# SPEAKER INDEPENDENT RECOGNITION OF CONNECTED DIGITS

L. R. Rabiner and M. R. Sambur
Bell Laboratories
Murray Hill, New Jersey 07974

## 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 inter-related 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. 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%, and the recognition accuracy was about 91% across 10 speakers (5 male, 5 female). With recordings made in a noisy computer room the segmentation accuracy remained close to 99%, and the recognition accuracy was about 87% across another group of 10 speakers (5 male, 5 female).

## I. OVERALL RECOGNITION SYSTEM

Figure 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 speech waveform, as described by Rabiner and Sambur.[1] Following endpoint alignment, the speech signal is analyzed 100 times per second, giving the following parameters:

1. Zero crossings
2. Log energy
3. 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.[2] The segmentation of the digits is based on the voiced-unvoiced analysis, and also uses information about the location and amplitude or minima in the energy contour to aid in locating the digit boundaries. 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 without affecting any of the results to be presented. However, it is required that the number of digits in the string be specified.

## II. SEGMENTATION

The structure of the basic segmentation algorithm is illustrated in Fig. 2. For the segmentation of the digits, we take advantage of the fact that there are no internal unvoiced or silence regions within the 10

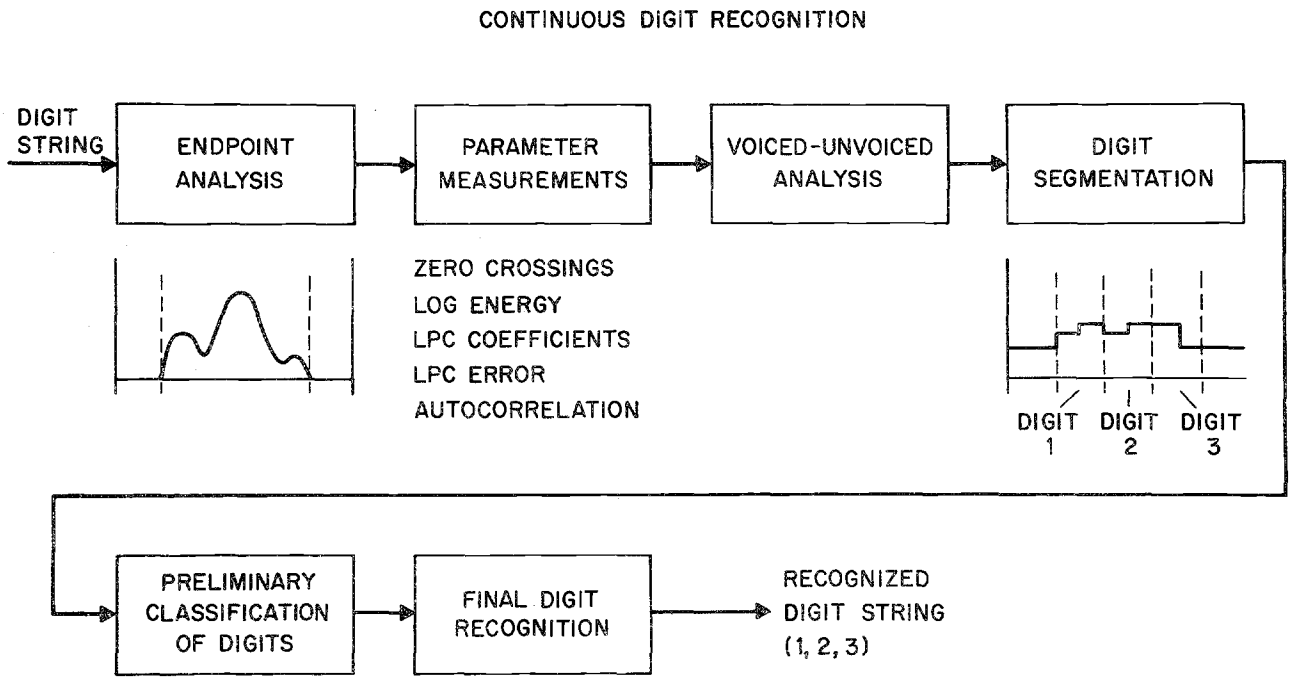


*Figure 1: Block diagram of the overall digit recognition system.*

digits. In addition, with a couple of exceptions, the energy contours of the digits have no internal local minima. Thus the occurrence of a local minima in the energy contour is usually a strong indication of a digit boundary.

