METHODS OF ENDPOINT DETECTION FOR ISOLATED WORD RECOGNITION

by

Lori F. Lamel

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREES OF

BACHELOR OF SCIENCE

and

MASTER OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 1980

Signature of Author ..... Lori Faith Lamel
Department of Electrical Engineering and Computer Science, February 4, 1980.

Certified by ..... U. V. Glave
Thesis Supervisor (Academic)

Certified by ..... Lawrence Rabiner
Company Supervisor (VI-A Cooperating Company)

Certified by ..... Aaron E. Rosenberg
Company Supervisor (VI-A Cooperating Company)

Accepted by ..... Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

METHODS OF ENDPOINT DETECTION FOR ISOLATED WORD RECOGNITION

by

Lori F. Lamel

Submitted to the Department of Electrical Engineering and Computer Science on February 4, 1980, in partial fulfillment of the requirements for the Degrees of Bachelor of Science and Master of Science.

ABSTRACT

Accurate location of the endpoints of an isolated word is important for reliable and robust word recognition. The endpoint detection problem for isolated words where the background noise is stationary and low level is relatively easy to solve. However, often the beginning and end of an isolated word is obscured by speaker generated artifacts (such as mouth noises, e.g., clicks, pops, lip smackings and breathiness) as well as those introduced by the recording environment and transmission system. Several techniques for endpoint detection of isolated words recorded over a dialed-up telephone line were studied. The techniques investigated are classified as either explicit, implicit, or hybrid. The class of explicit techniques are those in which the endpoints are located prior to and independent of the recognition and decision stages of the system. For implicit endpoint estimation, the endpoints of the isolated word are determined 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 provided the best estimate of both endpoints and correspondingly the highest recognition accuracy of all three classes that were studied. A new hybrid technique for endpoint detection is proposed which reduces the recording rejection rate by 30% while maintaining the same recognition accuracy as obtained using earlier techniques with clean speech (i.e., no artifacts).

THESIS SUPERVISOR: Victor W. Zue
TITLE: Research Associate and Lecturer, Electrical Engineering and Computer Science

THESIS SUPERVISOR: Lawrence R. Rabiner
TITLE: Member of Technical Staff at Bell Telephone Laboratories

THESIS SUPERVISOR: Aaron E. Rosenberg
TITLE: Member of Technical Staff at Bell Telephone Laboratories

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to my two advisors at Bell Laboratories, Dr. Lawrence R. Rabiner and Dr. Aaron E. Rosenberg for their guidance, enthusiasm and encouragement throughout the work of this thesis. I am grateful to Dr. James L. Flanagan of Bell Laboratories for inviting me to work in his department on this project and for providing me with all the necessary facilities. I also wish to thank Dr. Victor W. Zue for being my academic thesis supervisor and for his helpful suggestions.

Thanks are also due Cory Myers for providing several dynamic time warping algorithms, and Jay Wilpon for his help in implementing a new endpoint detector.

Additional thanks are due Dr. A. E. Rosenberg, Dr. L. R. Rabiner, and Mrs. Kristin Grecco for their invaluable assistance in the final preparations of this thesis.

Special thanks are due to the Word Processing and Drafting Departments of Bell laboratories for the production of this document.

Finally, I am thankful to my parents, Howard and Sheila Lamel, for their continuing love, understanding and encouragement for all my endeavors.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

FIGURES

TABLES

Table

CHAPTER 1

INTRODUCTION

1.1 The Need for Endpoint Detection in Isolated Word Recognition Systems

Automatic speech recognition technology has advanced to the point where it is now practical to interface a human to a machine with limited capability for voice input and voice response output. Several commercial systems of this type have been developed and successfully applied to applications such as voice entry of data, voice controlled sortation, and voice control of equipment, [1]. More experimental laboratory systems have been developed for use in applications such as directory assistance [2], voice activated airline reservation and information systems [3], and repertory dialing [4].

The heart of most of these systems is an isolated word recognition system. Figure 1.1 shows a block diagram of such an isolated word recognition system based on pattern recognition techniques. In the preprocessor the input speech is bandpass filtered, digitized, and blocked into frames. The digitized waveform is then analyzed on a frame-by-frame basis to locate the beginning and end of the spoken isolated word. Accurate location of the endpoints of an isolated 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.

Further discussion of the problems associated with endpoint detection is deferred to the next section. The next step in the recognition process is the measurement of recognition features for each frame within the detected word. Typical features chosen to represent the speech waveform include both time and frequency domain measurements. Common parametrizations are energy, zero crossing rates, minimum and maximum amplitude excursions, spectral and cepstral coefficients, and linear prediction (LPC) parameters and their transformations. The parameters selected must be relatively invariant to intra-speaker differences for a system that is

trained to the individual talker, and to both intra and inter-speaker variability for a speaker independent system. The parametric representation of the speech must provide enough information to characterize the words and to differentiate between acoustically similar words. The time sequence of the features, computed once per frame, is defined as the test pattern. Recognition is accomplished by comparing the test pattern to a set of previously stored reference templates and selecting the "closest match" as the recognized word. In general the time scales of the test pattern and reference pattern are not in alignment. Hence some type of time warping algorithm is needed to compensate for the temporal variations in spoken words by optimally aligning the test pattern to each reference pattern during the comparison. A distance measure is computed for the comparison between the test pattern and each reference. A decision rule is then used to determine an ordered list of recognition candidates for the spoken word.

The recognition system of Figure 1.1 will be used extensively in the work described in this thesis. It provides a quantitative measure of the success with which the endpoints are detected through the recognition accuracies. Feedback may be provided to the endpoint detector if no good recognition candidates are found and an alternate set of endpoints located. These issues will be discussed more fully in the next section.

1.2 The Problems Associated with Endpoint Detection

Isolated word recognition is based on the premise that if an utterance consists of an isolated word, preceded and followed by silence or other background noise, the speech segments can be reliably separated from the nonspeech segments. The separation of the speech segments of an utterance from the background; i.e. the nonspeech segments obtained during the recording process, is known as endpoint detection. As mentioned above, accurate detection of the endpoints of an isolated utterance is critical for reliable recognition and minimizes the amount of computation necessary for recognition.

The location of the beginning and end of an isolated utterance in a clean environment may be determined by the use of a simple energy threshold test and some related tests based on temporal constraints of the words [5]. A clean environment is one in which the energy of the

weakest speech sounds, such as weak fricatives, is significantly greater than the background energy and the recorded interval does not contain any artifacts. However, such recording environments are not common for most of the practical applications of automatic speech recognition. Thus the more realistic problem of detecting endpoints in an environment in which the background noise levels are at least commensurate with the energy levels of weak speech sounds will now be discussed.

In such cases the problem is to find the best estimate for the endpoints of the spoken word in the presence of any background noise. In the literature there have been documented several approaches to this problem. For example Rabiner and Sambur [6] proposed a double thresholding technique which used energy and zero crossing measurements to locate the endpoints of isolated utterances in a background with a signal to noise ratio of at least 30 dB. In this method the energy measurements provided an initial estimate of the endpoints, and the zero crossing measurements were used to detect the presence of initial or final fricatives and to adjust the endpoints accordingly. Other suggested endpoint detection schemes include voiced-unvoiced-silence discrimination [7], and the location of energy dips [8].

The techniques for determining the endpoints of isolated utterances mentioned above assume reasonably clean isolated words with reliable statistical properties and stationary, low level background noise. This thesis will be concerned with transient background noise problems associated with the speaker and/or the transmission environment. This type of background noise complicates the endpoint detection problem considerably. For example, the beginning or end of an isolated word is often obscured by speaker generated artifacts such as mouth noises, e.g. clicks, pops, lip smackings and breathiness, as well as those introduced by the transmission system, e.g. transients in the telephone system. Sometimes nonspeech events occur close enough to the actual word and are of sufficient energy to be included as part of the utterance. The types of problems encountered are illustrated in Figure 1.2 which shows plots of the log energy of a recorded signal versus time for four examples. Figure 1.2a shows the case of a relatively clean utterance (no user or system generated artifacts) with the endpoints correctly located (by an energy endpoint detector to be described in Section 2.3.1) as shown by

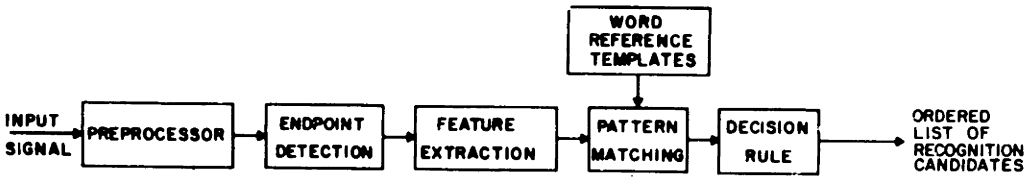


Figure 1.1 Block diagram of an isolated word speech recognition system.

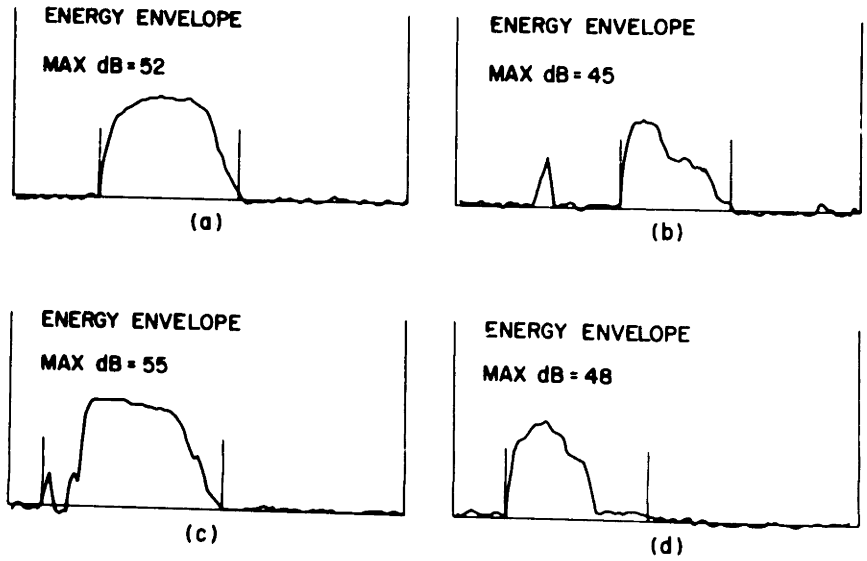


Figure 1.2 Examples of endpoint detection problems.

the vertical lines. Despite the artifact prior to the word in Figure 1.2b the endpoints have been accurately located. In Figure 1.2c the artifact prior to the spoken word is included as part of the word. Finally, the breathiness at the end of the word in Figure 1.2d has also been mistakenly included within the endpoints. As illustrated in this figure, a number of problems can occur to make accurate detection of the endpoints of an isolated word a difficult task for realistic recording conditions. The purpose of this thesis is to investigate a variety of approaches to this problem, and to determine the relative performance of several techniques in the presence of the types of problems shown in Figure 1.2.

Figure 1.3 illustrates the approaches to the endpoint detection problem that were considered. The approaches may be broadly classified as either implicit, explicit or hybrid. The class of explicit techniques, as shown in Figure 1.3a, consists of systems in which the endpoint detection precedes and is *independent* of the recognition and decision stages of the recognizer. Typically a set of features are extracted from the input data and the beginning and ending points of the speech are determined and sent in a feed forward manner to the next stage of the system. Various feature sets and endpoint detection schemes of the explicit type were implemented and are discussed in Chapter 2.

Figure 1.3b shows the canonic form for the class of implicit techniques for endpoint detection. In a purely implicit technique, the endpoints of the isolated word are determined by the recognition and decision phases of the automatic speech recognition system, i.e. there is no separate stage for endpoint detection. An algorithm of this type would attempt recognition of the test from all (or possibly a large set of) possible starting and ending locations. The use of implicit analysis techniques for endpoint detection is discussed further in Chapter 3.

The canonic form for the class of hybrid techniques for endpoint detection is shown in Figure 1.3c. The hybrid techniques incorporate aspects from both the explicit and implicit methods discussed above. In these techniques, one or several estimates of the endpoints are obtained from the temporal variation of the features derived from the speech, and passed to the recognition stages. Based on information obtained during the recognition stage, a set of end-

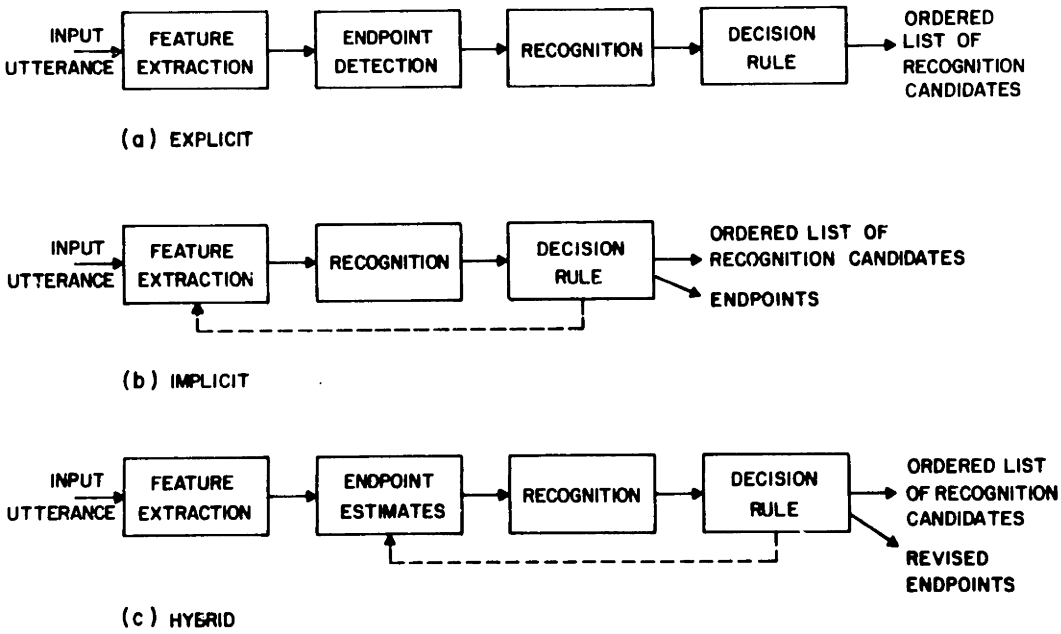


Figure 1.3 Techniques for endpoint detection.

points are selected from the initial estimates or revisions of the initial estimates. The hybrid techniques may be implemented as a feedback loop within the overall recognition system. The use of hybrid techniques for endpoint detection are discussed in Chapter 4.

1.3 Organization of the Thesis

In Chapter 2, the use of explicit techniques for endpoint detection are described. The data base used for testing the various algorithms is defined. Also defined are the notions of test, reference and feature vectors. The use of implicit methods of locating the endpoints are discussed in Chapter 3. Chapter 4 deals with the use of hybrid techniques for endpoint detection. Included in each chapter are performance measures of the endpoint detection schemes discussed along with the associated recognition accuracies. A summary as well as suggestions for further investigation in isolated word endpoint detection are presented in Chapter 5.

CHAPTER 2

EXPLICIT ANALYSIS TECHNIQUES FOR ENDPOINT DETECTION

2.1 Introduction

An explicit technique for estimation of the endpoints of an isolated word is one in which the beginning and end of the test word are determined prior to and independent of the succeeding stages of the recognition system. As shown in Figure 1.3a, the canonic representation of a recognition system using explicit endpoint detection is strictly that of a feed forward system. The speech and nonspeech portions of the recorded data are distinguished using some explicit measure, and the endpoints of the isolated word are located so as to include the speech events. A *single* set of endpoints are determined and supplied to the recognizer. In this chapter several separate but related techniques for the explicit detection of endpoints are proposed and evaluated.

The organization of this chapter is as follows. In Section 2.2 the temporal and the spectral properties of humanly generated artifacts (such as mouth clicks, pops) are examined and their similarity to the properties of normal speech sounds is shown. In Section 2.3 a consistent and uniform notation for the endpoint detection problem is introduced. Sections 2.4 and 2.5 describe several techniques for explicit endpoint estimation. The data base created for testing the endpoint detection algorithms presented in this chapter, and in Chapters 3 and 4, is described in Section 2.6. In Section 2.7 the recognition results for the endpoint detectors described in this chapter are given, along with some representative examples showing how the various detectors estimated the endpoints. Finally, in Section 2.8 the utility of explicit analysis techniques for endpoint detection is discussed.

2.2 Spectral and Temporal Properties of Artifacts

Accurate detection of the endpoints of isolated words can become a difficult task in the presence of high levels of background noise and/or transients or artifacts. The artifacts may be humanly generated, e.g., mouth clicks and breathiness, or introduced by the transmission

system, e.g., line transients. In order to distinguish the nonspeech sounds from the speech sounds, it is necessary to examine the temporal and spectral properties of the artifacts, and compare them to the temporal and spectral properties of typical speech sounds. In this section some examples of the types of artifacts encountered in word recognition tasks are given.

Several examples of recorded utterances containing both an isolated word and an artifact are shown in Figures 2.1-2.3, which give plots of the log energy contour and the time waveform for these examples. The log energy contour for a recording of the letter "K" is shown in Figure 2.1a. As shown in the figure, prior to the actual word there is a click. The corresponding acoustic waveform for this example is shown in Figure 2.1b. A sampling rate of 6.67 kHz is used in all of the examples. Hence, each line of the waveform corresponds to 1000 samples or 150 msec of data. Figure 2.2a shows another example of the log energy contour of an isolated word with an artifact. In this case the energy contour is for the letter "E". The breathiness following the word is denoted by the shaded region. The time waveform for this recording is shown in Figure 2.2b with the breathy portion appropriately labeled. As a final example the log energy contour of the word "three" followed by a mouth click is shown in Figure 2.3a. The acoustic waveform for this recording is given in Figure 2.3b. Figures 2.1-2.3 illustrate the types of artifacts commonly encountered in practical recording situations with inexperienced talkers. The artifacts may be roughly divided into two categories; short, mid-to-high energy artifacts such as clicks and pops, and longer duration, lower energy artifacts such as breath noise. The duration of a click is seen to be on the order of 10 to 30 ms, whereas the breath noise may last for 100 to 200 ms and even longer. The peak energy for the click is seen to be on the order of mid-energy speech sounds. The breath noise is generally of low amplitude, consisting mainly of energy above 2000 Hz.

Since it is necessary to be able to distinguish the artifacts from normal speech sounds some of the spectral properties of typical speech sounds are shown. Figures 2.4 and 2.5 show plots of the time waveforms and spectra of portions of voiced and unvoiced speech respectively. The amplitude of the waveforms have been scaled, thus it is important to note the magnitude given in the plots. The top plot in each part of the figure is of a 300 sample Hamming windowed

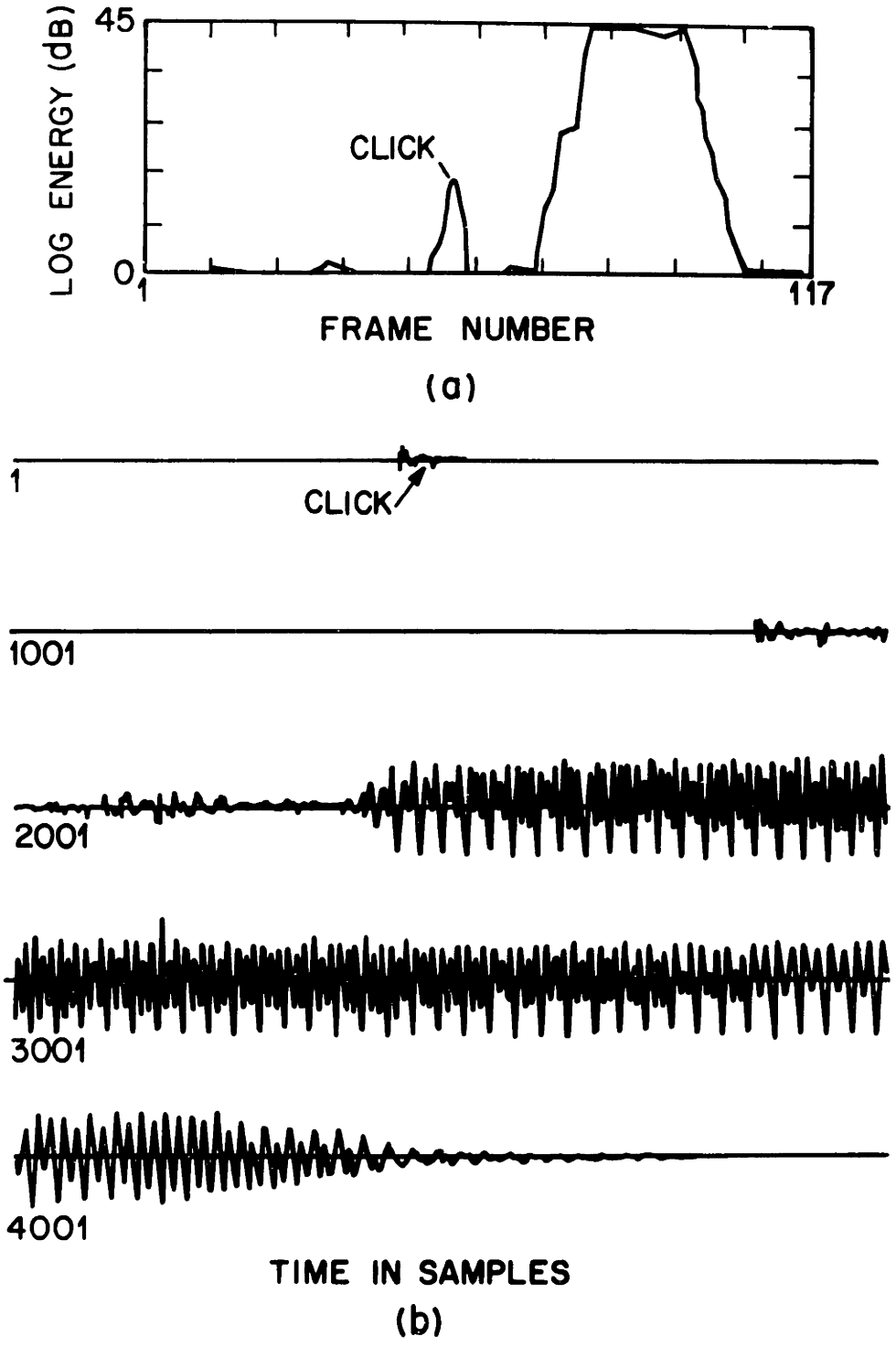
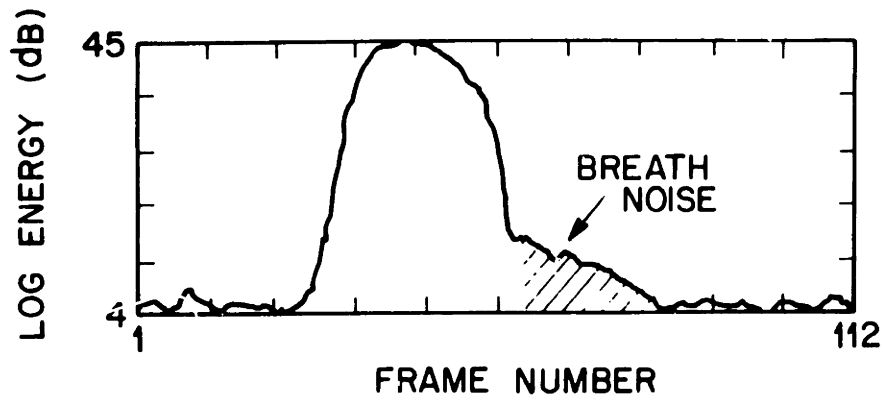
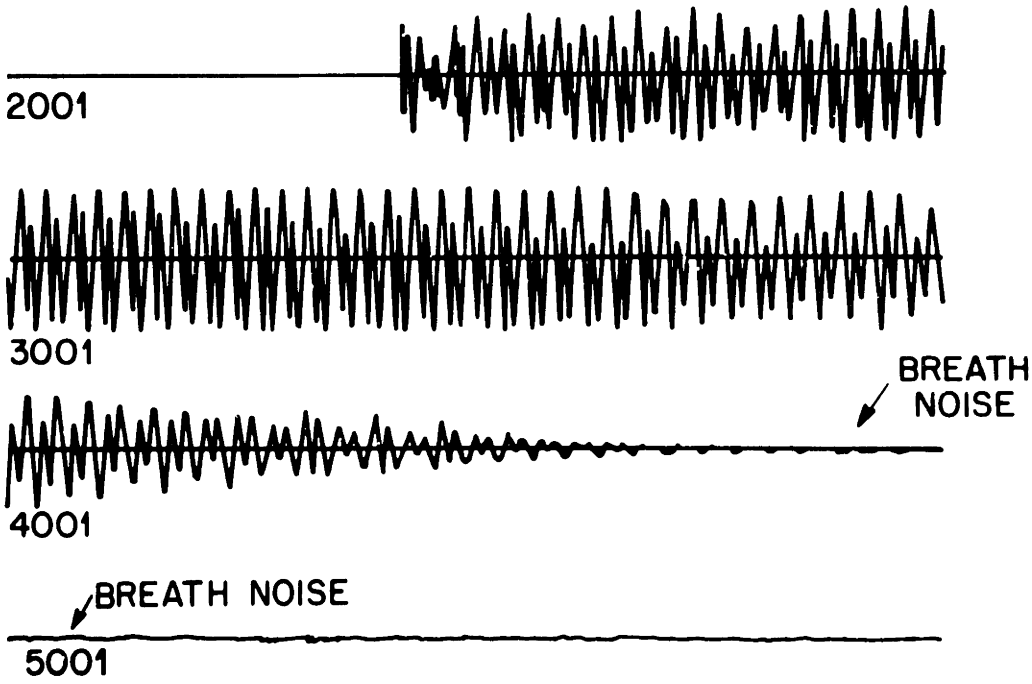


Figure 2.1 Example of click prior to isolated word.

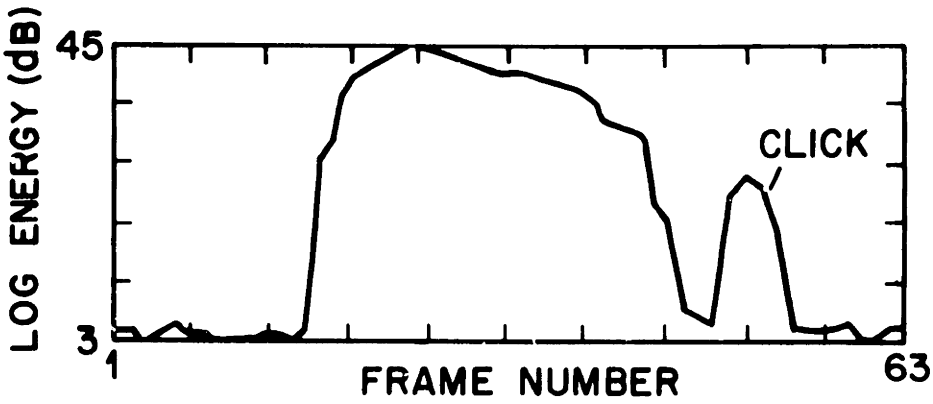


(a)

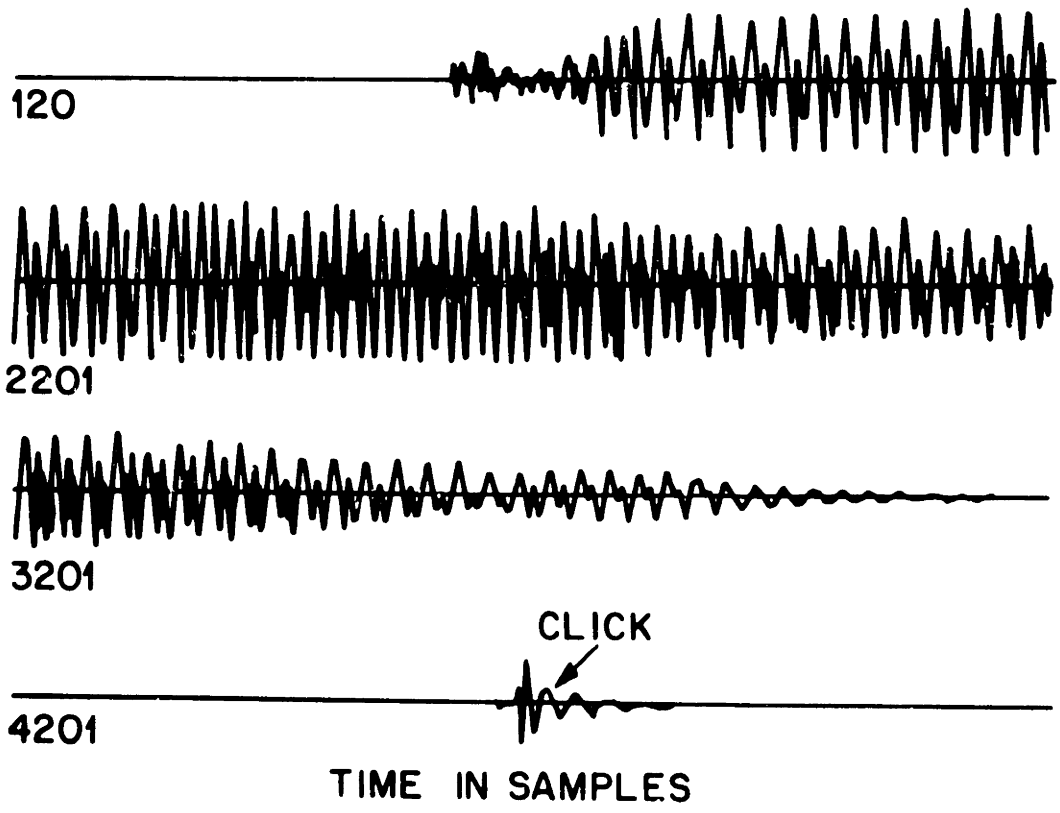


(b)

Figure 2.2 Example of breath noise.



(a)



(b)

Figure 2.3 Example of click following isolated word.

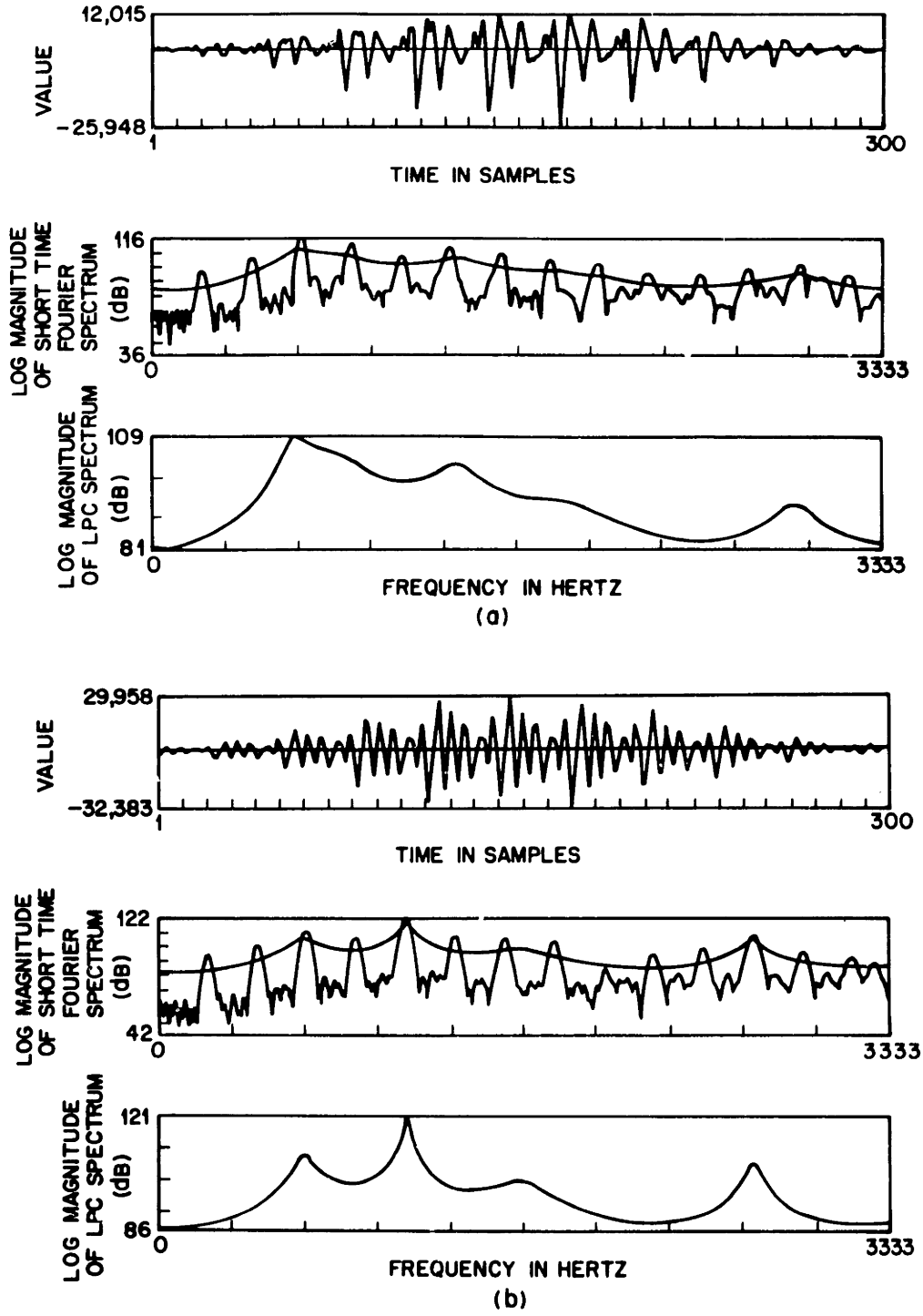


Figure 2.4 Spectral properties of examples of voiced speech.

segment of speech. The middle plot is the short-time Fourier spectrum of the speech segment after it has been pre-emphasized by a simple first order network. Superimposed on the short time spectrum is the 12 pole linear predictive coding (LPC) spectrum [9]. The bottom plot in each group is the LPC spectrum plotted on an expanded scale. In Figure 2.4a the voiced speech is a segment of the phoneme /*l*/ from the letter "L"; in Figure 2.4b it is a segment of the diphthong /*a*^{I}/ from the letter "I". The periodic structure of the voiced speech is seen clearly in both the time domain and also in the short-time spectra. It can be seen that the frequency components in the range 500-1000 Hz of the speech are approximately 20 dB above the higher frequency components (those above 1500 Hz). Figure 2.5 shows a similar set of plots for a segment of the unvoiced fricative /*s*/. The lack of a regular periodic structure is seen in both the time and frequency domain and the high frequency components (above 2000 Hz) are approximately 10 dB *above* the lower frequency components (below 1500 Hz). In Figure 2.6 the corresponding plots are shown for a portion of silence, where silence is defined to be the background noise of the recording environment. The maximum amplitude of the silence is approximately 30 dB below the unvoiced speech, and 30 to 60 dB below the voiced speech. Figure 2.7 gives a similar set of plots for a typical click. The peak amplitude of the click is on the order of the amplitude of the unvoiced fricative /*s*/; however, the duration of the click is less than 100 ms. The short-time spectrum of the click (Figure 2.7b) is extremely smooth and the LPC spectrum has a predominance of frequency components above 2000 Hz. As a final example, Figure 2.8 shows the corresponding plots for a typical example of breath noise. The spectra are seen to be comparable to those of the unvoiced fricative /*s*/ of Figure 2.5.

The similarity between the speech and nonspeech sounds complicates the problem of endpoint detection. From the above examples it can be seen that the spectra and amplitude variations of the artifacts can be very close to those of unvoiced fricatives or the releases of stop consonants. Hence, the explicit techniques discussed in this chapter must attempt to use fairly sophisticated methods relying on information about both the temporal and spectral properties of the artifacts to reliably distinguish them from normal speech sounds.

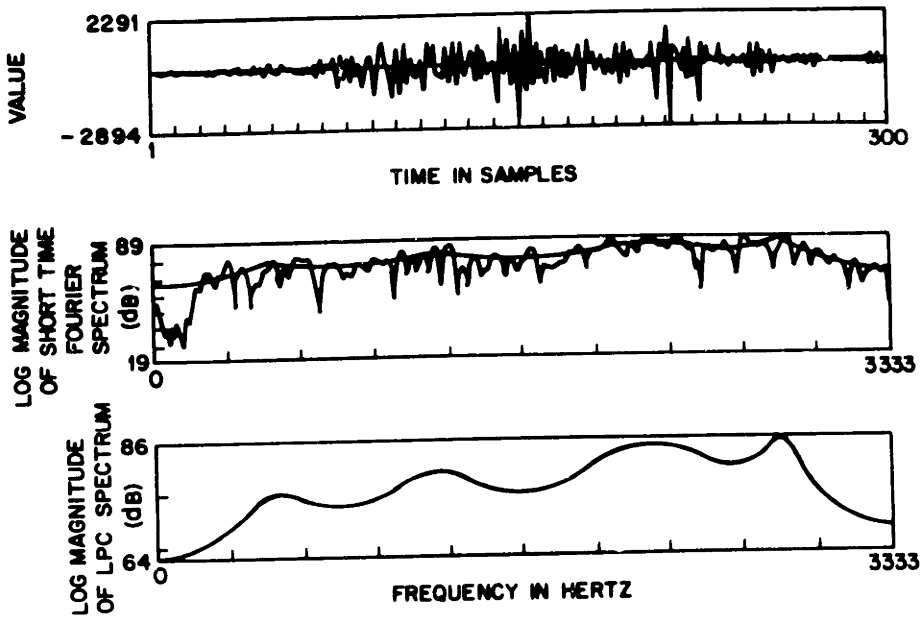


Figure 2.5 Spectral properties of an example of unvoiced speech.

