An Improved Endpoint Detector for Isolated Word Recognition

LORI F. LAMEL, STUDENT MEMBER, IEEE, LAWRENCE R. RABINER, FELLOW, IEEE, AARON E. ROSENBERG, MEMBER, IEEE, AND JAY G. WILPON

Abstract—Accurate location of the endpoints of an isolated word is important for reliable and robust word recognition. The endpoint detection problem is nontrivial for nonstationary backgrounds where artifacts (i.e., nonspeech events) may be introduced by the speaker, the recording environment, and the transmission system. Several techniques for the detection of the endpoints of isolated words recorded over a dialed-up telephone line were studied. The techniques were broadly classified as either explicit, implicit, or hybrid in concept. The explicit techniques for endpoint detection locate the endpoints prior to and independent of the recognition and decision stages of the system. For the implicit methods, the endpoints are determined solely by the recognition and decision stages of the system, i.e., there is no separate stage for endpoint detection. The hybrid techniques incorporate aspects from both the explicit and implicit methods. Investigations showed that the hybrid techniques consistently provided the best estimates for both of the word endpoints and, correspondingly, the highest recognition accuracy of the three classes studied. A hybrid endpoint detector is proposed which gives a rejection rate of less than 0.5 percent, while providing recognition accuracy close to that obtained from hand-edited endpoints.

I. INTRODUCTION

ISOLATED word recognition is based on the premise that the signal in a prescribed recording interval consists of an isolated word, preceded and followed by silence or other background noise. Thus, when a word is actually spoken, it is assumed that the speech segments can be reliably separated from the nonspeech segments. (Clearly, in the case when there is no speech in the recording interval, a request to repeat the spoken word must be made.) The process of separating the speech segments of an utterance from the background, i.e., the nonspeech segments obtained during the recording process, is called endpoint detection. In isolated word recognition systems, accurate detection of the endpoints of a spoken word is important for two reasons, namely:

1) reliable word recognition is critically dependent on accurate endpoint detection

2) the computation for processing the speech is minimum when the endpoints are accurately located.

Manuscript received September 8, 1980; revised January 26, 1981. This work is based on the M.S. thesis "Methods of endpoint detection for isolated word recognition," by L. F. Lamel, Massachusetts Institute of Technology, Cambridge, June 1980.

L. F. Lamel was with Bell Laboratories, Murray Hill, NJ 07974. She is now with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139.

L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon are with Bell Laboratories, Murray Hill, NJ 07974.

This paper discusses the problem of accurately locating the endpoints of isolated words for recordings made over dialed-up telephone lines. Problems in endpoint detection arise from transients associated with the speaker and/or the transmission system (i.e., the telephone system). This type of background noise complicates the endpoint detection problem considerably. For example, often the beginning or end of an isolated word is obscured by speaker generated artifacts such as mouth noises, e.g., clicks, pops, lip smackings, and heavy breathing. Similar types of artifacts may be introduced by the telephone transmission system. In many applications, the problem is further complicated by nonstationary backgrounds where there may be concurrent background conversations and noises due to movements of chairs, door slams, etc. One way of minimizing the effects of such transient backgrounds is to use a close-talking, noise cancelling microphone for recording the speech signal; however, this approach is not feasible for transmission over telephone lines. Hence, an accurate endpoint detection method is an essential component of an isolated word recognizer which operates over dialed-up telephone lines.

The endpoint detection techniques described in this paper assume that the desired spoken word is present in a given recording interval. This type of processing is reasonable for "nonreal-time" speech recognition systems. For "real-time" applications, the beginning of the spoken word must be detected before the word has ended (or else a large buffer storage is required). Many of the techniques to be described in this paper can be readily modified for such real time applications.

The importance of accurate endpoint detection was noted by Martin [1], who showed that recognition performance was directly related to endpoint accuracy. Although an endpoint detector is an essential component in all speech recognition systems, there has been very little published about specific algorithms for performing this task [2]. The reason for this is that most laboratory systems use reasonably clean recordings and, hence, there is no problem in finding endpoints from a simple heuristic, whereas commercial manufacturers, who have to worry both about real-time response and difficult recording conditions, are reluctant to publish their successful, working algorithms. As such, one purpose of this paper is to establish a framework for endpoint detection algorithms, and another is to provide an improved, heuristically conceived and carefully tested endpoint detection method.

The essential components of a speech recognition system are feature extraction, pattern comparison, and a decision rule.

0096-3518/81/0800-0777$00.75

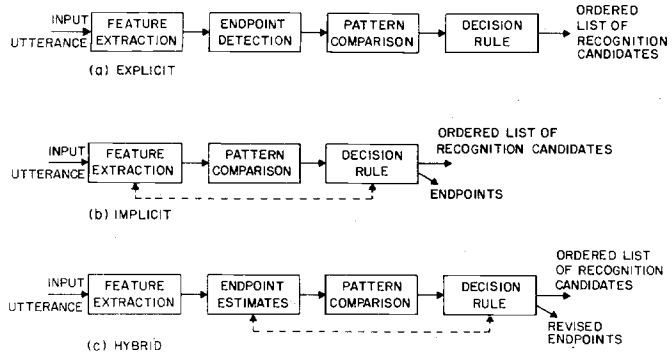


Fig. 1. Block diagrams of canonic forms of the explicit, implicit, and hybrid endpoint detectors.

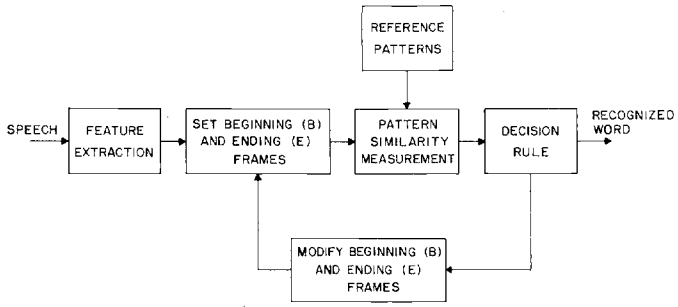


Fig. 2. Block diagram of a DTW-based implicit endpoint detector.