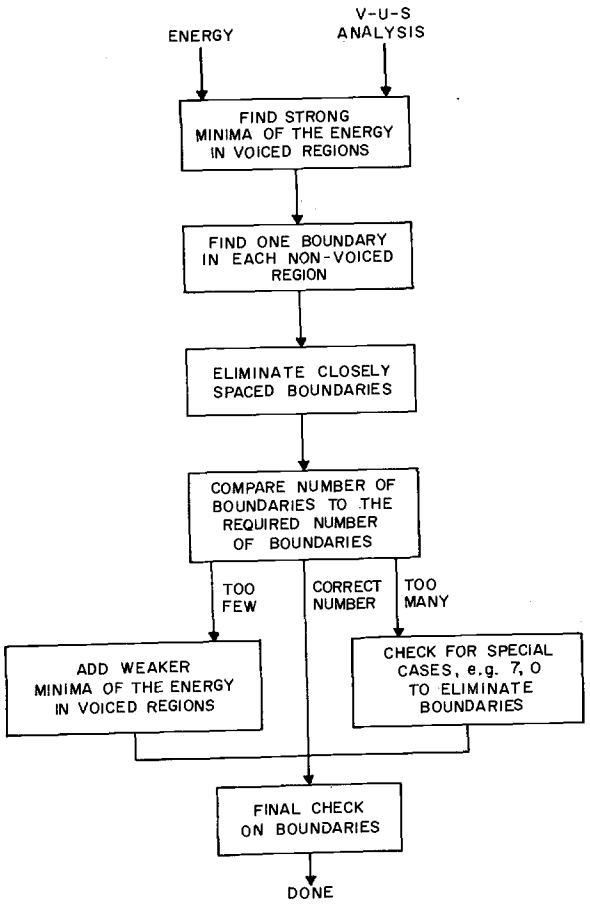


*Figure 2: Flow diagram for the digit segmentation algorithm.*

Figure 3 illustrates the operation of the segmentation rules 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. Although the exact 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 this figure 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 on the basis of durational considerations and instead choose the minimum in the second voiced region.

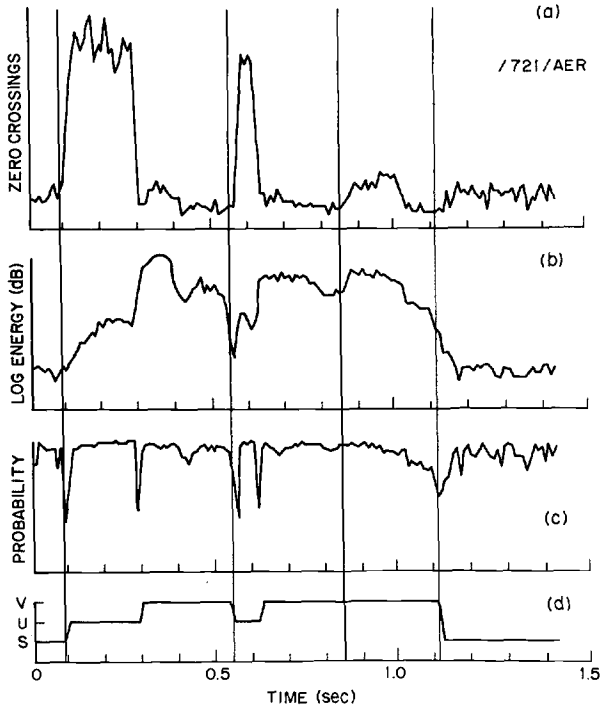


*Figure 3: Typical measurement contours and the resulting boundary locations for the utterance /721/.*

## III. RECOGNITION

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. An example of this procedure is given in Fig. 4. The digit string in this example is 650. For this case three unvoiced regions were found and in accordance with the algorithm, the boundaries were located at the beginning of each region. However, the second boundary should actually be within the unvoiced region - not at the beginning. 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.

