THE BELL SYSTEM TECHNICAL JOURNAL
Vol. 54, No. 1, January 1975
Printed in U.S.A.

A Speaker-Independent Digit-Recognition System

By M. R. SAMBUR and L. R. RABINER

(Manuscript received July 9, 1974)

This paper describes an implementation of a speaker-independent digit-recognition system. The digit classification scheme is based on segmenting the unknown word into three regions and then making categorical judgments as to which of six broad acoustic classes each segment falls into. The measurements made on the speech waveform include energy, zero crossings, two-pole linear predictive coding analysis, and normalized error of the linear predictive coding analysis. A formal evaluation of the systems showed an error rate of 2.7 percent for a carefully controlled recording environment and a 5.6 percent error rate for on-line recordings in a noisy computer room.

I. INTRODUCTION

With the widespread growth in the use of digital computers, there has been an increasing need for man to be able to communicate with machines in a manner more naturally suited to humans. The realization of this need has motivated a great deal of research in automatic recognition of speech by computer.[1–3] Although only a moderate degree of success has been obtained in solving the problems associated with machine recognition of continuous speech,[4] a greater degree of success has been obtained in recognition of isolated words from a fixed vocabulary. The performance of these systems range from about 92 percent correct decisions for 561 isolated words by an individual for which the system has been carefully trained[5] to nearly error-free performance for the recognition of a limited vocabulary (e.g., the digits) also spoken by a speaker for which the system has been trained.[6] However, performance of many of these word-recognition algorithms is radically degraded when the system has not been tuned to the speech characteristics of the individual user. The subject of this paper is an isolated-word, digit-recognition system that achieves high accuracy without having to be trained every time a different speaker wishes to use the system.

The development of a speaker-independent limited-vocabulary word recognizer is inherently more difficult than a speaker-adaptive system that can use comparatively simple pattern-matching algorithms to recognize the input words. It has been argued that the extended effort needed to design a speaker-independent system is unnecessary in view of the relative ease of training an adaptive scheme to learn to recognize the speech of a new user. There are two major reasons why such arguments are invalid. For small vocabulary systems (e.g., digit recognizers) with a *large number* of potential users, it is not feasible to store training data for every possible user. Furthermore, most systems cannot train themselves on new speakers very rapidly. Thus, the turnaround time for new users is often a major factor limiting the use of speaker-dependent systems. For a large vocabulary (250 words), the time required for a new speaker to form reference patterns for all the words in the vocabulary can be prohibitive. In addition, the variation with time of a speaker's voice characteristics may necessitate frequent updating of his reference patterns. Finally, the design of a speaker-adaptive word-recognition algorithm is so dependent on the uniqueness of each talker that very little insight is gained in the actual problem of recognizing speech. On the other hand, it is hoped that the development of a speaker-independent scheme will contribute to an understanding of the acoustic attributes of speech that reliably distinguish the various sounds. Without such an understanding, it would be difficult to duplicate the human capacity of recognizing the speech of a wide variety of speakers.

This paper discusses a speaker-independent digit-recognition system that was implemented on the computer facility of the acoustics research department at Bell Laboratories. Section II discusses the basic speech parameters that are measured and shows how the digits can be classified from these features in a speaker-independent manner. This section includes a discussion of the various signal-processing techniques that are heavily relied on throughout the classification process. Section III discusses the digit-classification scheme. The classification procedure is a tree-like decision algorithm for which backwards tracing is allowed when one of the parallel-decision algorithms indicates a high probability of error. Section IV gives the results of a formal evaluation of the recognition system. Finally, the paper concludes with a discussion of the strong and weak points of the system and suggestions for how it can be improved.

II. FRAMEWORK OF THE RECOGNITION SYSTEM

Figure 1 is a block diagram of the overall digit recognition system that was implemented. Following endpoint alignment in which the

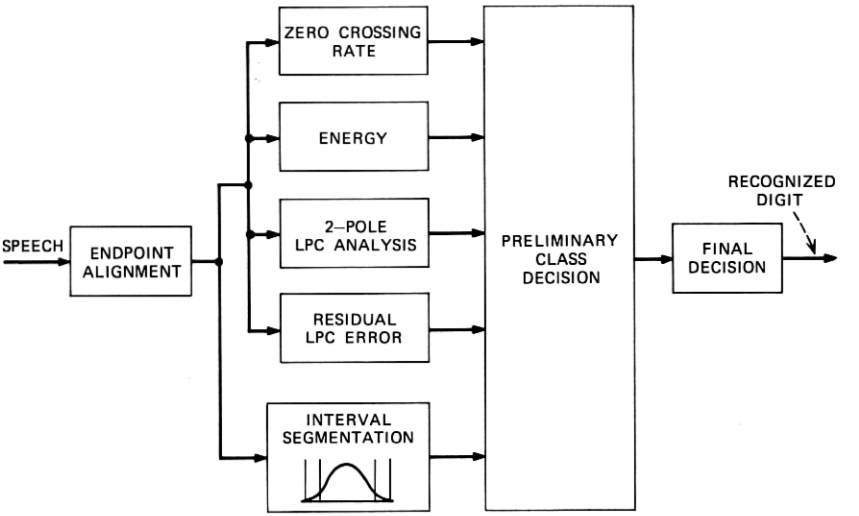

Fig. 1—Block diagram of the overall digit recognition system.

interval containing the word to be recognized is carefully determined, the speech is analyzed every 10 ms to obtain zero-crossing rate, energy, two-pole model linear-predictive-coding (LPC) coefficients, and the residual LPC estimation error. To aid in making preliminary classification decisions, the speech interval is segmented into three well-defined regions. All the speech information is fed in parallel into a preliminary decision-making algorithm that chooses one of several possible digit classes for the input utterance—e.g., one class contains the digits 1 and 9. A final decision is then made based on the presence or absence of certain key features in the input speech.

In this section, we show how the various digits can be characterized in terms of certain acoustic features. Then we discuss some key signal-processing functions that are heavily relied on in the decision algorithms and that contribute strongly to making the system speaker-independent.

2.1 Characterization of the digits

The elemental speech units (phonemes) that comprise English words can be classified into two broad categories, vowels and consonants. The vowels can be further classified into front (/i/, /I/, /e/, /ɛ/, and /ae/), middle (/ɜ/, /ʌ/), and back vowels (/u/, /U/, /ɔ/, and /o/). It is also convenient to subdivide the consonants into the categories noise-like (fricatives, plosives) and vowel-like (nasals, glides). Table I gives a list of the sequence of phoneme categories for each of the ten

Table I — Sound classes characteristic of the digits (from Ref. 6)

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.

digits, 0 through 9.[6] Our approach toward speaker-independent recognition of the digits is to use a set of robust measurements to classify the phonemes into the six broad categories listed in Table I. By robust measurements, we mean acoustic parameters that give a general indication of the gross nature of each phoneme without being too dependent on the speaker's voice characteristics. Through a combination of parallel processing and self-normalization, the phoneme categories are determined and the spoken digit is recognized. We now discuss the criteria for the selection of the robust measurements that are used, the technique of self-normalization of measurements, and finally the method of parallel-processing of the data to give a speaker independent classification of the digits.