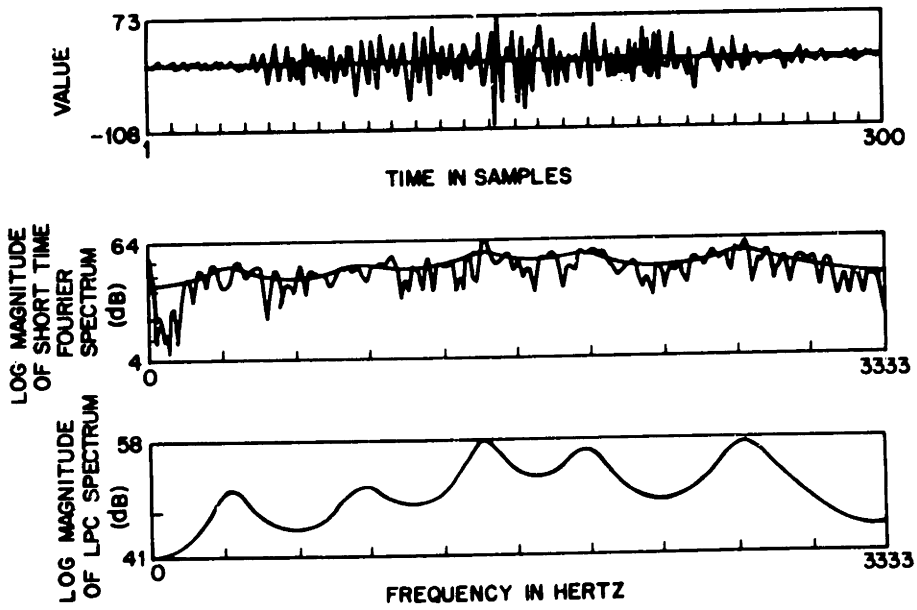


Figure 2.6 Spectral properties of an example of background silence.

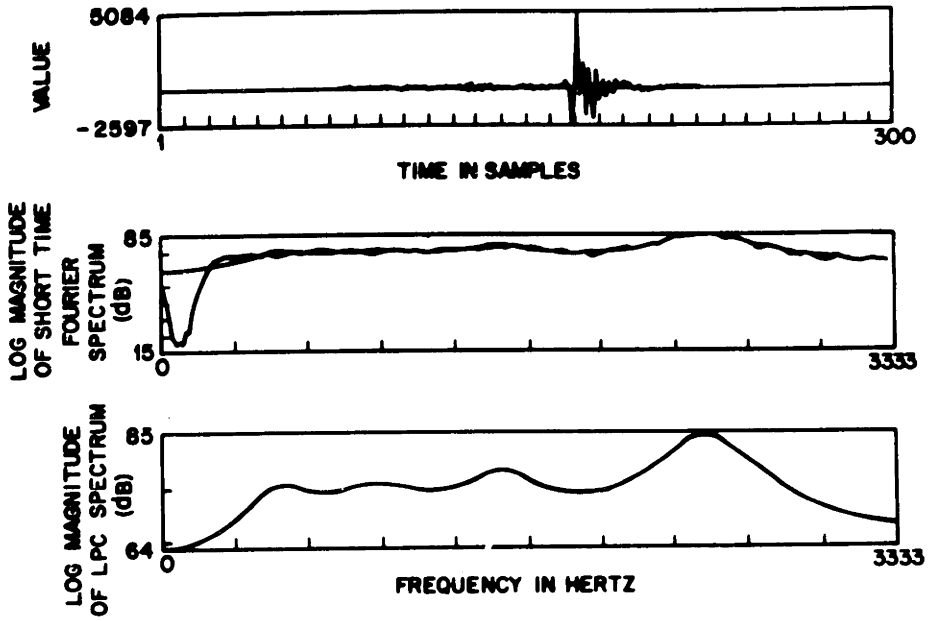


Figure 2.7 Spectral properties of a typical click.

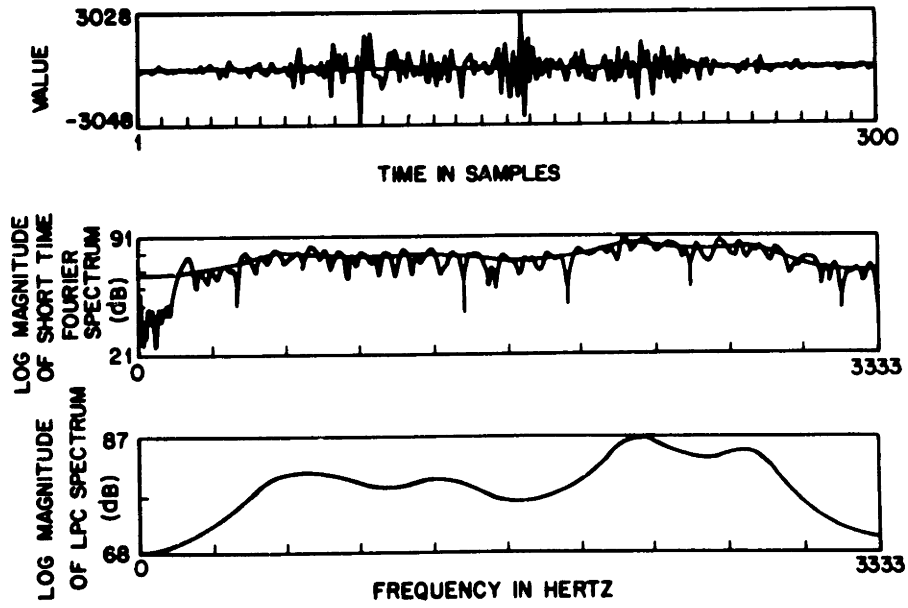


Figure 2.8 Spectral properties of an example of breath noise.

2.3 Notation for Endpoint Detection

In this section the notation and concepts to be used throughout the thesis are defined. However, first the operations denoted by the *preprocessor* in Figure 1.1 will be described. A detailed block diagram of the preprocessor is shown in Figure 2.9a. The input speech, recorded using a dialed-up telephone line, is bandpass filtered from 100 to 3200 Hz and sampled at a 6.67 kHz rate. The digitized speech is pre-emphasized using a first order digital filter with a system function.

$$H(z) = 1 - az^{-1}, \quad a = 0.95 \tag{2.1}$$

and blocked into overlapping frames of $N = 300$ samples (45 ms) with an offset between frames of 100 samples (15 ms). Each frame is then weighted by a Hamming window given by,

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N}\right), \quad 0 \leq n \leq N-1 \tag{2.2}$$

(where N is the number of samples per frame) to reduce the effects of the blocking. These operations are common to most of the systems to be discussed and when used will be denoted by the block labeled preprocessor in Figure 2.9b.

At the output of the preprocessor, the speech has been pre-emphasized and blocked into overlapping frames. A frame-by-frame sequence of features is measured from the blocked speech and denoted by the feature vector

$$T_l, \quad l = 1,L \tag{2.3}$$

where L is the length of the recording interval in frames. T_l is the feature vector of the l^{th} frame, and the time course of T_l for the appropriate range of l will be called the test pattern. The test pattern used during the recognition process is the set of normalized *pth* order auto-correlations of the speech calculated for each frame. The test pattern is discussed further in Section 3.2. The reference set of words is denoted as

$$\hat{R}_k(w), \quad w = 1,W, \quad k = 1,K_w \tag{2.4}$$

where W is the number of words in the reference set and K_w is the number of frames in the

w^{th} word. The reference set used consists of speaker independent reference templates created by statistical clustering techniques [14]. The output of the endpoint detector is the feature vector $\hat{T}_{\hat{l}}$, a subset of the feature vector T_l, such that:

$$\hat{T}_{\hat{l}} = T_{B+\hat{l}-1} \quad \text{for } \hat{l} = 1,\hat{L},\ \hat{L} \leq L \tag{2.5}$$

where $\hat{L} = E - B + 1$ and B and E are the beginning and ending frames, shown in Figure 2.10, as determined by the endpoint detector.

The endpoint detector is evaluated on the basis of two performance 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 determined from the time sequence of samples, the log energy contour, or some other measurement is subject to errors, even given the knowledge of what word was spoken. Fortunately, the continuity of speech and the inherent redundancy in speech allow the endpoints to vary by small amounts (perhaps 1 to 3 frames), while retaining the necessary acoustic information. The performance measures of the endpoint detectors are discussed in the last section of each chapter.

2.4 Detectors for Explicit Estimation of Endpoints

In this section two explicit techniques for the estimation of the endpoints of isolated words are presented. In Section 2.4.1 the energy based endpoint detector in use at Bell Laboratories is described. In Section 2.4.2 a new endpoint detector is proposed in which the word endpoints are obtained based on comparisons of the test pattern to a prescribed silence pattern.

2.4.1 Description of the Current Energy Endpoint Detector

This section describes the energy based endpoint detector in use at Bell laboratories

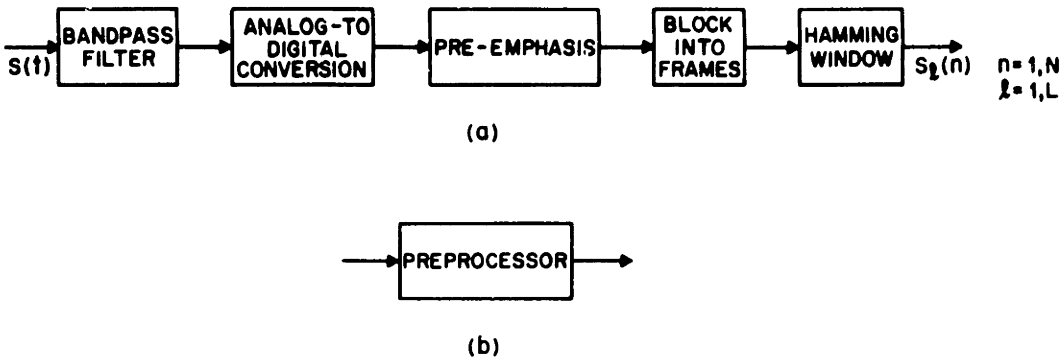


Figure 2.9 Canonic representation of preprocessor.

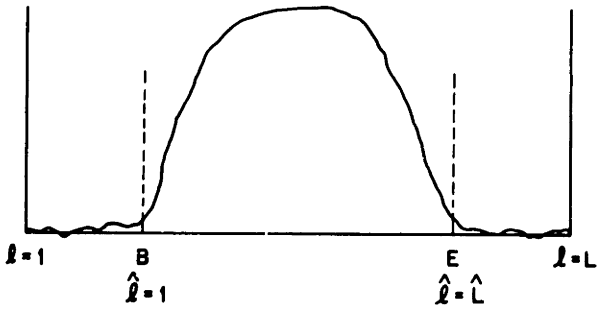


Figure 2.10 Specification of notation for endpoint detection.

[10] at the onset of this investigation. Although the development and implementation of the current energy endpoint detector were not part of the work of this thesis, it is presented in detail in this section for several reasons. First, the current energy endpoint detector is used as the basis of comparison for performance of all the other endpoint detectors investigated in this work. The goal was to develop a robust, reliable endpoint detector with performance better than the current energy endpoint detector. Second, the operation of the current energy endpoint detector is not documented in the literature; thus, the comprehensive description given here is necessary and appropriate. In addition, a formal evaluation of the current energy endpoint detector was included within the thesis work. The current energy endpoint detector was tested on the standard-test set used in this investigation. Finally, the modified energy endpont detector presented in Chapter 4 was developed from the study of the current energy detector. Some of the operations performed by the modified energy detector are simply modifications of the current energy endpoint detector. The overall conceptual basis for the two endpoint detectors is similar. However, the realizations are different.

Following the preprocessing operations discussed in Section 2.2, an eighth order autocorrelation analysis is performed for each frame of data, $s_l(n)$, $0 \leq n \leq N-1$, $1 \leq l \leq L$, giving the autocorrelation vector

$$R_l(m) = \sum_{n=0}^{N-1-m} s_l(n)s_l(n+m), \quad l=1,L,\ m=0,p,\ p=8,\ N=300 \tag{2.6}$$

where $s_l(n)$ is the pre-emphasized, windowed speech of frame l, m is the correlation delay, N is the number of samples per frame, and l is the frame number.

The current energy detector determines the endpoints of the word from the time sequence of the speech energy, as determined by the time sequence of the zeroth order autocorrelation. A block diagram of the current energy endpoint detector is given in Figure 2.11. The endpoint detector consists of three basic units; a second level preprocessor, an energy pulse detector, and a discriminator.

The second level preprocessor, as shown in Figure 2.12, creates a normalized level array of

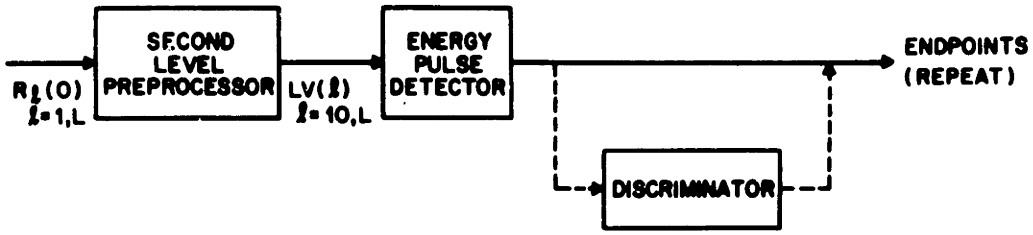


Figure 2.11 Block diagram of current energy endpoint detector.

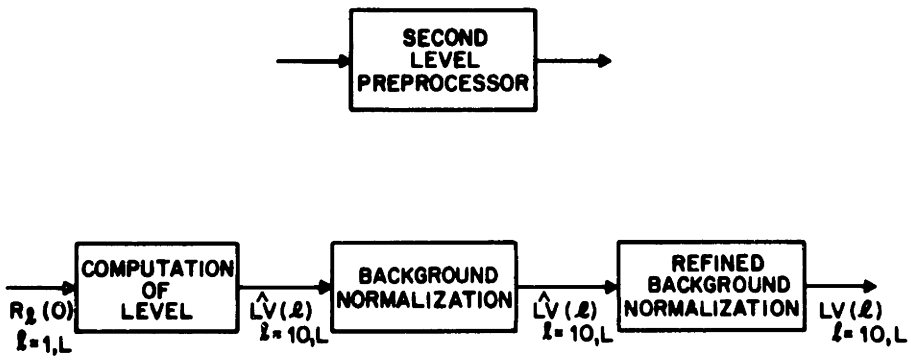


Figure 2.12 Block diagram of second level preprocessor.

the log energy of the signal as a function of time (frame number). The first nine frames of the recorded interval are ignored to eliminate any onset transients in the recording system. An integer valued level array is created from the zeroth order autocorrelation according to the equation

$$\hat{LV}(l) = \lfloor 10\log_{10} R_l(0) + 0.5 \rfloor, \quad l = 10,L \tag{2.7}$$

where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x. The level array $\hat{LV}(l)$ is ordered on a *dB* scale. Hence a 3 *dB* change in level represents a 2 to 1 change in intensity. The level array is normalized by *LVMIN*, where

$$LVMIN = \min_{l=10,L} \hat{LV}(l) \tag{2.8}$$

giving

$$\tilde{LV}(l) = \hat{LV}(l) - LVMIN, \quad l = 10,L \tag{2.9}$$

The effect of subtracting the level *LVMIN* in Eq. (2.9) is to normalize the LEVEL values with respect to the background level. The last step in the second level preprocessor is to refine the estimate of the average background noise by computing a histogram of the lowest ten values of the level array (values 0-9) and normalizing the level array by the mode of the (three point median) smoothed histogram. The final output is given as

$$LV(l) = \tilde{LV}(l) - LVMODE, \quad l = 10,L \tag{2.10}$$

where *LVMODE* is the mode of the smooth histogram.

The reason for using the computation of Eq. (2.10) is illustrated in Figure 2.13 in which the level array $\tilde{LV}(l)$ is shown as a function of l for 2 examples. In the first case (Fig. 2.13a) a single isolated frame has very low level causing *LVMIN* to be too small. In the second case (Fig. 2.13b) a small region has low level, again causing *LVMIN* to be too small. For such cases, the computation of Eq. (2.10) provides a reliable correction to the background level of the level array.

Following the normalization, a double thresholding technique [6], is used to locate all

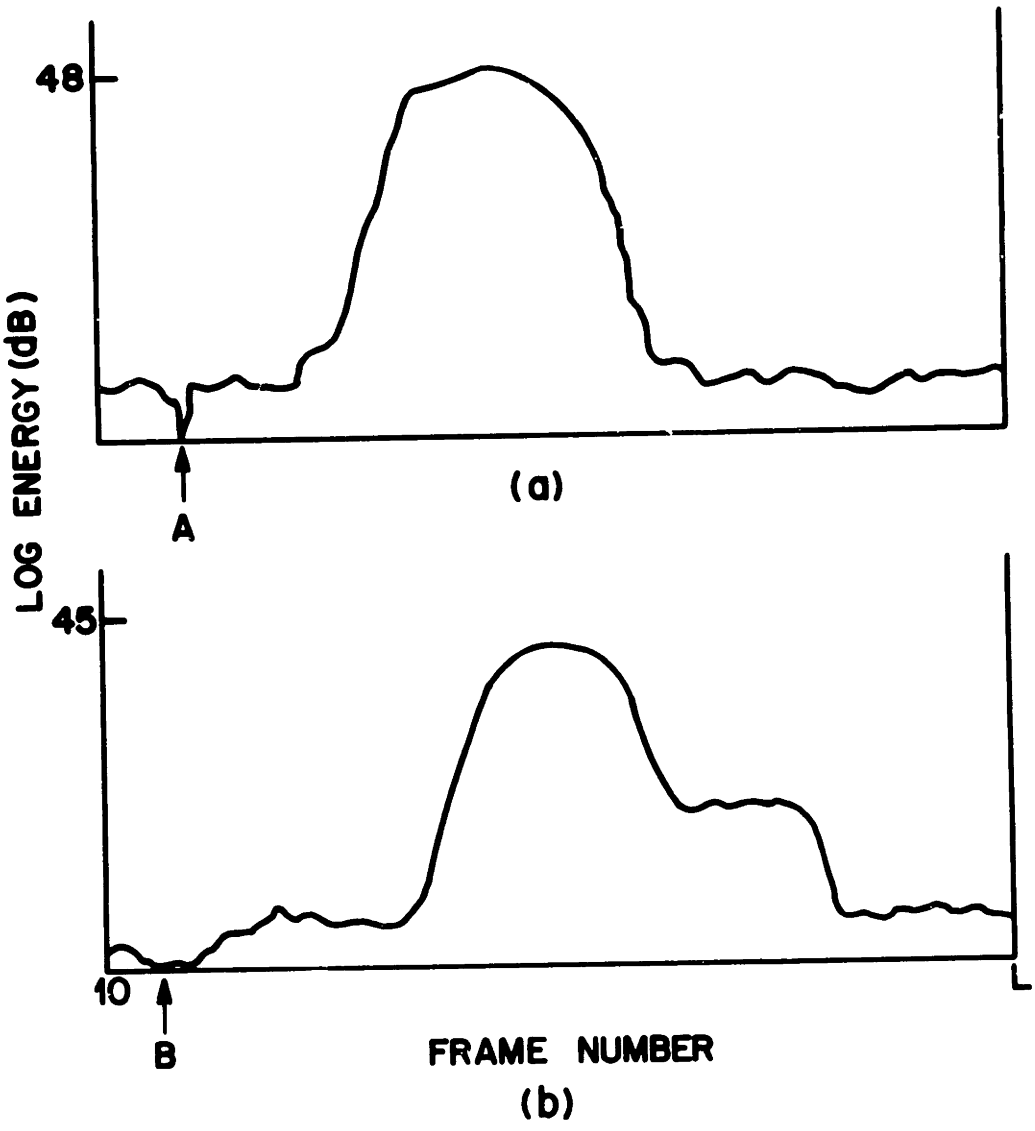


Figure 2.13 Illustration of the need for refined background normalization.

potential word boundaries. Figure 2.14a illustrates the double thresholding technique for an isolated word spoken in a relatively clean recording environment (i.e., low background noise). The thresholds K_1 and K_2 have been empirically set to 3 and 8 (dB) (above the threshold level of 0 dB) respectively. The beginning point of the word is defined to be the frame prior to the frame in which the level first exceeds the lower threshold K_1, followed by one or more frames in which the level value exceeds the higher threshold K_2, before falling below K_1 again. The ending frame is that frame first falling below K_1 following frames with level values exceeding K_2. Thus, in Figure 2.14a the word endpoints would be located at A_1 and A_4 as denoted by the heavy vertical lines. The region between A_1 and A_4 is defined as an energy pulse.

A flow chart for the double thresholding technique for locating the beginning point is shown in Figure 2.14b. The same algorithm may be applied working backwards from the ending frame to locate the endpoint of the word. The following set of relations mathematically summarize the technique.

Let B be defined as the beginning frame. Then $B = l$, if for some l, $l < L$

$$LV(B) < K_1 \tag{2.11}$$

and for some C, $B < C < L$,

$$LV(C) > K_2$$

and for all l, $B + 1 \leq l < C-1$,

$$K_1 < LV(l) < K_2$$

The corresponding ending point E is defined as $E = l$, if for some l, $l \leq L$

$$LV(E) < K_1 \tag{2.12}$$

and for some D, $D < E$,

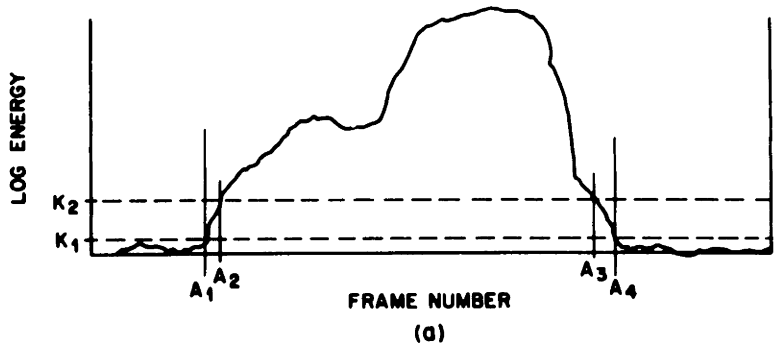
$$LV(D) > K_2$$

and, for all l, $D < l < E$,

$$K_1 < LV(l) < K_2.$$

The state diagram shown in Figure 2.14c summarizes these constraints.

The energy pulse detector iteratively applies the double thresholding technique described above as shown in the simplified flow chart of Figure 2.15a. In this manner all possible sets of beginning and ending frames based on the energy contour are found. The flow chart of the actual implementation of the energy pulse detector [10] is given in Figure 2.15b. The energy pulse detector searches the level array from frame ten to frame L and creates an array of the



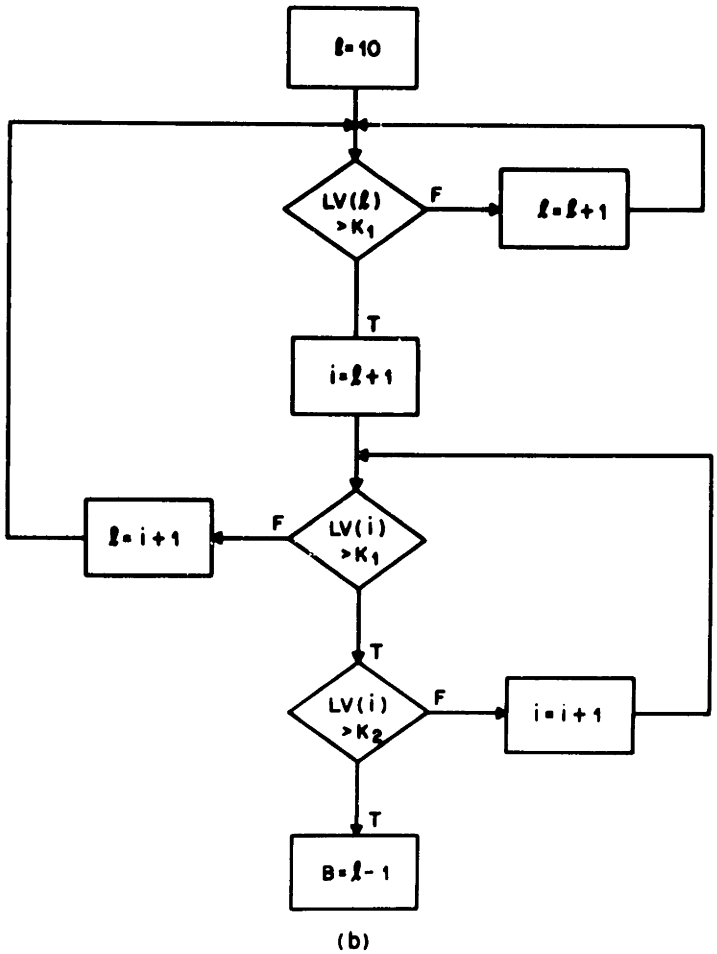


Figure 2.14 Illustration of double thresholding technique for isolated word endpoint detection.

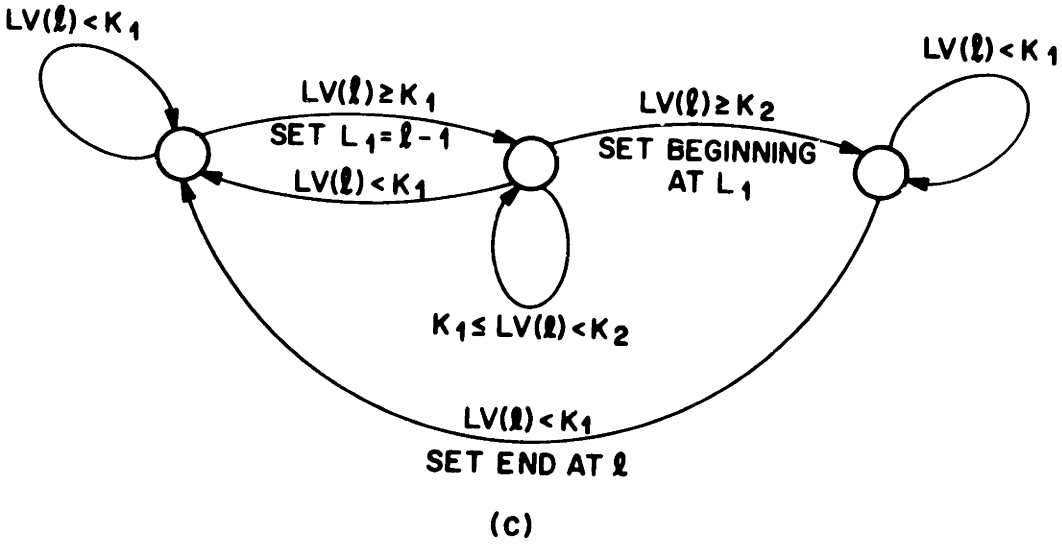


Figure 2.14 continued

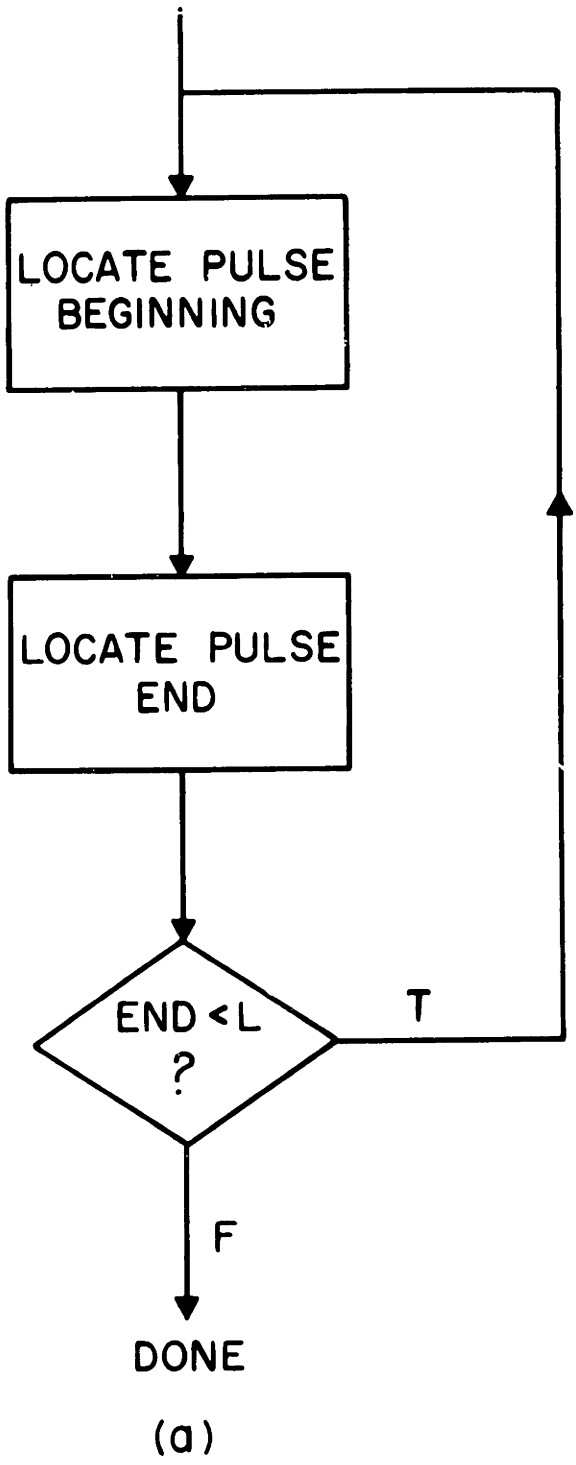
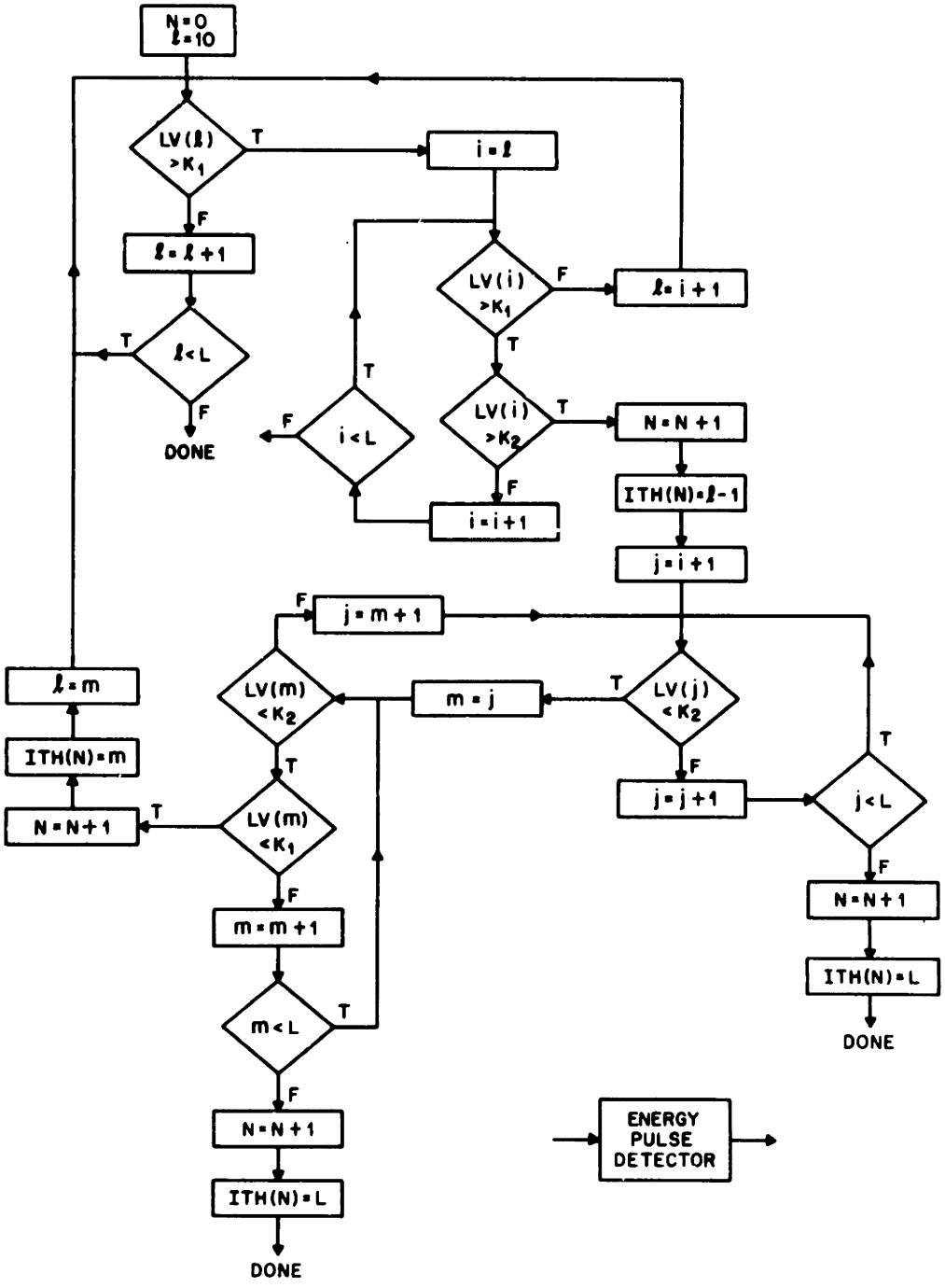


Figure 2.15 Flow chart of pulse detector.



(b)

Figure 2.15 continued

endpoints of all energy pulses. The endpoints of energy pulses are required to be found in pairs. Thus first a starting frame is found and then an ending frame. Figure 2.16 gives several examples of the energy pulses located by the energy pulse detector. In Figure 2.16a there is one clean pulse and the endpoints are located at A_1 and A_2. Two energy pulses are found in Figure 2.16b, one with endpoints A_1 and A_2, the other with endpoints A_3 and A_4. If the detected beginning or ending point occurs within LTOL frames (empirically set at 2) of the first or last frame of the recorded interval, the recording is rejected and a repeat requested. These conditions, called beginning fault and end fault are illustrated in Figure 2.16c and Figure 2.16d respectively. These cases represent recordings where the spoken data (or possibly some interference) is too close to the beginning or ending of the recording interval to reliably find the endpoints.

If a single pulse is detected by the algorithm described above, the endpoints are passed to the recognition stage and the endpoint detection is finished. In the case where multiple energy pulses have been found further discrimination must be performed. In the example of Figure 2.16b two energy pulses have been found. Visually it is apparent that the first pulse with endpoints A_1 and A_2 is an artifact and that the spoken word has endpoints A_3 and A_4. However, multiple pulses also appear in words like "stop", "repeat", and "eight". The discriminator, shown in Figure 2.11 must determine which pulses comprise the word and which are due to other factors such as artifacts.

The discriminator estimates the starting and ending frames of the isolated word from the array of energy pulses formed by the energy pulse detector. For each pair of pulses in the array, the variables *PULS*1, *PULS*2 and *GAP* are computed, as defined in Figure 2.17. The discriminator applies heuristics on the durations of the pulses and the gaps in order to determine the appropriate set of endpoints. A flow chart of the discriminator is given in Figure 2.18 [10]. The portion denoted Q is used to discriminate between two pulses separated by a medium sized gap (a gap of 6 to 9 frames). This condition occurs often in words like "H" and "X". The heuristics applied are summarized in Table 2.1, where *PULS*1, *PULS*2 and *GAP* are as defined in Figure 2.17.