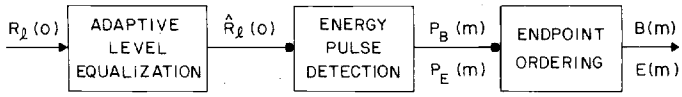


Fig. 3. Block diagram of the feed-forward processing of the hybrid endpoint detector.

Endpoint detection must be performed somewhere in the processing. The processing for finding word endpoints can be done explicitly, implicitly, or in a hybrid manner (i.e., a combination). These three approaches are illustrated in Fig. 1. For explicit endpoint detection methods, as shown in Fig. 1(a), the endpoint detection precedes and is *independent* of the recognition and decision stages of the recognizer. Typically, the endpoints of the spoken word are estimated from measurements made on the input speech and sent in a feed-forward manner to the next stage of the system.

In a purely implicit approach to endpoint detection [Fig. 1(b)], the endpoints of the isolated word are determined *solely* by the recognition and decision phases of the word recognition system; i.e., there is no separate stage for endpoint detection. An implicit endpoint detection method would attempt recognition using all (or possible a large set of all) possible endpoint sets.

The hybrid techniques [Fig. 1(c)] for endpoint detection incorporate ideas from both the explicit and implicit methods. Similar to the explicit approach, one or more estimates of *each* of the endpoints are obtained from features measured from the input utterance. Based on feedback from the recognition scores, alternative endpoint sets are considered. We consider all three types of endpoint detectors in this paper.

The organization of the paper is as follows. In Section II, a brief review of an explicit and an implicit endpoint detector is given. In Section III, a hybrid endpoint detector is described. In Section IV, an experimental evaluation of the performance of the endpoint detectors in an isolated word recognition system is presented and discussed. A final summary is given in Section V.

II. Explicit and Implicit Endpoint Detection

An example of an explicit endpoint detector is the energy-based approach as described by Rabiner and Sambur [2]. Using the energy contour of the recorded signal and an appropriate set of thresholds, the "beginning" and "ending" of the word are estimated. In [2], the zero-crossing contour was used to refine the word endpoints for words with fricative beginnings and endings. For telephone line recordings (with a 3 kHz bandwidth), the use of zero crossings is not effective. Hence, this feature is not used in the explicit endpoint detector which was evaluated in this paper. Only one endpoint set was obtained from this method, and a rejection occured whenever a consistent set of endpoints could not be found due to poor recordings, line transients, etc.

An example of an implicit endpoint detector is given in Fig. 2 [3]. For this system, all (reasonable) combinations of beginning points (B) and ending points (E) are used and the best output from the pattern similarity stage and decision rule (lowest distance) is used to implicitly define the word endpoint as well as the recognized word.

III. An Improved Hybrid Endpoint Detector

A block diagram of the proposed hybrid endpoint detector is given in Fig. 3 [4]. The input to the detector is the energy array $R_l(0)$, $l = 1, 2, \cdots, L$, where L is the number of frames in the recording interval. There are three blocks in the processing, namely, adaptive level equalization, energy pulse detection, and endpoint ordering. The function of each of these blocks is explained in subsequent sections. The output of the endpoint detector is the ordered set of beginning points $B(m)$ and ending points $E(m)$, where each set defines a word endpoint pair. For each endpoint pair, the pattern similarity and decision stages of the recognizer find the word with the smallest distance. If the distance obtained from one endpoint pair is sufficiently small, no other endpoint pairs are tried. Otherwise, the next pair of endpoints is tried, and the process is repeated. We will see later that the endpoint ordering algorithm is biased to include short events occurring prior to or following the main body of the word (e.g., stop releases, etc.) in the early endpoint pairs. Hence, the proposed method is applicable to vocabularies with similar words such as "for" and "afore," "tore" and "store," etc.

A. The Adaptive Level Equalizer

The first stage of the hybrid endpoint detector is the adaptive level equalizer which normalizes the (log) energy array to the background noise level. The equalized energy array $\hat{R}_l(0)$ is determined as

$$\hat{R}_l(0) = \log [R_l(0)] - Q, \qquad l = 1, 2, \cdots, L$$

where Q is the "averaged" noise background level which is

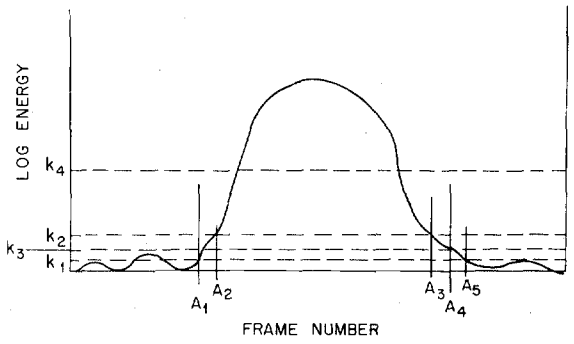


Fig. 4. Example illustrating the use of energy thresholds to find beginning and ending frames of energy pulses.

obtained as follows. First, minimum energy $E_{\min}$ is obtained as

$$E_{\min} = \min_{1 \leq l \leq L} \{\log[R_l(0)]\}.$$

Then a histogram is taken of the low 10 dB of the log energy levels from the values of $\log [R_l(0)]$ versus l. A three-point averaging of the histogram is made, and the peak of the histogram is found. Q is chosen as the peak of the smoothed noise level histogram.

The level equalized energy array has the property that during silence it fluctuates around the 0 dB level, and during speech it is considerably larger. Thus, absolute energy thresholds can be defined for detection of the presence of speech-like signals, as described in the next section.

B. Energy Pulse Detection

Based on the output of the adaptive level equalizer $\hat{R}_l(0)$, a set of four energy thresholds k_1, k_2, k_3, and k_4 are defined as illustrated in Fig. 4. The purpose of these thresholds is to define the presence of an "energy pulse," i.e., a speech-like burst of energy during the recording interval. The assumption is made that the spoken word contains a sequence of one or more such energy pulses, and the only problem is to find those pulses and to determine which ones belong to the spoken word.