2.2 Robust measurements for digit recognition

The requirements for a recognition parameter to be selected as being a robust measurement are:

(*i*) The parameter can be simply and unambiguously measured.
(*ii*) The parameter can be used to grossly characterize a large proportion of speech sounds.
(*iii*) The parameter can be conveniently interpreted in a speaker-independent manner.

Based on the above requirements, the zero-crossing-rate and spectral-energy parameters are excellent candidates for robust measurements. These parameters can be used to effectively characterize the general acoustic properties of the sound categories listed in Table I. For example, noise-like sounds have a relatively high zero-crossing rate,

relatively low energy, and a relatively high concentration of high-frequency energy. Thus, the noise-like sounds of any speaker can be characterized quite accurately based on these measurements. The term "relatively," in the above classification of noise-like sounds, can be conveniently interpreted for a given speaker by a simple self-normalization technique discussed later in this paper

To measure the distribution of spectral energy, a two-pole LPC analysis has been suggested by Makhoul and Wolf[7] as an excellent means of representing the gross features of the spectrum. Figure 2 (from Makhoul and Wolf[7]) shows the results of applying a two-pole model to a variety of speech sounds. This figure is a comparison of the spectra of several speech sounds obtained directly from FFT spectrum measurements compared with the spectra of the best two-pole LPC fit to the spectrum. For a two-pole LPC analysis, there is either one complex-conjugate pole or two real poles. In Fig. 2a, the spectra for the sound /sh/ as in the word "short" are plotted. For this example, the two-pole LPC analysis gives a complex conjugate pole at about 3000

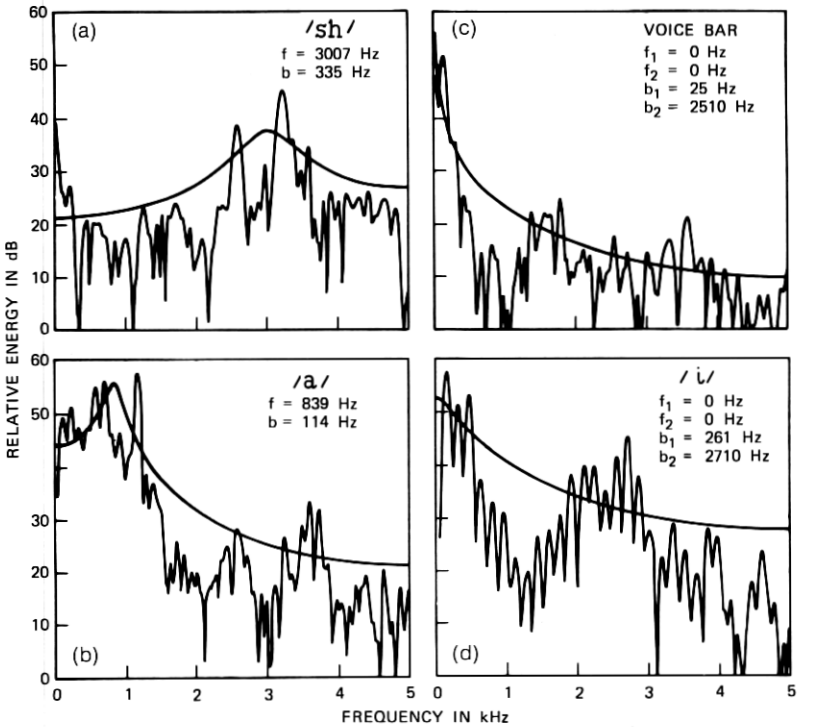

Fig. 2—Comparison of FFT spectra and two-pole LPC spectra for several speech sounds.

Hz—i.e., the region of maximum energy concentration in the spectrum. In Fig. 2b, similar results are shown for the vowel /a/ where the major concentration of energy in the spectrum is around 800 Hz. In the examples of Figs. 2c and 2d (a voice bar and the vowel /i/), the major concentration of spectral energy is around 0 Hz; thus, the two-pole LPC analysis gives two real poles in the right-half z-plane. From Fig. 2, it can be seen that the computed pole frequency gives a good indication of the location of the dominant portion of the spectral energy of the sound and can thus be effectively used to characterize sounds with relatively high-frequency or low-frequency concentrations of energy. For example, noise-like sounds are characterized by a relatively high-frequency spectral concentration of energy, while nasals and vowels generally have a much lower frequency for the energy concentration.

Figure 3 (also from Makhoul and Wolf[7]) illustrates the dynamic behavior of the computed pole frequency of the two-pole model and the corresponding spectrogram of the utterance, "Has anyone measured nickel concentrations · · · ." Examination of Figs. 2 and 3 shows that, for vowel-like sounds, the computed pole frequency is invariably situated somewhere between the first and second formants. In general, when F_1 and F_2 are not too far apart and have comparable amplitudes, the pole frequency falls almost midway between the two resonances. Since F_1 does not usually have as much dynamic movement as F_2, the computed pole frequency tends to follow the motion of the second formant. Since the motion of F_2 is quite important in the characteriza-

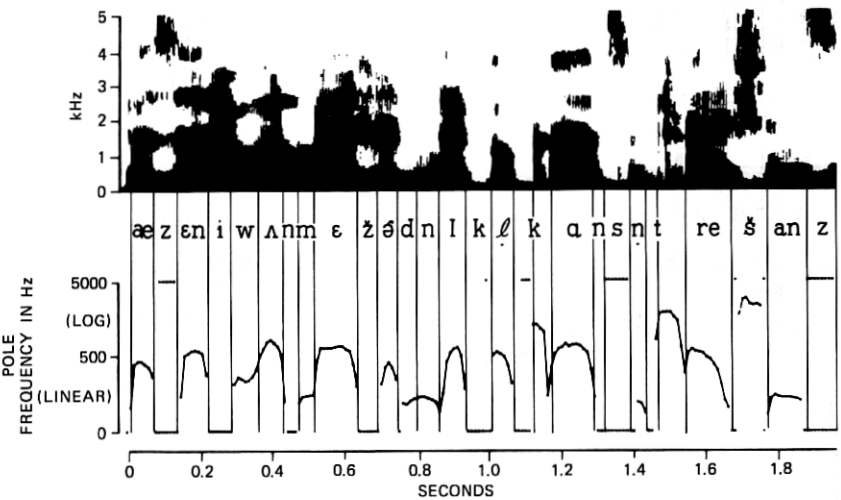

Fig. 3—Spectrogram and computed pole-frequency of two-pole LPC model for the utterance, "Has anyone measured nickel concentrations · · · ."

tion of the digits, the ability of the two-pole LPC analysis to track this motion has been used in the classification phase of the digit-recognition system.