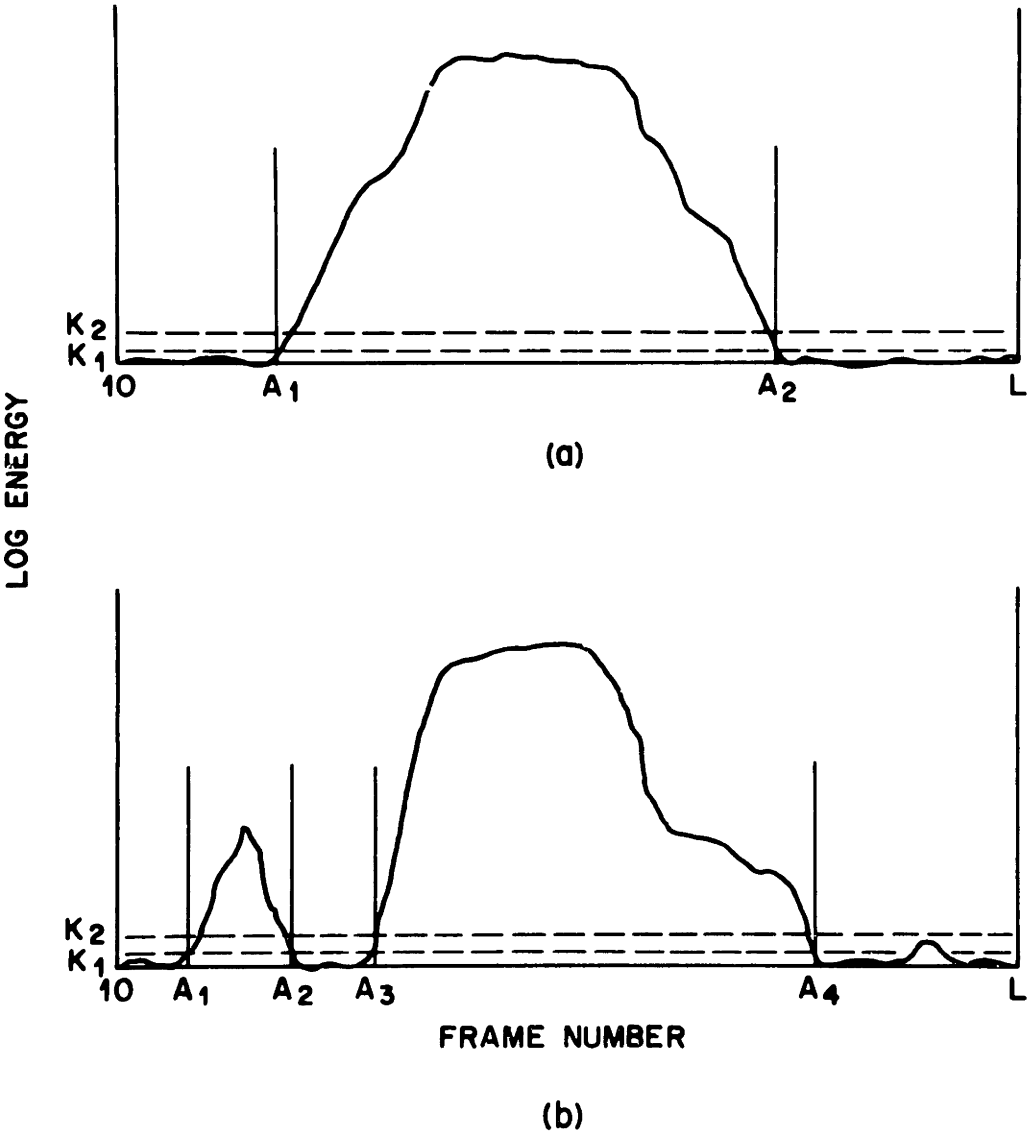


Figure 2.16 Examples of detected pulses.

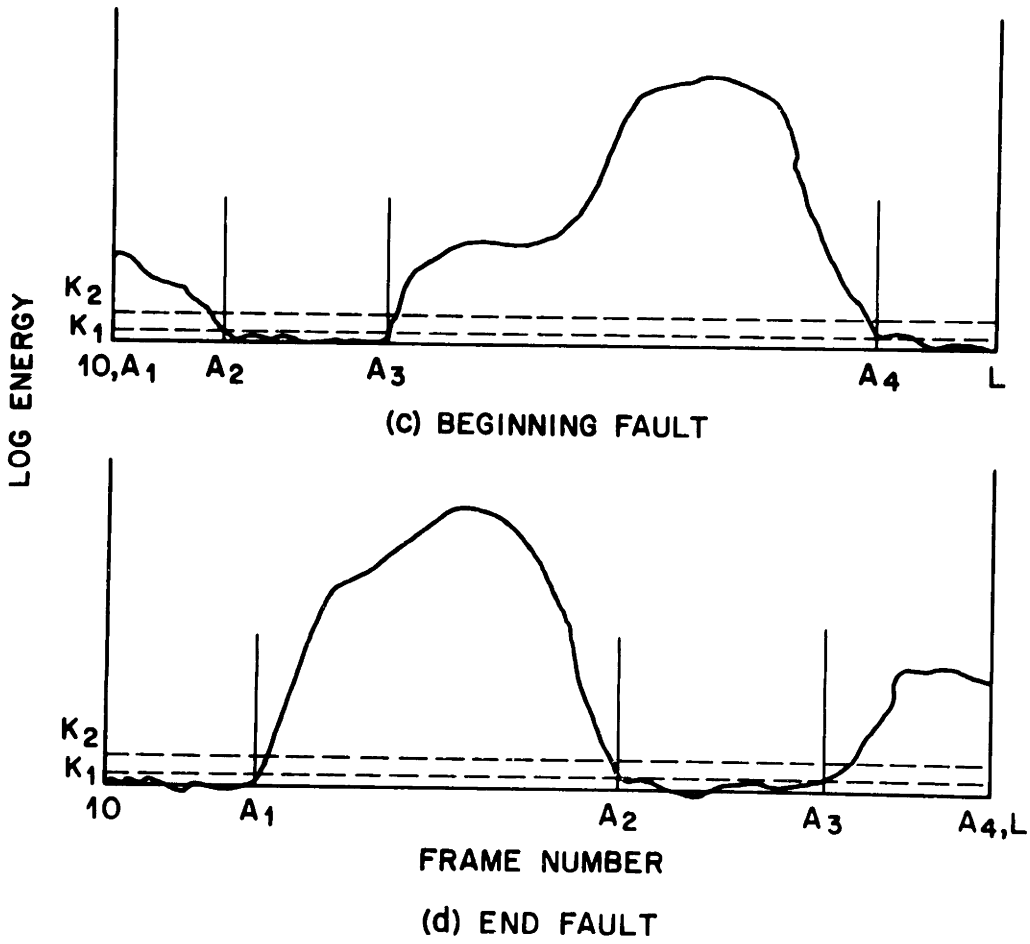


Figure 2.16 continued

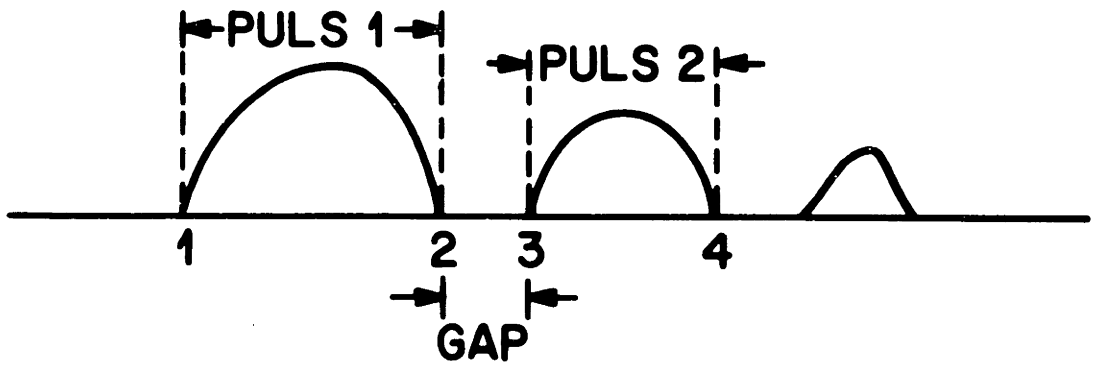


Figure 2.17 Definition of pulse and gap variables.

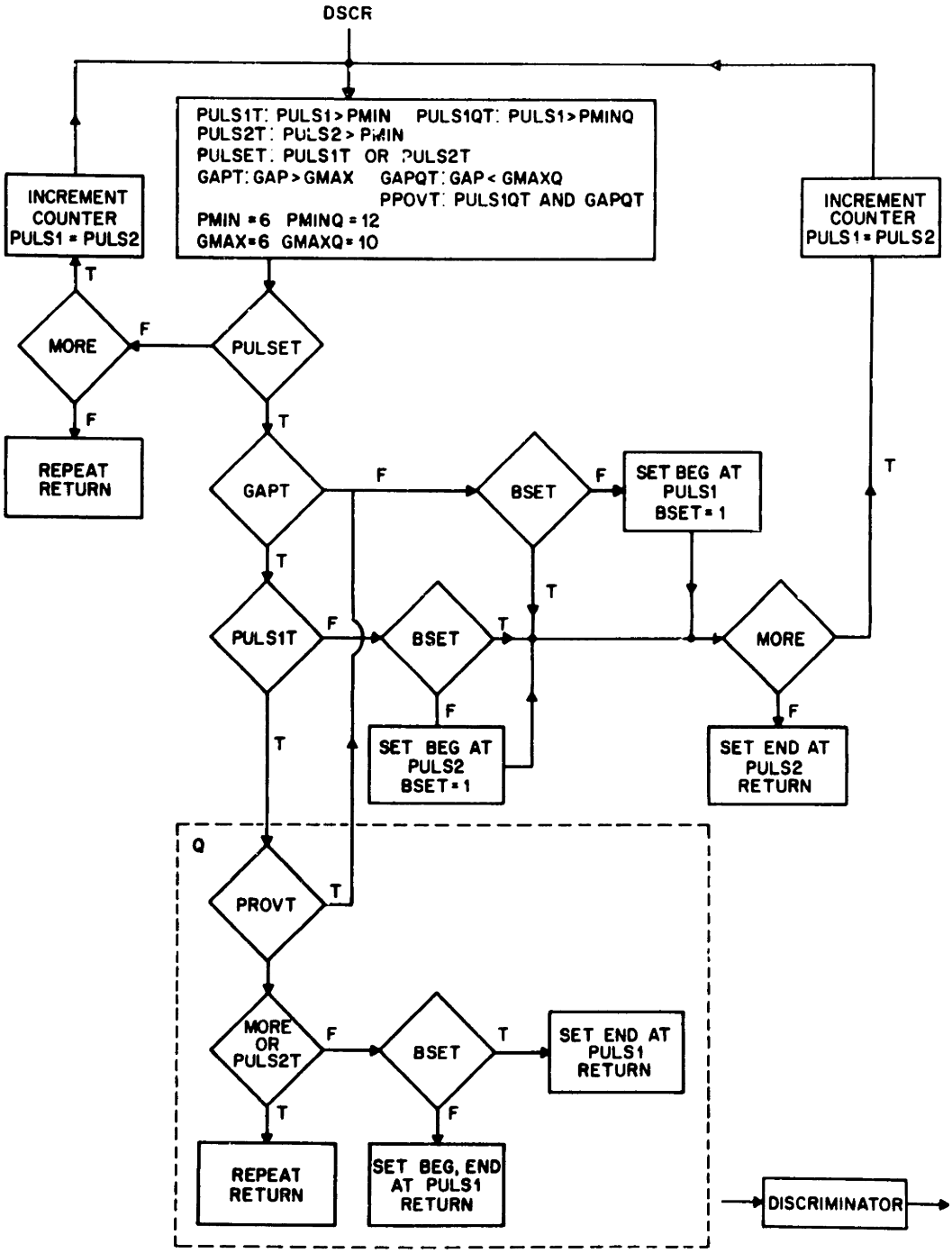


Figure 2.18 Flow chart of discrimnator.

As shown in Figures 2.15b and 2.18, the endpoint detector has the option of requesting a repeat when it cannot reliably determine an appropriate set of endpoints. The beginning fault and end fault conditions mentioned earlier are one possible cause of a repeat. A request for a repeat also occurs if two or more pulses of at least six frames duration are separated by greater than six frames. These conditions are denoted by the "repeat-return" in Figure 2.18 and the "repeat" in Table 2.1.

PULS1	PULS2	PULSET	GAPT	PULS1QT	GAPQT	PROVT	EXAMPLE	
0 0 1 1	0 1 0 1	0 1 1 1	0	0	0	0	REPEAT	LITTLE GAPS
1 1	0 1	1 1	0	1	0	0		
0 0 1 1	0 1 0 1	0 1 1 1	1	0	0	0	REPEAT REPEAT	BIG GAPS
1 1			1	1	0	0	REPEAT	
0 0 1 1	0 1 0 1	0 1 1 1	1	0	1	0	REPEAT REPEAT	MIDSIZE GAPS
1 1	0 1	1 1	1	1	1	1		

TABLE 2.1 Heuristics of the discriminator

Statistics were computed for the repeat request rate, the rate at which the discriminator was used, and the rate at which the branch of the discriminator denoted Q was used for a typical set of isolated word recordings. These statistics will be presented along with other performance measures in Section 2.7.1.

2.1.2 Endpoint Detection Based on Explicit Silence Measurements

The basic premise of endpoint detection for isolated word recognition is that the recording consists of an isolated word in a background of silence possibly accompanied by extraneous artifacts. In this section a method of endpoint detection is proposed in which the endpoints are located by finding regions of the recording that match the background silence poorly, and assuming that the word is located within these regions. This is accomplished by

computing a distance between each frame of the recording and a defined "silence reference template" and then using a thresholding method to determine the beginning and the end of the word.

The distance between frame i of the recording (called the test) and frame j of the silence reference is defined as the LPC log likelihood distance [11],

$$d(i,j) = \log\left[\sum_{m=0}^{p} T_i(m)\hat{R}_j(m)\right] \tag{2.13}$$

where $\underline{T}_i(m)$ is the test pattern for frame i, and $\underline{\hat{R}}_j(m)$ is the reference pattern for frame j of the silence reference. The distance $d(i,j)$ is a measure of the spectral similarity between $\underline{T}_i(m)$ and $\underline{\hat{R}}_j(m)$, as well as their energy difference. In general there are large energy and spectral differences between silence and speech sounds; thus the distance between a frame of silence and a frame of speech (e.g., a vowel) is expected to be large.

The canonic form for the proposed endpoint detector is given in Figure 2.19. For this implementation, the silence reference templates are obtained directly from the test, thereby dynamically and automatically accounting for different recording backgrounds. In addition, for statistical stability, the distances between the test and silence frames are averaged over three adjacent frames, thereby reducing the effects of short isolated artifacts. The average distance over three consecutive frames is computed for all possible starting and ending frames using the first and last three frames of the recording as silence templates. The average distance for a shift of s frames is given by

$$\bar{d}(s) = [d(s,1) + d(s,2) + d(s,3)]/3,\ s = 1,L \tag{2.14}$$

where $d(i,j)$ is defined in Eq. (2.13) and s is the shift relative to the first frame. A double threshold is used, as described in Section 2.4.1, with the lower threshold, T_L and the higher threshold, T_H, given as

$$T_L = [\underset{s}{Max}(\bar{d}(s)) - \underset{s<s_{Max}}{Min}(\bar{d}(s))]/C_1 \tag{2.15a}$$

$$T_H = [\underset{s}{Max}(\bar{d}(s)) - \underset{s<s_{max}}{Min}(\bar{d}(s))]/C_2 \tag{2.15b}$$

where s_{max} is the shift corresponding to the maximum distance, and C_1 and C_2 are empirically derived constants. Figure 2.20 shows some examples of the performance of the "silence-based" endpoint detector. In Figure 2.20a, the top plot is the log energy contour for the word "two". The endpoints, as determined by the initial silence distance computation with $C_1 = 5$, $C_2 = 2$, are denoted by the heavy vertical lines. The middle plot is the contour of average distance between the test and silence reference, using the first three frames of the test as the silence. The detected beginning point of the word has been forced to occur at least 20 frames prior to the last recorded frame L to insure a minimum word duration of 300 ms. The bottom plot shows the average distance contour for the ending frame, where the final three frames of the test were used as the reference template for silence. The locations of the endpoints are seen to be quite accurate for this example. For this case, the artifact prior to the word has not been included within the detected endpoints. Figure 2.20b shows a similar set of plots for the word "four". Once again the endpoints have been accurately located for this word. Figure 2.20c shows a final example for the letter "s". Here the endpoint detector has failed since the final fricative /*s*/ matches the silence template well enough to exclude a significant portion of the /*s*/. In general, the endpoint detector based on silence comparisons either accurately located the endpoints of the words or made gross mistakes of the type illustrated in Figure 2.20c. For such cases the endpoint detector failed in the manner expected, i.e., by excluding regions of low amplitude fricatives, which matched the background "silence" fairly closely. Also, in some cases, the smoothing of the distance scores caused the endpoints to be improperly estimated for sharply sloped energy contours.

A modification to this technique has also been implemented. This modification was to use only the final three frames of the recording (rather than both the initial and final frames) as the silence template based on the assumption that fewer isolated artifacts tend to occur at the ends of utterances than at the beginnings. Performance results for this endpoint detector are given in Section 2.7.2.

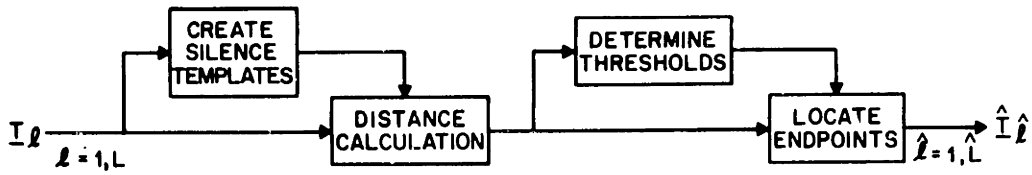


Figure 2.19 Canonic form for silence endpoint detector.

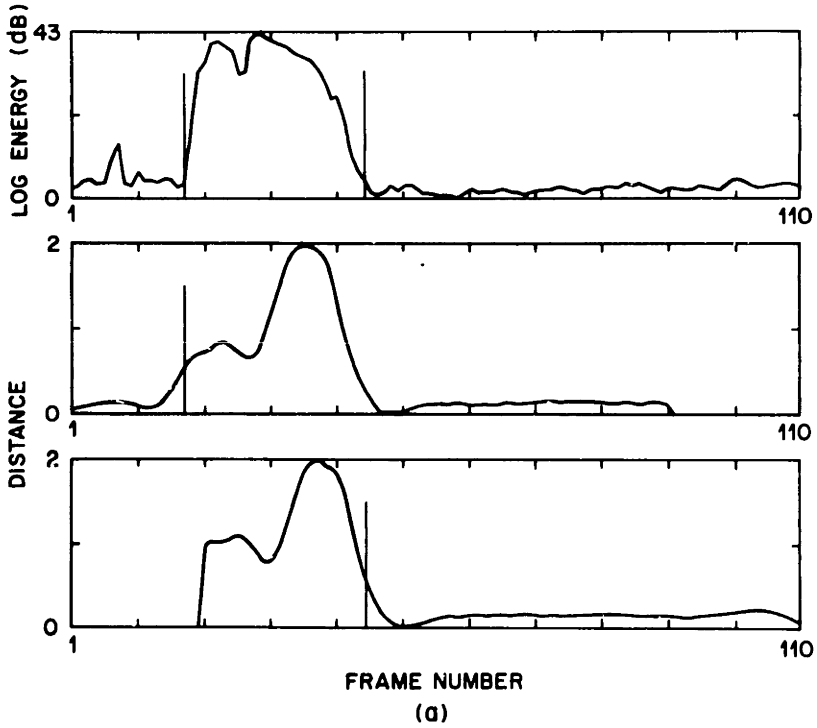
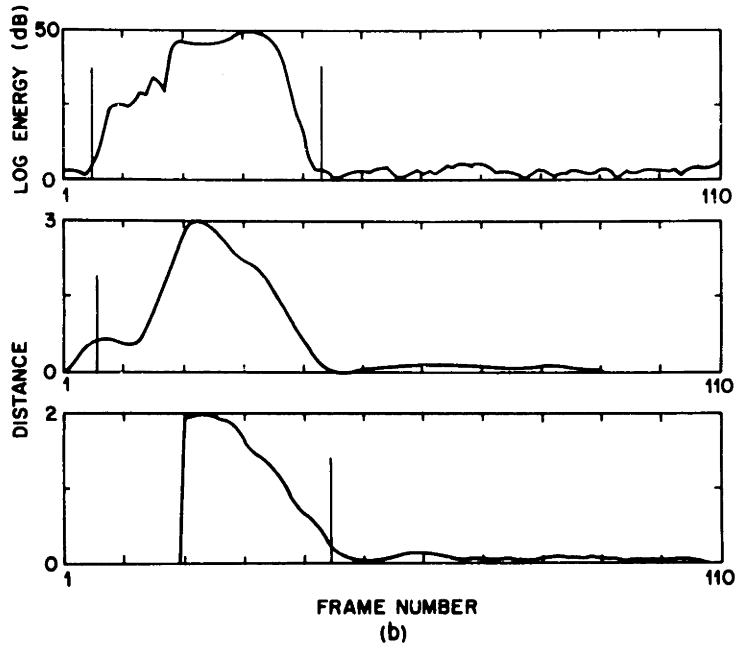


Figure 2.20 Examples of decisions of silence endpoint detector.



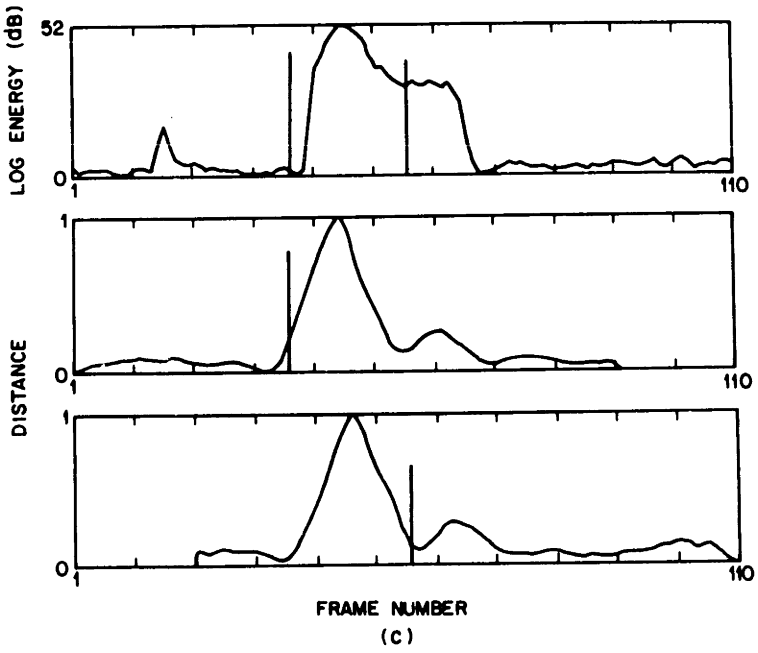


Figure 2.20 continued

2.5 Alternate Techniques for Explicit Estimation of Endpoints

An intuitive solution to the problem of endpoint detection is to classify each frame of the recording as either speech or nonspeech, and from this information locate the word endpoints. This procedure is essentially a segmentation and classification approach to endpoint detection. In order to implement such a procedure, a set of features for reliably classifying the test on a frame-by-frame basis must be found. In addition, some constraints on the temporal characteristics of the classes must be used to provide smooth, continuous transitions between speech and nonspeech regions. This approach to endpoint detection was studied in the following way. Six classes of speech and nonspeech sounds were defined. The speech sounds included voiced speech, unvoiced speech, and plosive releases. The nonspeech sounds, or artifacts, were classified as either silence, clicks, or breathiness. Features were measured for typical examples of each of the above classes in an attempt to find a feature set to reliably distinguish among the six classes. Preliminary investigations showed that, for telephone quality speech, the six classes could *not* be reliably distinguished. No feature sets were found that could statistically separate breathiness and unvoiced speech, or clicks and plosive releases. However, based on experimentation, it is believed that a set of features could be found to *pairwise* conditionally distinguish between the classes. A simple combinatorial approach could then be used to distinguish among the six classes.

A simpler segmentation approach to the problem of endpoint detection is the use of voiced-unvoiced-silence (VUS) classification, suggested by Rabiner, Schmidt, and Atal [7]. The use of the VUS classification is attractive for two reasons. First, VUS segmentation has been successfully used for telephone quality speech in clean recording environments [7]. Second, the time sequence of a smoothed version of the VUS contour can be used to aid in the determination of word endpoints for many practical cases.

An algorithm was developed to classify a speech waveform on a frame-by-frame basis into the three classes; voiced speech, unvoiced speech, and silence. For a given set of features, $\underline{X}$, the classification distance between $\underline{X}$ and the i^{th} class with mean feature vector $\underline{M}_i$ and covari-

ance matrix Λ_i is given by [7]

$$d_i = (\underline{X} - \underline{M}_i)\, \Lambda_i^{-1}\, (\underline{X} - \underline{M}_i)^t,\ i = 1,2,\ldots,Q \tag{2.16}$$

where Q is the total number of classes. The distance of Eq. (2.16) is the covariance weighted distance between class i and the unknown feature vector X. The unknown frame is generally classified as a member of the class i with the minimum distance, d_i. The means $\underline{M}_i$ and covariance matrices Λ_i for each of the possible classes were determined from a training set of data. The training data consisted of two sentences recorded by each of five speakers, three female and two male. The sentences were hand classified into voiced, unvoiced, and silence regions. Transition regions between classes were denoted and eliminated from the training data. Two features, namely zero crossing rate and log energy, were measured, and the mean and standard deviation, averaged over all speakers, were calculated for each class.

An example of the VUS classification for a sentence spoken by a speaker not used in the training is given in Figure 2.21. Figure 2.21a shows the hand classification of the sentence "Every salt breeze comes from the sea." The frames having value "T" are transition regions. The VUS contour in Figure 2.21b is the output of the minimum distance classifier of Eq. (2.16). The distance versus frame contour for voiced speech, unvoiced speech, and silence are shown in Figure 2.21c,d and e. The errors made by the VUS classifier are shown in Figure 2.21f. As can be seen, the majority of errors made by the classifier are isolated and can be eliminated by a simple smoothing procedure [12].

An endpoint detector was implemented using the VUS classifier and some simple temporal constraints. A block diagram of the endpoint detector is given in Figure 2.22. The first stage of the endpoint detector measures the features for each frame of the test. For this implementation, the features measured are the log energy and zero crossing rate. VUS analysis is performed for each frame, and each frame is classified as belonging to the class with the smallest covariance weighted distance according to Eq. (2.16). The resulting time sequence of the classified frames is smoothed to prevent single frames from being interpreted as events. Following the smoothing all runs of each class, where runs are defined as a sequence of frames of

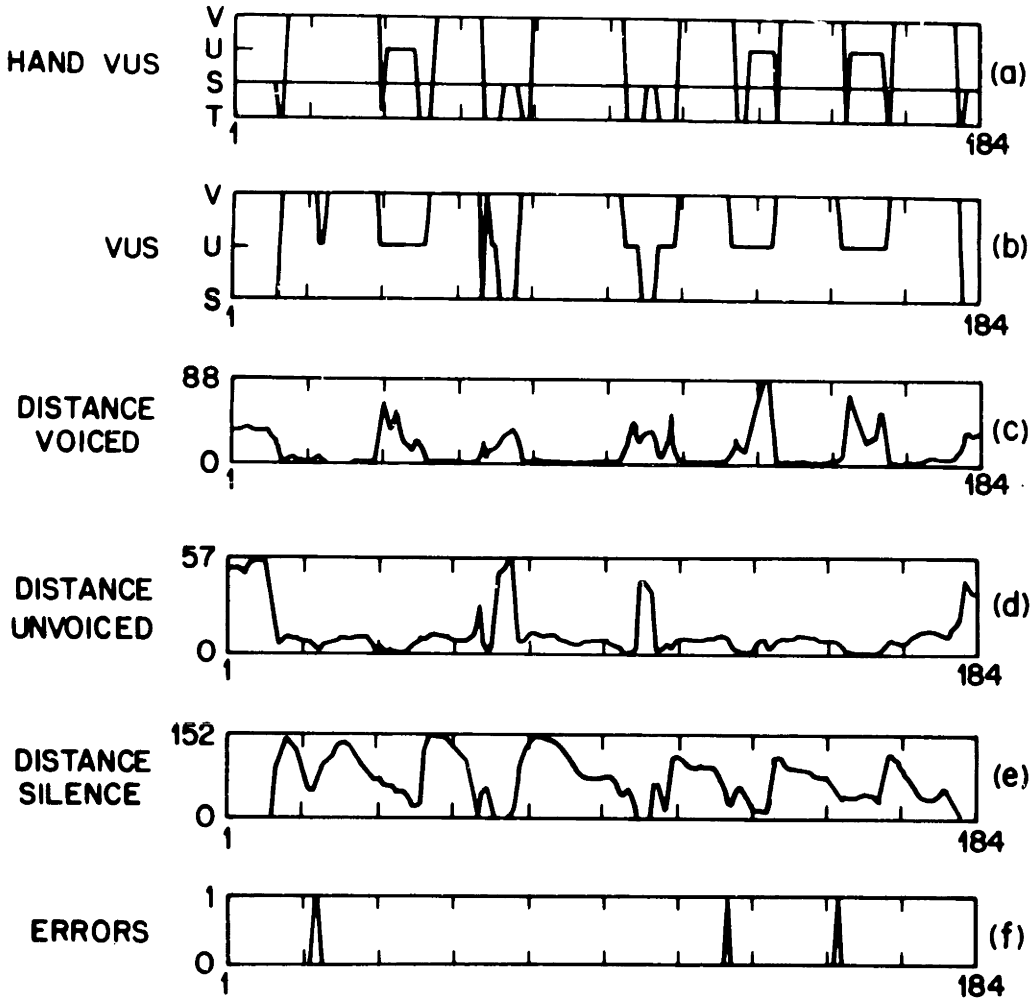


Figure 2.21 Example of VUS contours.

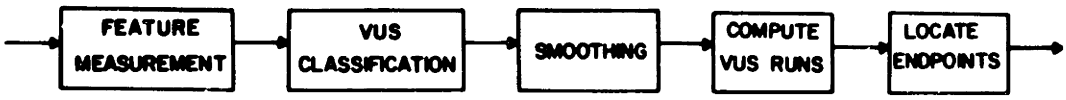


Figure 2.22 VUS endpoint detector.

the same class, are located. The endpoints are located from the sequence of the runs of voiced, unvoiced, and silence so as to satisfy temporal constraints. The constraints require the minimum duration of a word to be 300 ms. Also required is that there be at least 70 ms of voiced speech within the detected word, and the longest silence gap between parts of the word is 120 ms. The endpoint locations obtained by the VUS detector were compared to those of manual classification. These results, along with the resulting recognition accuracies, are given in Section 2.7.3.

2.6 Description of the Data Set Used for Testing the Endpoint Detectors

A testing data set was created to provide a common basis for comparison of all the endpoint detection algorithms discussed in this thesis. Three repetitions of a 39 word vocabulary were recorded by each of 10 speakers, (4 female and 6 male). The speakers were chosen because of known difficulties with endpoint detection using the current energy endpoint detector. The 39 word vocabulary [13] consisted of the letters of the alphabet, the digits, and the words "stop", "error", and "repeat" in the "randomized" order shown in Table 2.2. The resulting data set, denoted as TS1, included 1161 utterances of isolated words (9 words were lost due to manual recording errors). Of the recorded utterances, approximately 40 to 60 percent included extraneous artifacts such as clicks or breathiness. Unless otherwise stated, all recognition accuracies presented refer to tests on this data set.

1-26 A-Z
27 STOP
28 ERROR
29 REPEAT
30-39 ZERO-NINE

"randomized" order

19 S	37 SEVEN	35 FIVE
34 FOUR	23 W	39 NINE
11 K	17 Q	36 SIX
38 EIGHT	7 G	2 B
24 X	29 REPEAT	1 A
21 U	16 P	33 THREE
10 J	31 ONE	26 Z
30 ZERO	27 STOP	6 F
32 TWO	5 E	18 R
22 V	12 L	28 ERROR
4 D	13 M	20 T
9 I	3 C	8 H
25 Y	14 N	15 0

TABLE 2.2 Words in the vocabulary

An earlier data set, denoted as TS0 was created consisting of 207 isolated words, from the 39 word vocabulary, spoken by 6 speakers (2 female and 4 male). This data set was used for preliminary testing of several of the endpoint detectors.

2.7 Performance Results for the Explicit Techniques for Endpoint Detection

In this section the recognition accuracies and performance measurements for the various endpoint detectors presented in this chapter are discussed. Section 2.7.1 gives the results for the current energy endpoint detector. In Section 2.7.2, the results for the silence based endpoint detector are given. Finally, in Section 2.7.3 the results of the VUS based endpoint detector are presented.

2.7.1 Accuracy of the Current Energy Endpoint Detector

Several different measures will be used to assess the performance of the various endpoint detectors studied in this thesis. For the current energy based endpoint detector, it is of interest to examine the statistics of the usage of the various components of the detector - i.e., the overall rejection rate due to lack of reliable endpoints, the rate at which the discriminator is used, and the rate at which the Q function of the discriminator is used. These results are summarized in Table 2.3 and Figure 2.23. Table 2.3 gives, for each talker, statistics on

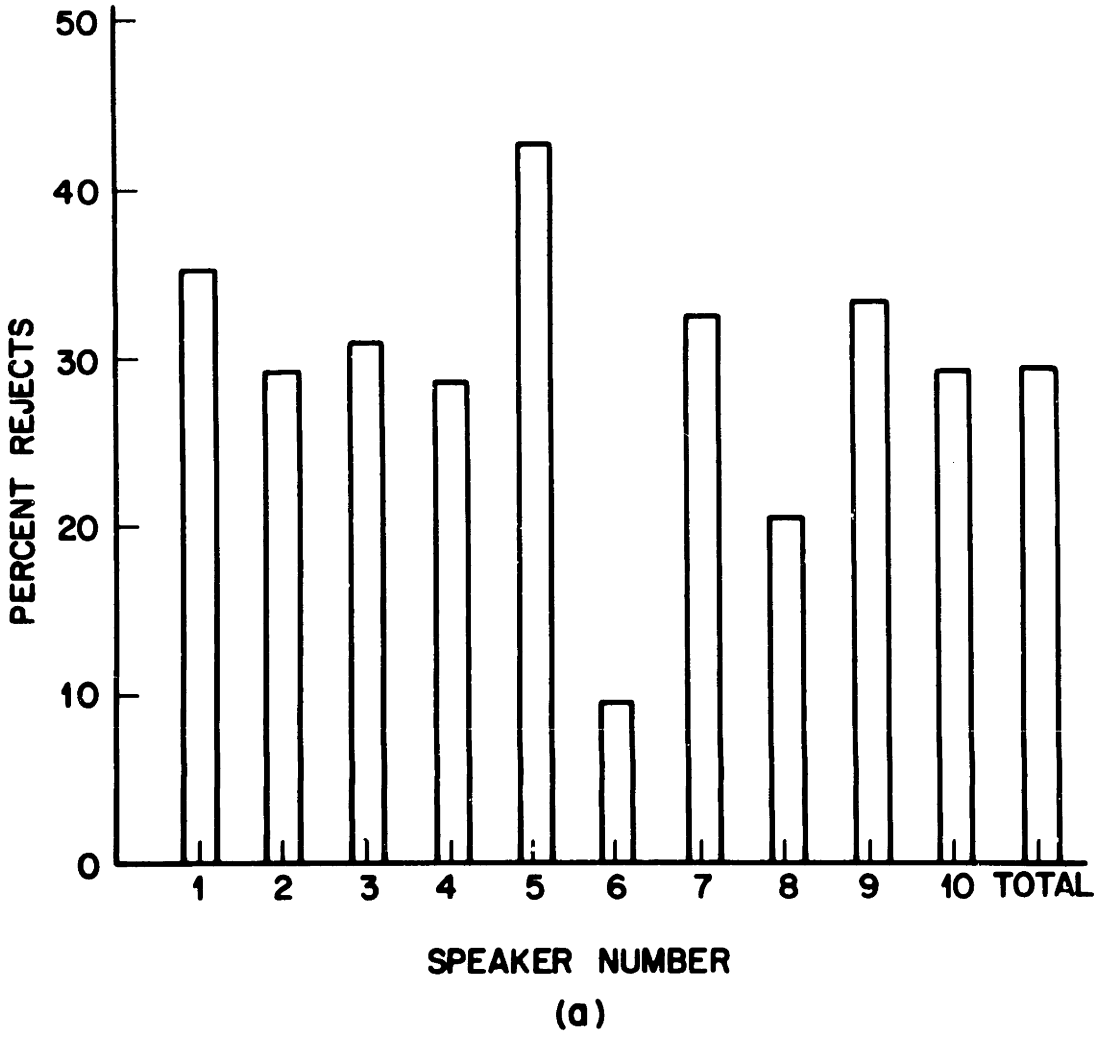
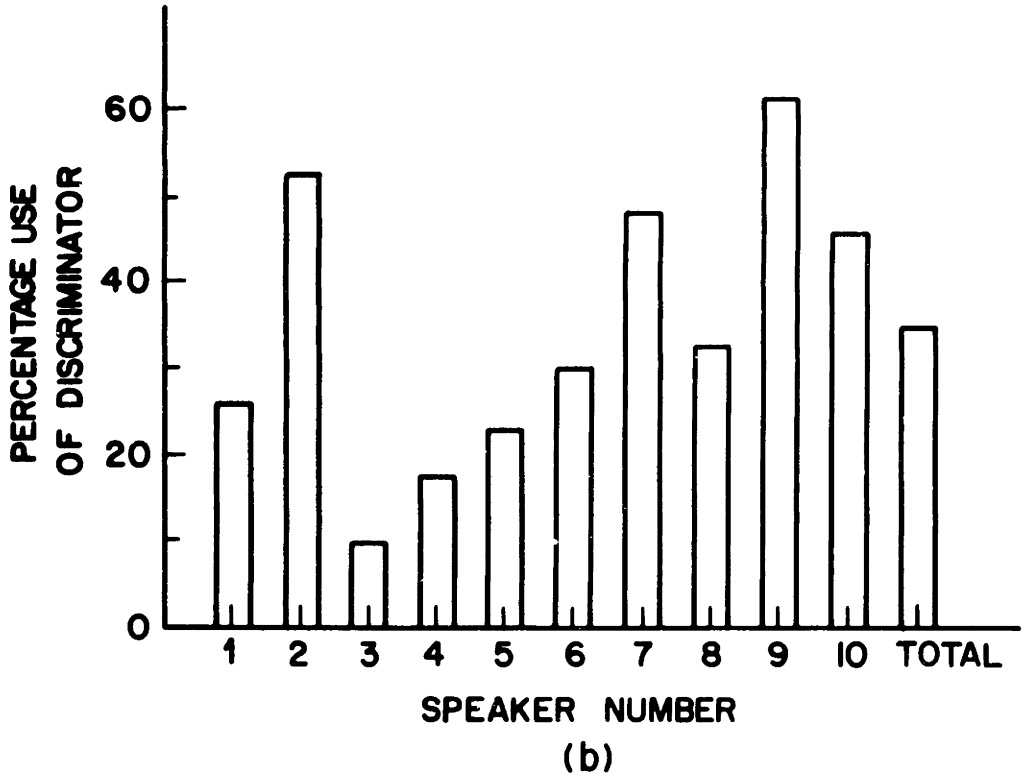
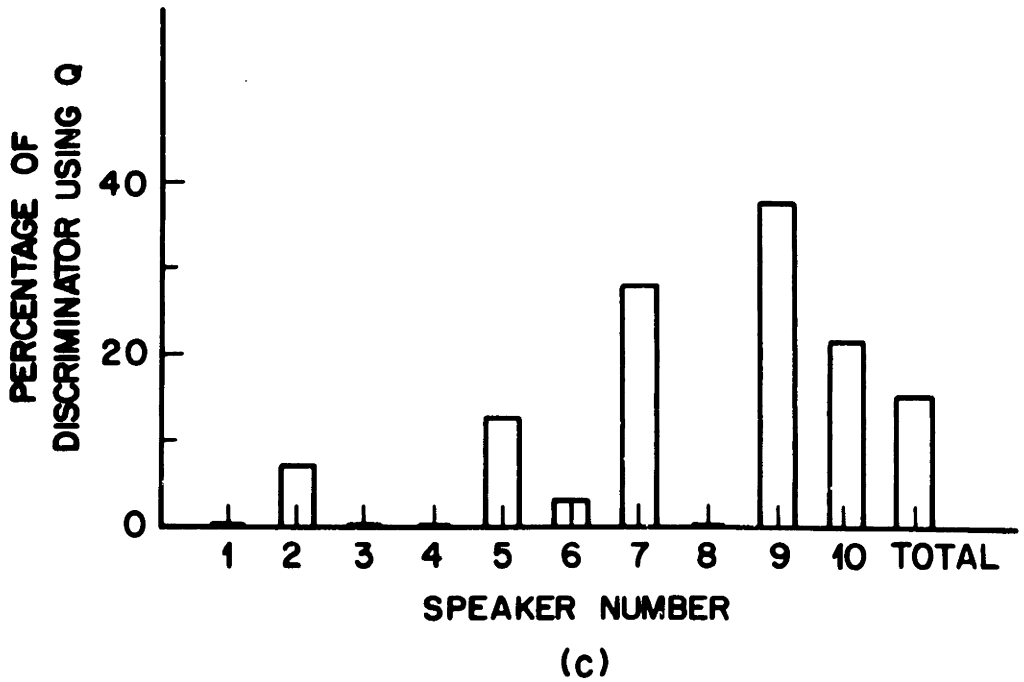


Figure 2.23 Current energy endpoint detector - reject rates, use of discriminator.