The detection of energy pulses proceeds from left to right. Values of $\hat{R}_l(0)$ are scanned (as l varies) and when $\hat{R}_l(0)$ exceeds the first threshold k_1, the frame number (A_1) is recorded. If $\hat{R}_l(0)$ exceeds the higher threshold k_2 before falling below k_1, the beginning of an energy pulse is detected. The beginning point is nominally chosen as frame A_1, unless the rise time (from A_1 to A_2) is too long, in which case the beginning point is chosen as frame A_2. The ending frame is detected in a manner similar to the starting frame using thresholds k_2 and k_3. However, if the duration from A_3 to A_4 is too long (this typically indicates breathing at the end of the word), the frame A_3 is used as the ending frame of the energy pulse.

Two further tests are made on each detected energy pulse. The peak energy of the pulse is measured, and if it falls below the level threshold k_4, the energy pulse is rejected as being part of the word. Also, the overall pulse duration is measured, and if it is too short (less than five frames, i.e., 75 ms), the energy pulse is rejected.

The output of the energy pulse detector is a series of pulse beginning points $P_B(m)$ and pulse ending points $P_E(m)$, $m = 1, 2, \cdots, M$ for M detected pulses in the recording interval. When $M = 0$ (i.e., no detected pulses), the recording is rejected and no endpoints are found. Checks are also made on whether pulses of significant energy occur at the boundaries of the recording interval. If so, the recording is again rejected. A flow diagram of the energy pulse detector is given in Fig. 5(a).

C. Pulse Endpoint Ordering

The purpose of the pulse endpoint ordering box is to determine, in order of likelihood (as defined below), the possible sets of word endpoint pairs from the set of pulse endpoints. The ordering logic is based on the following assumptions.

1) The isolated word whose endpoints are to be determined consists of one or more energy pulses.

2) The frame in the log energy contour with the maximum energy will always be included within the spoken word.

3) The larger the stopgap between two energy pulses, the less likely that they come from one multiple-pulse word.

4) Energy pulses separated from the pulse containing the maximum energy by a stopgap of greater than 150 ms are unlikely to be part of the word.

Based on these assumptions, the energy pulses are grouped into combinations of word-endpoint pairs and ordered. A flow diagram of the ordering procedure is given in Fig. 5(b), and Fig. 6 illustrates the method. In Fig. 6 we see three detected energy pulses, P_1, P_2, and P_3, with pulse separations X_1 and X_2 frames. If both X_1 and X_2 are less than 150 ms, then the first pair of endpoints is chosen as A_1 and A_6, the second pair is chosen as A_3 and A_6 (assuming $X_1 > X_2$), the third pair is chosen as A_1 and A_4, and the fourth pair is chosen as A_3 and A_4. If $X_1 > 150$ ms and $X_2 < 150$ ms, the ordered endpoint pairs are (A_3, A_6), (A_3, A_4), and (A_1, A_4). Finally, if both X_1 and X_2 are greater than 150 ms, the ordered pairs are (A_3, A_4), (A_3, A_6) and (A_1, A_4). This simple example illustrates how the ordering is sensitive to pulse separation.

D. Feedback in the Hybrid Endpoint Detector

As shown above, the output of the feed-forward portion of the hybrid endpoint detector is an ordered set of estimates of word endpoints. Each endpoint set is used in the recognizer to determine the word with the lowest distance score. Whenever the resulting distance score is too large, the recognizer decision stage requests the next set of endpoints and, if available, repeats the recognition comparisons. This process continues until all endpoint sets are used or until a reliable distance score is obtained.

IV. Experimental Evaluation of the Endpoint Detection Methods

The performance of each of the three endpoint detection algorithms was evaluated using a single-testing data set to provide a common basis for comparison. This testing set consisted of three repetitions of a 39 word vocabulary, recorded by each of ten talkers (four female and six male). The talkers were problematic ones, i.e., they were known to generate