It should be noted, however, that when either F_1 and F_2 are sufficiently far apart or the amplitude of F_1 is significantly greater than the amplitude of F_2, the pole frequency will either follow F_1 or result in two positive real poles. Figure 3 shows that, during the /i/ in the word "anyone," the pole frequency begins to dip sharply as the separation between F_1 and F_2 grows greater and finally results in positive real axis poles as the separation reaches some critical threshold. For nasal sounds, the energy is so highly concentrated near the first resonance that the two-pole model usually results in positive real axis poles, as seen in Fig. 3.

In addition to using the computed pole frequency as a measure of the location of dominant spectral energy and a characterization of the dynamic movement of F_2, the normalized error of the two-pole model contains important information about the spread of spectral energy. The normalized or residual error is defined as

$$V = 1 - a_1 r_1 - a_2 r_2,$$

where a_1 and a_2 are the two-pole LPC coefficients and r_1 and r_2 are the normalized autocorrelation coefficients. It can be shown that the more concentrated the energy spectrum, the lower the normalized error.[8] For speech sounds, 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 the pole frequency and normalized error, important information about the structure of the voiced region of the word can be obtained. An example of the usefulness of the pole parameter in specifying the speech sounds comprising the digits is given in the next section.

In summary, a reasonable set of robust measurements that have been implemented for this digit recognition algorithm is as follows:

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

2.3 Self-normalization of parameters

Almost all classification algorithms use some set of threshold levels in the decision process. Using a fixed set of thresholds leads to a large number of problems for speech recognition in that many of the thresholds are speaker- or time-dependent. To eliminate this difficulty, the technique of self-normalization of parameters was used in which many of the most significant thresholds were obtained from measurements made directly on the speech sample being recognized. Thus, for example, in the case of setting thresholds on zero-crossing rate to determine whether a sound is noise-like or nasal, a statistical description of the zero-crossing rate (ZCR) was made for the entire utterance. 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. Figure 4 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.

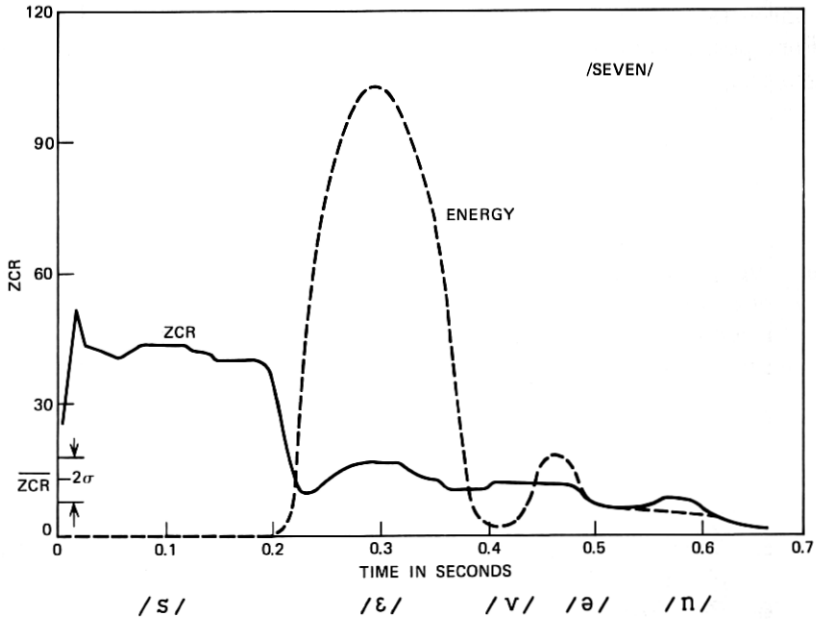

Fig. 4—Energy and ZCR for one example of the word "seven."

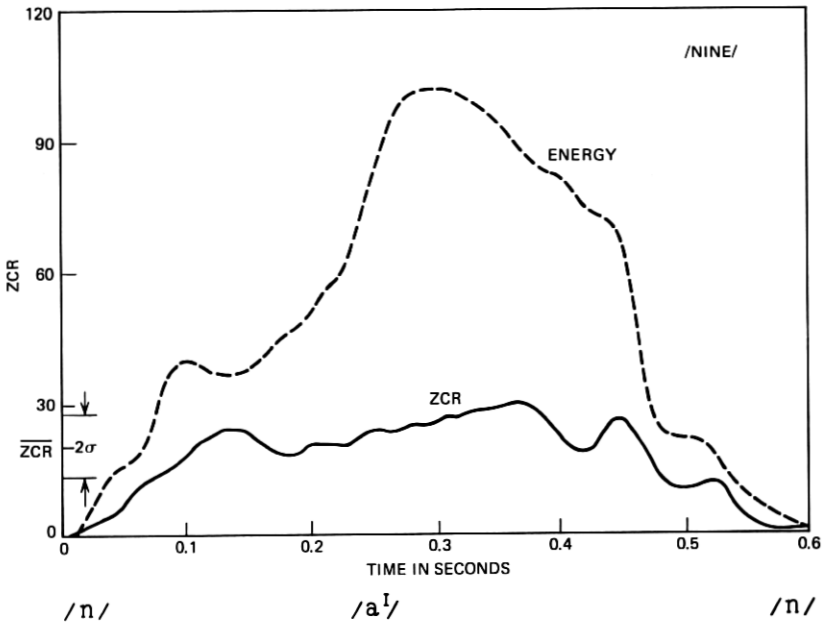

Fig. 5—Energy and ZCR for one example of the word "nine."

In much the same manner, a self-normalized ZCR threshold can be set for classifying a segment as a nasal-like sound. In such cases, the ZCR generally falls below a level one standard deviation *below* the average ZCR for the utterance. As an example, Fig. 5 shows the measured ZCR for the word "nine" and gives the spread of ZCR around the average value. For the initial and final nasals, the ZCR is much lower than the average and thus is a good indication of the nasals.

The idea of determining thresholds based on measurements made during the course of the utterance being recognized can be used for any or all measurements described in the preceding section. For example, Fig. 6 shows the two-pole model normalized error and the pole frequency for the word "nine." The nasal sections are clearly characterized by low normalized error and a zero-Hertz pole frequency. In contrast, Fig. 7 shows the same measurements for the word "six." Again, the noise-like sections are clearly depicted by the relatively high values of normalized error, pole frequency, and ZCR.