(b)



(c)

Figure 2.23 continued

rejection rate, rate of using the discriminator (DSCR), and rate of using the Q function. These same data are plotted in Figure 2.23. It can be seen from Figure 2.23a that the highest rejection rate, for any talker, was approximately 43%, while the lowest was about 10%.

speaker #	# words spoken	% rejected	% using DSCR	% of DSCR Using Q
1	114	35.1	25.7	0.0
2	117	29.1	53.0	6.8
3	116	31.0	10.0	0.0
4	112	28.5	17.5	0.0
5	117	42.7	23.9	12.5
6	117	9.4	30.2	3.1
7	117	32.5	49.4	28.2
8	117	20.5	32.3	0.0
9	117	33.3	61.5	37.5
10	117	29.1	45.8	23.7
TOTAL	1161	29.1	35.0	15.3

TABLE 2.3 Statistics for current energy detector

Of the 1161 spoken words by the ten speakers, 338, i.e., almost 30% were rejected and a repeat was requested. Figure 2.24 shows two examples for which repeats were requested. For both of these cases, the actual word endpoints are easily located visually at frames denoted by A_1 and A_2. The results of Table 2.3 show that the current energy endpoint detector failed on a high percentage (almost 30%) of trials and requested a repeat.

For the recordings which were not rejected by the energy endpoint detector, approximately 35% had multiple energy pulses and required the use of the discriminator. Of these cases, about 15% were analyzed by the Q function. Hence, the logic of the Q function was not required in a significant percentage of the cases in which multiple energy pulses were detected.

For those cases in which the endpoints were determined by the current energy detector, recognition scores were measured by the recognition system described previously, using a set of speaker independent word templates. Figure 2.25 shows the percent correct recognition as a function of the number of best candidates for these cases. The number of best candidates at the l^{th} position is a measure of how often the word actually spoken occurred among the top recognition candidates. The percent correct recognition curves, averaged over all talkers, are shown for a K-nearest neighbor (KNN) decision rule for $K = 1$ and $K = 2$ [14]. It can be

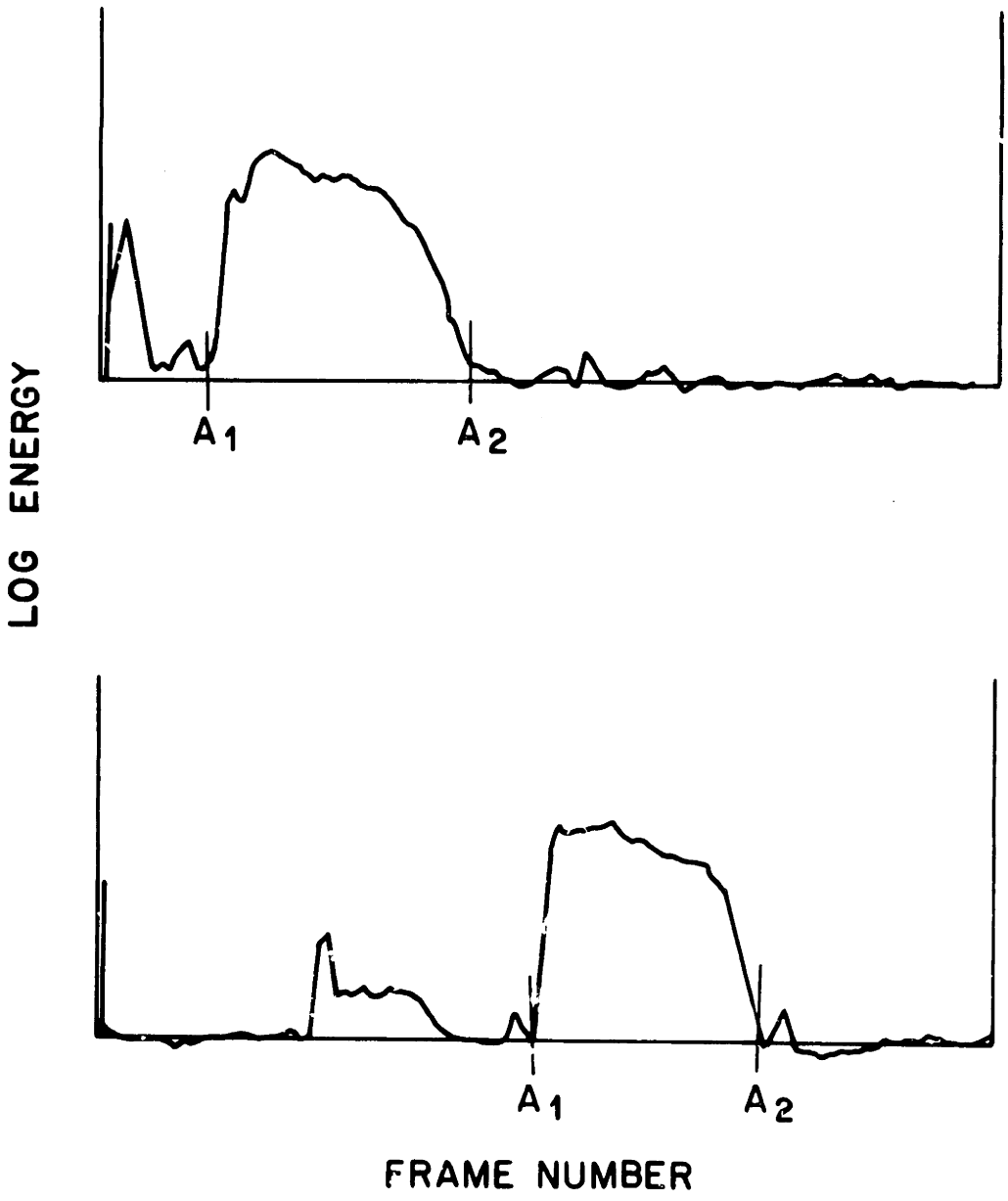


Figure 2.24 Examples of utterances rejected by current energy detector.

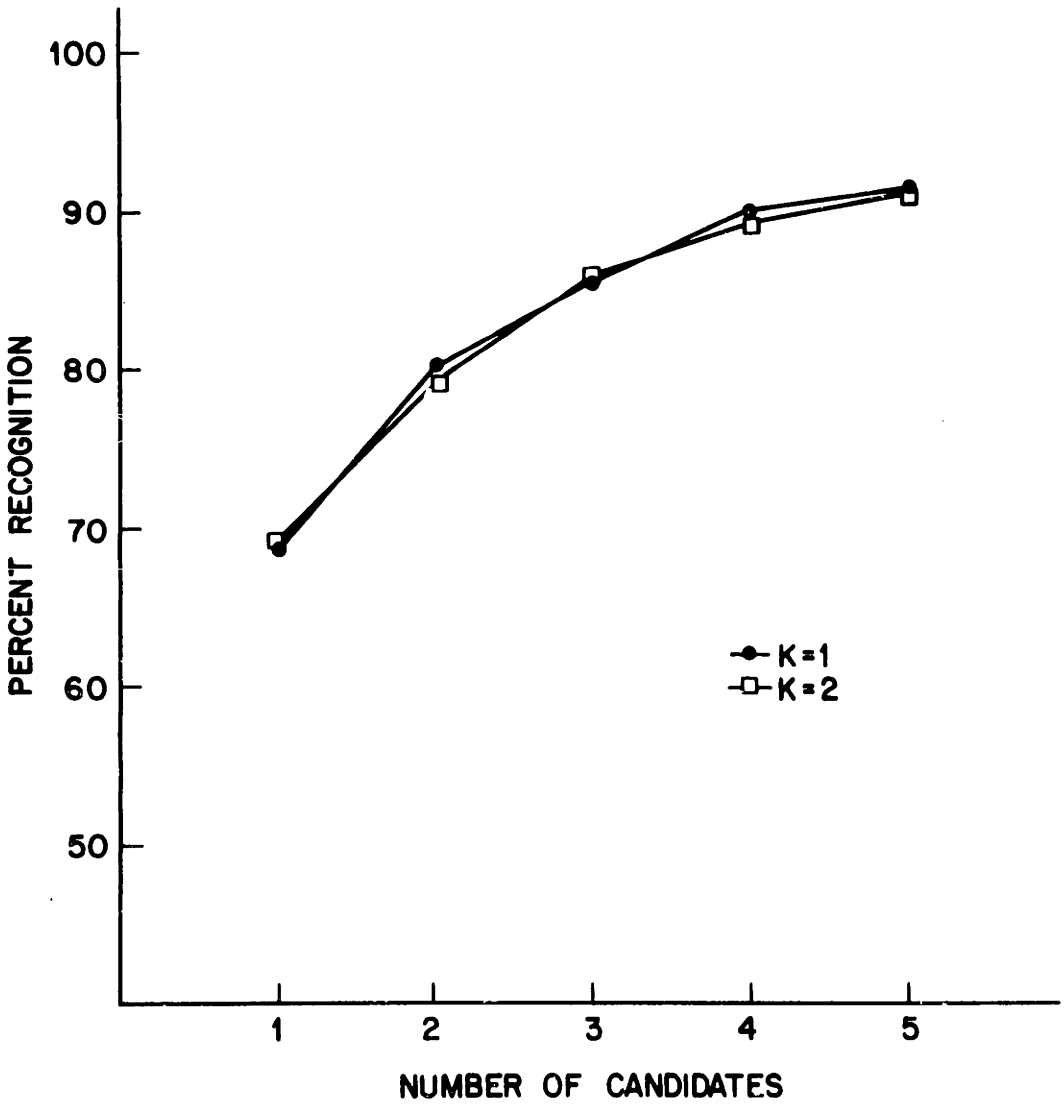


Figure 2.25 Recognition scores using current energy detector - averaged for all speakers.

seen from Figure 2.25 that the $K = 1$ and $K = 2$ recognition results are almost identical. Correct recognition for the top candidate is approximately 70%. (For the individual talkers this rate varied from 50% to 83%.) The overall recognition score for the top five candidates, ranged from 83% to 97%, with an average of 91%. The recognition rates for the individual talkers are given in Table 2.4a and plotted in Figure 2.26a for the $K = 1$ decision rule and in Table 2.4b and Figure 2.26b for the $K = 2$ decision rule.

speaker #	# tests	% rejects	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

Table 2.4a k = 1

speaker #	# tests	% rejects	candidate position 1	2	3	4	5
1	114	35.1	77.0	90.5	94.6	95.9	95.9
2	117	29.1	61.4	75.9	85.5	88.0	90.4
3	116	31.0	85.0	92.5	95.0	95.0	98.7
4	112	28.5	82.5	86.2	92.5	95.0	95.0
5	117	42.7	59.7	68.7	74.6	82.1	82.1
6	117	9.4	67.9	81.1	85.8	88.7	88.7
7	117	32.5	74.7	83.5	91.1	92.4	94.9
8	117	20.5	62.4	72.0	81.7	87.1	91.4
9	117	33.3	56.4	67.9	80.8	84.6	85.9
10	117	29.1	68.7	77.1	83.1	88.0	90.4
TOTAL	1161	29.1	69.1	79.5	85.9	89.2	91.1

Table 2.4b k = 2

TABLE 2.4 - Recognition results for current energy endpoint detector
12 speaker independent templates/word
(percent correct in candidate position i, i=1, 2, 3, 4, 5)

Some examples of the locations of the endpoints found by the current energy detector are given in Figure 2.27, which shows plots of the log energy contour versus time in frames. In Figure 2.27a the endpoints, as denoted by the vertical lines, are accurately located. In Figure 2.27b and c the beginning point of the word is incorrect due to initial artifacts. The correct

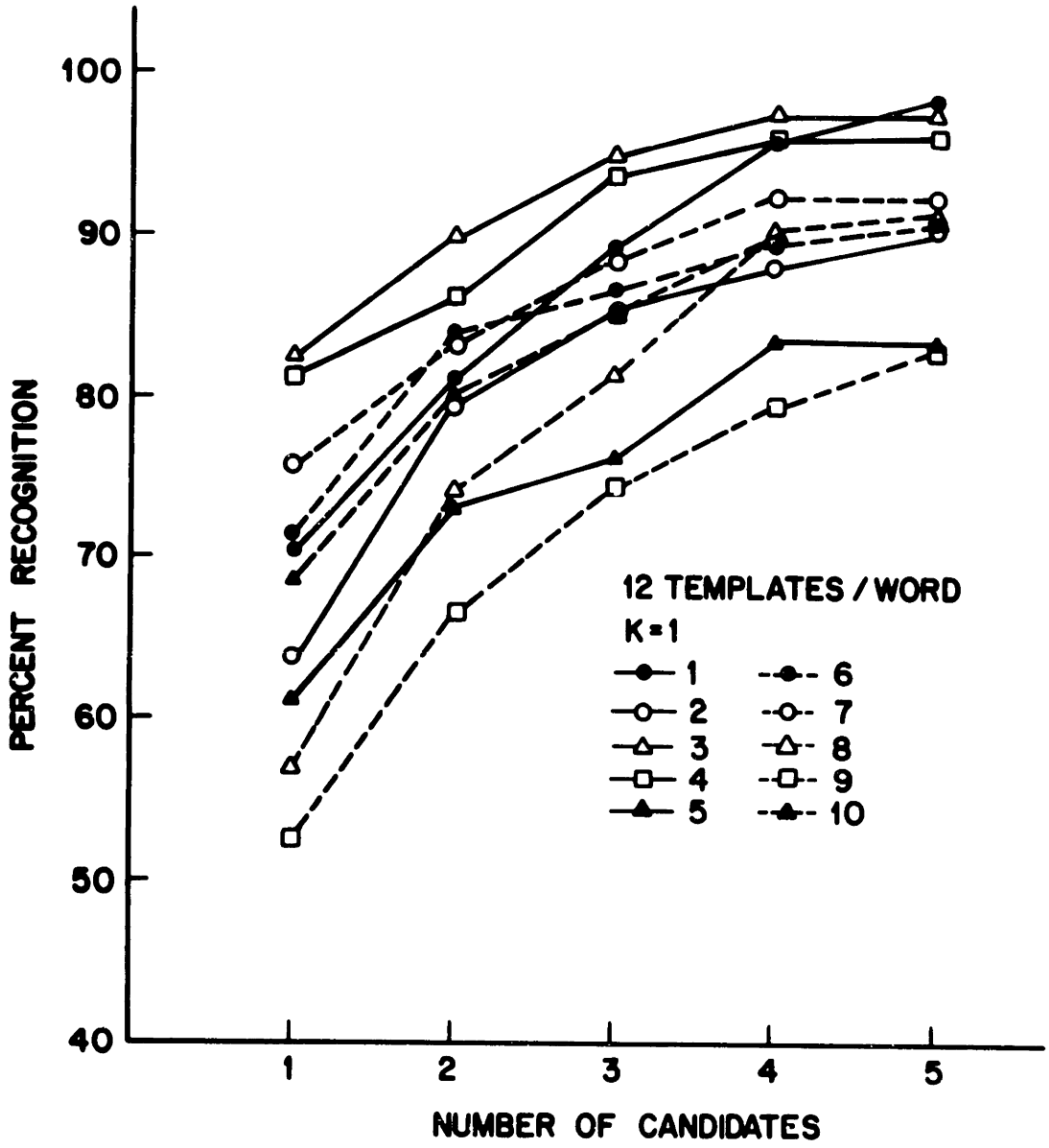


Figure 2.26 Recognition scores using current energy detector for individual speakers.

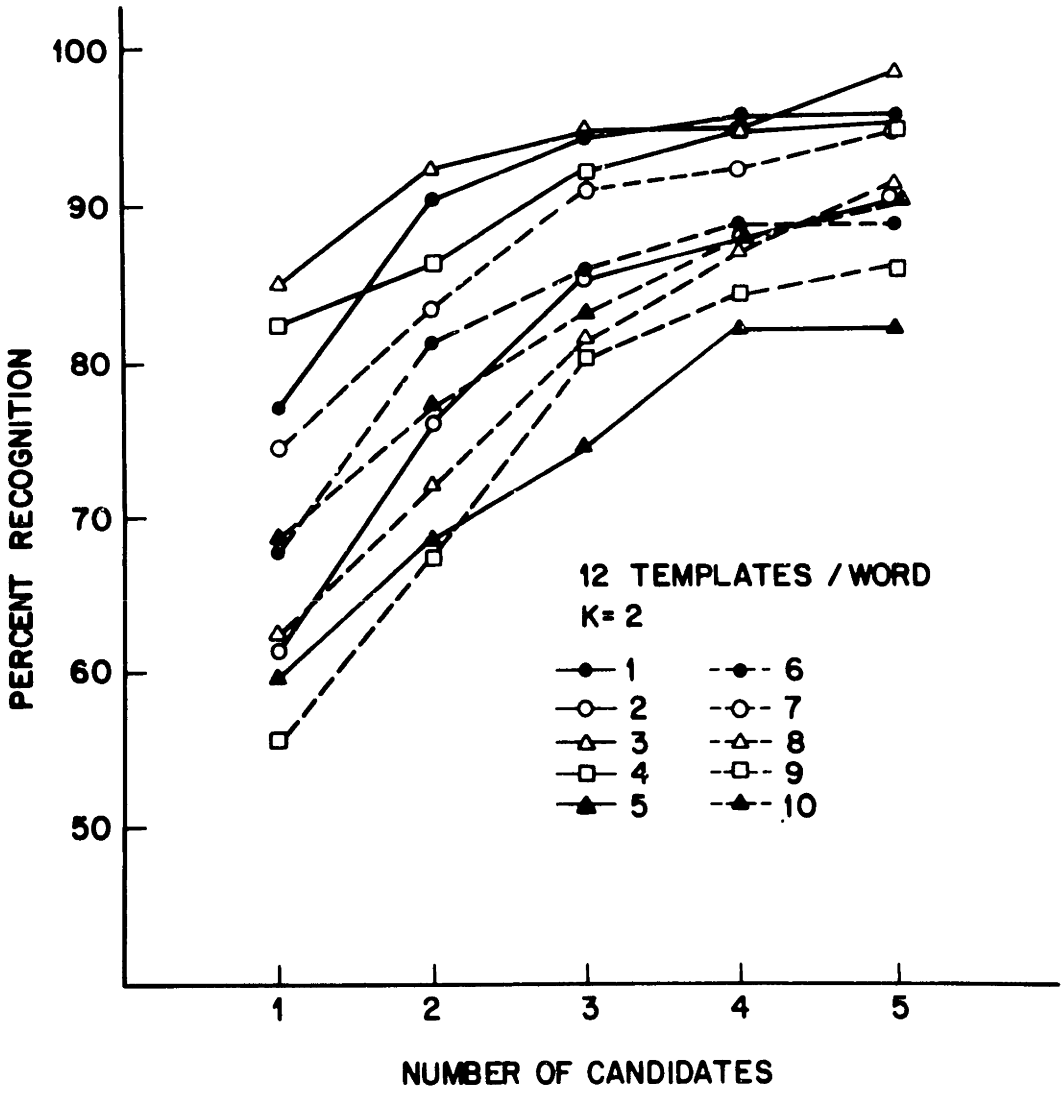


Figure 2.26 continued

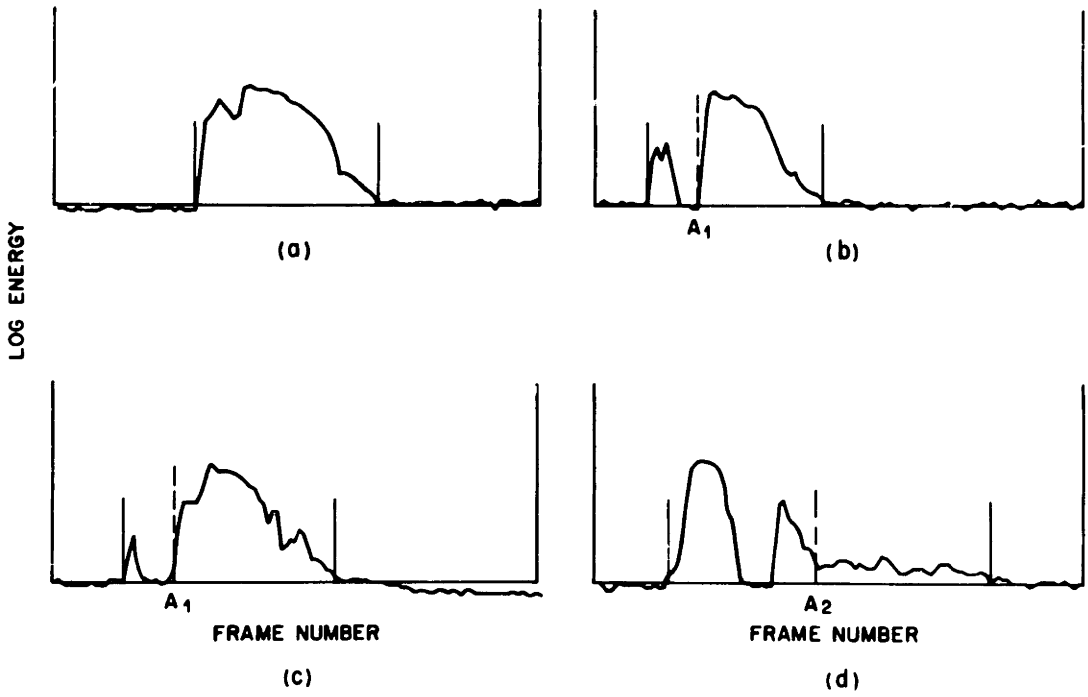


Figure 2.27 Examples of endpoints as estimated by current energy detector.

beginning point is denoted by the dashed line at frame A_1. For the example of Figure 2.27d, the ending point is improperly found because of breath noise included at the end of the word. The correct ending frame is denoted by the dashed line at frame A_2. In general, the current energy endpoint detector tended to incorrectly estimate the endpoints by including clicks and breath noise. However, for those cases in which the endpoints are found, the overall recognition accuracies of the current energy endpoint detector are comparable to those achieved previously on speech with no artifacts [14]. Thus, the major problem with the energy endpoint detector is its high rejection rate for utterances with any recording problems.

2.7.2 Accuracy of the Silence Based Endpoint Detector

In this section the recognition results for the silence based endpoint detector proposed in Section 2.4.2 are presented. The two versions of the endpoint detector were tested on the data base described in Section 2.6. In the first, which created separate silence templates from the first and last three frames, the template from the first three frames was used to locate the beginning point. Similarly, the template from the final three frames was used to locate the ending frame. The test utterances were divided into those that were accepted by the current energy endpoint detector (i.e., a set of endpoints was found) and those that were rejected. The recognition results for the silence based detector are plotted in Figure 2.28. The recognition accuracies are averaged over all speakers for values of $K = 1$ and $K = 2$ of the KNN decision rule. The recognition accuracy obtained for the words accepted by the current energy detector is 5 to 10 percent higher than for words that were rejected. However, the overall average recognition score is only about 60% for the top candidate, and about 80% for the top five candidates. Hence, the accuracies of the word endpoints obtained from the silence based analysis are significantly worse than the recognition accuracies obtained by the current energy detector.

The evaluation of an endpoint detector is dependent on the performance criterion. As presented before, the recognition scores for the current energy endpoint detector are given as a percentage of the accepted recordings. The resulting recognition scores are based on 70% of the test utterances rather than the complete data base. On the other hand, the recognition scores obtained by the silence based endpoint detector reflect the accuracy achieved on the

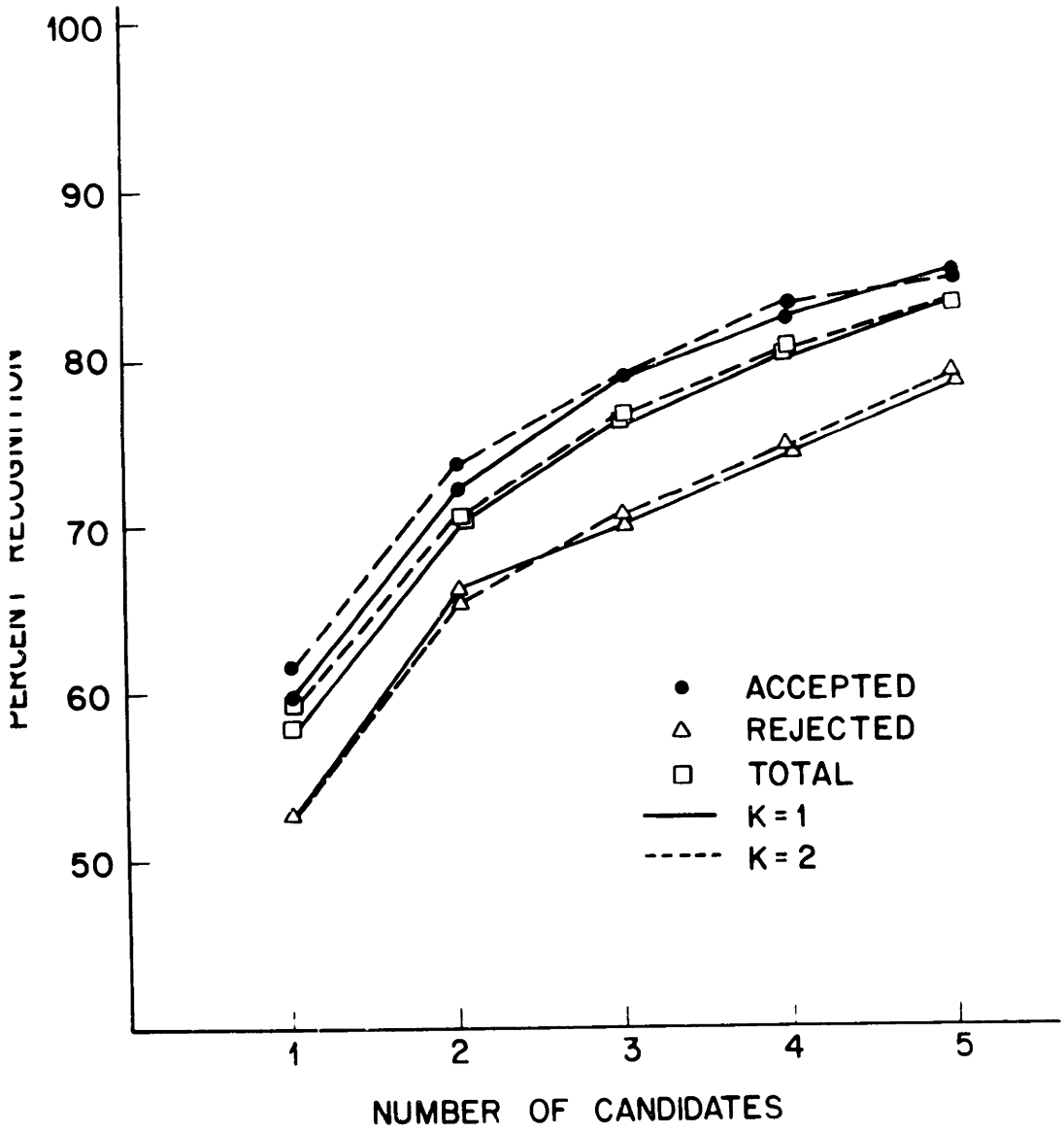


Figure 2.28 Recognition scores for silence endpoint detector - beginning and ending frames used as silence templates.

entire data base. This includes the tests both accepted and rejected by the current energy endpoint detector. Comparing the endpoint detectors solely on the basis of the percent of words correctly recognized does not account for the differences in the associated rejetion rates.

Alternatively the data may be presented in terms of the actual number of words recognized from the entire data set rather than as percentages relative to the number of words tested. These results are given in Table 2.5.

		candidate position				
	# tests	1	2	3	4	5
current energy						
accepted	823	569	654	707	734	750
silence-based						
accepted	823	505	607	650	683	698
rejected	338	182	222	239	252	267
total	1161	687	829	880	935	965

TABLE 2.5 Recognition scores for current energy and silence-based endpoint detectors (number correct in candidate position i, i=1,2,3,4,5)

Comparing the data in Table 2.5 it is seen that the current energy endpoint detector locates the endpoints more accurately than the silence-based endpoint detector for the accepted words. In fact, the current energy endpoint detector achieves correct recognition of 50 to 60 more words than the silence-based detector. However, these results comprise only 70% of the tests. Of the 338 tests rejected by the current energy detector, correct recognition is achieved for the top candidate for 182 of the tests, greater than 50%. The correct word appeared within the top 5 candidates for 267, almost 80% of the rejected tests. Using the complete data set, the silence based endpoint detector correctly locates the endpoints for 120 (in the top candidate position) to over 200 (within the top 5 candidates positions) more tests than the current energy detector. Presentation of the data in terms of the number of words correctly recognized illustrates some of the tradeoffs between rejection rate and recognition accuracy. For most applications a high rejection rate is more tolerable than a low recognition accuracy. Rejection causes inconvenience to the user, but as long as the reject rate remains within reasonable bounds it is tolerable. Inaccurate recognition with a low rejection rate is generally quite costly in a recognition system. Thus, the presentation of the recognition scores obtained reflect only the accepted tests, with

the associated rejection rates given.

The second version of this endpoint detector used only the last three frames of the test as the silence template. The recognition results using this modification are shown in Figure 2.29. The recognition accuracy scores for both accepted words and rejected words are lower than those shown in Figure 2.28 in which two silence templates were used. Hence it is concluded that this modification is not an improvement in the silence based endpoint detector.

Figure 2.30 compares the locations of the endpoints for several isolated words as estimated by the silence based endpoint detector and the current energy endpoint detector. Each plot shows the log energy contour with the endpoints denoted by the heavy, vertical lines. The top plot shows the endpoints as estimated by the current energy endpoint detector. For cases when the current energy endpoint detector failed, i.e., requested a repeat, no endpoints are drawn (see Figure 2.30b-d). The second plot shows the estimates of the endpoints as determined by the silence based endpoint detector, using both the beginning and ending frames as silence references. The third plot has the endpoints as estimated by the silence based method using only the ending frames as the silence template. In Figure 2.30a, all of the endpoint detectors have successfully excluded the initial click, and accurately located the endpoints. The current energy detector failed to find endpoints for the word in Figure 2.30b. Despite the final breathiness in the recording, the silence distance based endpoint detectors locate a reasonable set of endpoints. The best estimates of the endpoints are those using the beginning and ending frames for the silence reference. Figures 2.30c and d show two more examples in which the current energy endpoint detector failed to find the endpoints. In part c the background is relatively uniform and the silence based distance endpoint detectors both give good estimates for the endpoints. A problem associated with these silence based endpoint detectors is illustrated in Figure 2.30d. The background following the spoken word has higher energy than normal in the final few frames; thus, a large distance is obtained when the final frames are compared to the nominal energy background. This results in gross errors in the endpoints as seen in Figure 2.30d. For words whose energy contour has a slow onset, the silence based endpoint detector was generally able to accurately locate the endpoints. However, for words whose energy

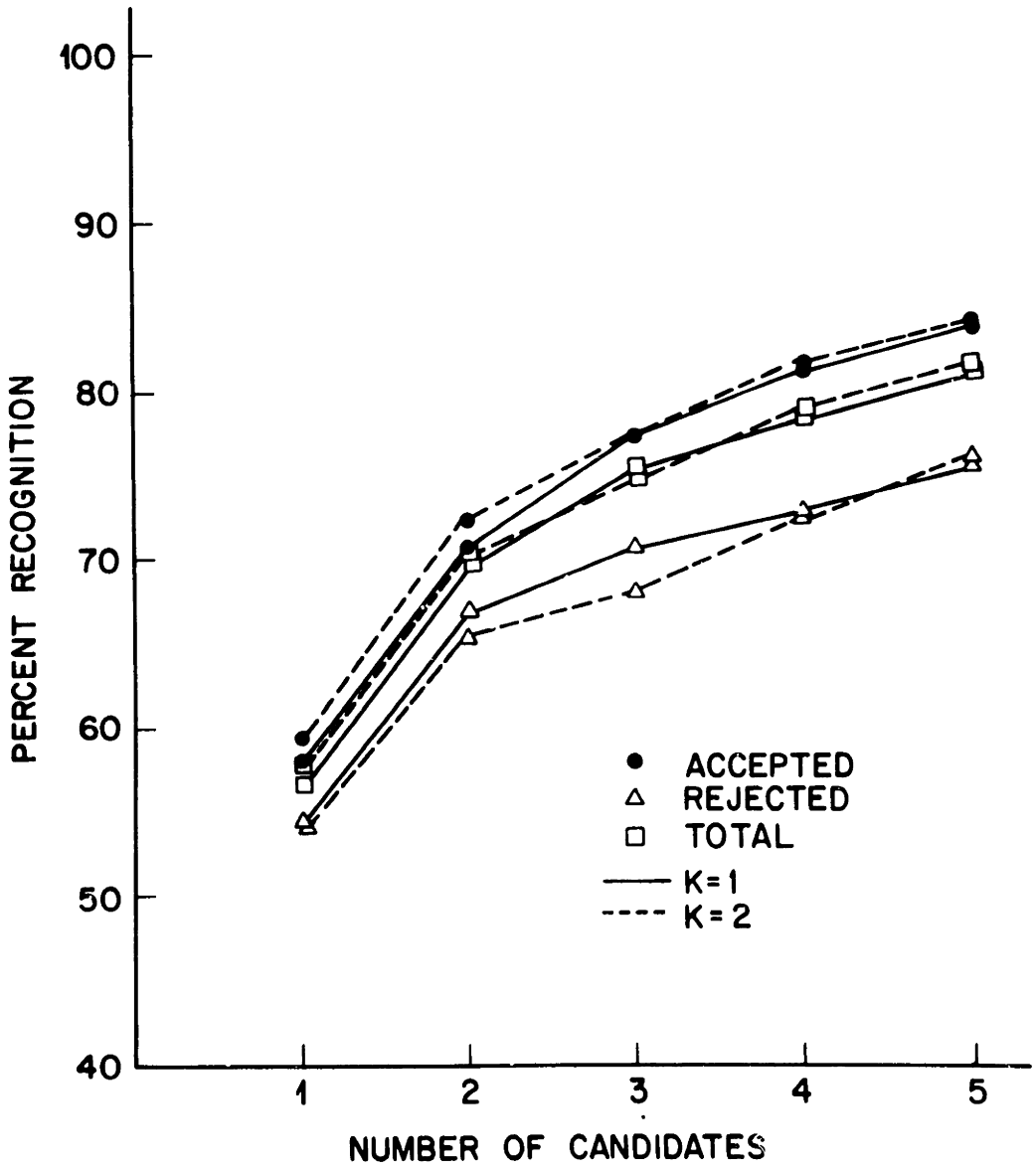


Figure 2.29 Recognition scores for silence endpoint detector - ending frames only used as silence reference.