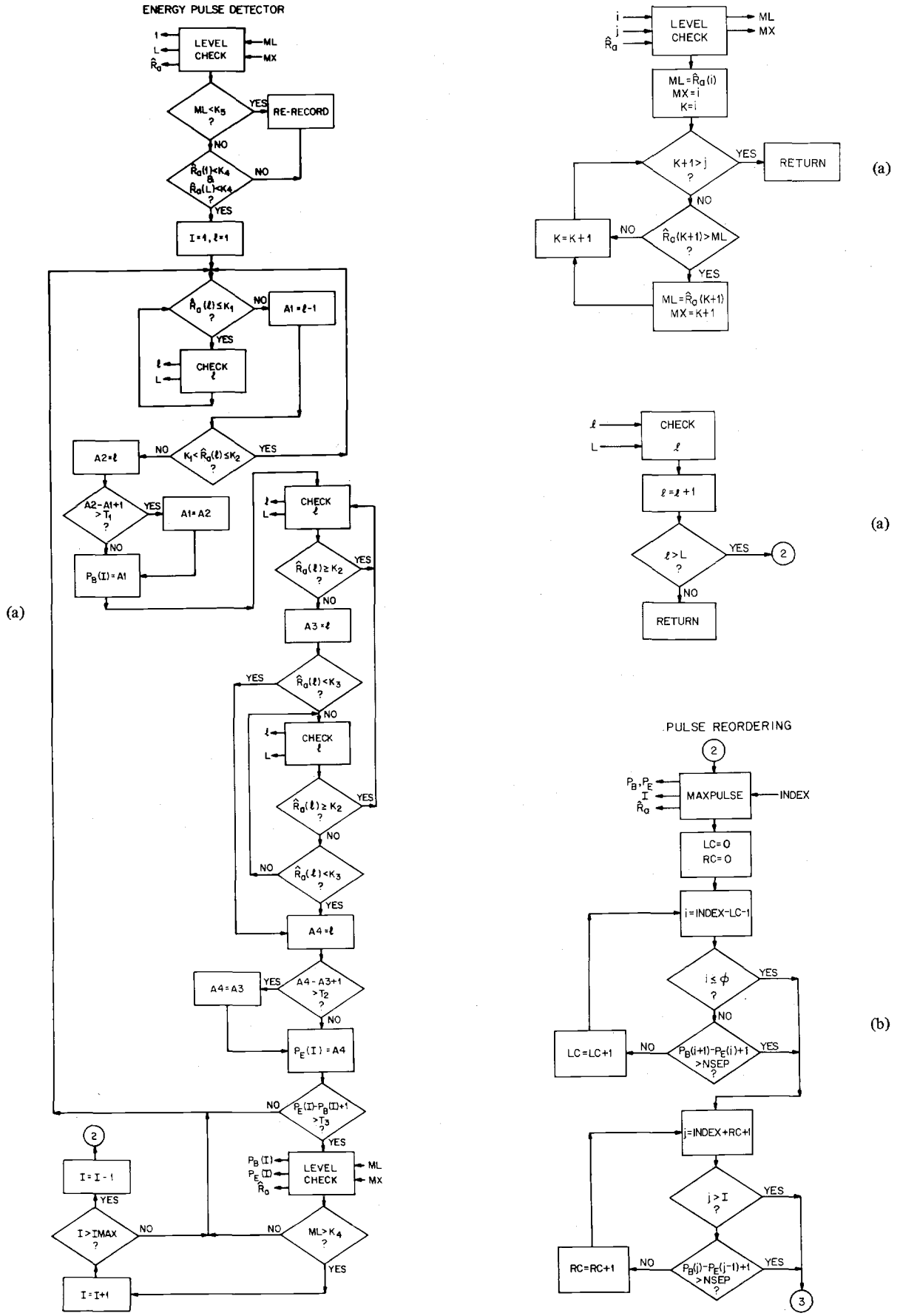


Fig. 5. (a) Flow charts of the energy pulse detector. (b) Flow chart of the pulse ordering procedure.

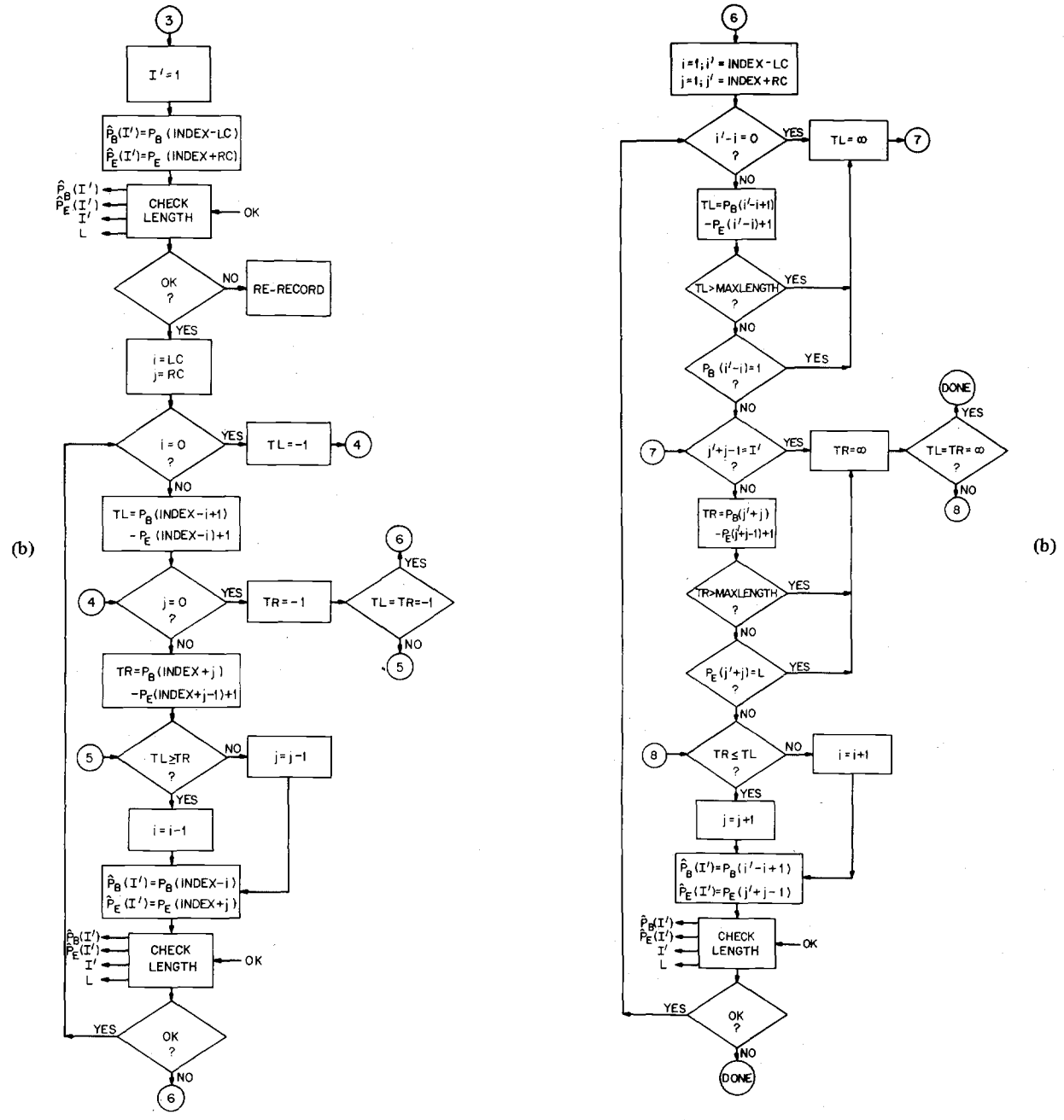


Fig. 5 continued. (b) Flow chart of the pulse ordering procedure.

many of the artifacts discussed in the introduction while speaking isolated words. The 39 word vocabulary consisted of the letters of the alphabet, the digits, and the words /STOP/, /ERROR/, and /REPEAT/ spoken in a randomized sequence. The resulting test set included 1161 utterances (nine words were lost due to manual recording errors). Of the recorded utterances, 40-60 percent included artifacts of some type.