The transitional nature of the normalized error and pole frequency can be used to classify the vowels into types high, middle, and back. Figure 8 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

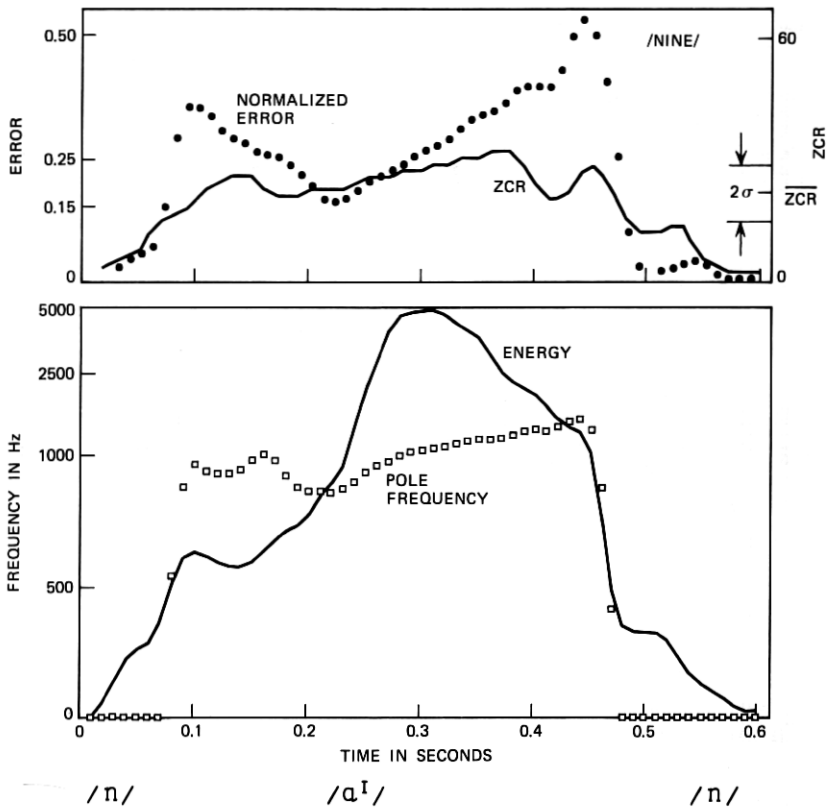

Fig. 6—Complete set of measurements for one example of the word "nine."

decreases. The decrease in the normalized error is due to the fact that the vowel nature changes from front (because of the /t/) to back. Thus, without specifying any absolute thresholds, the constituent structure of the voiced section of the word can be obtained by noting the relative changes in the normalized error. As described earlier, changes in the pole frequency can also be used to indicate the constituent vocalic structure. As seen in Fig. 8, the pole frequency is continually decreasing throughout the voiced region in the word "two," thereby indicating a continually decreasing second formant. As another example, Fig. 9 shows the parameters for the word "four." The gradually increasing normalized error and pole frequency are indicative of a progression from a back vowel to a middle vowel.

2.4 Parallel processing

Using the self-normalization technique, each robust measurement can, by itself, classify a speech sound into one of the six broad categories

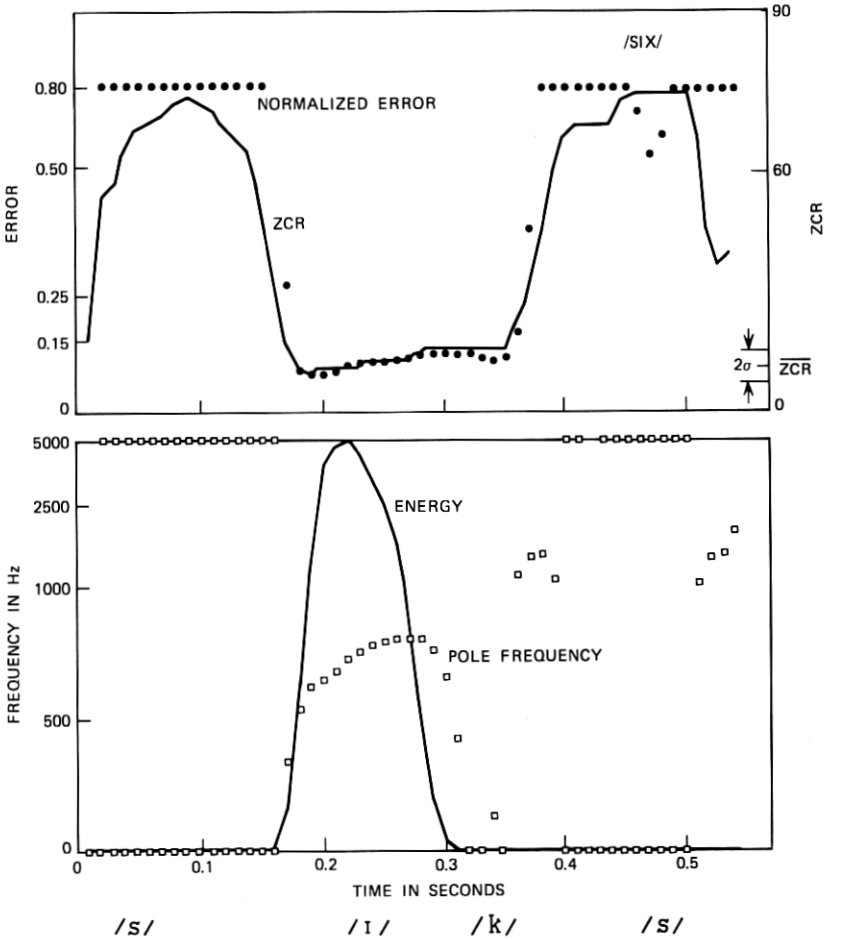

Fig. 7—Complete set of measurements for one example of the word "six."

of interest. Unfortunately, the classification will not be error-free; but if the results of all the measurements are "intelligently" pooled together, then the classification performance can be significantly enhanced. The operation of combining the measurements is termed parallel processing. Parallel-processing ideas have met with good success in other areas of speech processing.[9]

The idea of parallel processing as it is used here not only involves a suitable combination of the robust measurements but also the incorporation of certain structural constraints of the lexicon as additional input. For example, as seen in Fig. 9, the initial section of the word shows only a slight indication of the initial fricative /f/ (i.e., the high zero crossing and normalized error for the initial 10 to 20 ms of the