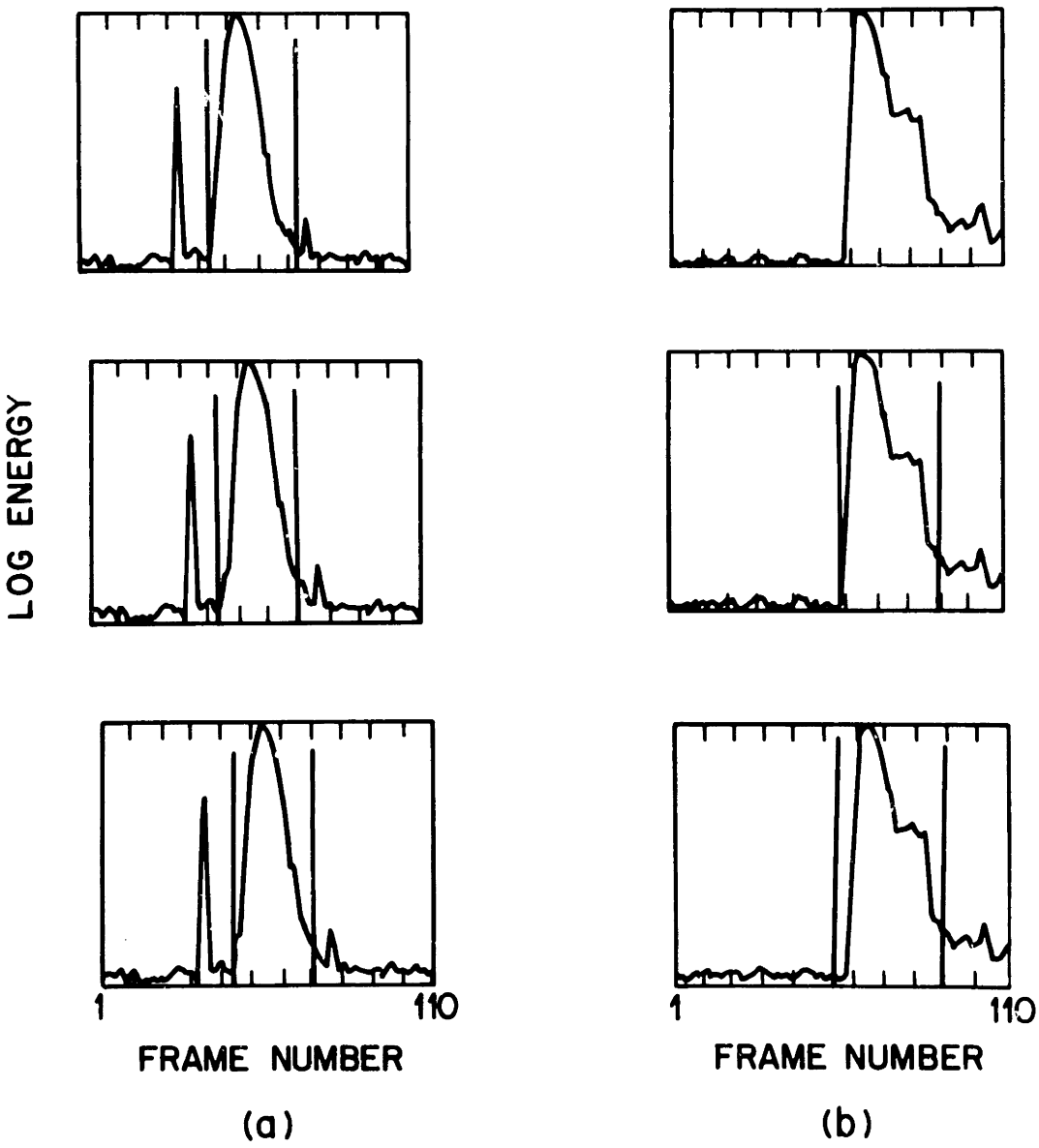


Figure 2.30 Comparison of endpoint locations for explicit detectors.

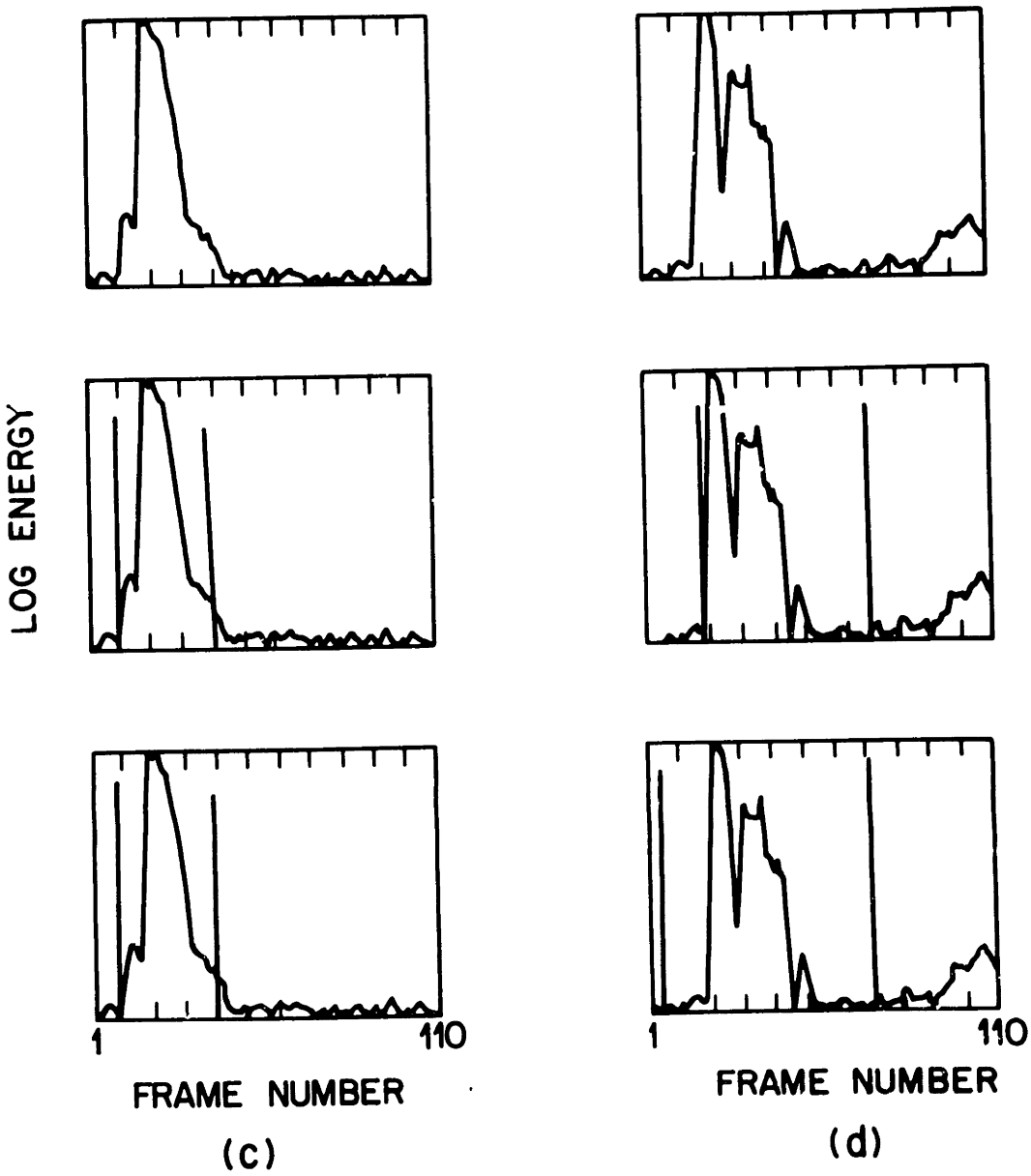


Figure 2.30 continued

contours rose rapidly, the silence based endpoint detector smoothed the energy contour and improperly (by several frames) located the endpoints. Another problem was that the silence based endpoint detectors often omitted significant portion of low amplitude fricatives. In summary, the silence based endpoint detectors do not provide performance scores comparable to those of the current energy based endpoint detector.

2.7.3 Results Using the VUS Endpoint Detector

Due to time limitations, the VUS endpoint detector was only tested using the preliminary data set, TSO described in Section 2.6. The endpoint locations obtained by the current energy endpoint detector were compared to those of the VUS detector, and to those obtained via hand classification. In addition, the recognition accuracies obtained using each of the endpoint detectors were compared.

Figure 2.31 and 2.32 shows histograms of the difference between the location of the endpoints obtained using the endpoint detectors, and those obtained by hand classification. Figures 2.31a and c are for the energy endpoint detector and Figure 2.31b and d are for the VUS detector. Figure 2.31a and b show histograms for the starting point locations for "clean" utterances, i.e., those classified by hand of being free from artifacts. Figure 2.31c and d show the corresponding histograms for utterances having extraneous artifacts. Figure 2.32a-d show the same set of histograms for the ending points.

From the histograms it can be seen that

(1) The ending point locations exhibit more variation than the starting points

(2) The current energy endpoint detector tends to make errors by estimating the beginning frame too early, and the ending frame too late.

(3) Both the current energy endpoint detector and the VUS endpoint detector make a large number of gross endpoint errors for recordings with artifacts.

The recognition accuracies obtained using the VUS endpoint detector for the preliminary data set are shown in Table 2.5a for the case of two speaker independent templates per word.

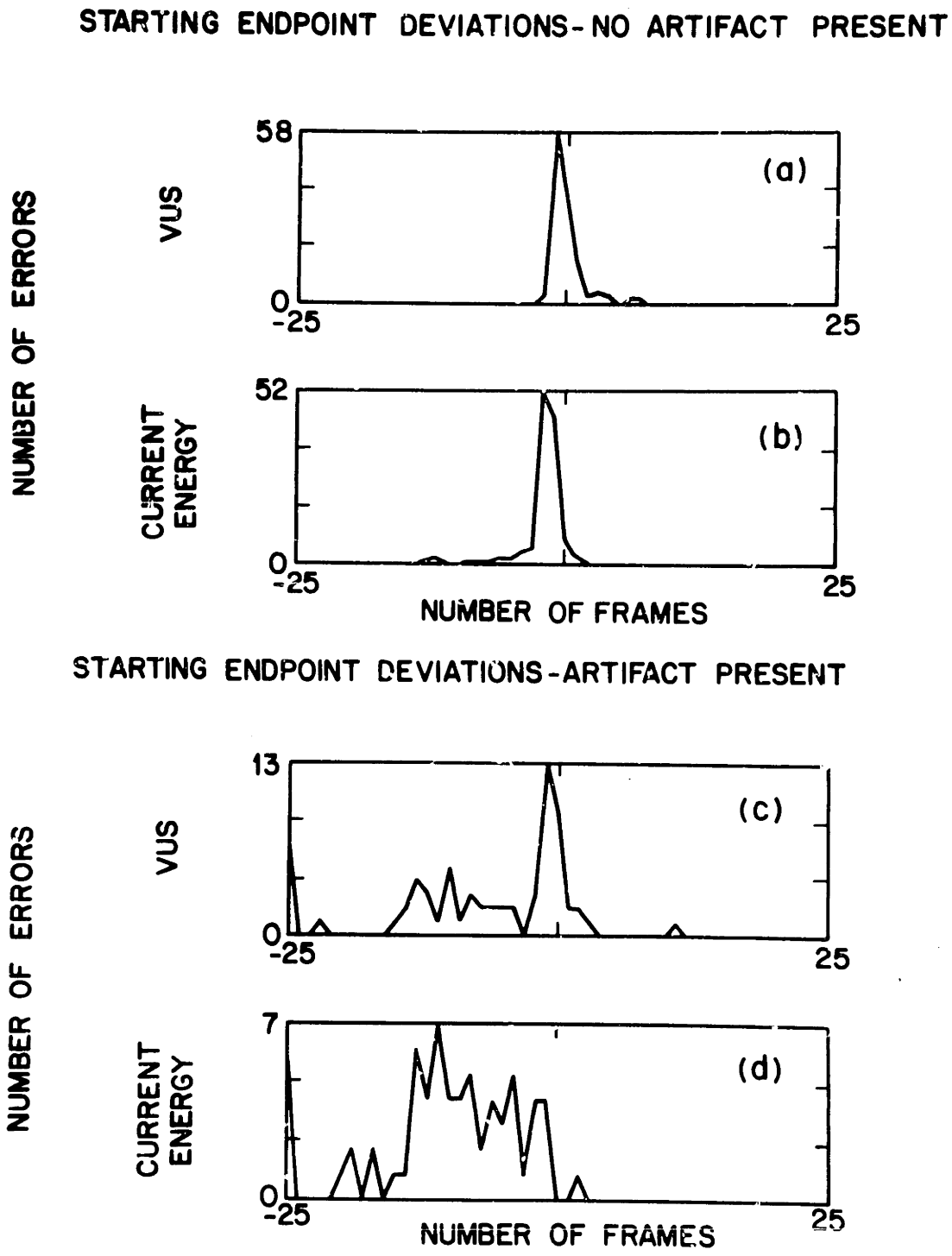


Figure 2.31 Histograms of starting frame deviations from hand location of endpoints for VUS and current energy detectors.

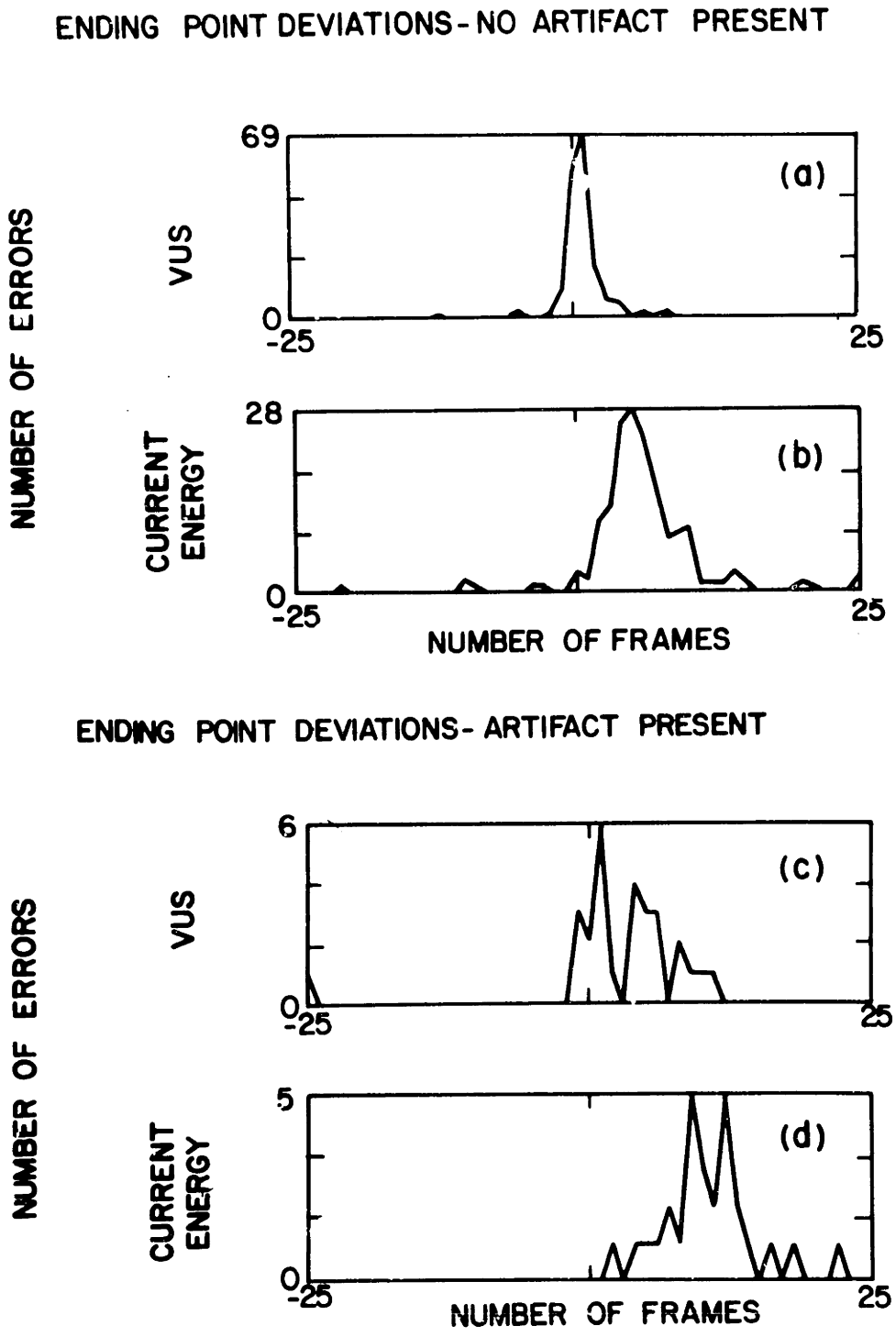


Figure 2.32 Histograms of ending frame deviations from hand location of endpoints for VUS and current energy detectors.

Both the VUS and current energy detector have recognition accuracies comparable to those obtained from hand classification. Table 2.6b gives the recognition accuracies for twelve speaker independent templates per word, using a $K = 1$ *KNN* decision rule.

	candidate position				
	1	2	3	4	5
Current Energy	39.6	57.5	67.1	71.5	74.4
VUS	44.0	58.5	66.8	72.0	76.7
Hand	46.1	61.7	70.5	75.1	81.3

Table 2.6(a) 2 speaker independent templates/word

	candidate position				
	1	2	3	4	5
Current Energy	56.0	70.5	76.8	78.7	81.2
VUS	58.5	74.6	79.8	82.4	86.0
Hand	58.3	76.2	87.0	92.2	93.8

Table 2.6(b) 12 speaker independent templates/word

TABLE 2.6 - Comparison of Recognition Scores for Hand, VUS, and Current Energy Endpoint Estimation

The recognition accuracies are plotted in Figure 2.33 as a function of the number of best candidates. The solid lines correspond to the twelve templates per word results and the dashed lines correspond to the two templates per word results. As anticipated, the results show that the best recognition scores were achieved with the hand classified endpoints. The recognition scores using the VUS endpoint detector is about 10% less for the top five candidates, and the scores for the current energy endpoint detector are approximately 15% less. For test set TS0 the VUS endpoint detector achieves recognition scores approximately 5% better than the current energy detector.

2.8 Conclusions on the Utility of Explicit Techniques for Endpoint Detection

The results presented in the previous section show that the explicit endpoint detectors do not achieve acceptable recognition rates for all recording conditions. There are several ways in

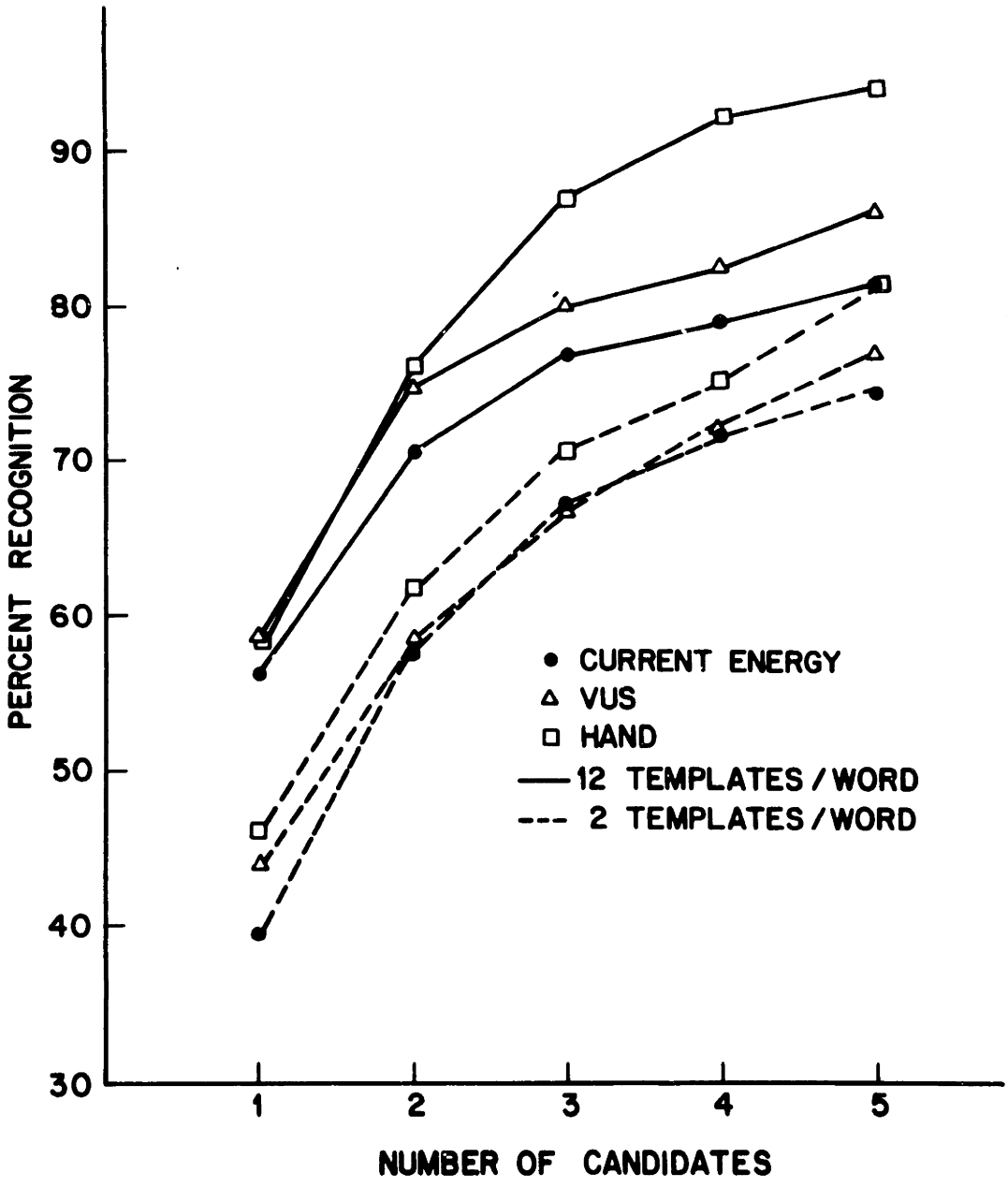


Figure 2.33 Recognition scores for VUS, current, and hand estimation of endpoints for test set TS0.

which these endpoint detectors have been seen to fail. Often they exclude low energy fricatives and include surrounding silence or artifacts. The results show that the explicit endpoint detectors are not capable of handling all of the situations encountered in a practical recording environment.

CHAPTER 3

IMPLICIT ANALYSIS TECHNIQUES FOR ENDPOINT DETECTION

3.1 Introduction

The explicit techniques for endpoint detection, discussed in the last chapter, all had in common the property that the endpoints were determined prior to, and independent of, the actual recognition process. Conversely, the implicit techniques for endpoint detection have the property that the endpoints are determined by the recognition and decision stages of the recognition system - i.e., there is no separate series of computations to locate the endpoints. Thus, the output of the recognizer consists of both an ordered list of recognition candidates from the decision rule, and a list of the associated word endpoints. In this chapter the class of implicit analysis techniques for endpoint detection are discussed.

In Section 3.2 an implementation of an implicit endpoint detector is proposed. The implicit algorithm is based on an unconstrained beginning and ending point dynamic time warp alignment procedure used to simultaneously determine the recognition candidates and the word endpoints. Results of the performance of the endpoint detector are given in Section 3.3. A discussion of the utility of implicit analysis for endpoint detection is given in Section 3.4.

3.2 The Use of Dynamic Time Warping For Implicit Endpoint Detection

One approach to endpoint detection of isolated words is to attempt recognition for every possible combination of beginning and ending points and then select as the recognized word the one which gives the smallest distance score. Although this approach would seem to require a prohibitive amount of computation, the use of an unconstrained beginning and ending point dynamic time warping (DTW) algorithm [15] substantially reduces the computation. A block diagram of an isolated word recognition system using an unconstrained beginning and ending point DTW algorithm is given in Figure 3.1. The input speech is preprocessed as discussed in Section 2.2 and a p^{th} order (p=8) autocorrelation analysis is performed, to give a feature representation of the signal, according to Eq. (2.6). Linear prediction (LPC) analysis [13] is

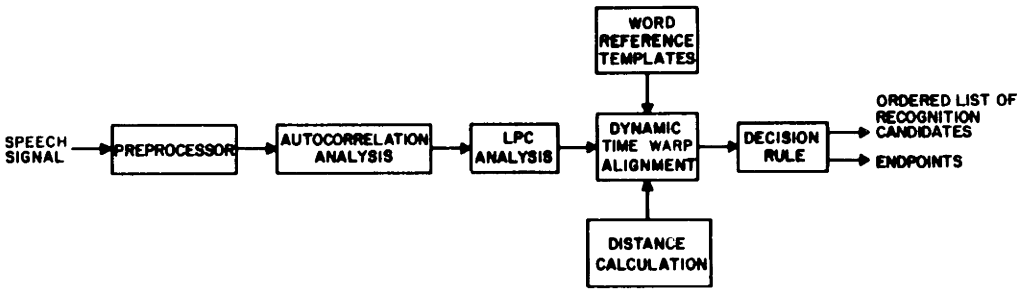


Figure 3.1 Block diagram of implicit recognition system using an unconstrained beginning and ending point DTW.

performed for each frame, and the autocorrelations of that frame are normalized by the LPC residual error according to

$$T_l(m) = R_l(m)/E_l \quad m = o,p, \quad l = l,L \tag{3.1}$$

where $\underline{T}_l$ is the normalized test pattern, $\underline{R}_l$ is the unnormalized test pattern (where each frame is a vector of autocorrelation coefficients), E_l is the linear prediction residual error for the l^{th} frame, and L is the total number of frames in the recording interval. The next stage of the recognition process uses an unconstrained beginning and ending point DTW algorithm to determine the best time alignment for each test-reference comparison. The DTW algorithm simultaneously computes both the optimal path and the associated distance for every combination of starting and ending points for a given test and reference as discussed by Myers [15]. The dynamic time warping algorithm used here is one originally proposed by Itakura [11]. The range of possible paths or time alignment regions is shown in Figure 3.2. The beginning point is assumed to be within a region of width B frames. Similarly, the ending point is assumed to be within a region of width E frames. With these assumptions, the global path constraints form the region enclosed by the heavy lines in Figure 3.2. The global path constraints serve to limit the maximum compression or expansion of a test pattern relative to a reference pattern to a range of 1/2 to 2. The extended parallelogram formed by these global constraints defines the region of allowable time alignment paths.

By extending the range of allowable starting and ending points to encompass the entire range of the test (i.e., essentially all L frames), the region of possible paths increases markedly as shown in Figure 3.3. Since there is no a priori knowledge of where within the test the beginning or ending frames lie, it is necessary to search a region of the type shown in Figure 3.3. If the length of the w^{th} reference word in frames is denoted as K_w, then the beginning region is defined for

$$b_1 = 1 \leq b \leq L - \frac{K_w - 1}{2} = b_N \tag{3.2a}$$

and similarly the ending region is defined for

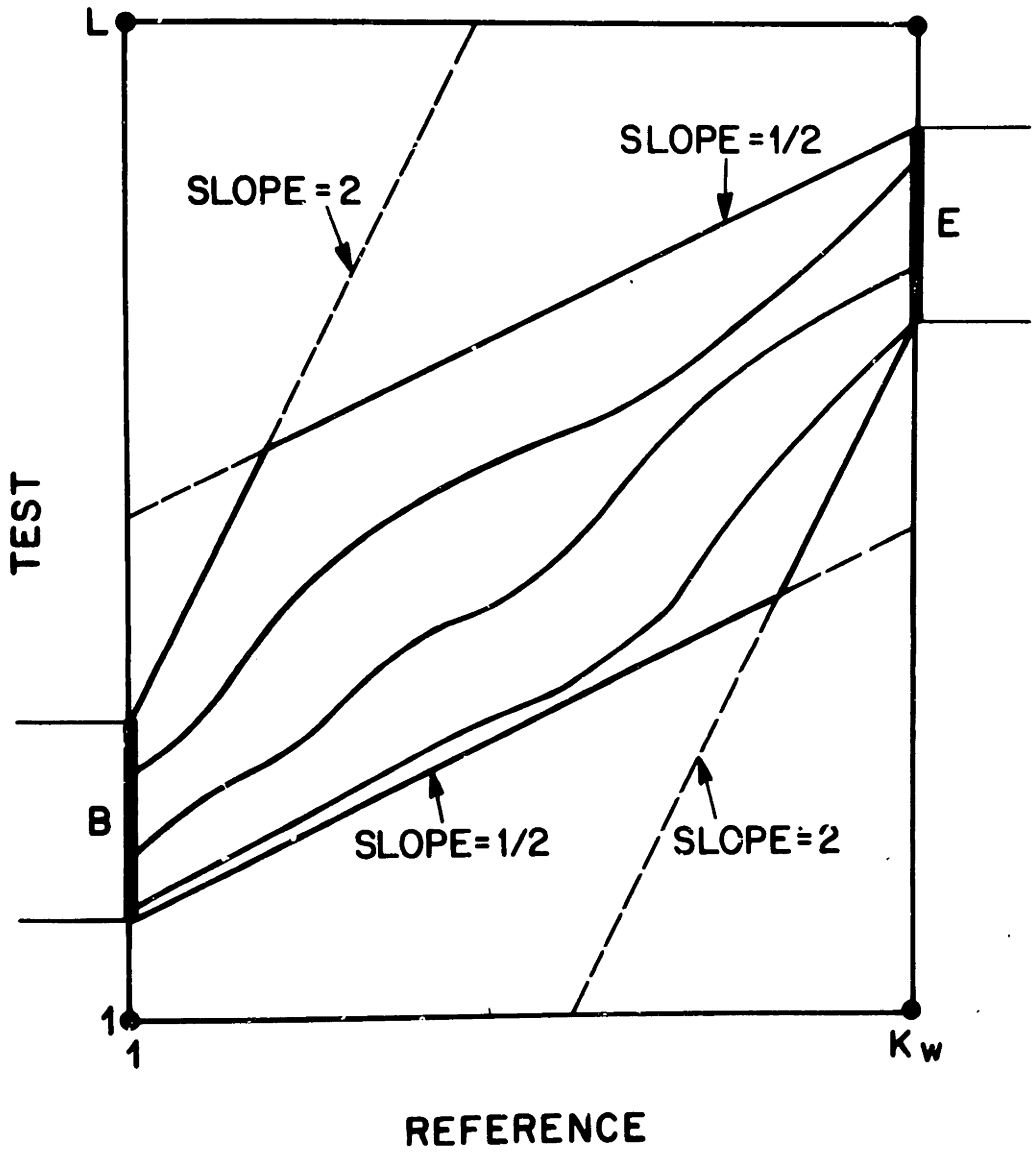


Figure 3.2 Global constraints on paths for unconstrained endpoint DTW.

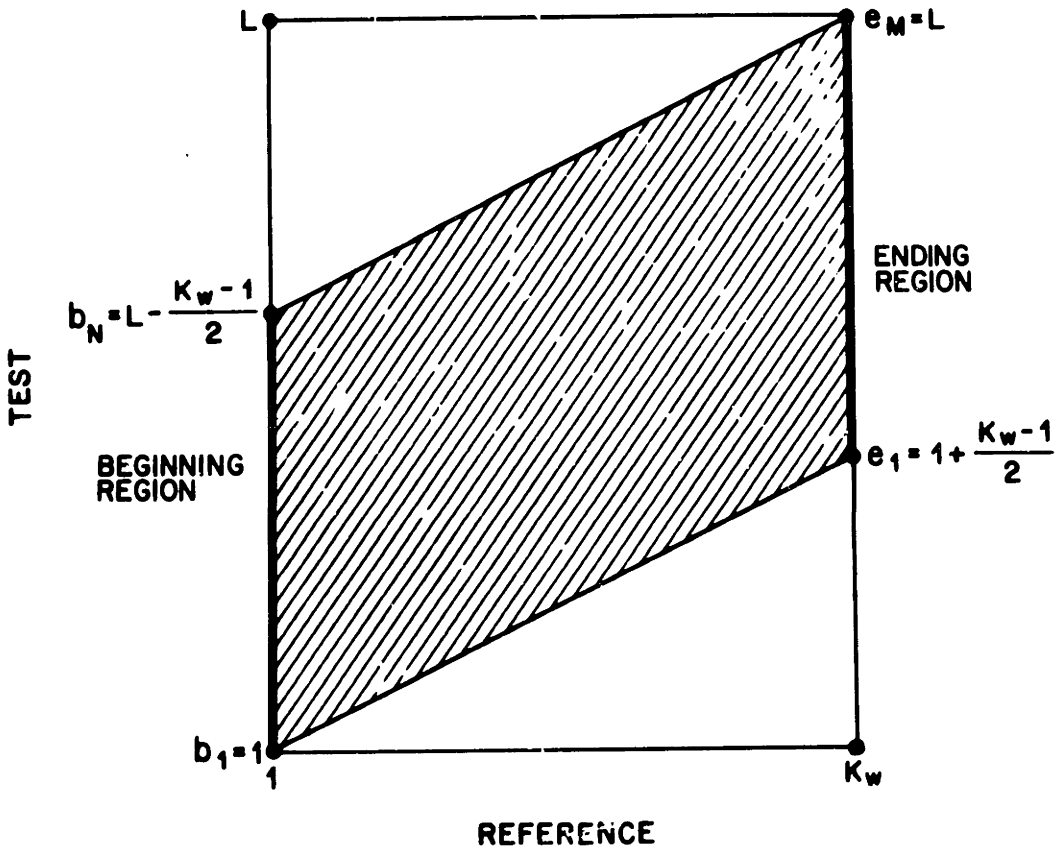


Figure 3.3 Region of allowable paths for implicit endpoint implementation.

$$e_1 = 1 + \frac{K_w - 1}{2} \leq e \leq L = e_M \tag{3.2b}$$

The beginning and ending regions, as expressed by Eq. (3.2), are determined by the global path constraints. For each reference template, the entire region of Figure 3.3 with beginning and ending regions of Eqs. (3.2) must be searched to find the optimal time alignment path and the associated word distance. The endpoints associated with the optimal path and the corresponding distance for each reference are passed to the decision stage of the recognizer. The distance associated with the w^{th} reference word is given by

$$D(w) = min \quad \hat{D}(w, B(w), E(w)) \tag{3.3}$$
$$B(w) \in \{b_1 ... b_N\}$$
$$E(w) \in \{e_1 ... e_M\}$$

where $D(w)$ is the minimum distance for the w^{th} reference, $B(w)$ and $E(w)$ are the associated beginning and ending points, and $\hat{D}(w, B(w), E(w))$ is the matrix of the minimum accumulated distances associated with beginning points $B(w) \in \{b_1 ... b_N\}$ and ending points $E(w) \in \{e_1 ... e_M\}$ for reference word w.

The decision rule determines an ordered list of recognition candidates from the generated distances, $D(w)$, associated with all the reference words. The candidates are ordered according to the distance scores of Eq. (3.3), i.e., the reference with the smallest distance is chosen as the first candidate, the reference with the next smallest distance is chosen as the second candidate. Using the above approach it has been found that the distance between the test and the correct reference template is generally small and in the range one would expect for a correct reference-test word recognition distance. By way of example Figure 3.4a shows a histogram of distances obtained when the test and reference words were the same for one of the talkers who recorded the test material. The data plotted includes all three repetitions of the vocabulary for that speaker. The solid curve is the measured histogram of distances and the dashed curve is a Gaussian fit to the data. It can be seen from Figure 3.4a that the average distance for correct words is about 0.4 with a standard deviation of about 0.1. Thus, the dynamic time warping alignment procedure just described does provide reasonably small distances when the test is

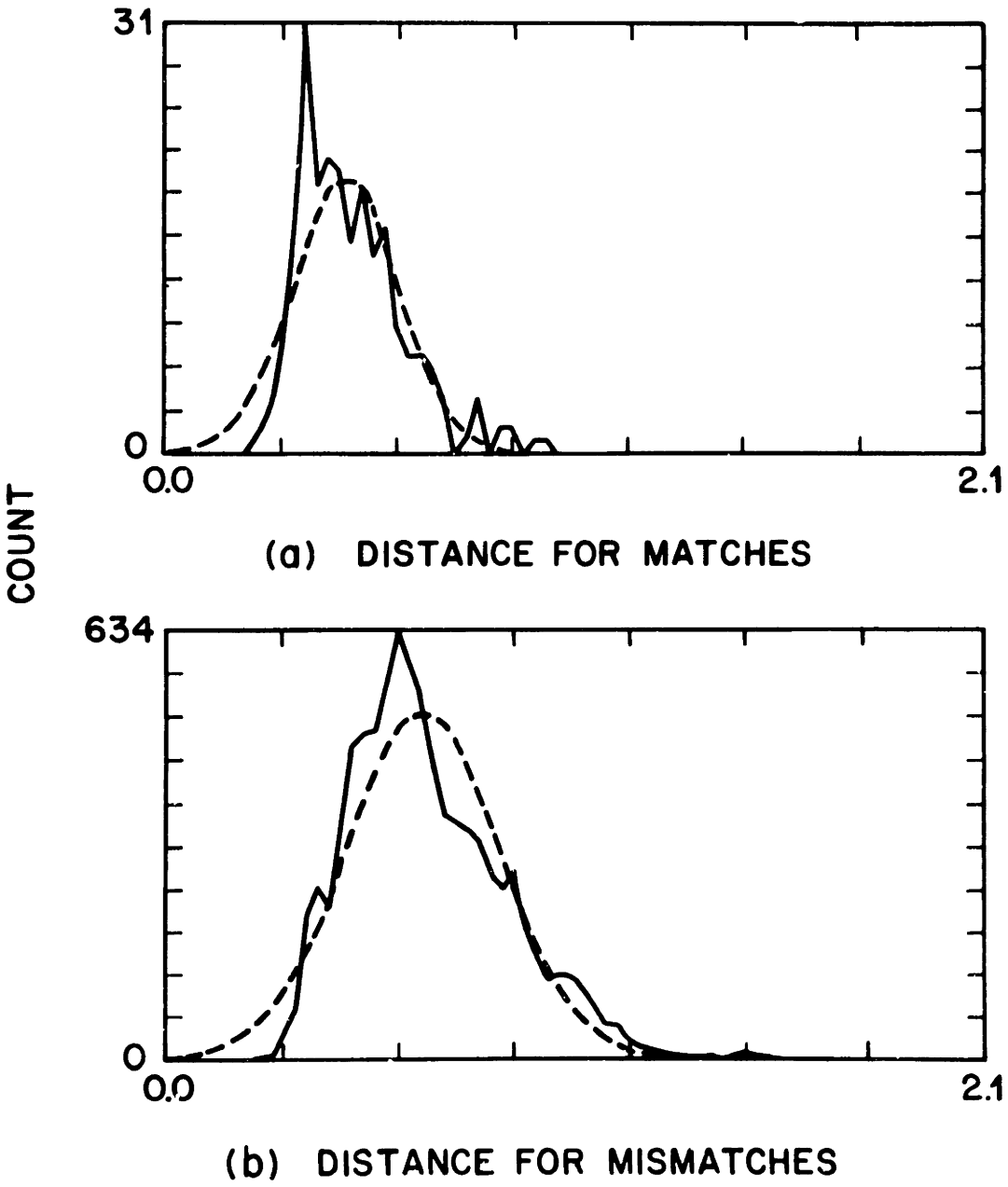


Figure 3.4 Histograms of distances for correct and incorrect references for good speaker.