Each endpoint detector was evaluated on the basis of two criteria, namely, 1) the recognition accuracies achieved by the recognition system using the selected endpoints, and 2) the goodness or accuracy of the locations of the endpoints. The first criterion is a well-defined, quantitative measure of the actual performance of the endpoint detector within the recognition system. The second criterion, however, is highly subjective, as the accuracy of the endpoint detector is determined relative to a humanly defined standard. Manual location of the endpoints of an isolated word, determined from the time sequence of the samples, the log energy contour, or some other measurement, is subject to error, even given the knowledge of what word was spoken. Fortunately, the continuity of speech and the inherent redundancy in the speech allow the endpoints to vary by small amounts (perhaps one to three frames) without strongly affecting recognition accuracy. The rejection rate associated with each method was also measured. There is clearly a tradeoff between recognition accuracy and rejection rate. The object is to simultaneously obtain the highest recognition accuracy and the lowest rejection rate. The recognition accuracy is upper bounded by the recognition

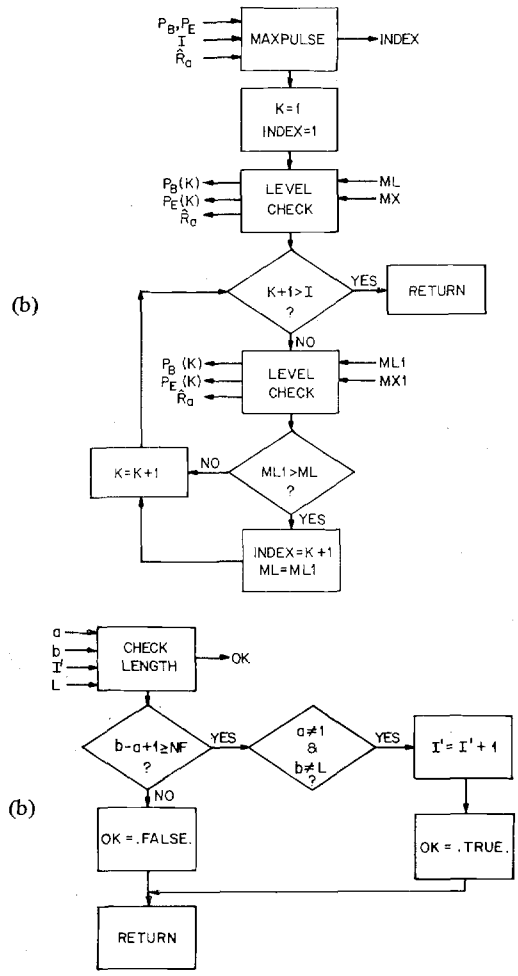


Fig. 5 continued. (b) Flow chart of the pulse ordering procedure.

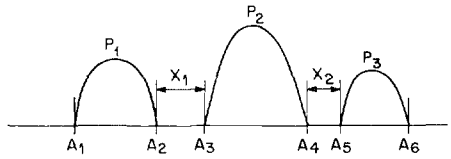


Fig. 6. Example illustrating the use of temporal constraints in discriminating between adjacent energy pulses.

scores obtained using clean speech (no artifacts) with manually obtained endpoints. For most applications, the recognition accuracy must attain a minimum performance level to be useful. The cost of a high rejection rate is an inconvenience to the user; if the rejection rate gets too large, the user will not be willing to use the system.

A. The Isolated Word Recognizer

The isolated word recognizer used in the evaluation is the one described in [5]. Recognition is achieved by comparing the test pattern to a set of previously stored reference templates (using a DTW alignment procedure) and selecting the "closest match" (i.e., the reference with the smallest average distance) as the recognized word. The reference set of words consisted of twelve speaker-independent reference templates per word created using statistical clustering techniques [5].[1]

[1]Clearly, any template and test set can be used to evaluate an endpoint detection method.

TABLE I
PERFORMANCE RESULTS FOR ENERGY-BASED EXPLICIT ENDPOINT DETECTOR

Talker	Number of words spoken	% rejected	Candidate Position 1	2	3	4	5
1	114	35.1	71.6	81.1	89.2	95.9	97.3
2	117	29.1	63.9	79.5	85.5	88.0	90.4
3	116	31.0	82.5	90.0	95.0	97.5	97.5
4	112	28.5	81.3	86.2	93.8	95.0	95.0
5	117	42.7	61.2	73.1	76.1	83.6	83.6
6	117	9.4	71.7	84.0	86.6	89.6	91.5
7	117	32.5	75.9	83.5	88.6	92.4	92.4
8	117	20.5	57.0	74.2	81.7	90.3	91.4
9	117	33.3	52.6	66.7	74.4	79.5	83.3
10	117	29.1	69.9	80.7	83.1	86.0	89.2
TOTAL	1161	29.1	68.8	80.1	85.5	90.0	91.4

An average distance score is computed for the comparison between the test pattern and each time aligned reference pattern. The decision rule of the recognition system creates an ordered list of recognition candidates from the average distance scores; i.e., the reference word with the smallest average distance is the top candidate, the next smallest average distance is the second candidate, etc.

B. Endpoint Detector Performance

For the explicit endpoint detector, the performance criteria used were the recognition accuracy and the rejection rate. In evaluating the implicit endpoint detector, the performance criteria were the recognition accuracy, the rejection rate, and the relative distribution of distance scores (for correct and incorrect references). For the hybrid endpoint detector, the criteria were the number of endpoint pairs found, the recognition accuracy, and the rejection rate.