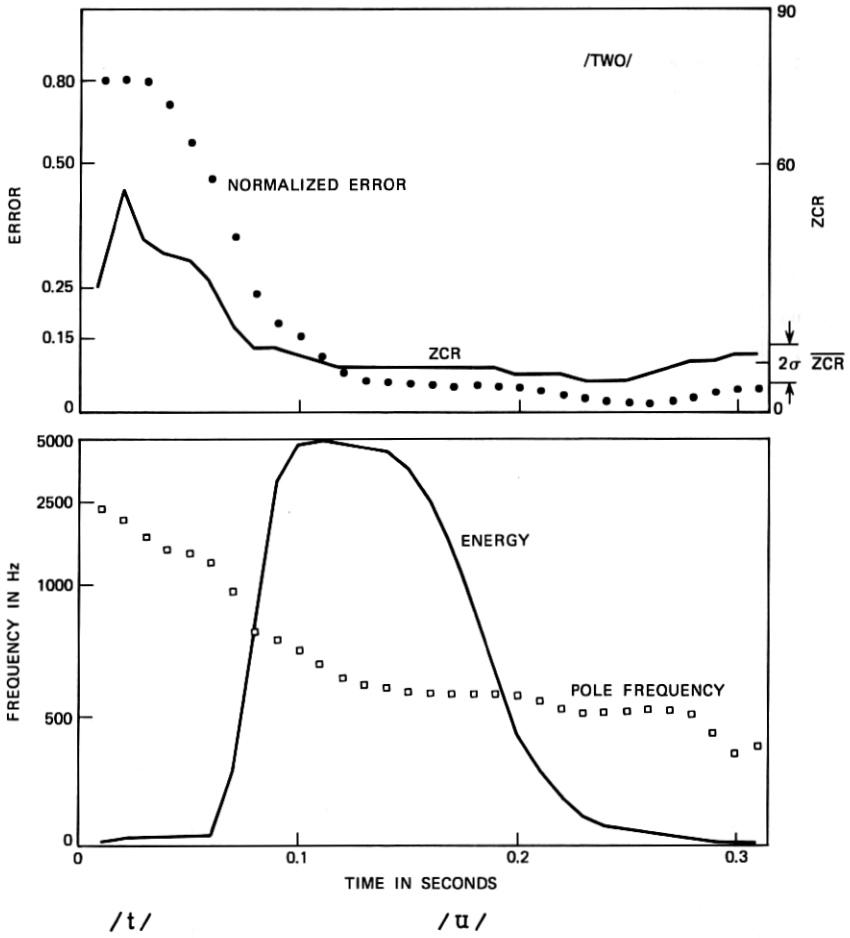

Fig. 8—Complete set of measurements for one example of the word "two."

word), and one might conclude that there is no frication. However, for the digits, it is known that only 1, 9, and 8 do not normally begin with frication. Since the ZCR, normalized error, and pole frequency are relatively too high for the digits 1 or 9, these digits can be safely omitted from consideration. In addition, a combination of the facts that the normalized error is low and increasing and that the pole frequency is increasing indicates that the voiced region in the word is more than likely composed of a back-type voiced sound followed by a middle-type voiced sound. Since the voiced section of the word "eight" is a front vowel sound, the odds are quite high that the word is not "eight." Additional evidence that the word is probably not "eight" can be

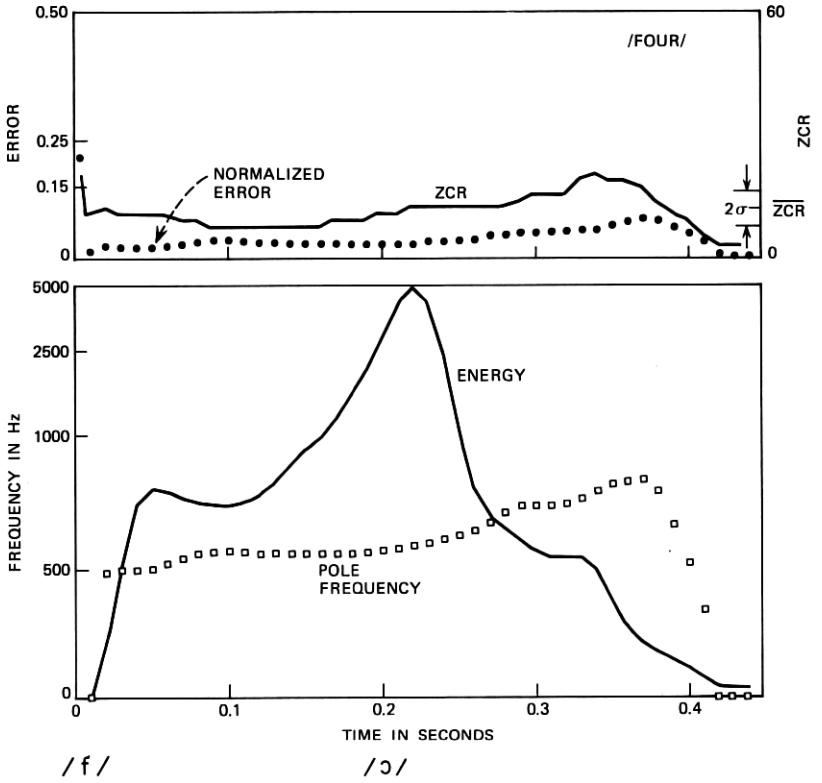

Fig. 9—Complete set of measurements for one example of the word "four."

obtained from the lack of a burst at the end of the spoken word. By pooling knowledge from the individual measurements with information about the structure of the words in the lexicon, the weak frication in the spoken "four" can be recognized. Thus, the major feature of parallel decisions is the ability to arrive at a correct decision even if one or more of the parallel inputs is in error. In the next section, we discuss the organization of the digit recognizer and the specific nature of the logic rules.

III. DIGIT RECOGNIZER

As seen from Fig. 1, the first step in the recognition scheme is the important problem of endpoint alignment (determining the location of the spoken word during the recording interval). The algorithm used in this scheme has been described by Rabiner and Sambur[10] and has been shown to give reasonably good results over a wide variety of speakers and background levels. However, the algorithm sometimes

has trouble finding the end of the word when the speaker sighs or puffs after reciting the word. To compensate for this problem, the decision algorithm does not place too much dependence on the end region of the word. The "end region" is defined as the region from the end of the word to the point at which the energy first exceeds 10 percent of the maximum energy. Equivalently, an "initial region" is defined from the beginning of the word to the point at which the energy first exceeds 10 percent of the maximum. The remaining section is termed the "middle region." The process of determining the three regions of the word is labeled "interval segmentation" in Fig. 1.

Throughout the duration of the detected word, the four robust parameters discussed in Section 2.2 are measured once every 10 ms (i.e., every 100 points for a 10-kHz sampling rate) and smoothed using a nonlinear smoothing algorithm proposed by Tukey.[11] In addition, the first two formant frequencies are computed using a 12-pole LPC analysis at three points during the middle region. These include the point of maximum energy, the beginning of the middle region, and the end of the middle region. Since formant frequencies are quite speaker-dependent, they were used in the decision process only as a *supporting* measurement to discriminate between sounds that were quite dissimilar when viewed in the $F_1 - F_2$ plane (/i/ and /a/ are examples). The supporting nature of the formant measurement is also necessitated by the fact that the extraction of formants is not a simple and unambiguous task, and too great a reliance on these parameters is fraught with danger.

Following the measurement phase, a preliminary class decision is made for the utterance. An expanded view of the preliminary decision box is given in Fig. 10. The decision algorithm is in the form of a tree structure that traces down the most probable branch to arrive at the decided digit. However, it should be noted that there are provisions in the algorithm for back-tracking if some measurement strongly suggests that an error has been made.