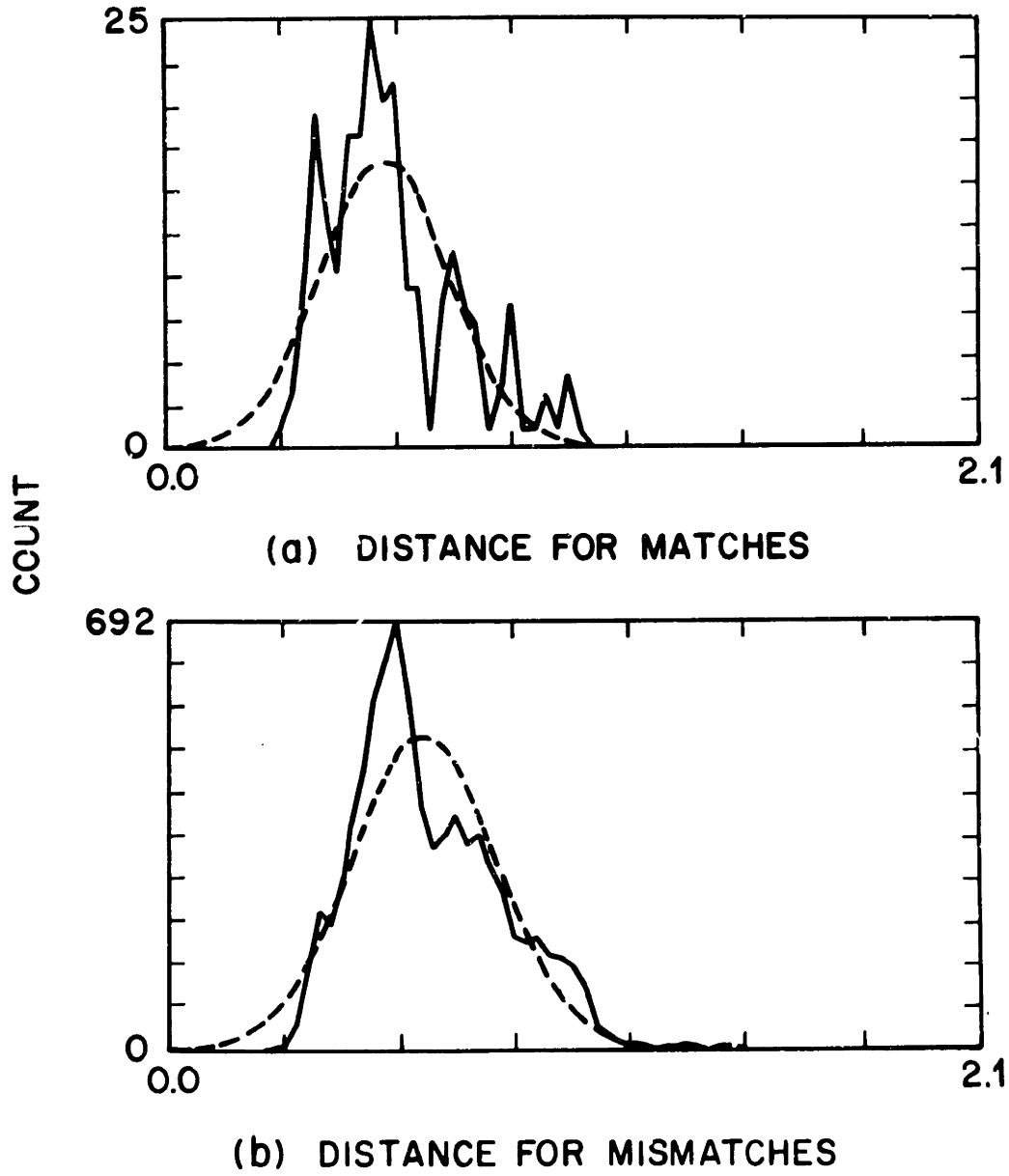


Figure 3.5 Histograms of distances for correct and incorrect references for poor speaker.

compared to the correct reference pattern. However, this is not the whole picture since what is equally, or perhaps more important is how well the test pattern matches incorrect reference words. This result is illustrated in the distance histograms of Figure 3.4b for the same speaker where the test and reference words were different. It can be seen that there is substantial overlap in the distributions, thereby implying a substantial error rate in recognition. Based on a simple model, and using the Gaussian fits to the data [16], the estimated error probability for the data of Figure 3.4 was 28.2%.

A similar set of histograms of word distance is shown in Figure 3.5 for a different speaker. In this case there is almost total overlap in the distributions, thereby implying an even higher error rate than in the previous example. The conclusion that is drawn from these figures is that by allowing the time alignment path to lie anywhere within a large, unconstrained region, the DTW algorithm has been provided with too much freedom to find the best path for each reference word. Thus, although it will generally find the correct path for the matching template, it may find an equally good path for an incorrect template.

A formal evaluation of this method was made using the standard data set presented in Section 2.6. The results of the tests are given in Section 3.3.

3.3 Performance Evaluation of the Dynamic Time Warping Algorithm for Endpoint Detection

The implicit endpoint detector based on the DTW algorithm, as described in Section 3.2, was tested on the data set TS1. Recognition accuracies were determined using both two and twelve speaker independent reference templates per word. The recognition scores, averaged over all of the speakers, are plotted as a function of the recognition candidate in Figure 3.6 for both the two and twelve templates per reference word. The recognition scores for the individual speakers are given in Table 3.1. The table shows the wide range of recognition scores among the speakers. However, even for the best speaker, the implicit algorithm has about 10% - 20% lower recognition accuracy than the explicit techniques of Chapter 2. Figure 3.7 shows some examples of the endpoints as located by this endpoint detector. In parts a and b, the endpoints are seen to be located fairly accurately. However, parts c and d show cases of grossly

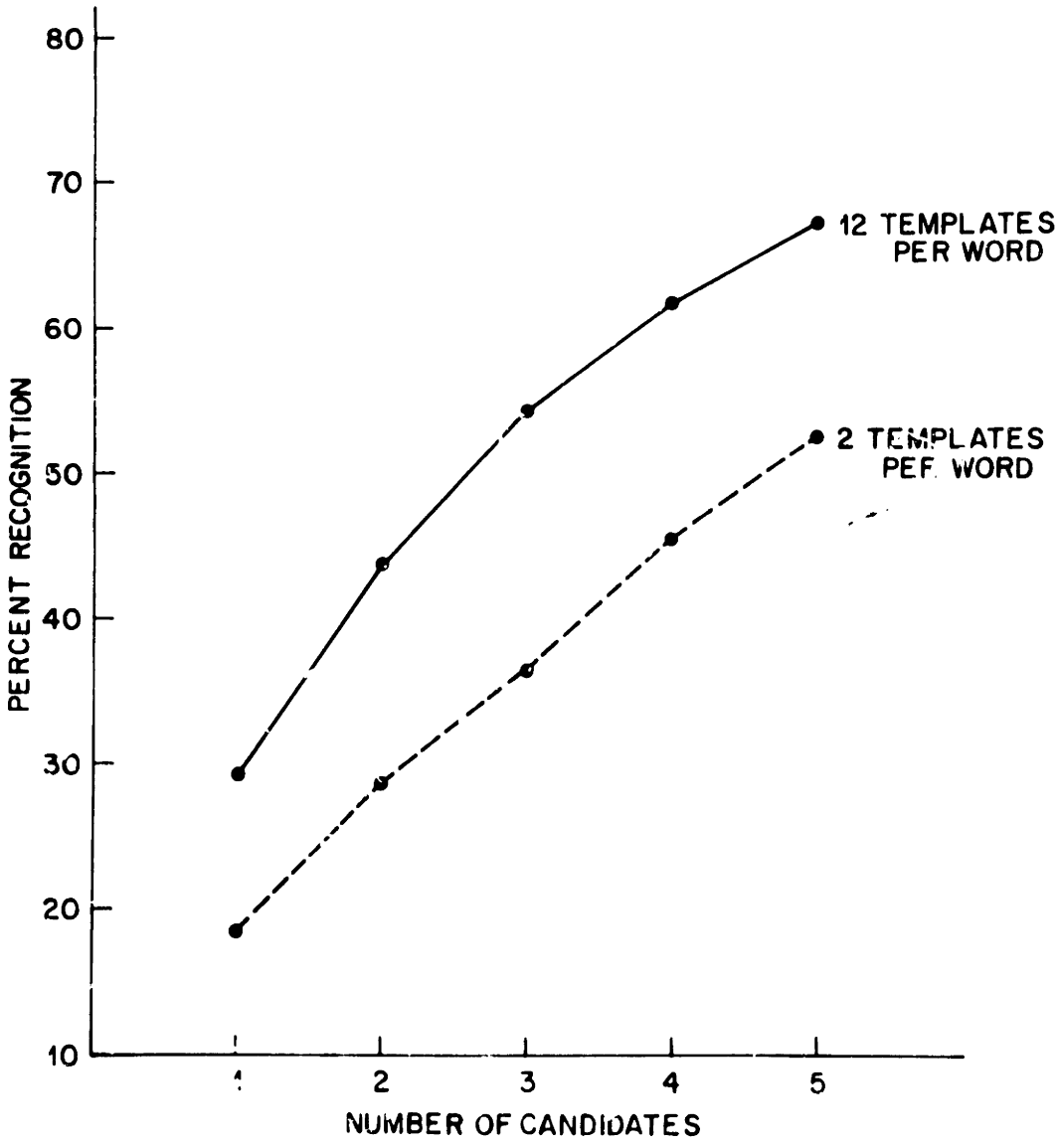


Figure 3.6 Recognition scores for unconstrained beginning and ending point DTW.

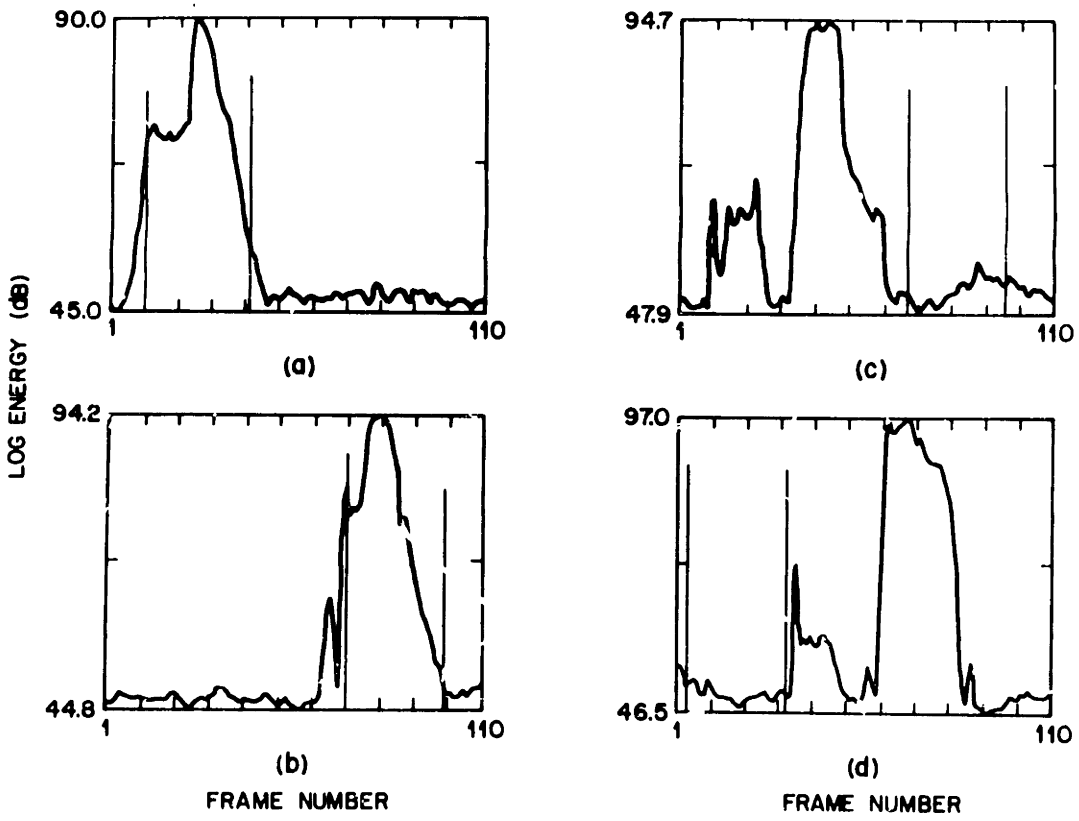


Figure 3.7 Examples of endpoint locations for unconstrained beginning and ending point DTW.

incorrectly located endpoints. The endpoint detector is seen to perform extremely poorly, on average.

		candidate position				
speaker	# tests	1	2	3	4	5
1	114	17.5	31.6	41.2	55.3	59.7
2	117	17.1	19.7	24.8	33.3	43.6
3	116	30.2	46.6	50.0	57.7	62.9
4	112	28.6	34.8	41.7	49.1	52.7
5	117	12.8	28.2	41.0	49.6	55.6
6	117	7.7	12.8	19.7	23.9	32.5
7	117	15.4	29.9	39.3	52.1	60.7
8	117	18.8	31.6	41.0	47.8	60.7
9	117	16.2	20.5	26.5	36.8	44.4
10	117	22.2	34.2	41.0	51.3	57.3
TOTAL	117	18.6	28.9	36.5	45.7	52.7

a. 2 speaker independent reference templates per word

		candidate position				
speaker	# tests	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

b. 12 speaker independent reference templates per word

TABLE 3.1 Recognition scores for DTW based endpoint detector (percent correct in candidate position i, i=1,2,3,4,5)

3.4 Conclusions on the Utility of Implicit Algorithms for Endpoint Detection

The implicit endpoint detector has been seen to perform extremely poorly. There is a tendency for it to make gross errors. Such errors are not permissible for a word recognition system. Figure 3.8 shows the locations of the endpoints as determined by the implicit endpoint detector as compared to those located by the current energy endpoint detector for the same words. In each part, the top plot is the log energy contour with the endpoints as located by the current energy endpoint detector. When no vertical lines have been drawn, the current energy

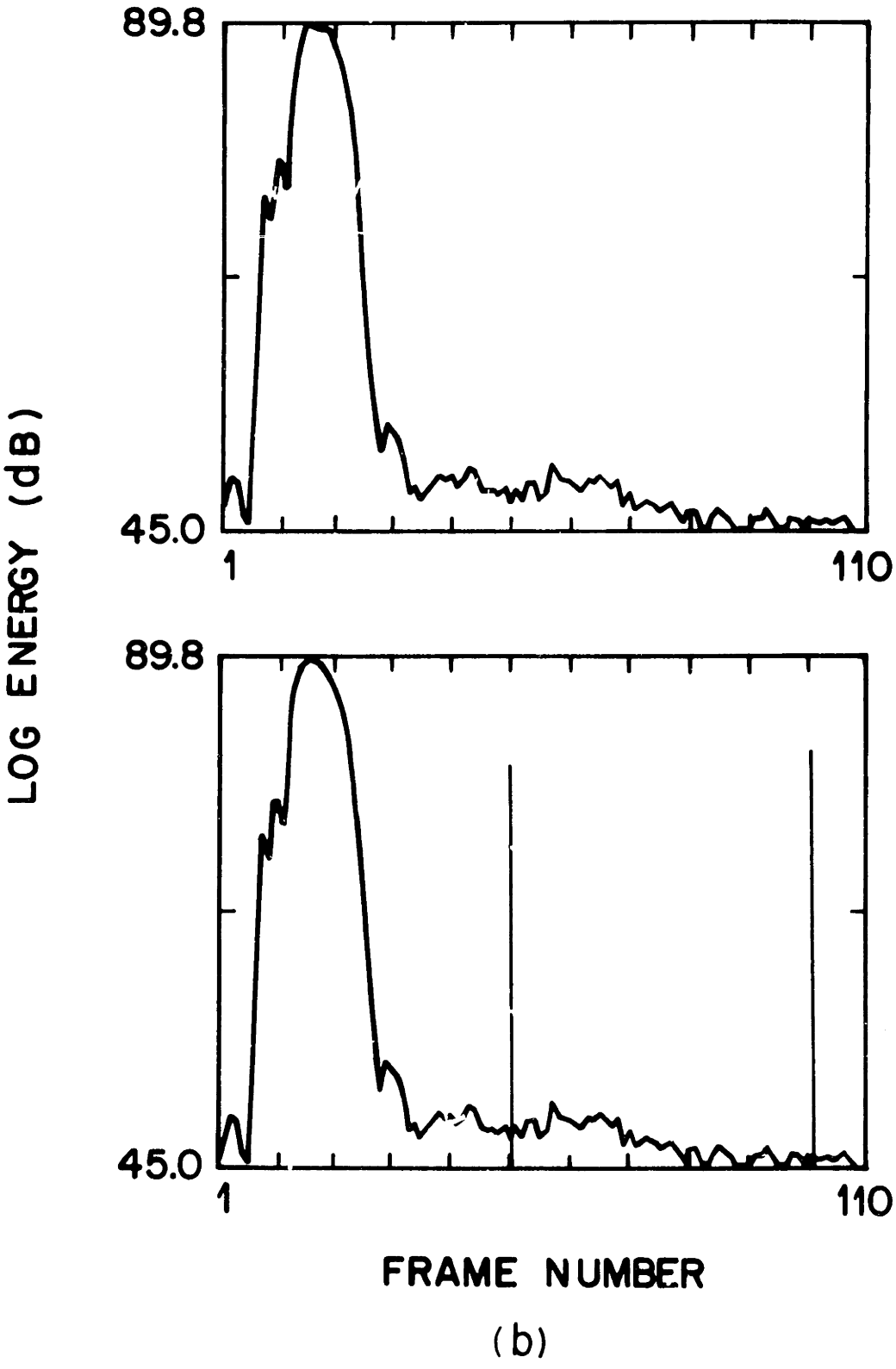


Figure 3.8 continued

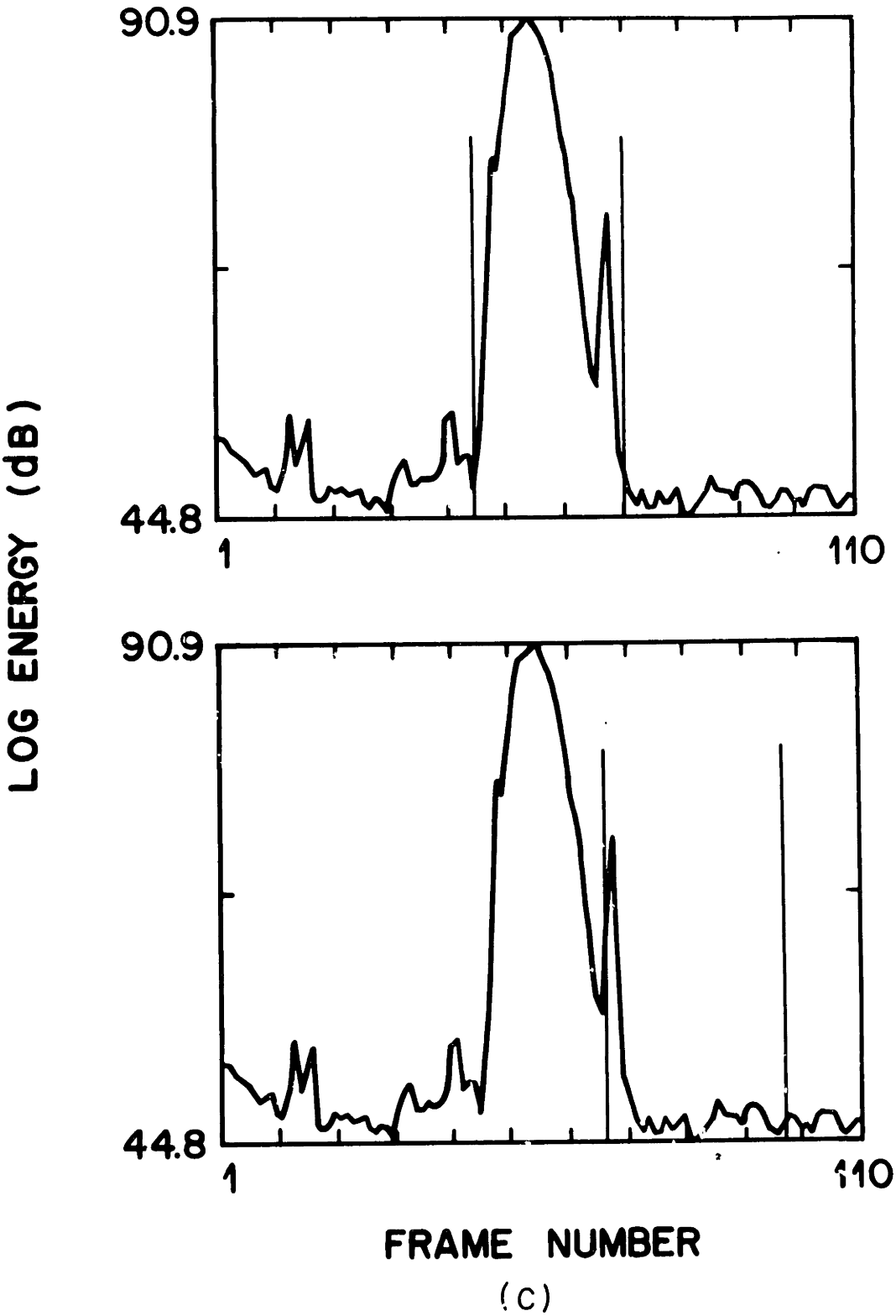


Figure 3.8 continued

detector has failed to locate the endpoints. The second plot corresponds to the endpoint locations as determined by the unconstrained endpoint DTW. In Figure 3.8a the current energy endpoint detector has included an artifact prior to the actual word within the endpoints. The DTW algorithm has accurately located the endpoints for the same word. In Figure 3.8b both the current energy and DTW endpoint detectors have failed to locate the endpoints of the word. As shown in Figure 3.8c, often when the current energy endpoint detector is able to locate the endpoints accurately, the implicit endpoint detector is apt to fail.

The method chosen to illustrate the implicit approach to endpoint detection is not suitable for use in practical isolated word recognition systems for several reasons. Due to the high computational requirements the endpoint detector could not be readily implemented in a real-time environment. The required computation is too great even when apriori knowledge of a region in which the spoken word occurred is available. In addition, the results of this investigation show that the totally unconstrained beginning and ending point DTW allowed the recognition algorithm too much freedom in the determination of the endpoints. The capability of the recognizer to throw away portions of the test utterance enabled it to match the incorrect references to the test as well or better than the matches to the correct references.

CHAPTER 4

HYBRID ANALYSIS TECHNIQUES FOR ENDPOINT DETECTION

4.1 Introduction

The hybrid techniques for estimation of the endpoints of an isolated word incorporate aspects of both the explicit and implicit methods. The canonic form for a recognition system using a hybrid endpoint detector is shown in Figure 1.3c. As for the explicit endpoint detection methods, the hybrid endpoint detection methods have a separate stage for endpoint detection while providing some of the freedom afforded by the implicit approach. The endpoint detector may supply a single estimate or several estimates for each of the endpoints to the recognition stage, with the final decision as to the "correct" endpoints being made by the recognition stage. Alternatively, a hybrid system may use feedback from the recognition or decision stage of the system to obtain a revised estimate of the endpoints. In this chapter hybrid techniques for endpoint detection are investigated.

In Section 4.2 two implementations of hybrid endpoint detectors are proposed. Results of performance evaluations are presented in Section 4.3. The utility of hybrid techniques for endpoint detection is discussed in Section 4.4.

4.2 Hybrid Techniques for Endpoint Estimation

In this section two hybrid techniques for endpoint detection are proposed. In Section 4.2.1 a modified version of the current energy endpoint detector presented in Chapter 2 is suggested. In Section 4.2.2 an endpoint detector which locates the endpoints based on a comparison of the test pattern and each reference pattern is described. However, it should be noted that these endpoint detectors are not simply combinations of the explicit and implicit techniques described in Chapters 2 and 3.

4.2.1 Modified Energy Endpoint Detector

The endpoint detector presented in this section is derived from the current energy endpoint detector described in Section 2.4.1. The main problems with the current energy detector are:

1. The reject rate is high, approximately 30%.
2. Extraneous, relatively isolated artifacts occurring close to the spoken word are sometimes included within the endpoints.
3. There is often substantial breath noise included within the endpoints if breath noise is present.

In addition the discriminator described in Section 2.4.1 is based on complicated heuristics.

The endpoint detector described in this section was developed to eliminate the aforementioned problems and to provide an algorithmic decision rule. A block diagram for the modified energy endpoint detector is shown in Figure 4.1. The endpoints of the isolated word are located from the time sequence of the signal energy (the zeroth delay autocorrelation) as given by Eq. (2.6). The modified energy endpoint detector consists of three subunits; the second level preprocessor, the energy pulse detector, and the decision rule. The second level preprocessor performs the operations shown in Figure 2.12. However, for this endpoint detector the first 9 frames of the recording interval are not ignored in an attempt to eliminate transient effects. Thus the normalized, integer valued level array is created for all frames of the recorded interval, frames 1 to L.

The energy pulse detector is based on the same principles as the energy pulse detector described in Chapter 2. The energy pulse detector incorporated in this endpoint detector uses on a triple thresholding technique instead of the double thresholding technique described in Section 2.4.1. The triple thresholding technique is illustrated in Figure 4.2. The third threshold, K_3, is used in locating the end of the pulse instead of the threshold K_1. The threshold for the ending frame has been set higher than the threshold for the beginning frame as the energy at the end of the word tends to have a more gradual slope than the energy at the word

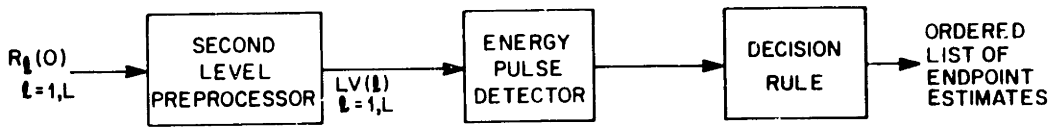
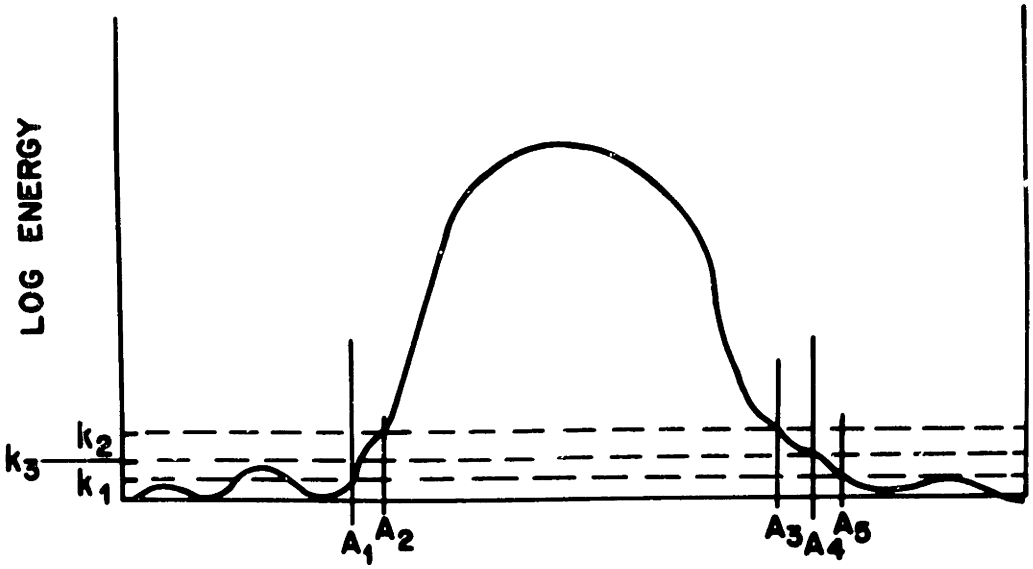
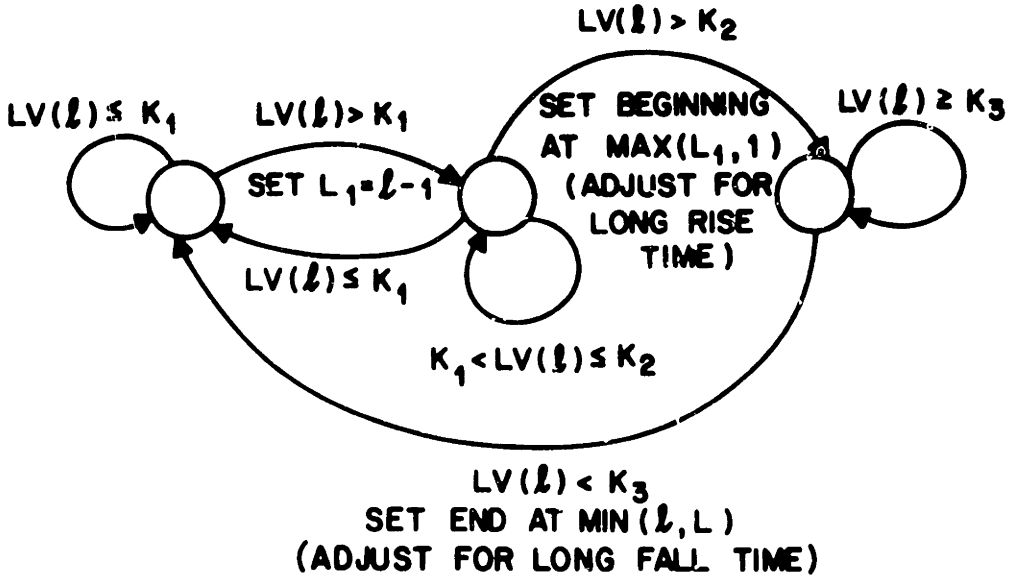


Figure 4.1 Block diagram of modified energy endpoint detector.



(a)



(b)

Figure 4.2 Triple threshold technique.

beginning. Using the triple thresholding technique the endpoints for the word in Figure 4.2a would be located at A_1 and A_4 whereas the double thresholding technique would have located the endpoints at A_1 and A_5. The corresponding state diagram for the triple thresholding technique is given in Figure 4.2b.

Included within the energy pulse detector are internal backup frame counters to detect potential breath noise and to automatically adjust the pulse boundary accordingly. The pulse boundaries are cut in (decreasing the total duration of the pulse) if the rise or fall time of the pulse energy is too long. These counters eliminate low energy breath noise during the detection of the energy pulses.

The use of the backup frame counters for the energy pulses is illustrated in Figure 4.3. The first time the log energy of the signal exceeds the lower threshold, K_1, is the frame A_1. The log energy contour continues to rise and exceeds K_2 at frame A_2. At the end of the pulse the log energy falls below K_2 for the first time at A_3, and continues to decrease until it falls below K_1 at frame A_4. If the rise time, A_2-A_1, is larger than threshold, MAXLONG frames, or the fall time, A_4-A_3, is larger than MAXLONG frames, the beginning or ending frame location is adjusted according to:

$$\text{If } A_2-A_1 > MAXLONG, \text{ set the beginningframe at } A_2-NSUB. \tag{4.1a}$$

$$\text{If } A_4-A_3 > MAXLONG, \text{set the ending frame at } A_3. \tag{4.1b}$$

where for the implementation MAXLONG = 5 frames and NSUB = 3 frames.

The triple thresholding technique, including the backup counters, is iteratively applied from left to right, that is in increasing time, for the duration of the recorded interval. The endpoints of the pulses are required to be found in pairs. Thus, first a beginning frame is located and then the associated ending frame. A flow chart fort the location of the beginning frame is shown in Figure 4.4a. The corresponding flow chart for the ending frame is shown in figure 4.4b. The flow chart shows the use of the backup counters for the beginning and ending frames. Not shown in the flow chart is an internal check for the end of the recording interval.

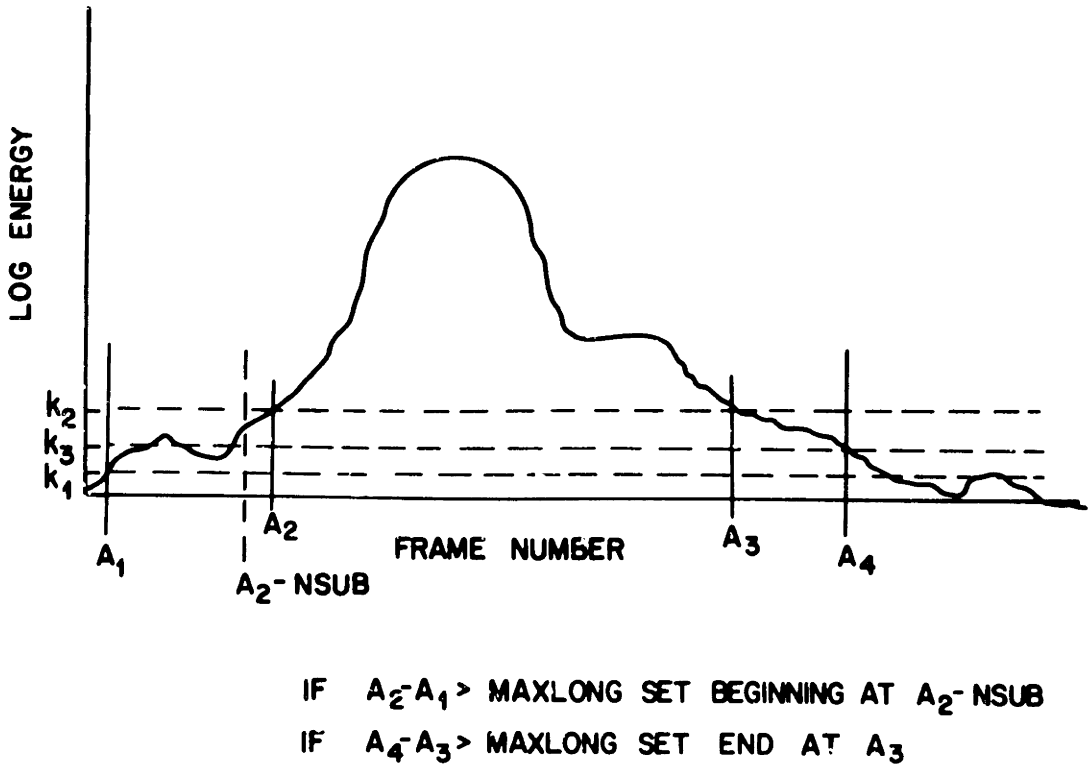


Figure 4.3 Energy pulse backup counters.