1) Performance of the Energy-Based Explicit Endpoint Detector: The results of the recognition tests on the energy-based explicit endpoint detector are shown in Table I, which gives, for each talker, the rejection percentage and the recognition accuracy as a function of candidate position, i.e., the percentage of words that were correct in the top n positions of the ordered candidate list. As can be seen in the column labeled "% rejected," of the 1161 test words, 338, i.e., almost 30 percent, were rejected and a repeat requested. The rejection rates range across talkers from about 10 to 42 percent. These results show that the energy-based endpoint detector failed in a large percentage of trials. The average recognition accuracy for the top candidate is about 70 percent, ranging across talkers from a low of 50 percent to a high of over 80 percent. The overall recognition for the top five candidates ranges from 83 to 98 percent with an average of 91 percent. These results indicate that the explicit endpoint detector is reasonably accurate when the recording is relatively clean, but it tends to reject the recording in the presence of artifacts.

2) Performance of the DTW-Based Implicit Endpoint Detector: For the implicit endpoint detector based on DTW matching, an important performance indicator is the distribution of the distance scores for both the correct and the incorrect reference templates. It was found that, using a totally unconstrained beginning and ending point DTW algorithm, the distances between the tests and the correct references are generally small and in the range one would expect for a correct reference-test word recognition distance. By way of example,

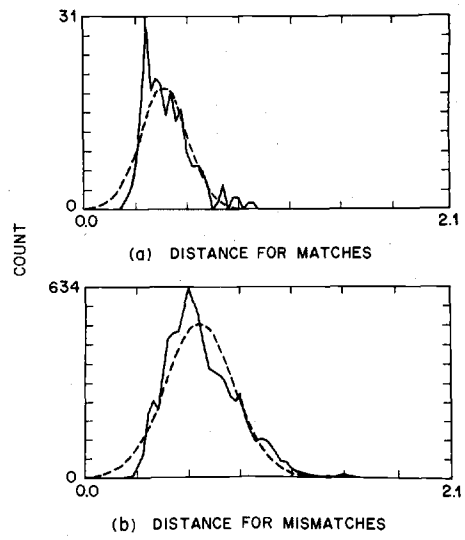


Fig. 7. Histograms of average word distance for matches (a) and mismatches (b) for one talker using an unconstrained endpoint DTW algorithm.

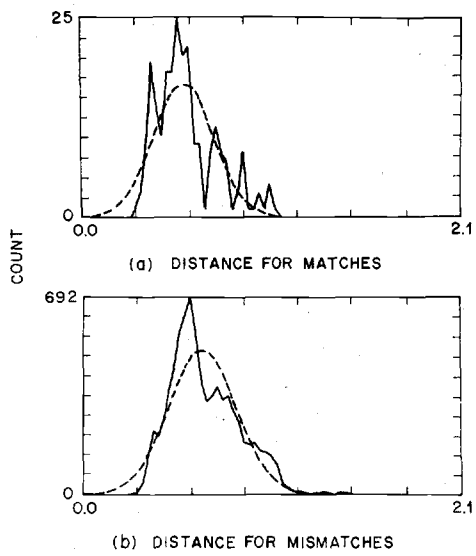


Fig. 8. Histograms of average word distance for matches (a) and mismatches (b) for another talker.

Fig. 7(a) shows a histogram of distances obtained when the test and reference words were the same for one of the talkers. The data plotted include all the tokens recorded for that talker. The solid curve is the measured histogram of distances and the dashed curve is a Gaussian fit to the data. The average distance for the correct words is seen to be about 0.4 with a standard deviation of about 0.1. Thus, the distances obtained using the test and the correct reference pattern are seen to be reasonably small. What is equally or, perhaps, more important is the distribution of distances when the test pattern is compared to incorrect reference templates. This result is illustrated in Fig. 7(b) for the same talker where the test and reference words were different. It can be seen that there is substantial overlap in these two distributions, thereby implying a substantial error in recognition. A similar set of histograms are shown in Fig. 8 for a different speaker. Here, there is almost total overlap in the distributions, implying an even larger error rate that in the previous example.

TABLE II
PERFORMANCE RESULTS FOR DTW-BASED IMPLICIT ENDPOINT DETECTOR

Talker	Number words spoken	Recognition Accuracy candidate position 1	2	3	4	5
1	114	27.2	36.8	56.1	65.8	76.3
2	117	25.6	34.2	41.9	47.0	47.9
3	116	44.0	64.7	70.7	75.9	81.0
4	112	55.4	67.0	75.9	80.4	83.0
5	117	17.0	28.2	37.6	49.6	56.4
6	117	7.7	14.5	23.9	32.5	37.6
7	117	33.3	53.9	60.7	70.9	76.1
8	117	29.1	47.9	61.5	68.4	72.7
9	117	27.4	48.7	59.0	63.2	70.1
10	117	28.2	44.4	58.1	65.8	72.7
TOTAL	1161	29.4	43.9	54.4	61.8	67.3

TABLE III
STATISTICS ON THE NUMBER OF ENDPOINT SETS FOR THE HYBRID ENDPOINT DETECTOR

	Words Spoken	Number of endpoint sets 1	2	3	mean
1	114	94	19	0	1.17
2	117	93	23	1	1.21
3	116	100	16	0	1.14
4	112	102	10	0	1.09
5	117	97	20	0	1.17
6	117	99	18	0	1.15
7	117	99	18	0	1.15
8	117	102	14	0	1.12
9	117	92	22	1	1.21
10	117	94	22	0	1.19
TOTAL	1161	972	182	2	1.16

The conclusion drawn from these examples is that a totally unconstrained beginning and ending point DTW algorithm allowed the recognizer too much freedom in the determination of the endpoints. By allowing the algorithm to discard any portions of the test, it was able to match the test pattern as well or better to the incorrect reference patterns than to the correct reference patterns. These conclusions are verified in the recognition accuracies obtained using this implementation of an implicit endpoint detector, as shown in Table II, which shows recognition accuracy as a function of candidate position. Although the rejection rate is zero for all talkers, the recognition scores are too low for reliable recognition. For the top candidate, the recognition scores range from less than 10 to 55 percent correct across talkers. Even for the top five candidates, the recognition scores only range from 37 to 83 percent, with an average of less than 70 percent. Overall recognition for the implicit endpoint detector using five candidates is lower than for the top candidate using the explicit endpoint detector.