The final stage in the method is the digit recognition algorithm which is shown in Fig. 5. This algorithm is similar in philosophy but greatly different in details of implementation from the isolated digit recognition algorithm of Sambur and Rabiner.[3] The recognition algorithm is basically a tree search method in which the sequence of branches was designed to resolve the most obvious designs and then proceed to the more difficult decisions. The differences between this and the isolated digit recognition scheme are due primarily to the coarticulation

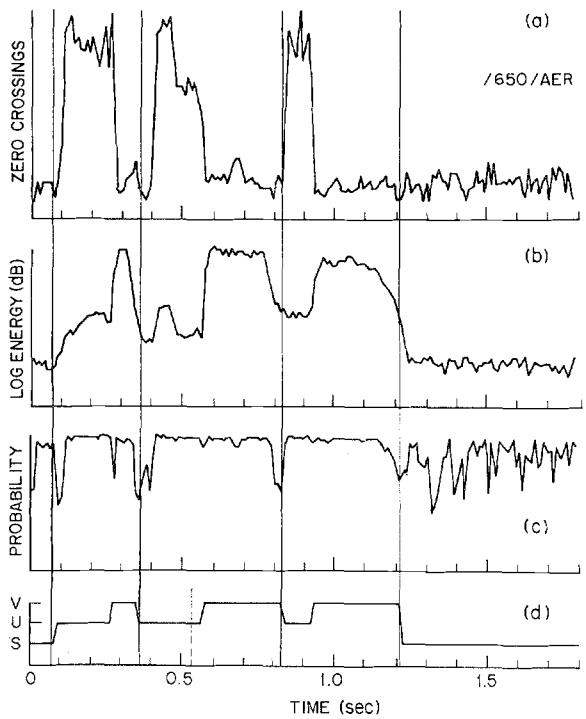


*Figure 4: Typical measurement contours and the resulting boundary locations for the utterance /650/.*

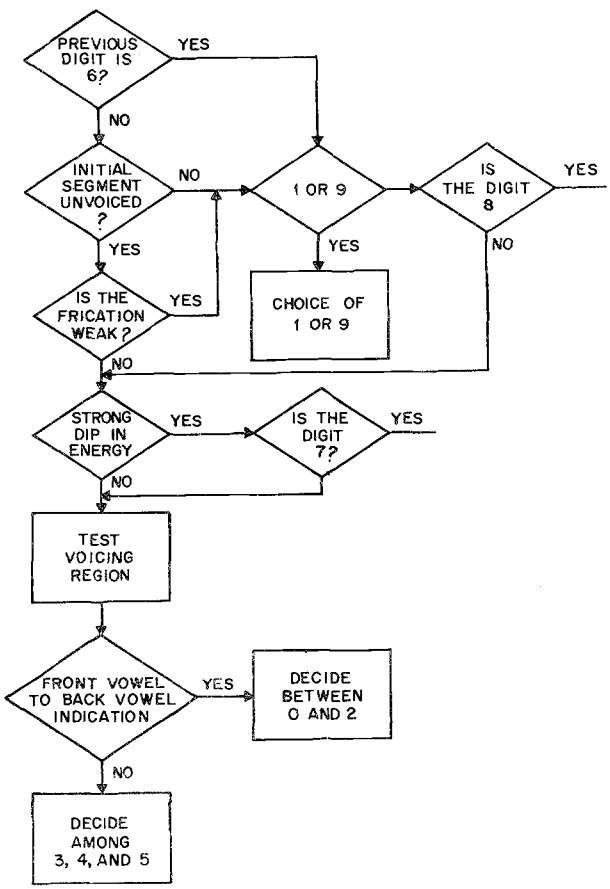


*Figure 5: The decision tree for the digit recognition algorithm.*

effects which are present with connected digits, but which are absent for isolated digits. In addition there is a great deal less preciseness in the location of the digit boundaries for connected digits than there was for isolated digits. Thus the final digit recognition algorithm is considerably more difficult than for isolated digits.