The first branch in the tree is to decide whether or not the initial portion is nasal-like. As we discussed previously, this decision is based upon the fact that nasal-like sounds have relatively low ZCR, low normalized error, and low pole frequency. If a nasal-like beginning is detected, the preliminary choice is between 1 and 9. As a further check on this preliminary 1, 9 decision, the ending region is checked for nasal-like characteristics. If there is no evidence of initial nasal-like sounds, the digits 1 and 9 are removed from further consideration. When the initial region is deemed nasal-like, a relatively simple decision can be used to decide between 1 or 9. The digit 9 can be distin-

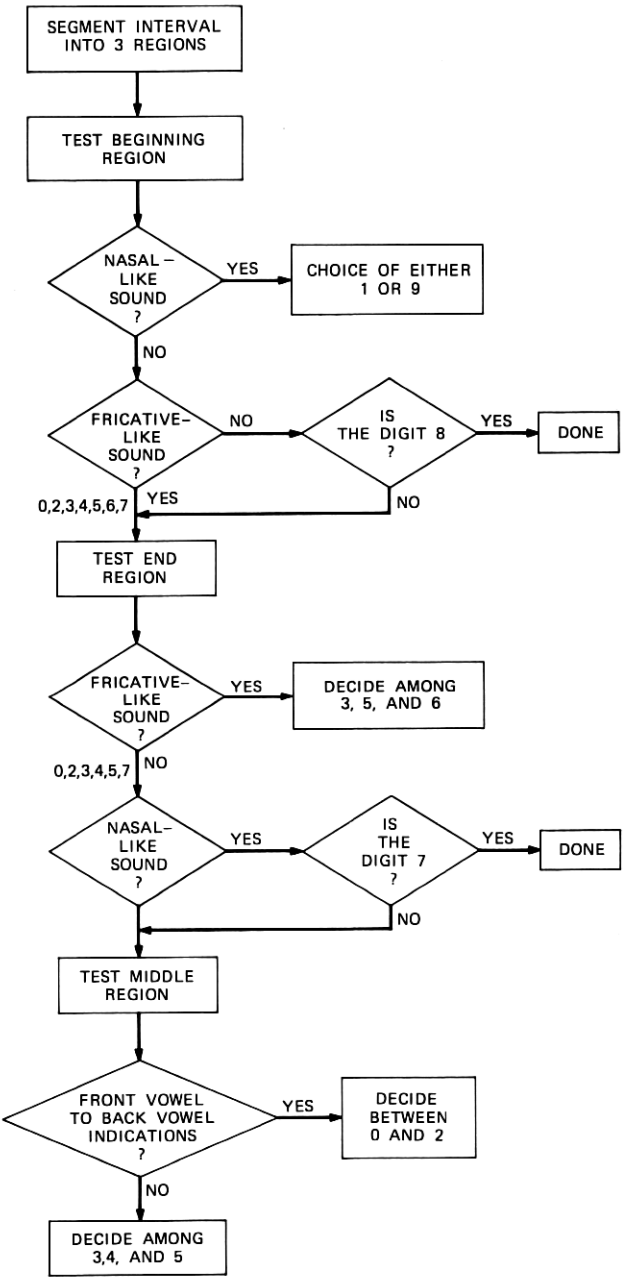

Fig. 10—Preliminary decision tree for digit classification algorithm.

guished by a sharp discontinuity in pole frequency (between the initial nasal and the vowel) and high normalized error during the transition from the nasal /n/ to the diphthong /aI/ (see, for example, Fig. 5). On the other hand, as shown in Fig. 11, there is no discontinuity at the end of the nasal-like section in 1.

Tracing along the decision tree, the next choice is to decide whether there is any definite indication of initial frication as shown by significantly high ZCR, normalized error, and pole frequency. The positive detection of frication eliminates the choice of the word "eight." If there is no definite frication, the end region is checked for a burst and the middle region is checked for front vowel-like characteristics. The formant parameters are also used to check if the middle region is composed of only front vowels. A suitable combination of the results of these tests is used to reject or accept the word "eight" as the digit. It should be noted that the decision process is in the form of a hypothesis test. In other words, we assume that the spoken word is "eight" and check to see if the acoustic parameters are consistent with this hypothesis. In fact, the basic structure of the entire digit recognizer is to first hypothesize and then test the acoustic consequences of this hypothesis. The parallel processing aspect of the decision assigns the appropriate weight to a particular test. For example, the detection of a burst at this point in the decision tree is an almost 100-percent indication that the word is "eight." However, the lack of a burst does not necessarily preclude the possibility of "eight," and this result should be weighed accordingly.

Assuming that we reject the spoken word as the digit 8, the remaining possibilities are 0, 2, 3, 4, 5, 6, and 7. The end region is then checked for fricative-like behavior. If frication is indicated, a hypothesis test on the digit 6 is made. The middle region is checked for front vowel characteristics, and the only timing measurement in the entire digit recognition program is performed. This measurement compares the relative duration of the initial frication plus ending frication to the length of the middle region. The frication duration is defined from the beginning (or ending) of the word until the point at which the ZCR remains within one standard deviation of the average for three time frames. This definition can be modified when any abrupt discontinuities in normalized error or two-pole frequency indicate a more probable location for frication. In addition, the extent of frication is not allowed to go beyond the 10-percent maximum energy points that form the boundaries of the middle region. For the digit 6, the timing ratio should be less than one and the middle region should be less than 250 ms. A combination of the results of the hypothesis test are used to verify that the spoken word was "six."

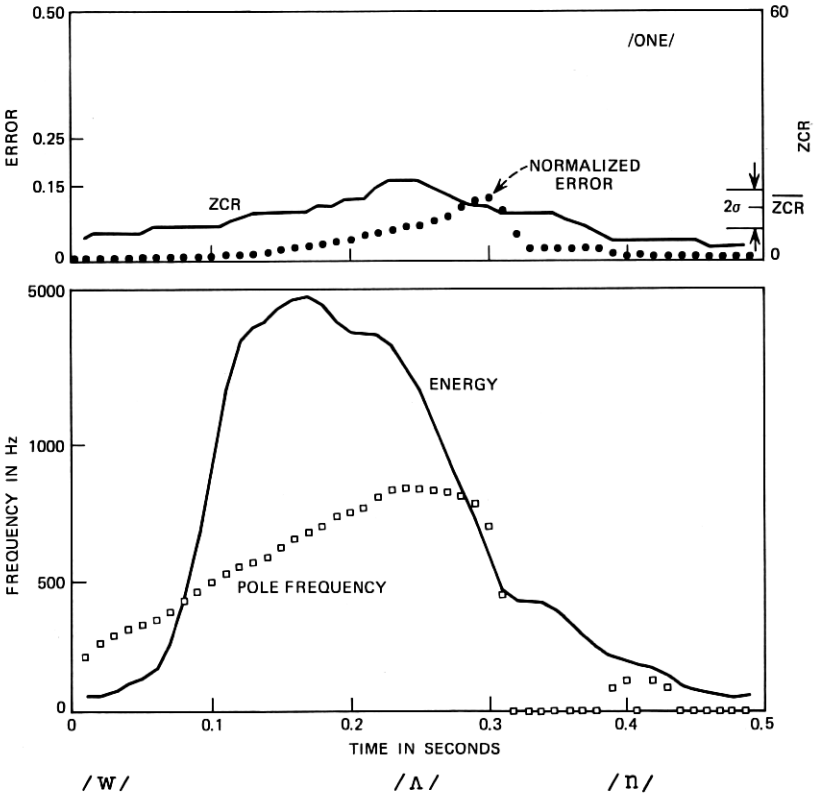

Fig. 11—Complete set of measurements for one example of the word "one."