3) Performance of the Energy Hybrid Endpoint Detector: For the hybrid endpoint detector, one important factor is the number of endpoint pairs located, as this is an indication of the potential gained by using feedback from the recognizer. The number of endpoint pairs located for each talker is given in Table III. The data show that, on average, only slightly more than one endpoint pair is found. The effectiveness of the screening operations of the endpoint detector is reflected in the small number of endpoint pairs found by eliminating energy pulses corresponding to extraneous artifacts. Since for the majority of the tests (84 percent), only one endpoint pair was located, it is reasonable to use this hybrid endpoint detec-

TABLE IV
RECOGNITION SCORES FOR THE HYBRID ENDPOINT DETECTOR

Talker	Number Words Spoken	Number of rejects	Recognition Accuracy 1	2	3	4	5
1	114	1	79.6	97.3	98.2	99.1	100.0
2	117	0	73.5	86.3	90.6	95.7	96.6
3	116	0	84.5	93.1	95.7	98.2	99.1
4	112	0	81.1	86.6	89.3	92.0	92.9
5	117	0	74.3	81.1	88.9	91.4	91.4
6	117	0	71.8	84.6	87.2	89.7	92.3
7	117	0	74.4	92.3	96.6	97.4	98.3
8	117	1	61.2	77.6	86.2	93.1	94.8
9	117	2	70.4	80.9	84.3	85.2	89.6
10	117	1	78.4	87.9	95.7	97.4	99.1
TOTAL	1161	5	74.9	86.8	91.3	93.9	95.4

tor with only the top estimated endpoint pair as an explicit endpoint detector. The recognition accuracies obtained using only the top endpoint pair are given in Table IV for the individual talkers. The overall recognition is about 75 percent for the top candidate position and 95 percent for the top five candidate positions. Comparing these results to those of the explicit endpoint detector shows there is a 10-15 percent increase in the recognition accuracy for the top candidate for most of the talkers. Averaged over all of the talkers, there is 4-6 percent improvement in recognition accuracy for all candidate positions.

The advantage of supplying several estimates of the endpoints is demonstrated by the use of a distance threshold for requesting additional endpoint pairs during the decision stage of the recognition process. If the recognition score obtained using the best estimated endpoint pair does not provide a sufficiently small distance, recognition with the second endpoint pair may be attempted. Setting a low distance threshold for reliable recognition will cause recognition to be attempted with successive endpoint pairs unless the first match is a very good one. It can be seen from the data in Table III that even if an extremely low threshold is used, only one or two endpoint pairs will be used for recognition. Setting a high distance threshold increases the likelihood that the recognized word with the first endpoint pair will have a recognition score below the threshold and, thus, be acceptable. The threshold should be set such that if the first set of endpoints are accurate, the corresponding recognition distance will fall within the acceptable range. At the same time, if the top endpoint pair is incorrect, it is desirable that the recognition score lie above the acceptable threshold. It is important to stress that the distance threshold does not affect the number of endpoint pairs located, but does determine how many of the endpoint candidates are used.

The optimum distance threshold was determined for each of the ten talkers. These results are summarized in Table V. A detailed examination of the recognition results shows that the optimum distance threshold should lie somewhere between 0.35 and 0.4 for highest recognition accuracy. For some of the talkers, the threshold was irrelevant. This implies that for these talkers, either the first candidate was always correct, or any improvement in the recognition of some words was counteracted by errors introduced for other words. Only for two of the talkers did the recognition improve significantly by using the distance thresholds. For these, talkers 2 and 9, there was a uniform increase of 2-3 percent in the recognition

TABLE V
OPTIMUM DISTANCE THRESHOLDS FOR THE HYBRID ENDPOINT DETECTOR

Talker	Distance optimum threshold
1	>0.3
2	<0.4
3	>0.35
4	irrelevant
5	irrelevant
6	>0.6
7	irrelevant
8	irrelevant
9	<0.4
10	<0.4

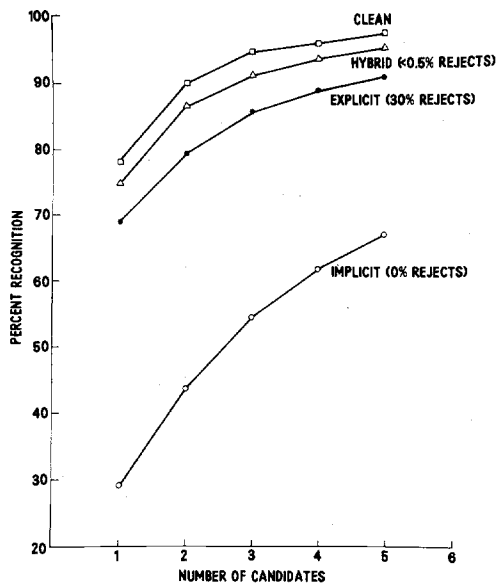


Fig. 9. Plots of recognition accuracy versus number of candidates for three endpoint detectors and for hand-edited words.

accuracy for all candidate positions. The data above suggest that the use of alternate sets of endpoints provides some improvement in recognition accuracy for problematic talkers. This capability adds little to the computational requirements of the system and has proved to be valuable during on-line recognition tests with untrained talkers.

C. Summary of Performance Scores

A comparison of the recognition scores obtained by all three methods of endpoint detection in the word recognition test is shown in Fig. 9. This figure shows plots of the recognition accuracy as a function of candidate position. It can be clearly seen that the hybrid endpoint detector yields significantly higher recognition results than either the explicit or implicit methods used here. An even stronger result is that the hybrid endpoint detector attains its recognition accuracy with a rejection rate of less than 0.5 percent, whereas the explicit endpoint detector has a rejection rate of almost 30 percent for the same test set. The implicit algorithm has a recognition accuracy far below that of the explicit and hybrid algorithms.