The fundamental aspect of the recognition procedure involves the concept of "self normalization" in which the decision process classifies sounds according to the transitional natura of the various measurements. For example, in the recognition algorithm we made use of the fact that the relative magnitude of the normalized error generally increases from sonorants to vowels and then to fricatives. 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 both the 2 pole LPC frequency and normalized error, important information about the structure of the voiced region of the word can be obtained. As an example Fig. 6 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.

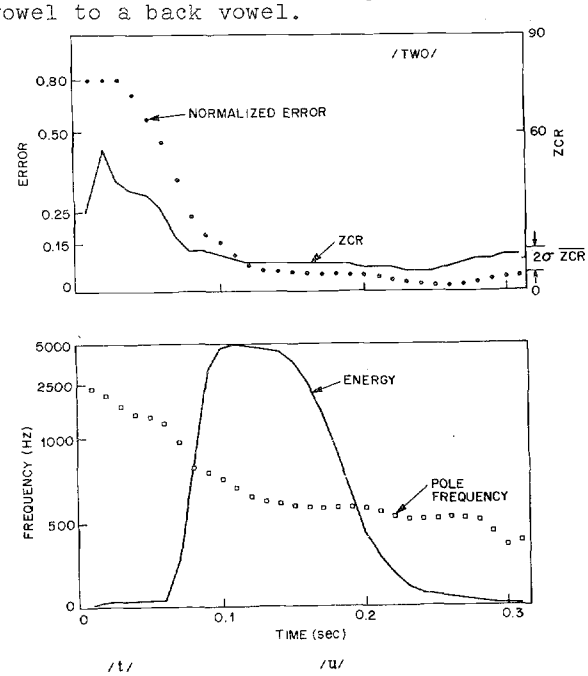


*Figure 6: Complete set of analysis measurements for the digit two.*

## IV. EVALUATION TESTS

The entire digit recognition scheme was experimentally evaluated in two separate experiments. In one experiment 10 speakers recorded a sequence of 100 seven digit telephone numbers read from a randomly generated list of telephone numbers in a low noise environment. From these data 20 telephone numbers were chosen at random as test data for each of the speakers. The results of this experiment are shown in Fig. 7. The overall recognition

accuracy was 91% across the 10 speakers. The worst digit confusions were found to occur between 3 and 8 and 1 and 9.

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 |
| 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 |

*Figure 7: Error Scores for Experiment #1.*

The second experiment consisted of an evaluation of the system using recordings made in a noisy computer room. For this experiment 10 speakers recorded 10 randomly selected groups of 3 digits each. The results of this experiment are shown in Fig. 8. The overall recognition accuracy was 87% for this recording environment.

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 |

*Figure 8: Error Scores for Experiment #2.*

## V. SUMMARY

In summary, the digit recognition system discussed in this paper shows considerable promise for applications where recognition of connected digits is required. Further experimentation is necessary to make the digit recognition rules sophisticated enough to reliably recognize the digits of all speakers.

## REFERENCES


1. L. R. Rabiner and M. R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances," Bell Syst. Tech. J., Vol. 54, No. 2, pp. 297-315, February 1975.

2. B. S. Atal and L. R. Rabiner, "A Pattern-Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition," submitted to IEEE Trans. on Acoustics, Speech, and Signal Processing.

3. M. R. Sambur and L. R. Rabiner, "A Speaker-Independent Digit-Recognition System," Bell Syst. Tech. J., Vol. 54, No. 1, pp. 81-102, January 1975.