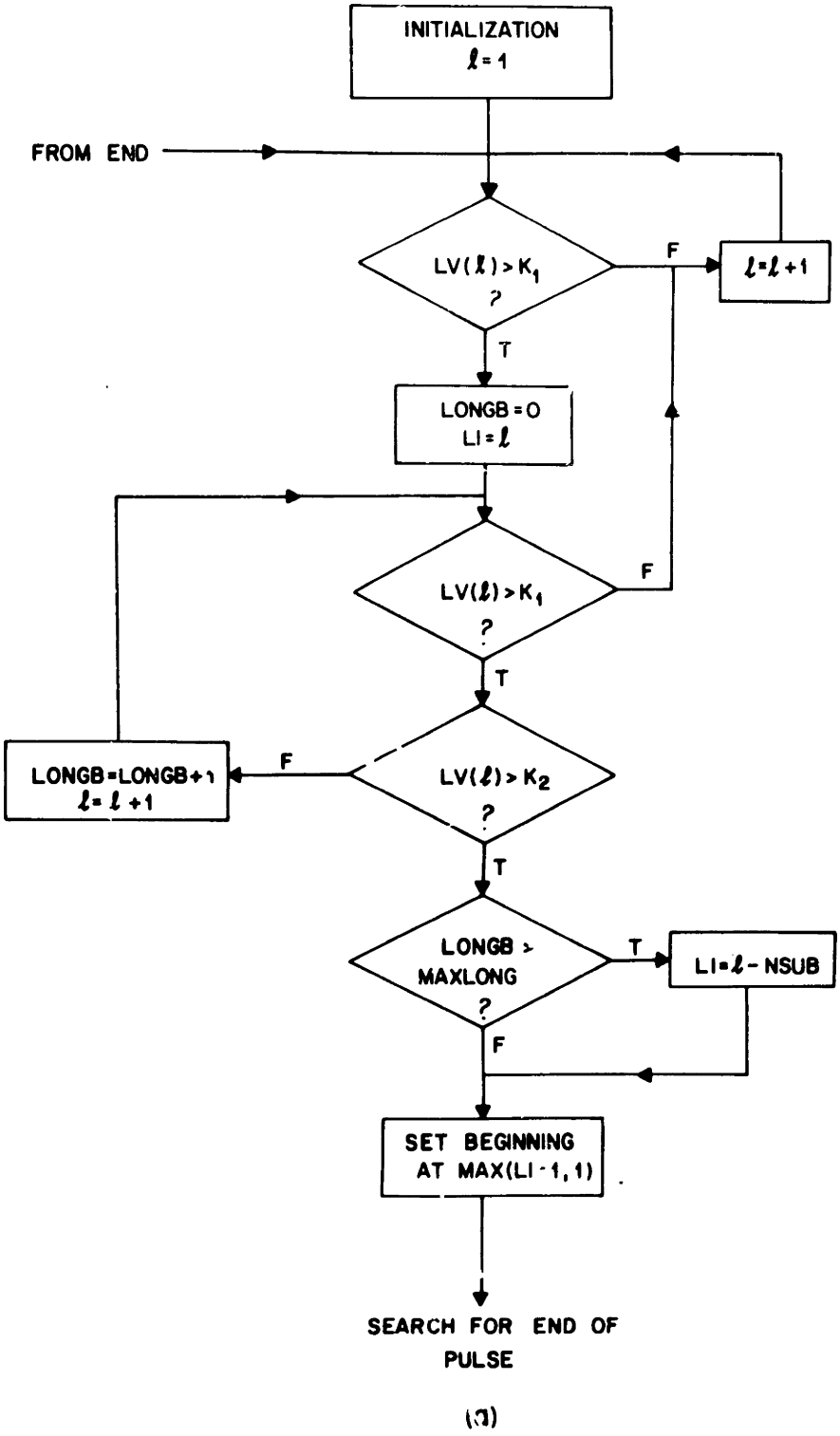


Figure 4.4 Energy pulse detector.

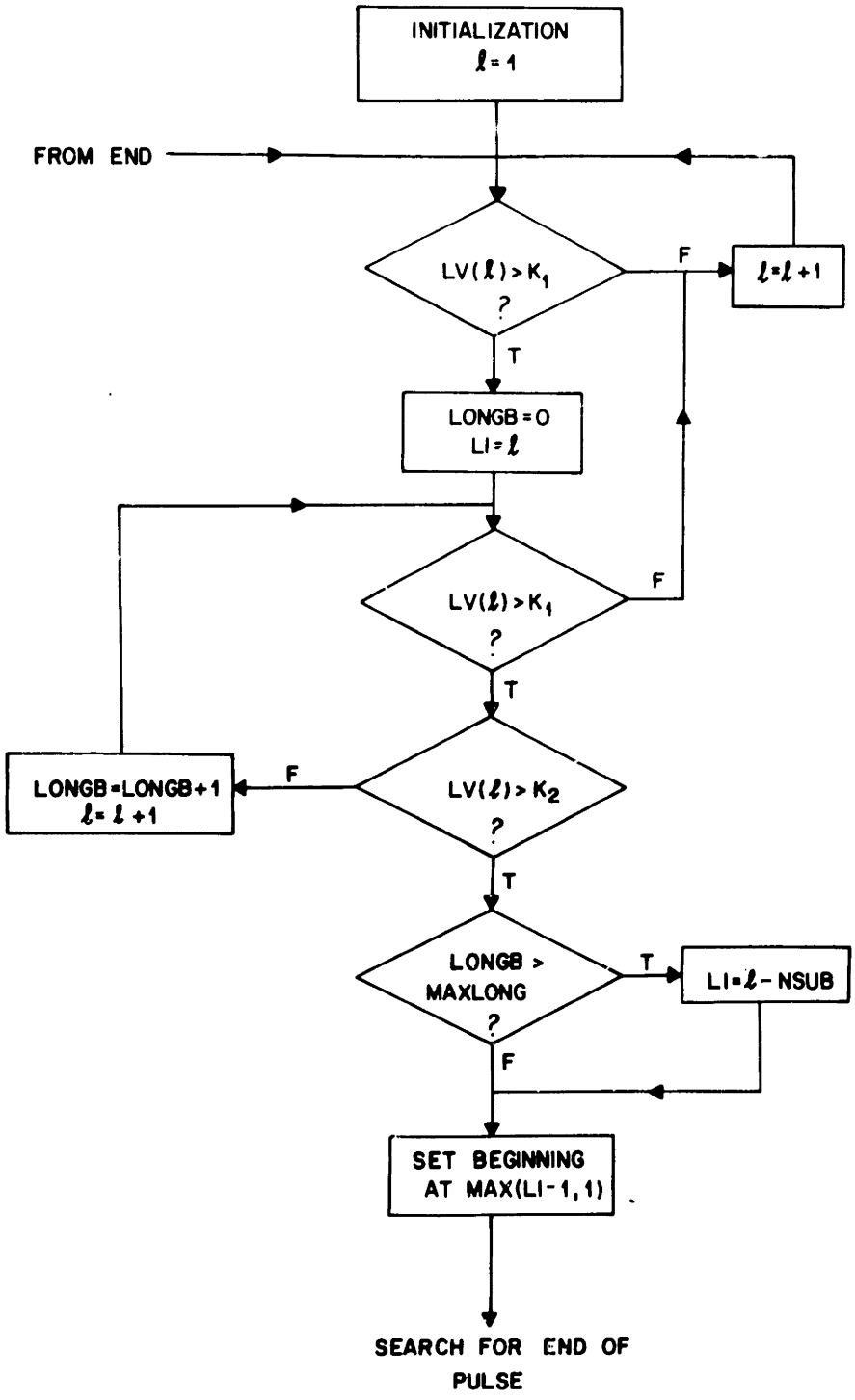
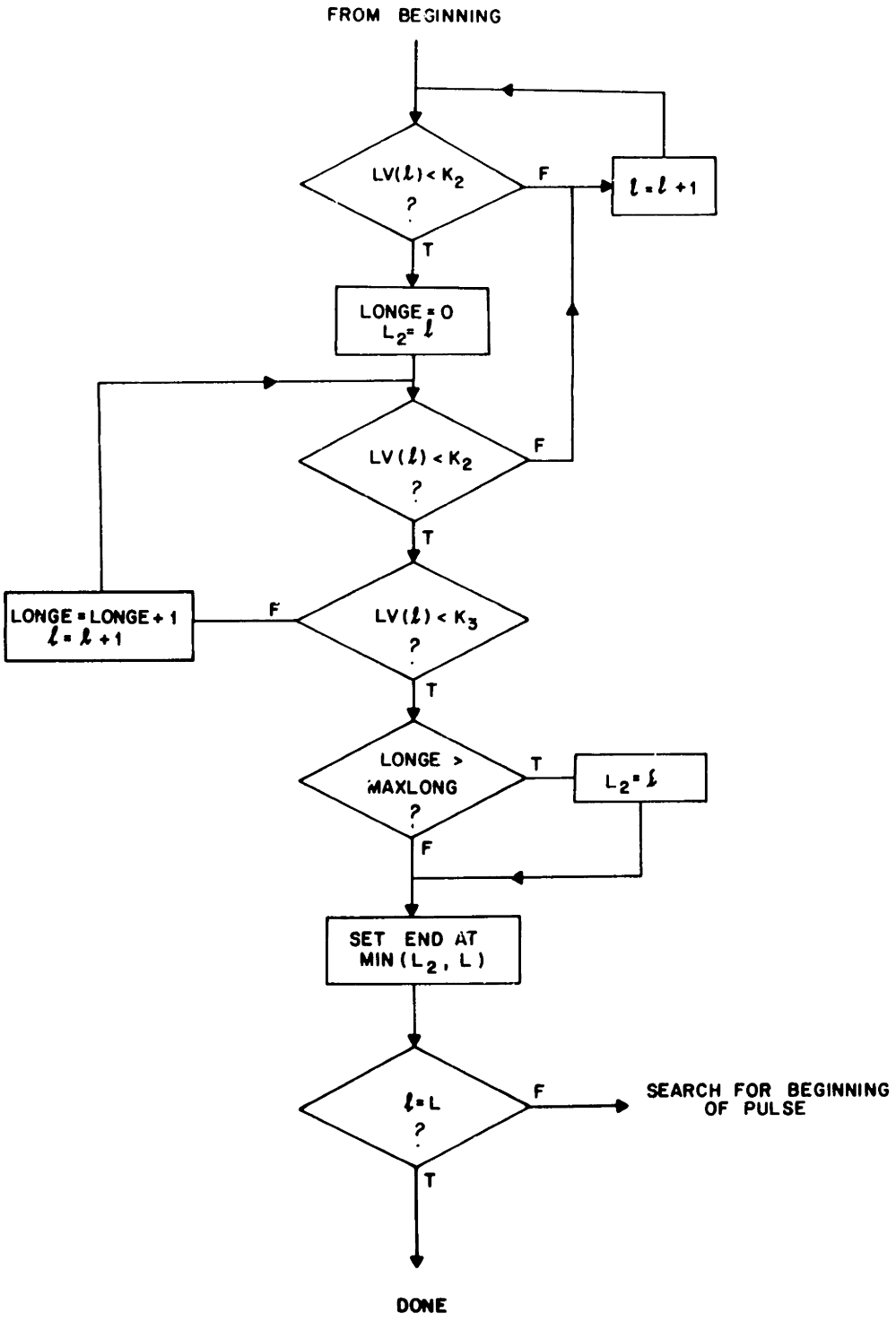


Figure 4.4 Energy pulse detector.



(b)

Figure 4.4 continued

If the beginning of a pulse has been located and no corresponding ending frame has been found prior to the last frame of the recording interval, L, the end of the pulse is forced to occur at L. If part of the speech or an artifact overlaps the boundary of the recording interval, the endpoint of the energy' pulse will occur on the boundary. The array containing the endpoints of the energy pulses located, along with the log energy contour (LV) are passed to the decision rule.

The decision rule determines an ordered list of endpoint candidates from the log energy contour and the energy pulse array. The decision rule is based on the following assumptions:

1. The isolated word whose endpoints are to be determined consists of one or several energy pulses.
2. The frame in the log energy contour with the maximum value will always be contained within the spoken word.
3. The larger the stop gap between two energy pulses the less likely that they form one multiple-pulse word.

The decision rule consists of three stages. In the first stage all easily detected artifacts are eliminated. In the second stage a smoother is applied to the array of energy pulses retaining both the smoothed and unsmoothed pulse endpoints. The third stage determines an ordered list of endpoint candidates from the smoothed and unsmoothed pulse arrays. Finally there is a check to insure that the detected word does not overlap the boundary of the recorded interval. The canonic representation of the decision rule is shown in Figure 4.5.

The first three blocks of the decision rule comprise the first stage of the decision process. The operations of the first stage are a major component in the reliable performance achieved by the modified energy endpoint detector. The block denoted the speech and boundary check determines whether or not speech is present in, or has overlapped, the boundary of the recorded interval. This is accomplished by requiring that the maximum normalized log energy of the signal be greater than a threshold value and that the log energy of the first and last frame of the recorded interval be less than that threshold value. If these requirements are not met there is substantial evidence that either there is no speech present in the test or the speech has

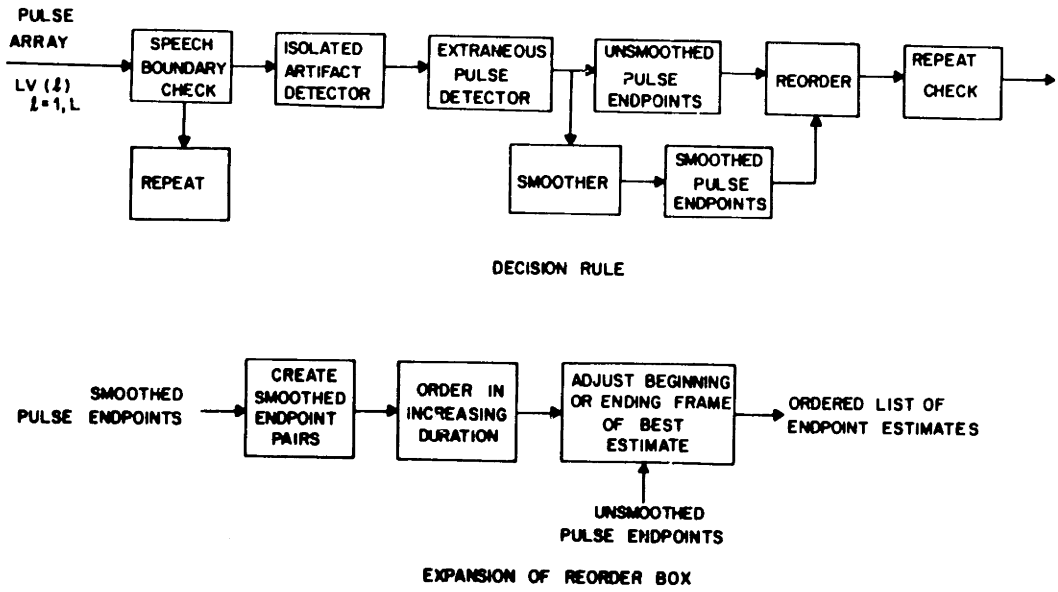


Figure 4.5 Decision rule.

overlapped the recording interval and the test is rejected.

The frame having the maximum log energy of the recording, L_{max} is located from the level array, LV, according to

$$L_{max} = argmax\ LV(l) \qquad l = 1,L \tag{4.2a}$$

where $argmax[f(x)]$ is the value of x maximizing $f(x)$. The maximum value of the level array is defined as

$$LV_{max} = LV(L_{max}) \tag{4.2b}$$

If LV_{max} is less than an empirically defined threshold, (MINDB, set at 30 dB) it is determined that no speech is present and a repeat is requested.

The boundary check requires that significant energy is not present on the boundary of the recorded interval. If $LV(1)$ or $LV(L)$ is of greater magnitude than MINDB it is assumed that the speech has overlapped the recording interval and a repeat is requested. It should be noted that most artifacts occurring on the boundary of the recording interval are not of sufficient energy to cause a rejection.

Next the energy pulse array is passed through the isolated artifact detector which eliminates all isolated pulses of less than a minimum amplitude or of less than a minimum duration. The parameters have been selected to eliminate all pulses of maximum (normalized) amplitude of less than 15 dB or of duration less than 5 frames. Pulses falling into either of the above categories are very likely to be artifacts.

Following the preliminary screening for artifacts the energy pulse array consists of one or several pulses, all of which are potentially speech. For a word containing stop gaps the energy pulses usually occur reasonably close together. It is assumed that the maximum stop gap within a word is 150 ms. Thus, the extraneous artifact detector eliminates all pulses separated from the pulse with the maximum energy by a stop gap longer than 150 ms.

Figure 4.6 illustrates the screening techniques for the artifacts. In Figure 4.6a, the repeat and boundary check are illustrated. The frame with the maximum log energy is denoted L_{max}.

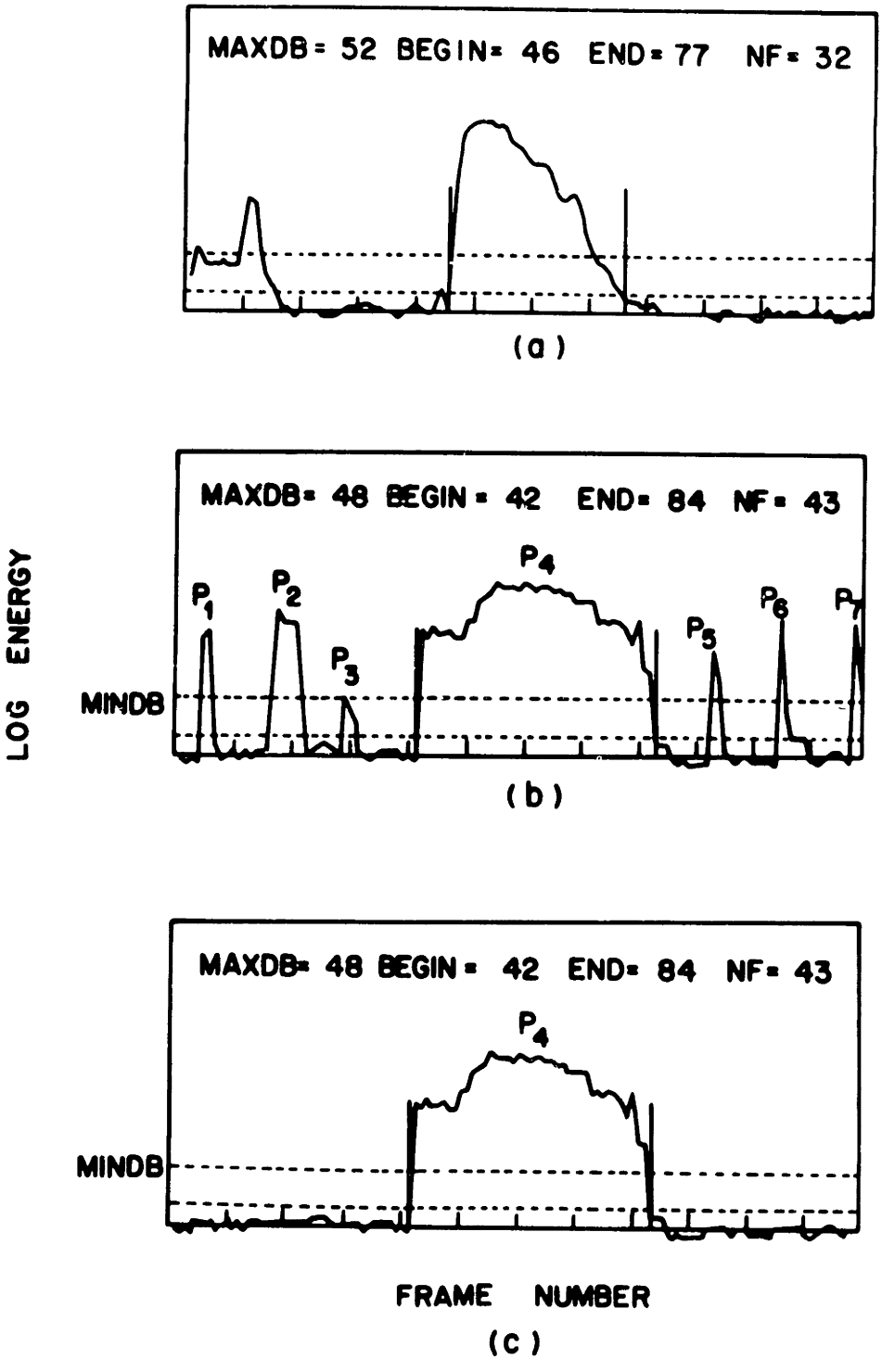


Figure 4.6 Illustration of artifact screening.

The value of the log energy contour at L_{max} is given by $LV(L_{max})$. If $LV(1)$ or $LV(L)$ is greater than MINDB, or if $LV(L_{max})$ is less than MINdB, a repeat is requested. Figure 4.6b illustrates the operations of the isolated artifact detector. Shown in the figure are seven pulses, denoted P_1 through P_7. For this example pulse P_4 contains the frame with the maximum log energy. Pulse P_3 is eliminated because its maximum amplitude is less than MINDB. The pulses denoted P_1, P_5 and P_7 have sufficient energy to potentially be speech, however, their duration falls below the minimum value of 75 ms. The array of energy pulses given to the extraneous energy pulse detector consists of the pulses P_2,P_4, and P_6. Let the gap between pulses P_2 and P_4 be denoted as X_1 and the gap between P_4 and P_6 be denoted as X_2. Then if $X_1 > MAXLONG$ pulse P_2 will be elimina ed. Similarly if $X_2 > MAXLONG$ pulse P_6 is eliminated. If both $X_1 \leq MAXLONG$ and $X_2 \leq MAXLONG$ neither P_2 nor P_6 are eliminated. For this example $X_1 > MAXLONG$ and $X_2 > MAXLONG$, thus the output of the extraneous pulse detector is the single pulse, P_4.

The refined energy pulse array and log energy contour are passed to the smoother in order to combine multiple pulses that form a single word. This array consists of all the pulses detected in the original array that are of sufficient amplitude and duration, and occur within a close enough proximity to the pulse containing the maximum energy, to be considered part of the speech. Whether or not two pulses are smoothed together is a function of the separation between the pulses. Figure 4.7 shows two energy pulses P_1 and P_2 separated by X frames. Clearly the larger of the value of X, the more likely that either P_1 or P_2 is the spoken word. As X decreases, the word may contain both P_1 and P_2. If the separation, X, is less than a defined maximum separation, NSEP, the two pulses are collapsed into one larger pulse with endpoints A_1 and A_4. If the separation X is larger than NSEP the pulses are kept distinct with the endpoints as shown. The value for NSEP was chosen as 90 ms, the largest stop gap expected within the word. Thus, the smoother collapses into one pulse any two pulses separated by less than 90 ms. If two pulses are separated by a stop gap larger than 90 ms they are treated as distinct pulses. The combining of pulses occurring close together is necessary for the accurate location of the endpoints of words with multiple energy pulses such as "eight",

"accumulate", and "repeat". However, an important feature of the endpoint detector is that the unsmoothed pulse endpoints are retained as well as the smoothed pulse endpoints.

The third stage of the decision rule is the formation and ordering of all the endpoint candidates. The operations performed are shown in the expansion of the block denoted reorder in Figure 4.5. The ordering is accomplished based on the durations of the pulses. The ordered list of endpoint candidates is simply given by all possible endpoint pairs determined from the smoothed pulse array, in increasing duration. All endpoint pairs must include the frame with the maximum energy and must be a minimum number of frames in duration. Initially the minimum duration corresponds to 300 ms and is dynamically decreased if the minimum duration requirements are not met. At this point it has been assumed that for most words comprised of multiple energy pulses the smoother has collapsed all of the pulses into one larger pulse. Thus words containing two or more smoothed pulses are likely to include an artifact within the detected word endpoints. However, all potential endpoint pairs are formed to handle the situation where a stop gap of a multiple pulsed word is larger than expected.

The beginning or the ending frame of the best endpoint pair is adjusted using the information present in the unsmoothed pulse array in order to provide a second endpoint estimate. If an adjusted endpoint pair exists it is inserted into the ordered list of endpoint estimates. The unsmoothed energy pulse array is used to adjust the beginning or ending frame of the best estimate as follows. The most general form for the best estimate is shown in Figure 4.8. The endpoints, A_1 and A_6 are those of a single, smoothed pulse. This pulse may be expressed as a pulse containing the maximum energy preceded and followed by a single pulse. This form is the most general as P_1 and P_3 may or may not exist, and P_2 may be comprised of several pulses that were smoothed together. If neither P_1 nor P_3 are present the beginning frame is A_3 and the ending frame is A_4. In this case the top candidate consists of only one unsmoothed pulse and no further action is taken. For either P_1 or P_3 present, but not both, the revised beginning or ending point is created by eliminating this pulse, i.e., the revised endpoints are A_3 and A_4. When both P_1 and P_3 are present, the pulse of shorter duration is eliminated. This assumes that, in general, an isolated artifact occurring close to the speech will be of shorter

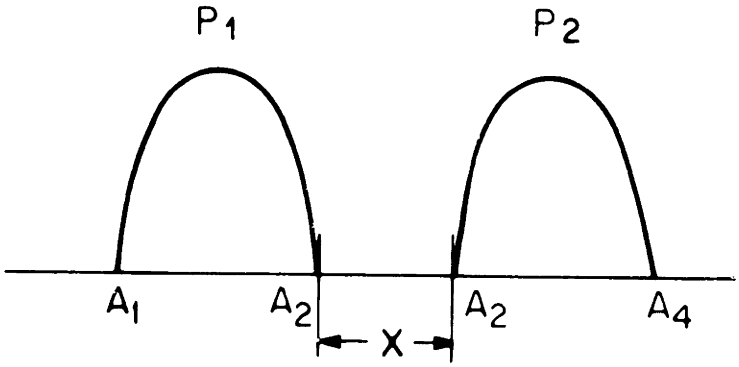


Figure 4.7 Definition of energy pulses and separation.

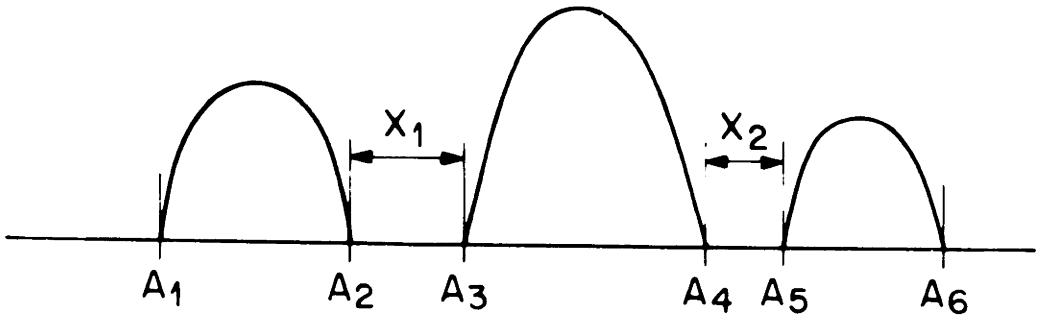


Figure 4.8 General form for endpoint estimate.

duration than pulses of speech. In the condition where the two pulses are of equal length, P_1 will be eliminated under the assumption that isolated artifacts are more likely to occur at the beginning than at the end of words. The endpoints obtained by adjusting those of the top candidate are inserted into the ordered list of estimates in the second position, shifting all the lower estimates down one position. Table 4.1 summarizes the operations of reorder for a smoothed pulse comprised of one, two, or three pulses in the unsmoothed array. This type of algorithm could be applied recursively, however, it has been observed that even with problematic speakers, using the algorithm described above, rarely are multiple artifacts included within the detected word. Additionally, further application would tend to bias the resulting endpoints against words with multiple pulses.

			Ordered Endpoint Candidates
I.	one pulse P_1		A_1A_2
II.	two pulses P_1,P_2 P_1 has maximum energy	$X_1 < NSEP$	A_1A_4 A_1A_2
		$X_1 > NSEP$	A_1A_2 A_1A_4
III.	three pulses P_1,P_2,P_3 P_2 has maximum energy	$X_1,X_2 < NSEP$	A_1A_6 A_3A_6 or A_1A_4
		$X_1 < NSEP, X_2 > NSEP$	A_1A_4 A_3A_4
		$X_1 > NSEP, X_2 < NSEP$	A_1A_6 A_3A_6 A_3A_4
		$X_1,X_2 > NSEP$	A_1A_6 A_3A_4 A_3A_6 A_1A_4 A_1A_6

X_1 - separation between P_1 and P_2 (A_3-A_2)
X_2 - separation between P_2 and P_3 (A_5-A_4)

TABLE 4.1 Ordered endpoint candidates
(refer to Figure 4.8)

The final stage of the endpoint detector is a check to insure that the detected word does not overlap the boundary of the recorded interval. If either of the top estimates for the endpoints of the detected word falls on the boundary of the recorded interval, a repeat is requested.

There are several aspects of this method of endpoint detection which are novel. One aspect is the creation of the smoothed pulse array while maintaining the original unsmoothed pulse information. The assumption that pulses occurring close together form a single word suggests the desirability of some form of smoothing. However, retaining the information in the unsmoothed pulse array it is possible to eliminate errors introduced by the smoothing procedure. Another desirable feature of the endpont detector is the capacity for providing several estimates of each of the endponts in preferential order. Thus an alternate endpont pair may be tried if the recognition scores obtained using the first set are not reliable. The justification for this feature is that if an artifact is included within the endponts of the detected word the recognition scores should be higher than a generally accepted threshold.

The endpoint detector described in this section was implemented and tested on the standard testing data set. The recognition scores, as well as some discussion of the features of the endpont detector are presented in Section 4.3.1.

4.2.2 Shifting Distance Endpoint Detector

In this section an endpoint detection algorithm is described which locates the endpoints of a test by comparing the test frames to the reference template. For each test-reference comparison a different set of endpoints may be found. A block diagram of the endpoint detector is shown in Figure 4.9. This method of endpoint detection is similar to the silence-based endpoint detector proposed in Section 2.4.2. However, instead of determining where the test pattern no longer matches silence, the frame where the test pattern best matches the beginning and the end of the reference pattern is found. To accomplish this, the distances for all possible starting and ending locations of the test as compared to the initial and final frames of the reference are computed. The distances are averaged over three frames for statistical stability. The distance $d(i,j)$ between frame i of the test and frame j of the reference is as given by Eq. (2.13). The average distance, $\bar{d}(s)$, for a given shift s along the test is given by Eq. (2.14) where the first and last three frames of the reference are used in the distance computation. The resulting distance as a function of shift, is lowpass filtered with a 5 Hz cutoff frequency,

digital filter, for further smoothing. The impulse and frequency response of the lowpass filter are shown in Figure 4.10. The lowpass filter has an attenuation of 40 dB in the stop band (i.e., for frequencies greater than about 8 Hz).

The endpoints are determined from the smoothed distance array as follows. The method is described for the beginning frame; however, the same method is applied to locate the ending frame by working backwards from the end of the recording interval. First the absolute minimum and maximum of the distance array are located according to

$$d_{\max} = \max_{s=1,L-20} \bar{d}(s) \tag{4.3a}$$

$$d_{\min} = \min_{s=1,s_{\max}} \bar{d}(s) \tag{4.3b}$$

and $s_{\min}$ and $s_{\max}$ are defined as the frames corresponding to the minimum and maximum distance. Next all local minima satisfying the condition

$$\bar{d}(s) \leq c^* d_{\min}, \quad s = 1, s_{\max} \tag{4.4}$$

are located. From Eqs. (4.3) and (4.4) it is seen that the beginning frame is required to occur at least 20 frames from the end of the recorded interval. Local constraints are applied to differentiate between local minima separated by less than a predetermined number of frames, f_m. If two local minima are separated by less than f_m frames, the minimum having the higher distance is discarded. The estimates of the endpoints are ordered with regard to proximity to the shift with the maximum distance, $s_{\max}$. Thus, the best endpoint is selected as the local minimum occurring closest, but prior to, $s_{\max}$. The second estimate is the second closest local minimum, etc. If no local minimum exists (within the threshold $c^* d_{\min}$), the default beginning frame is $s_{\min}$, the frame corresponding to the absolute minimum distance.

Figures 4.11-4.13 show some examples of the location of the endpoints by this detector as well as the distance versus shift contours. In these figures, the top plot is the log energy contour of the test. The solid dark lines correspond to the best estimate of the endpoints and the dashed vertical lines correspond to the second best estimate. The middle plot is the distance contour for the beginning point, with the estimates for the starting frame denoted. The bottom

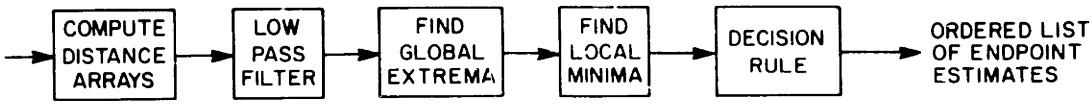


Figure 4.9 Block diagram of shifting distance endpoint detector.

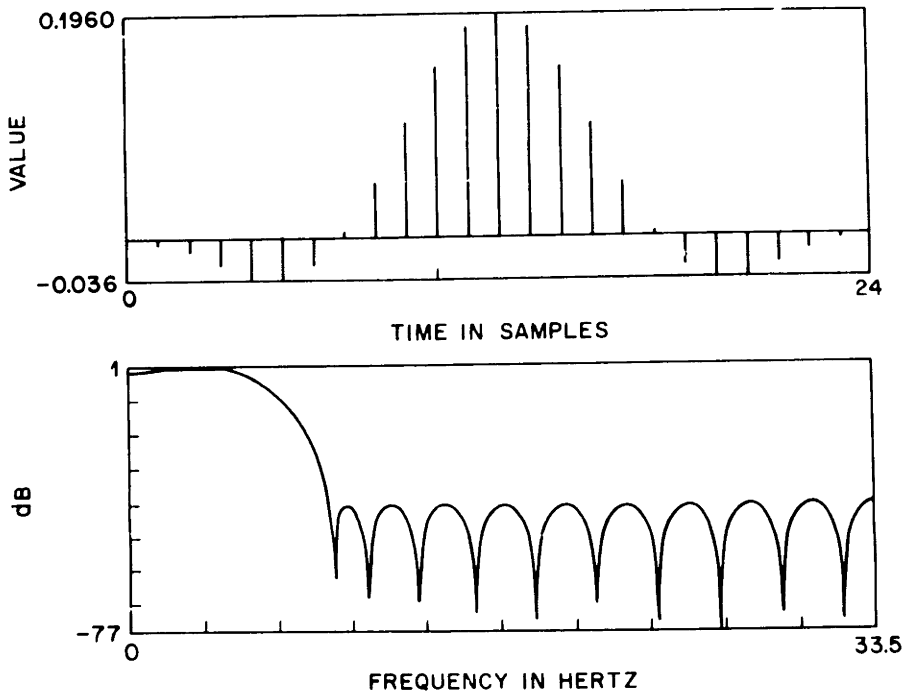


Figure 4.10 Impulse and frequency response of low pass filter.

plot is the corresponding distance contour for the ending point, with the endpoint estimates shown by the vertical lines. Figure 4.11a is an example of a comparison of the test with the correct reference template. The comparison is for the word "FOUR". Here the best estimates for the endpoints are seen to be accurate. Part b shows the plots for the same test when compared to an incorrect reference. In this case the reference template is for the word "TWO". Once again the endpoints are accurately located. Figure 4.12a shows the result of comparing a different test to its correct reference. The test and the reference patterns are for the letter "K". The first estimate of the endpoints is seen to be good. However, the same test compared to another matching template, shown in Figure 4.12b, has an accurately located beginning point, and a grossly incorrect ending point. The breathiness following the word has mistakenly been included within the endpoints. Figure 4.13 is an example of a case in which the best estimates for the endpoints are poor, but the second estimates are quite reasonable. The test for this example is the letter "S" whereas the reference is the letter "L". These examples illustrate some of the problems of this endpoint detector. An evaluation of the performance of this method for endpoint detection is given in Section 4.3.2.

4.3 Performance Results for the Hybrid Techniques For Endpoint Detection

The recognition accuracies and performance measurements for the hybrid endpoint detectors are presented in this section. Section 4.3.1 gives the results for the modified energy detector. The results for the shifting distance based endpoint detector are given in Section 4.3.2.

4.3.1 Results on the Performance of the Modified Energy Endpoint Detector In this section the performance of the modified energy endpoint detector is evaluated. The modified energy endpoint detector supplies the recognition stage of the system with an ordered list of candidate pairs for the endpoints of the test word. The number of endpoint pairs is determined both by the number of energy pulses located and by a set of temporal constraints on valid endpoints for isolated words. There is a tradeoff between the computational costs of attempting recognition with several sets of endpoints and the reduced reliability obtained by requiring the endpoint detector to supply only one set of endpoints.

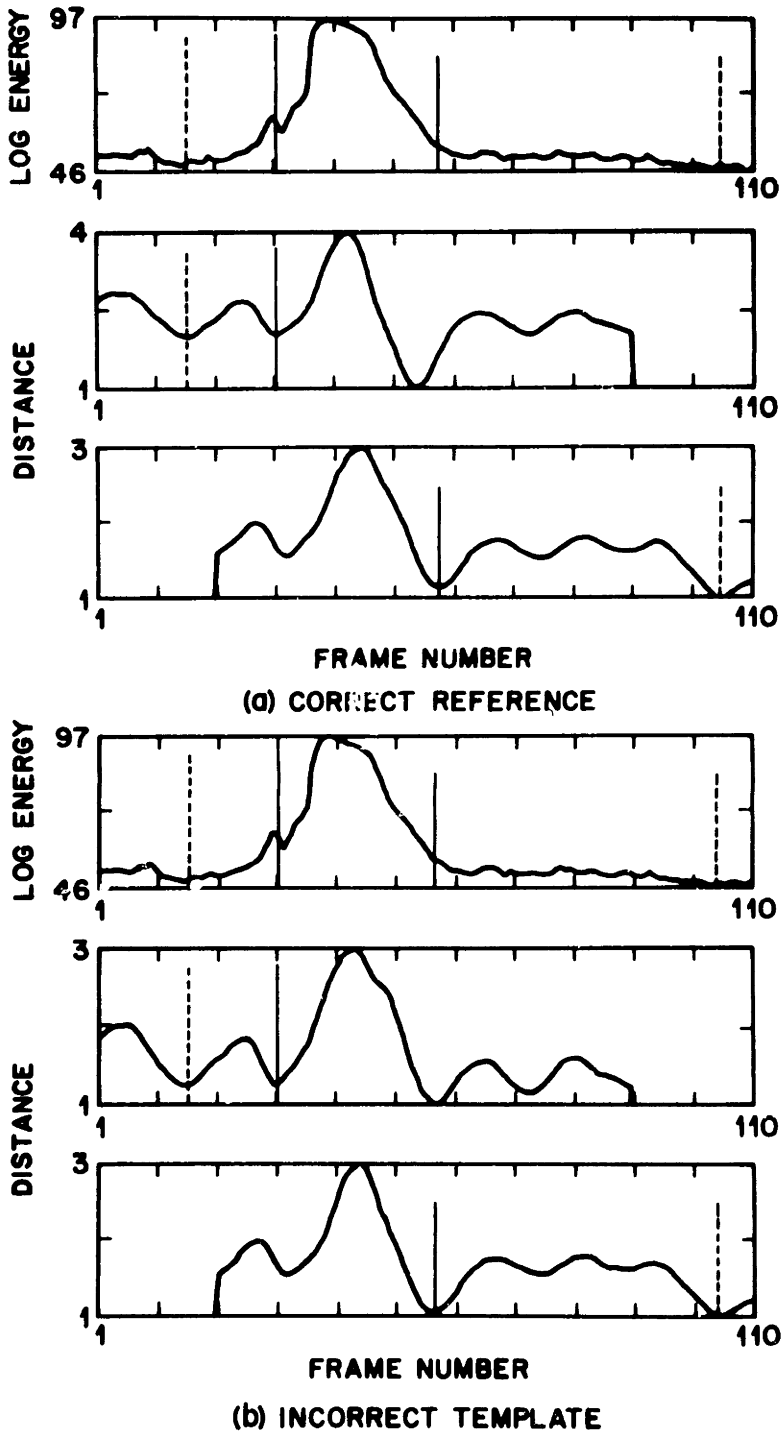


Figure 4.11 Endpoints accurately located by shifting distance detector using correct and incorrect references.

