							25	4
	3		1		34					2		37	3
Digit	4	1	1		1	29						32	3
Spoken	5				1	1	18				1	21	3
	6				1			36				37	1
	7					1			24			25	1
	8		1					1		21		23	2
	9	1	2		2	2	1				25	33	⑧
												294	39

band-limiting were particularly severe in the recognition stages which relied heavily on zero crossings and the two-pole analysis to classify the digits.

Although the general structure of the recognition stage could still be used on the telephone speech, the individual tests and decisions had to be greatly modified to account for the loss of information in the recognition parameters. Work on this aspect of the system is still in progress.

V. SUMMARY AND CONCLUSIONS

A system has been described for recognizing connected digits. The system is speaker independent and therefore can be used on speakers who have had no prior training in the use of the system. An evaluation of the recognition accuracy using strings of three connected digits showed scores of 91 percent correct recognition for high-quality speech, and 87 percent correct recognition for room-quality speech.

These recognition accuracies can be compared against similar scores obtained in other digit recognition schemes. Using isolated digits in a speaker independent scheme, Sambur and Rabiner [1] obtained digit recognition accuracies of 97 percent and 94 percent for high-quality and room-quality speech, respectively. The 6 percent degradation in the recognition accuracy scores is due primarily to the coarticulation effects which are present in connected digit strings which both makes it difficult to accurately locate the digit boundary, and makes the digit cues in the region of the digit boundaries somewhat unreliable for accurate recognition.

Another comparison which can be made is with the results reported by Martin [3]. Using 10 500 digits, Martin reported recognition accuracies of from 86.8 percent to 92.4 percent on strings of three digits. Although Martin used an order of magnitude more data than in the present study and, therefore, his recognition accuracies are statistically more reliable than those reported on here the results appear to be quite comparable. Additionally, the types of errors made in both studies are of a similar nature.

In summary, the digit recognition system discussed in this paper shows considerable promise for applications where recognition of connected digits is required. More sophistication in the digit recognition rules should help to raise the recognition accuracy. Additional flexibility is required to apply the system to telephone-quality speech due to the degradations caused by the band-limiting nature of the telephone line.

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, "Practical applications of voice input to machines," *Proc. IEEE*, vol. 64, pp. 481-501, Apr. 1976.

[3] ——, "Acoustic recognition of a limited vocabulary in continuous speech," Ph.D. dissertation, Univ. Pennsylvania, Philadelphia, 1970.

[4] T. G. Von Keller, "On-line recognition system for spoken digits," *J. Acoust. Soc. Amer.*, vol. 49, pp. 1288-1296, Apr. 1971.

[5] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," *J. Acoust. Soc. Amer.*, vol. 24, pp. 637-642, Nov. 1952.

[6] P. B. Denes and M. V. Mathews, "Spoken digit recognition using time-frequency pattern matching," *J. Acoust. Soc. Amer.*, vol. 32, pp. 1450-1455, Nov. 1960.

[7] J. Suzuki and K. Nakata, "Recognition of spoken digits," *J. Inst. Elec. Comm. Eng.* (Japan), vol. 45, pp. 303-309, Mar. 1962.

[8] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Syst. Tech. J.*, vol. 54, pp. 297-315, Feb. 1975.

[9] 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*, to be published, June 1976.

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

[11] G. E. Peterson and H. L. Barney, "Control methods used in a study of the vowels," *J. Acoust. Soc. Amer.*, vol. 24, pp. 175-184, Mar. 1952.