If no frication is indicated during the ending region, a test for nasal-like behavior is made. If this test is positive, then a hypothesis test of the digit 7 is made. From Fig. 12, we can see that ZCR, two-pole frequency, and energy dip sharply during the consonant /v/. The hypothesis test consists of verifying these dips and checking the vowel characteristics on either side of the dip. Again, the combined output of the test determines whether to reject or accept 7.

If the digits 6 and 7 are eliminated from consideration, the middle region is analyzed to ascertain its structure. The preliminary analysis is achieved by noting the relative change in normalized error. If the normalized error increases, then the structure is characterized as an initial back vowel to a middle or front vowel. Thus, for increasing normalized error, the digits 3, 4, and 5 are considered the most likely. The relative change in the three-second formant measurements are used as supporting evidence to confirm the structure. For the digits

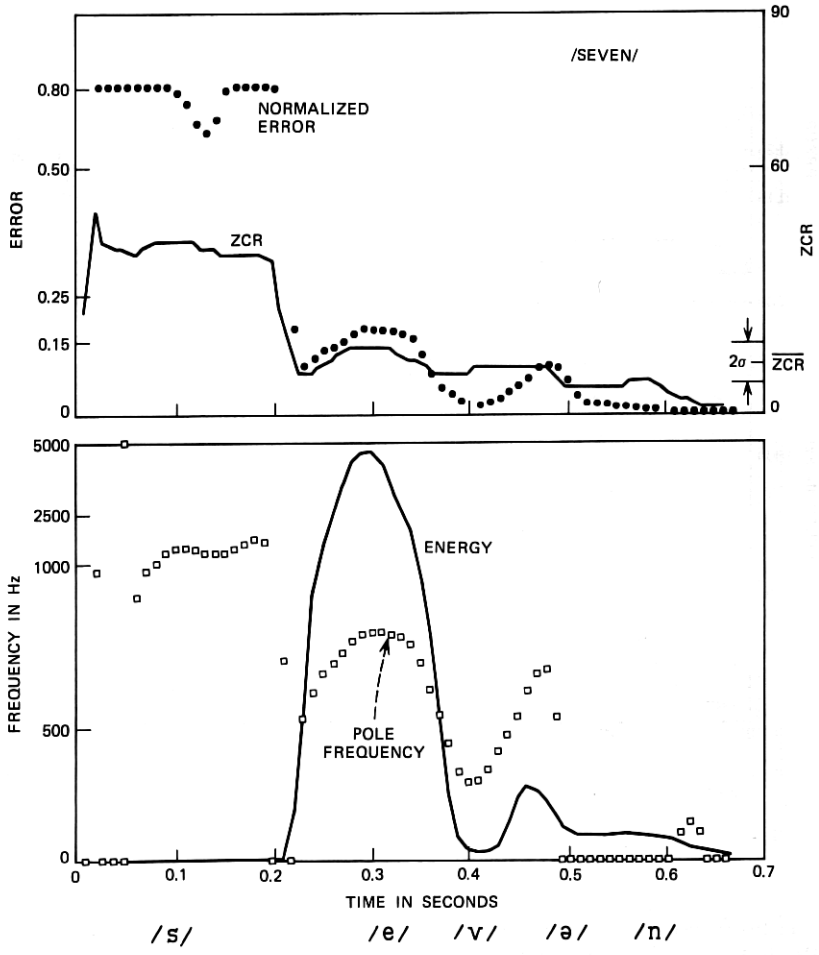

Fig. 12—Complete set of measurements for one example of the word "seven."

3, 4, and 5, F_2 should be increasing. The final decision among the possibilities 3, 4, and 5 is easily achieved on the basis of the robust parameters and formant measurements.

If the normalized error decreases during the middle region, the digits 0 and 2 are then the most probable choices. A decreasing F_2 helps support these choices. To decide between 0 and 2, a dip detector program is used to discover the presence of the sonorant /r/ as depicted by a slight dip in pole frequency, normalized error, and energy. Figure 13 shows the typical dip behavior for these parameters during the spoken 0. The presence (or absence) of a dip is checked with other measurements to verify the final decision.

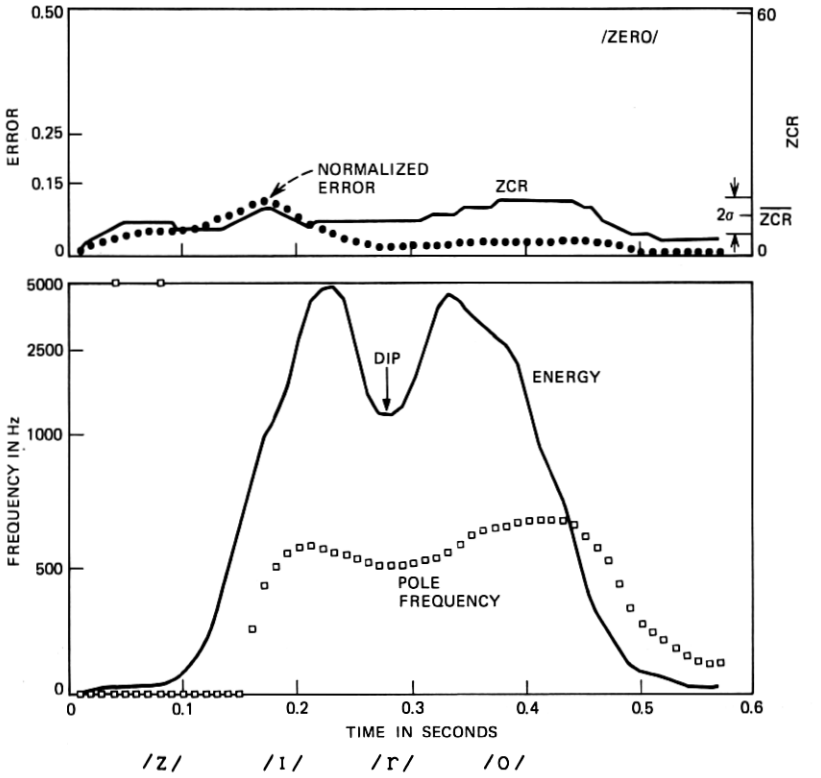

Fig. 13—Complete set of measurements for one example of the word "zero."