A final performance comparison is given in the top curve of Fig. 9, labeled "clean" speech, which shows the recognition accuracy obtained from hand-edited endpoints with highly trained talkers [5]. The vocabulary was the same as that used in this test. The hybrid endpoint detector is seen to achieve recognition accuracies within 3-5 percent of those obtained

with the clean speech. It is important to keep in mind that the results using the hybrid method were obtained using a testing data set in which 40-60 percent of the utterances contained some form of artifact, many of which would have been discarded (due to the uncertainty as to where the actual endpoints would be placed) even by hand-editing. Thus, the results obtained by the hybrid endpoint detector are seen to approach those of hand-edited clean speech.

V. Discussion and Summary

The results presented in the previous sections lead to the following general conclusions.

1) Simple approaches to endpoint detection are doomed to failure under some sets of recording conditions. These failures can be manifested as either rejections of the recording or recognition errors due to improper location of endpoints.

2) Pattern classification techniques are not readily applied to the endpoint detection problem since there is a strong overlap between "speech" sounds and "nonspeech" sounds, especially humanly produced transients.

3) Providing a great deal of latitude in the specification of the endpoint locations tends to degrade the recognition performance severely. Hence, accurate location of endpoints is a strong requirement for a practical recognition system.

As applied to the general endpoint detection classes, these conclusions imply that implicit methods of endpoint detection will perform poorly, whereas a sophisticated explicit technique can perform well. Clearly the hybrid methods, being a superset of the explicit techniques, should always perform as well as the explicit techniques.

One of the main results of this work was the development of an improved hybrid (explicit) endpoint detector. The improvements consisted of an optimized strategy for finding endpoints using a three-pass approach in which energy pulses were located, edited, and the endpoint pairs scored in order of most likely candidates. The performance of this improved technique approached the performance of the "ideal" endpoint detector—i.e., the recognition accuracy was comparable to the recognition accuracy obtained on high quality recordings, and the rejection rate was negligible (less than 0.5 percent).

One final issue is the capability of incorporating the hybrid endpoint detector into a real-time recognizer. The required modifications are simple. First, the adaptive level equalization is performed on any reasonably "quiet" interval—i.e., when no speech is present. The energy pulse detection stage is carried out in real time (i.e., all processing is left to right) and the pulse locations are stored in a temporary buffer. When no energy pulses are detected for, say, 200 ms, the word is assumed over and the endpoint ordering algorithm is then used to provide the beginning and ending point arrays required for recognition. The real-time version of the algorithm is currently in use in a hardware implementation of the recognizer of [5] (see [6]).

References

[1] T. Martin, "Applications of limited vocabulary recognition systems," in *Rec. 1974 Symp. Speech Recognition*, D. R. Reddy, Ed. New York: Academic, 1975, pp. 55-71.

[2] 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.

[3] C. S. Myers, "A comparative study of several dynamic time warping algorithms for speech recognition," M.S. thesis, Massachusetts Inst. Technol., Cambridge, Feb. 1980.

[4] L. F. Lamel, "Methods of endpoint detection for isolated word recognition," M.S. thesis, Massachusetts Inst. Technol., Cambridge, June 1980.

[5] L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon, "Speaker independent recognition of isolated words using clustering techniques," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 336-349, Aug. 1979.

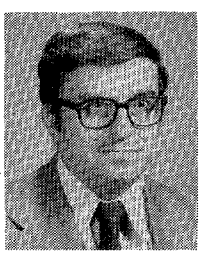
[6] J. G. Ackenhusen and L. R. Rabiner, "Microprocessor implementation of an LPC-based isolated word recognizer," in *Proc. 1980 BTL/WE Microprocessor Symp.*, Sept. 1980, pp. 35-42.

Lori F. Lamel (S'79-M'80-S'80) was born in New York on August 9, 1957. She received the B.S. and M.S. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, in June 1980.

From 1977 to 1980 she participated in a cooperative program for electrical engineering and computer science with Bell Laboratories, Holmdel and Murray Hill, NJ, where she worked on firmware for microprocessor implementation of modems, digital interface between high speed memories and a host computer for an information retrieval system, and endpoint detection for isolated word speech recognition. She has also been a participant in the Bell Laboratories Graduate Research Program for Women since 1978. She is currently a doctoral student in the Department of Electrical Engineering and Computer Science at M.I.T. and a Research Assistant in the Speech Communications Group. Her interests include speech processing and recognition, digital signal processing, and biomedical engineering.

Ms. Lamel is a member of Eta Kappa Nu and the Society for Women Engineers.

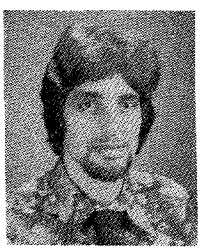
Lawrence R. Rabiner (S'62-M'67-SM'75-F'76), for a photograph and biography, see p. 297 of the April 1981 issue of this TRANSACTIONS.



Aaron E. Rosenberg (S'57-M'63) received the S.B. and S.M. degrees from the Massachusetts Institute of Technology, Cambridge, in 1960, and the Ph.D. degree from the University of Pennsylvania, Philadelphia, in 1964, all in electrical engineering.

Since 1964 he has been with Bell Laboratories, Murray Hill, NJ. He is presently engaged in studies of systems for man-machine communication-by-voice in the Acoustics Research Department at Bell Labs.

Dr. Rosenberg is a member of Eta Kappa Nu, Tau Beta Pi, and Sigma Xi, a Fellow in the Acoustical Society of America, and a member of the IEEE Acoustics, Speech, and Signal Processing Society's Technical Committee on Speech Processing.



Jay G. Wilpon was born in Newark, NJ, on February 28, 1955. He received the B.S. and A.B. degrees (cum laude) in mathematics and economics, respectively, from Lafayette College, Easton, PA, in 1977.

Since June 1977 he has been with the Acoustics Research Department at Bell Laboratories, Murray Hill, NJ. He has been engaged in speech communications research and is presently concentrating on problems of speech recognition.