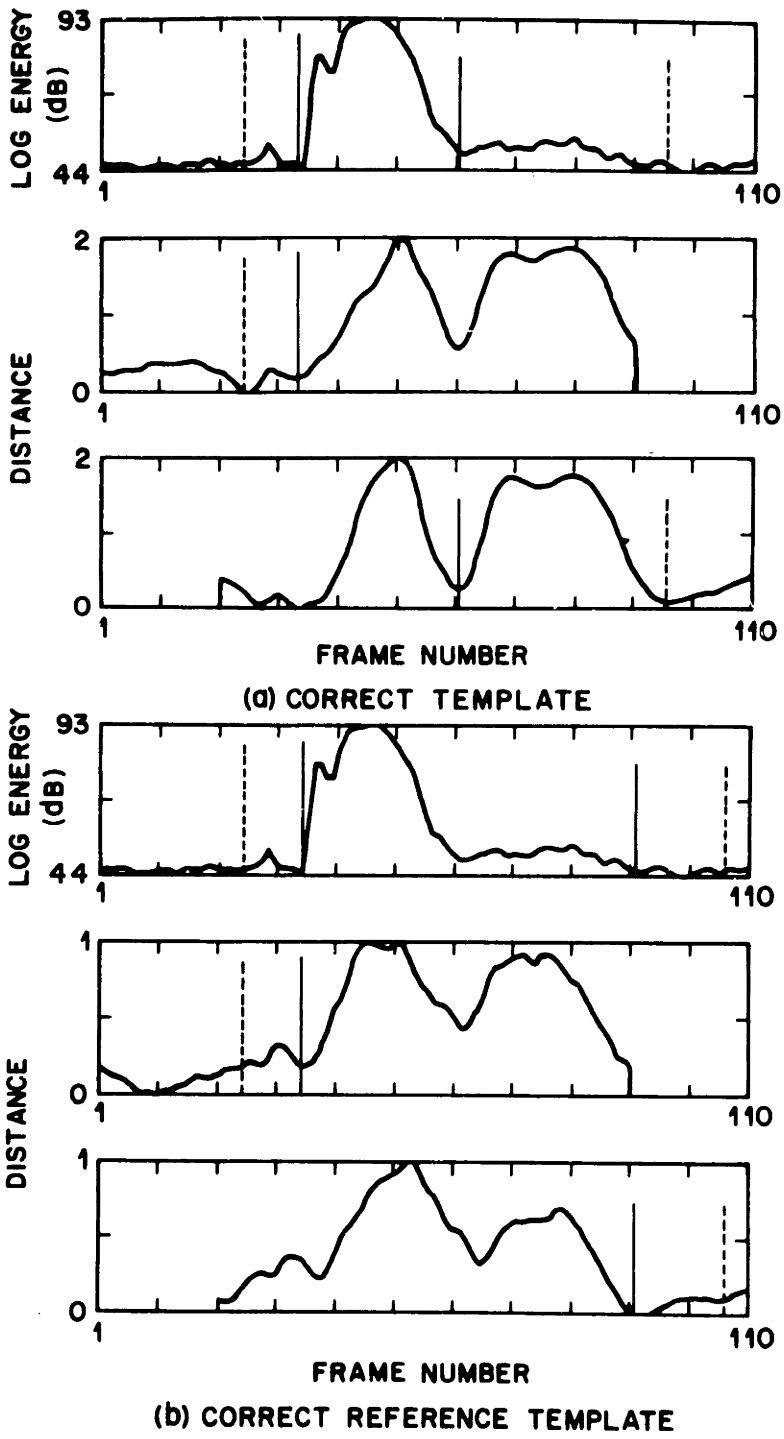


Figure 4.12 Endpoints incorrectly located by shifting distance detector using correct reference.

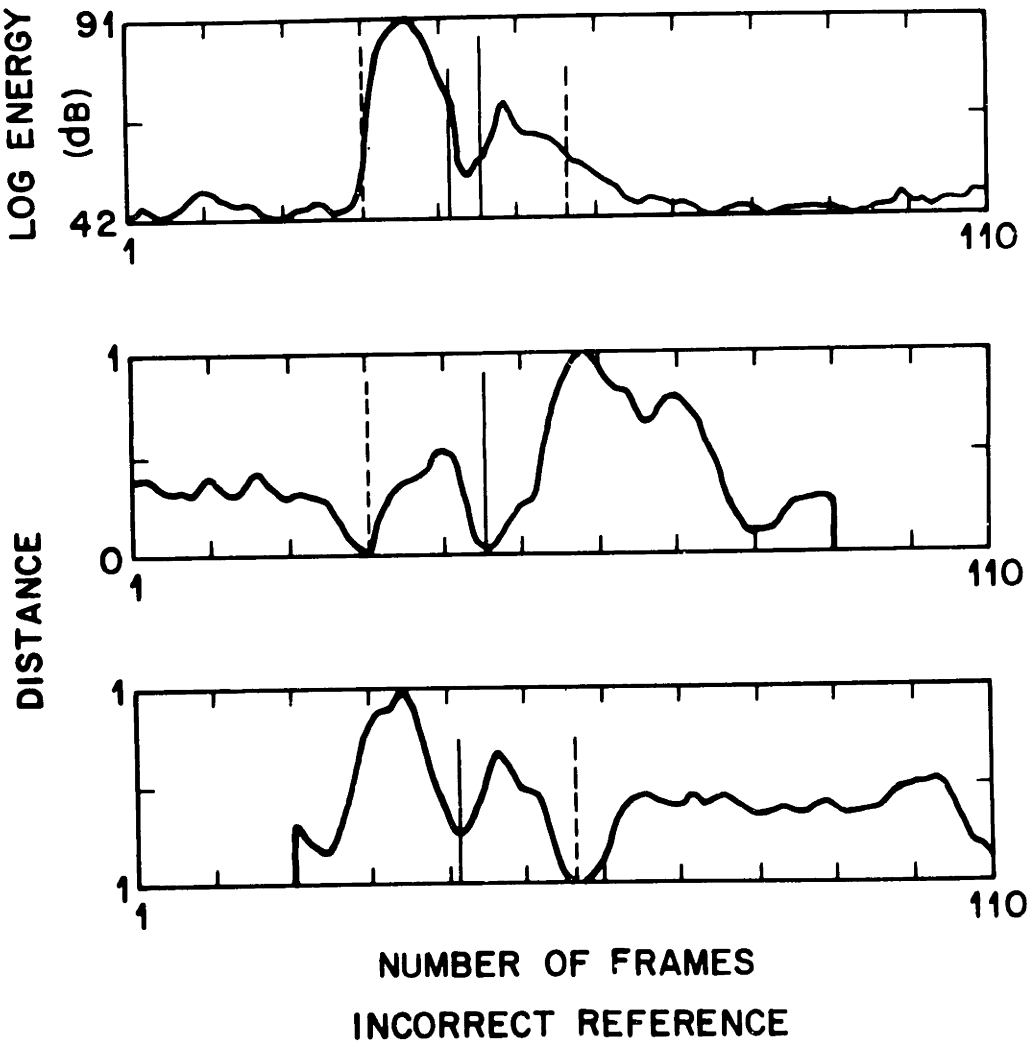


Figure 4.13 Second estimated endpoints more accurate than top candidate.

In the evaluation of the modified energy endpoint detector several considerations are addressed. The most important factor in the performance of the endpoint detector is the recognition accuracy achieved and the associated rejection rate. The endpoint detector was developed so as to obtain at least the recognition accuracy of the current energy endpoint detector while reducing the reject rate to a minimum. The test is rejected only if there is sufficient evidence that the speech has overlapped the boundary of the recorded interval, or that no speech is present within the recorded interval. In all other cases the endpoint detector supplies one or more estimates of each of the endpoints to the recognizer. Since the endpoint detector supplies several estimates of the endpoints it is important to determine the average number of endpoint pairs found and where within the ordered list of candidates the correct endpoints are situated. Given an ordered list of endpoint candidates the recognizer may attempt recognition using successive endpoint pairs until a reliable match is found. This may be implemented by use of a threshold on the recognition score obtained. Recognition is attempted until a reliable score is obtained or until all endpoint candidates have been used.

In addition a study was performed to determine the utility of the backup counters of the energy pulse detector. Statistics were tabulated as a measure of the slope of the log energy contour at the beginning and end of each detected pulse. A histogram was computed of the number of frames that occurred in the log energy contour in the interval from when the lower threshold K_1 was exceeded until the upper threshold K_2 was exceeded at the pulse beginning. Similarly a histogram was computed for the number of frames in the interval in which the log energy fell from the upper threshold K_2 to below the middle threshold K_3 at the pulse end. The resulting data are given in Table 4.2. In part a of Table 4.2 the data correspond to the pulse beginning. Part b gives the data for the pulse end. These data verify the observation that the energy at the end of a pulse tends to fall off more gradually than the energy rises at the onset. Based on the data the maximum allowable rise/fall time was selected as 5 frames (or equivalently 75 ms). Using this value the data show that the beginning frame was adjusted on 22 pulses and the ending frame on 18 pulses. In total about 2% of the pulses located by the energy pulse detector had an endpoint adjusted by use of the backup counters. Although the

number of pulses adjusted by the backup counters is small, the backup counters aid in locating the endpoints of difficult energy pulses, particularly those with breath noise.

speaker #	# tests	# frames 0	1	2	3	4	5	>5
1	114	118	50	9	1	0	0	0
2	117	151	74	10	5	1	1	0
3	116	88	55	8	3	0	1	1
4	112	107	43	2	1	0	1	0
5	117	103	64	17	2	2	0	4
6	117	92	62	14	5	2	0	0
7	117	161	58	8	6	1	0	3
8	117	110	57	5	3	3	0	3
9	117	140	71	14	3	4	2	5
10	117	131	57	15	6	4	3	6
TOTAL	1161	1201	591	102	35	17	8	22

a. Pulse beginning

speaker #	# tests	# frames 0	1	2	3	4	5	>5
1	114	89	63	14	9	2	1	0
2	117	122	87	17	12	2	0	2
3	116	63	71	18	1	1	1	1
4	112	50	70	27	5	2	0	0
5	117	62	77	23	16	5	6	3
6	117	59	65	30	13	6	1	1
7	117	93	102	29	9	3	0	1
8	117	64	54	26	14	13	3	7
9	117	113	101	17	2	3	1	2
10	117	71	76	42	17	11	4	1
TOTAL	1161	786	766	243	98	48	17	18

b. Pulse end

TABLE 4.2 Use of backup counters

The number of endpoint pairs found by the modified energy detector is given for each speaker in Table 4.3. The data show that for the majority of the tests only one endpoint pair was found. Averaged over all of the speakers 84% of the test words had only one estimated endpoint pair. In fact of the 1156 tests, only 2, less than 0.2% had more than two endpoint candidate pairs. For these two tests three endpoint candidates had been located. The number of endpoint pairs directly reflects the performance of the artifact screening operations of the endpoint detector. Reliable screening greatly reduces the potential number of endpoint pairs by eliminating energy pulses corresponding to extraneous artifacts. Averaged over the speakers the mean number of endpoint pairs found is 1.16 with a standard deviation of 0.37. These results indicate that the screening operations of the decision rule are effective in eliminating

most artifacts.

speaker #	# tests	# 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

TABLE 4.3 Number of endpoint candidates

Since for 84% of the test words only one endpoint pair was found, it is of interest to determine the recognition accuracy obtained using only the top pair of endpoint candidates for recognition. The results are given as percent recognition in Table 4.4. The average recognition scores are plotted in Figure 4.14. Comparing the recognition scores to those obtained by the current energy detector (refer to Table 2.4), it can be seen that the modified energy endpoint detector has higher recognition scores for all candidate positions. The improved recognition is obtained with a reject rate of less than 0.5%, whereas the current energy detector had a reject rate of almost 30%. For most of the speakers there is a 10 to 15% increase in the recognition accuracy of the top recognition candidate position. Averaged over all of the talkers there is a 4 to 6% improvement in recognition accuracy for all candidate positions. Comparative recognition curves are shown in Figure 5.1.

speaker #	# tests	# rejects	candidate position 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

TABLE 4.4 Recognition Scores using top endpoint pair

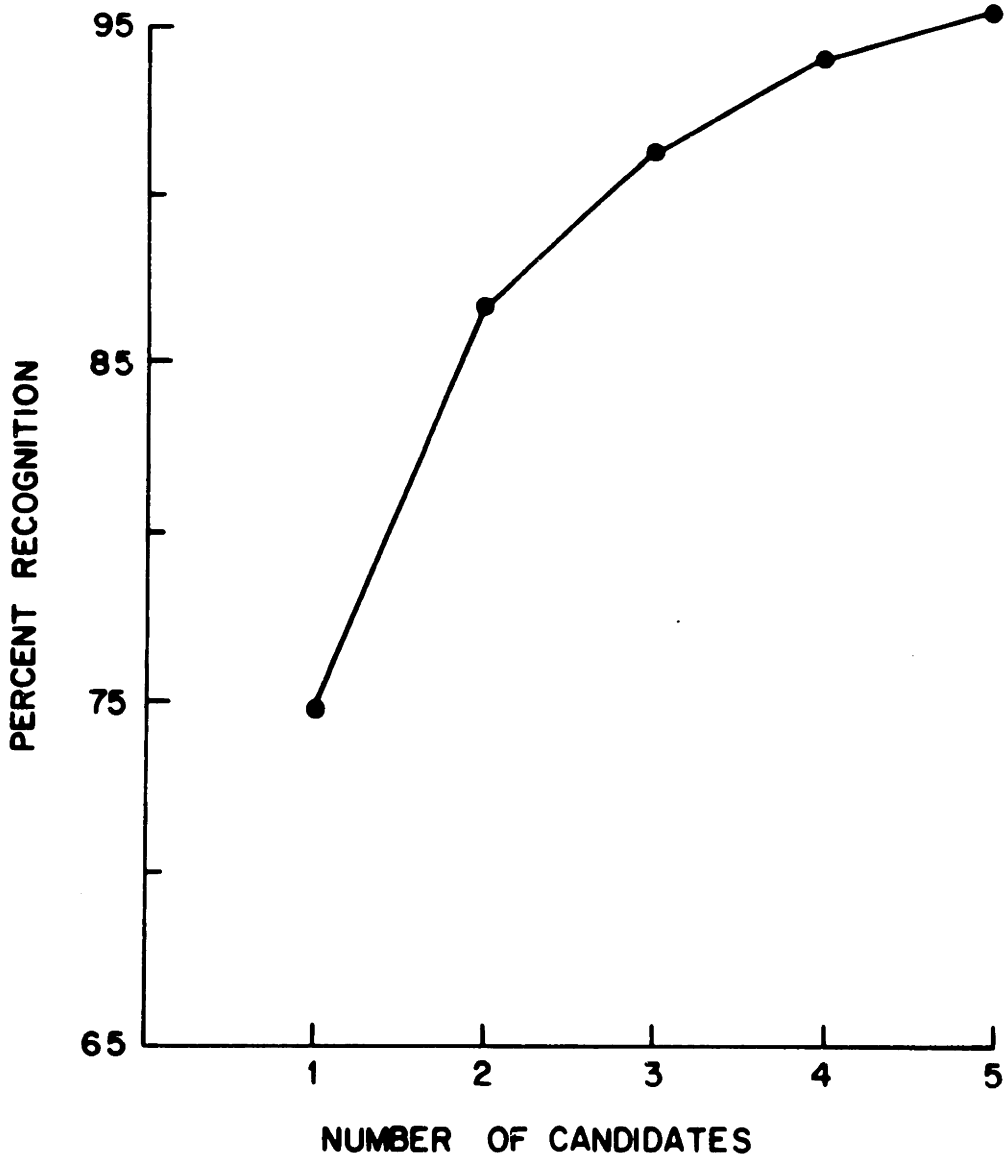


Figure 4.14 Recognition accuracy of modified energy.

The above results demonstrate that reliable recognition is obtained by the modified energy endpoint detector using only the best endpoint pair. Used in this manner the modified energy endpoint detector is appropriately classified as an explicit endpoint detector. However, the advantages of supplying several estimates of the endpoints is demonstrated by the use of a threshold during the decision stage of the recognition system. If the recognition score obtained using the best endpoint pair does not provide for reliable recognition, recognition with the second endpoint pair may be attempted. Setting a low threshold for "reliable" recognition will cause recognition to be attempted with alternate endpoints pairs unless the first match is a very good one. In the limiting case where the threshold is set at zero recognition will be attempted using all endpoint candidates. From the data in Table 4.3 it is seen that even in this case only one or two endpoint pairs will be used for recognition. However, setting a low threshold can cause potential problems when the test word contains multiple energy pulses. An example is the word "EIGHT". The second endpoint pair, not including the stop release, may match the reference for the letter "A" better than the first comparison matched the template for the word "EIGHT". Setting a high threshold increases the likelihood that the recognized word with the first endpoint pair will have a recognition score below the threshold and be acceptable. The threshold should be set such that if the first endpoints are accurate the corresponding recognition score will fall within the acceptable range. At the same time, if the first endpoint set is incorrect, it is desirable that the recognition score lie above the acceptable threshold. It is important to note that the threshold level does not affect the number of endpoint candidates located, but does determine how many of the endpoint candidates are used.

The optimum threshold was determined for each of the speakers. The results are summarized in Table 4.5. The data suggest that on average, the optimum threshold should be in the range 0.35 to 0.4 for reliable recognition.

speaker #	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

TABLE 4.5 Optimum recognition threshold

For four of the ten speakers the recognition accuracy did not vary for different threshold values. This implies that for these speakers either the first endpoint candidate was always the best, or that any improvement in the recognition of some tests was counteracted by errors introduced for other tests. Only for two of the speakers did the recognition accuracy improve significantly by using thresholds. For these speakers, speakers 2 and 9, the recognition rates increased uniformly by 2 to 3% for all candidate positions.

The data above suggest that the use of alternate sets of endpoints gains only little improvement in recognition accuracy for the testing data set. At the same time it should be noted that the provision for supplying alternate endpoint estimates adds little to the computational requirements of the system. The ability for the modified energy endpoint detector to supply alternate endpoint estimates has proved to be useful during on-line recognition tests. The data presented simply indicates that the alternate endpoint sets are only used for problematic speakers, such as speakers 2 and 9.

Figure 4.15 illustrates the performance of the modified energy endpoint detector for four examples. In part a of the figure the log energy contour corresponds to the word "EIGHT". The first estimate of the endpoints, as denoted by the solid vertical lines, is seen to be accurate. The second estimate of the ending frame, as shown by the dashed vertical line, does not include the stop release, and would not be used. The log energy contour shown in part b consists of the letter "R". Included within the first estimated endpoint pair is an artifact prior to the word. The second estimate for the beginning frame is seen to be correct. Part c of the figure is the log energy contour of the letter "K". Here the endpoints are accurately located despite the

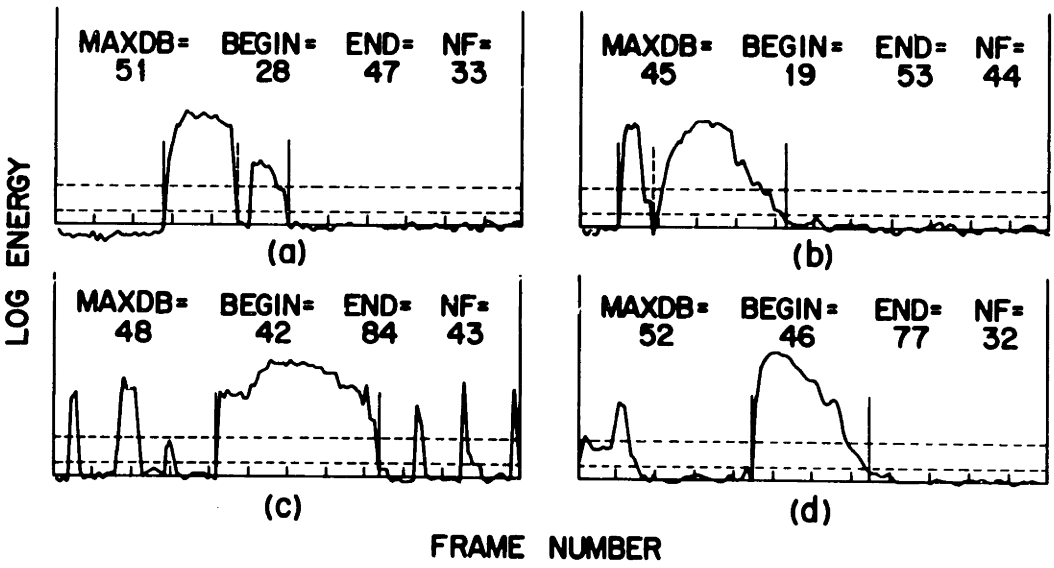


Figure 4.15 Endpoints located by modified energy detector.

spike-like artifacts included within the recording. Even though an artifact overlaps the boundary of the recorded interval in part d of the figure, the endpoints of the test word "ERROR" have been correctly located. The examples in Figure 4.15 show the modified energy endpoint detector to perform reliably in online testing conditions.

4.3.2 Results for the Shifting Distance Endpoint Detector

The shifting distance endpoint detector was tested on the data set TS1 described in Section 2.6. However, due to time constraints, the testing was performed using only two templates per reference. The recognition scores plotted in Figure 4.16 as a function of the number of candidates. For comparison, the recognition scores obtained by testing the modified energy endpoint detector on the same data set using two templates per reference are also plotted. The shifting distance endpoint detector is seen to have recognition scores approximately 10 percent lower than the modified energy endpoint detector. The recognition scores obtained by the modified energy endpoint detector are comparable to those obtained earlier by Rabiner et al. [14] for 2 speaker independent reference templates per word.

An inherent disadvantage of this method of endpoint detection is that the endpoints obtained for each reference-test comparison may be different, thus the endpoint locations are dependent on the reference word. This results in increased computation proportional to the number of reference templates and decreased releability.

4.4 Conclusions on the Utility of Hybrid Algorithms for Endpoint Detection

The hybrid endpoint detectors, (in particular the modified energy endpoint detector), are seen to be more reliable than both the explicit and implicit detectors. The endpoints are more accurately located by the hybrid methods than the other methods as is reflected by the improved recognition scores. In addition, the hybrid endpoint detectors are able to reduce the rejection rate for test utterances from approximately 30 percent to less than 0.5 percent. In part the performance improvement of the hybrid endpoint detection method is achieved by allowing the endpoint detector to supply several estimates for the endpoints of the test word. Thus, the hybrid endpoint detectors have successfully combined the increased freedom of the implicit

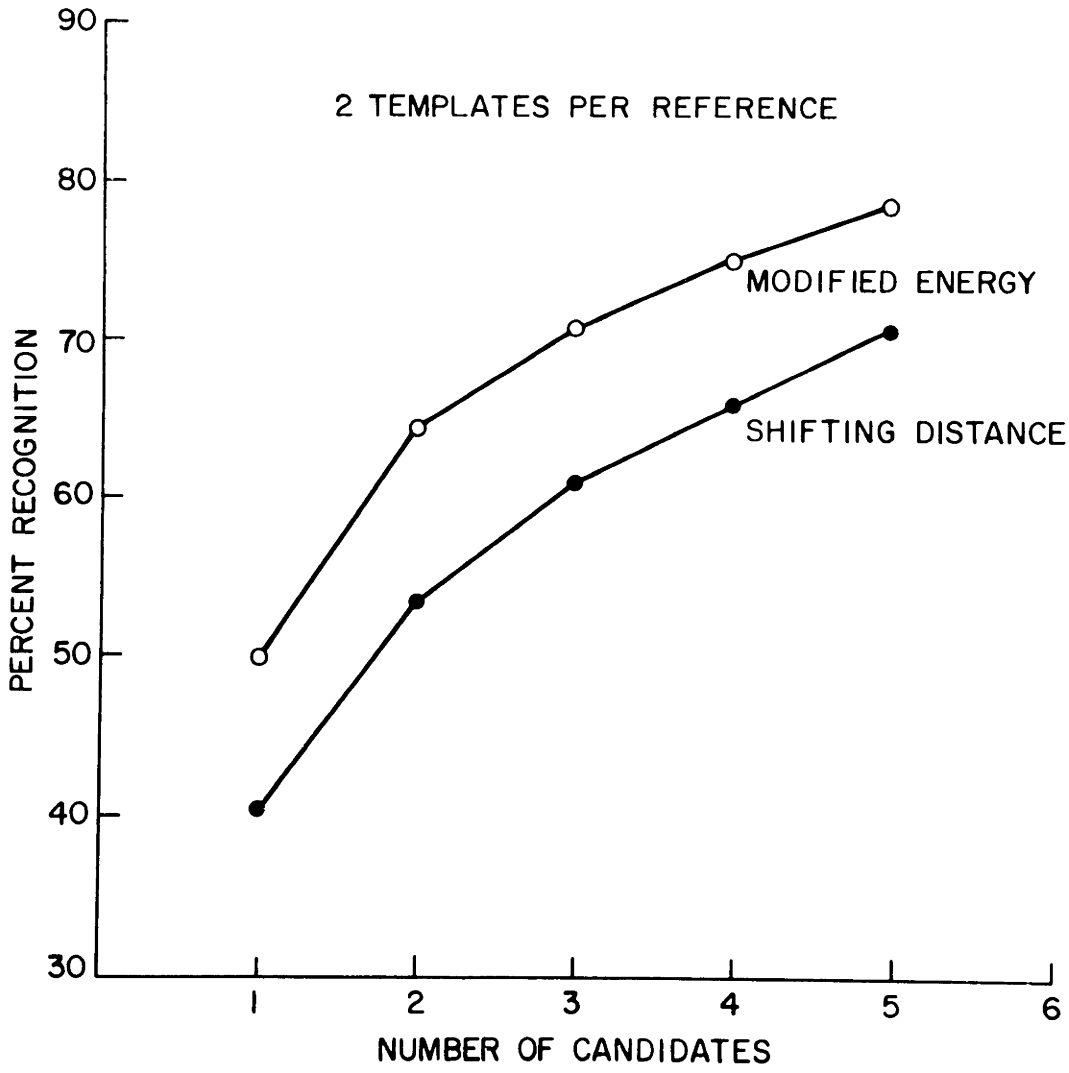


Figure 4.16 Recognition accuracy of shifting distance endpoint detector.

techniques for endpoint detection with the better accuracy and reliability of the explicit techniques.

CHAPTER 5

SUMMARY AND CONCLUSIONS

5.1 Introduction

In Chapters 2 through 4 three classes of endpoint detectors were defined for use in isolated word recognition systems. For each class one or more endpoint detectors were proposed, and their relative performances within each class have been compared. In this chapter the results presented in the previous chapters are compared and summarized.

5.2 Relative Performance of the Endpoint Detection Methods

In Chapter 2 the explicit endpoint detectors were compared. The results indicated that due to the high reject rate obtained by even the best of these methods, namely the current energy endpoint detector, these methods are not satisfactory for implementation in a practical isolated word speech recognition system. The explicit methods were shown to be reliable when the recordings were relatively clean; i.e., reasonable signal-to-noise ratio and no artifacts in the recordings [14].

The implicit algorithm presented in Chapter 3 was shown to be significantly less reliable than the explicit techniques. By allowing the dynamic time warping algorithm the freedom to find the best time alignment for each test-reference pair, the algorithm was able to find time alignments for the incorrect templates that were as good as those found for the correct templates. This algorithm was shown to be totally unreliable for use in practical speech recognition systems.

The hybrid techniques, specifically, the modified energy detector of Chapter 4, was shown to have higher reliability than the current energy detector for locating the endpoints, of all input recordings, i.e., both for clean recordings and for recordings with artifacts. This high reliability is achieved by allowing the endpoint detector to supply several sets of endpoints rather than the single set associated with the explicit techniques. The modified energy endpoint detector was shown to be robust with regard to the accuracy of the endpoint locations and the associated

recognition scores.

Figure 5.1 compares the recognition scores on the isolated word recognition tests using the best algorithm of each of the classes. In addition, the recognition rates obtained previously using hand edited endpoints [14] are also plotted. The hybrid modified energy endpoint detector provides the highest recognition accuracies of the three classes that were studied. The recognition accuracies achieved by the modified energy endpoint detector are about 3-5% below those using hand edited endpoints. However, the modified energy detector results are based on a testing set in which 40% to 60% of the utterances would have been discarded due to the uncertainty as to where the actual endpoints would be placed even by hand editing.

5.3 Suggestions for Future Research

The use of hybrid analysis techniques has provided the basis for robust endpoint detection. However, there are many problems associated with endpoint detection for isolated word recognition that need to be solved. While successfully eliminating isolated artifacts from the actual word, the hybrid techniques may still fail when the artifacts are not separated from the isolated word. These conditions exist with cases of heavy breathing at the end of a word and may possibly be solved by using a variable ending frame if the recognition score for the initial estimated endpoint is too large. One possible solution would be an endpoint detector which would use the energy contour to locate a *region* in which the endpoints lay. An unconstrained DTW algorithm could then be used to find the best path and associated endpoints within the beginning and ending region.

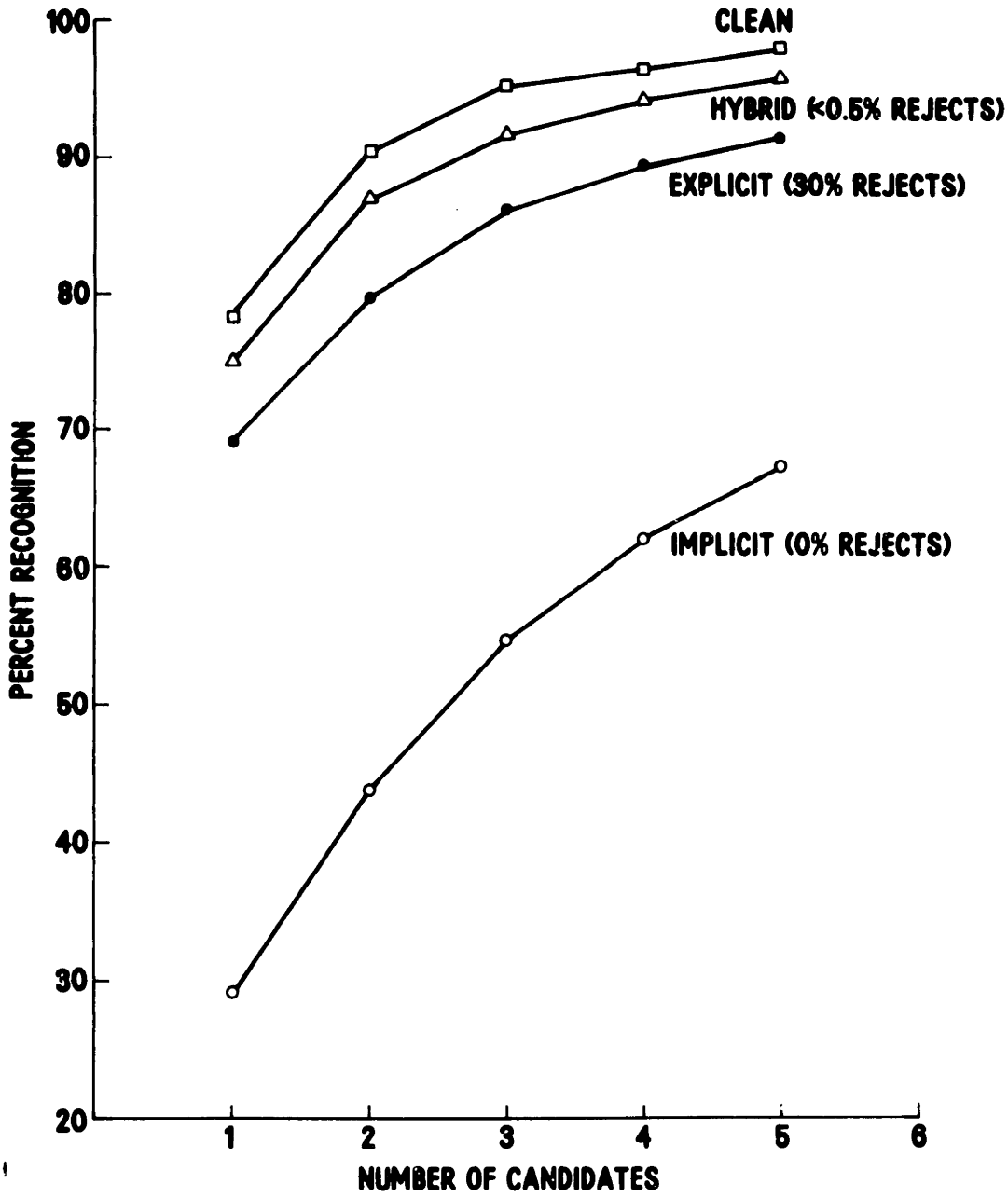


Figure 5.1 Comparison of recognition scores for explicit, implicit, hybrid and hand edited endpoints.

[1] T. B. Martin, "Practical Applications of Voice Input to Machines," Proc. of the IEEE, vol. 64, no. 4, pp. 487-502, April 1976.

[2] A. E. Rosenberg and C. E. Schmidt, "Directory Assistance by Means of Automatic Recognition of Spoken Spelled Names," Bell System Technical Journal, vol. 58, pp. 1797-1823, October 1979.

[3] S. E. Levinson, A. E. Rosenberg, and J. L. Flanagan, "Evaluation of a Word Recognition System Using Syntax Analysis," Bell System Technical Journal, vol. 57, no. 5, pp. 1619-1626, May-June 1978.

[4] L. R. Rabiner, J. G. Wilpon, and A. E. Rosenberg, "A Voice Controlled Repertory Dialer," submitted for publication.

[5] H. F. Silverman and N. R. Dixon, "A Parametrically Controlled Spectral Analysis System for Speech," IEEE Trans. on Acoustics, Speech and Signal Processing, vol. ASSP-22, no. 5, pp. 362-381, October 1974.

[6] L. R. Rabiner and M. R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances," Bell System Technical Journal, vol. 54, no. 2, pp. 297-315, Fib. 1975.

[7] L. R. Rabiner, C. E. Schmidt, and B. S. Atal," Evaluation of a Statistical Approach to Voiced-Unvoiced-Silence Analysis for Telephone-Quality Speech," BSTJ, vol. 56, no. 3, pp. 455-481, March 1977.

[8] T. B. Martin, "Applications of Limited Vocabulary Recognition systems," in *Speech Recognition: Invited Papers Presented at the 1974 IEEE Symposium.* New York: Academic Press, 1975, pp. 55-72.

[9] J. D. Markel and A. H. Gray, *Linear Prediction of Speech,* New York: Springer Verlag, 1976.

[10] A. E. Rosenberg, private correspondence.

[11] F. Itakura, 'Minimum Prediction Residual Principle Applied to Speech Recognition," IEEE Trans. on Acoustics, Speech and Signal Processing, vol. ASSP-23, no. 1, pp. 67-72, February 1975.

[12] L. R. Rabiner, M. R. Sambur, and C. E. Schmidt, "Applications of a Nonlinear Smoothing Algorithm to Speech Processing," IEEE Trans. on Acoustics, Speech and Signal Processing, vol. ASSP-23, no. 6, pp. 552-557, December 1975.

[13] L. R. Rabiner, "The Overall Structure of an Isolated Word Recognizer," unpublished paper.

[14] L. R. Rabiner, S. E. Levinson, A. E. Rosenberg and J. G. Wilpon, "Speaker Independent Recognition of Isolated Words Using Clustering Techniques," IEEE Trans. on Acoustics, Speech and Signal Processing, vol. ASSP-27, No. 4, August 1979.

[15] C. S. Myers, "A Comparative Study of Several Dynamic Time Warping Algorithms for Speech Recognition," Master's Thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, February 1980.

[16] L. Breiman, *Statistics: with a View Toward Applications, Boston:* Houghton Mifflin Company, 1973.