IV. EXPERIMENTAL RESULTS

The experimental test of the digit recognizer was conducted in two parts. The first part consisted of 10 speakers (five women and five men) who each made 10 complete recordings of the 10 digits. The recording sessions were spaced over a five-week period to include the effects of time variation in the testing. The recordings were made in a quiet room with a high-quality microphone. The decision algorithm was not designed for the characteristics of each particular speaker, so as to give a true test of the speaker-independent nature of the scheme. The results of this experiment are shown in Table II. The average error rate is 2.7 percent.

A confusion matrix for each of the 100 tests of each digit is presented in Table III. The confusion matrix indicates that all occurrences of initial frication were correctly detected by the decision algorithm. In only 6 out of 200 examples of the digits 1 and 9 was the initial nasal-

Table II — Error scores for first digit recognition experiment

	Correct	Wrong	Percent Correct
Women			
SK	97	3	97
KD	96	4	96
CMcG	96	4	96
BMcD	97	3	97
SP	97	3	97
Total			96.6
Men			
MS*	89	1	98.8
LR	100	0	100
RS	97	3	97
AR*	88	2	97.7
JH	97	3	97
Total			98.1
Sum	954	26	97.3

* Missed one recording session.

like consonant incorrectly determined. The confusion matrix also shows that most errors were made in the final detailed decision. More sophisticated processing would probably enhance the final decision and thereby make the system performance compatible with adaptive schemes.

Table III — Confusion matrix for first digit recognition experiment

		Word Spoken									
		0	1	2	3	4	5	6	7	8	9
Word Recognized	0	93	0	3	0	0	1	0	0	0	1
	1	0	96	0	0	0	0	0	0	0	0
	2	2	0	90	1	0	0	0	1	0	1
	3	2	0	2	97	0	0	1	0	0	2
	4	0	1	2	0	97	1	0	0	0	0
	5	0	0	0	0	0	96	0	0	0	1
	6	0	0	0	0	0	0	97	0	0	0
	7	1	0	0	0	1	0	0	97	0	0
	8	0	0	1	0	0	0	0	0	98	0
	9	0	1	0	0	0	0	0	0	0	93
Individual Errors		5	2	8	1	1	2	1	1	0	5

Total—26 errors out of 980 utterances

Table IV — Distribution of errors for the second digit recognition experiment

	Number of Digits Incorrect			Total Percent Correct
	0	1	2	
Female (30)	11	15	4	92.3
Male (25)	17	8	0	96.8
Total (55)	28	23	4	94.4

To ensure the validity of these experimental results, another more challenging test was conducted. In this experiment, 55 speakers (30 women and 25 men) were selected at random and asked to give one rendition of the 10 digits. Instead of using a high-quality microphone in a quiet environment, the input speech was taken from a close-talking microphone alongside a chattering Teletypewriter. The decision was performed on-line, and the speaker was only instructed when to say each digit. The results for this experiment show an average error rate of 5.6 percent. In addition, no speaker tested did worse than 2 out of 10 wrong. The distribution of errors for this experiment is shown in Table IV. The distribution matrix indicates the generally good performance of the system.

It should be noted that there was no effort in our experimentation to select speakers with good diction. The speakers represent dialects from most of the regions in the U.S. In informal on-line demonstrations of the system, many non-American speakers (French, Japanese, Indian, German) tried having their English-pronounced digits recognized. The informal results were in good agreement with the other experiments. In addition, an informal attempt to "beat" the system by holding one's nose or using falsetto also proved generally unsuccessful.

V. DISCUSSION

The digit-recognition system that has been described in this paper can be considered as a first pass in the direction of speaker-independent speech recognition. Our approach has been to describe a variety of speech sounds in terms of a set of robust measurements. We then devised a tree-structured decision algorithm that used these measurements to characterize the acoustic features of the presented word. The sequence of branches in the tree was designed to resolve the most obvious sounds and then proceed to the more difficult decisions. Thus, the relatively easy problem of distinguishing between noise-like sounds and nasal-like sounds was attacked first, and the determination of the vowel-like constituents was then determined. The output of the pre-

liminary decision tree was a small subset of the 10 possible digits that almost invariably included the spoken word. The major portion of the errors in the system were made in the box labeled "final decision" in Fig. 1.

The preliminary decision tree (Fig. 10) incorporated some original ideas about self-normalization that effectively eliminated the need for tuning the system to the characteristics of a given speaker. Such a decision tree can be extended to prune down a much larger lexicon and arrive at a small list of possible choices. Improvements in the method of selection within the list of possibilities could lead to speaker-independent systems that can truly compete with the performance of adaptive schemes. Such improvements could result by incorporating more sophisticated probabilistic methods into the framework of the "hypothesize-verify" technique proposed in this paper.

Our goal in the development of the digit-recognition system is to show that speaker-independent digit recognition is possible through an intelligent description of broad categories of speech sounds. This description uses what is known about the *necessary* characteristics of each category instead of blindly using pattern-matching algorithms to rigidly quantify the sounds. The later approach is doomed to failure for a large enough speaker population because it overlooks the fact that the information in the patterns contain as much personal information as linguistic information.

REFERENCES

1. Proceedings IEEE Symposium on Speech Recognition, Carnegie-Mellon University, April 1974.
2. Proceedings 1972 Conference on Speech Communication and Processing, April 1972.
3. A. Newell et al., *Speech Understanding Systems*, Springfield, Va.: National Technical Information Service, May 1971.
4. R. Reddy et al., "Working Papers in Speech Recognition," Carnegie-Mellon University, April 1972.
5. P. Vicens, "Aspects of Speech Recognition by Computer," Ph.D. thesis, Stanford University, April 1969.
6. T. B. Martin, "Acoustic Recognition of a Limited Vocabulary in Continuous Speech," Ph.D. thesis, University of Pennsylvania, 1970.
7. J. Makhoul and J. Wolf, "The Use of a Two-Pole Linear Prediction Model in Speech Recognition," Report 2537, Cambridge, Mass.: Bolt, Beranek and Newman, Inc., September 1973.
8. J. Makhoul and J. Wolf, "Linear Prediction and the Spectral Analysis of Speech," Report 2304, Cambridge, Mass.: Bolt, Beranek and Newman, Inc., August 1972.
9. B. Gold and L. R. Rabiner, "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain," J. Acoust. Soc. Am., *46*, No. 2 (August 1969), pp. 442–448.
10. L. R. Rabiner and M. R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances," to be published in B.S.T.J., *54*, No. 2 (February 1975).
11. J. W. Tukey, "Nonlinear (Nonsuperposable) Methods for Smoothing Data," 1974 EASCON Record, p. 673.