An Algorithm for Determining the Endpoints of Isolated Utterances

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

(Manuscript received June 10, 1974)

An important problem in speech processing is to detect the presence of speech in a background of noise. This problem is often referred to as the endpoint location problem. By accurately detecting the beginning and end of an utterance, the amount of processing of speech data can be kept to a minimum. The algorithm proposed for locating the endpoints of an utterance is based on two measures of the signal, zero crossing rate and energy. The algorithm is inherently capable of performing correctly in any reasonable acoustic environment in which the signal-to-noise ratio is on the order of 30 dB or better. The algorithm has been tested over a variety of recording conditions and for a large number of speakers and has been found to perform well across all tested conditions.

I. INTRODUCTION

The problem of locating the beginning and end of a speech utterance in an acoustic background of silence is important in many areas of speech processing. In particular, the problem of word recognition is inherently based on the assumption that one can locate the region of the speech utterance to be recognized. A further advantage of a good endpoint-locating algorithm is that proper location of regions of speech can substantially reduce the amount of processing required for the intended application.

The task of separating speech from background silence is not a trivial one except in the case of acoustic environments with extremely high signal-to-noise ratio, e.g., an anechoic chamber or a soundproof room in which high-quality recordings are made. For such high signal-to-noise ratio environments, the energy of the lowest-level speech sounds (e.g., weak fricatives, low-level voiced portions, etc.) exceeds the background noise energy and a simple energy measure suffices.[1] However, such ideal recording conditions are not practical for real-world applications of speech-processing systems. Thus, simple energy measures are not sufficient for separating weak fricatives (such as the /f/ in "four") from background silence. In this paper, we propose a fairly simple algorithm for locating the beginning and end of an utterance, which can be used in almost any background environment with a signal-to-noise ratio of at least 30 dB. The algorithm is based on two measures of speech: short-time energy and the zero crossing rate. The algorithm possesses the feature that is somewhat self-adapting to the background acoustic environment in that it obtains all the relevant thresholds on its decision criteria from measurements made directly on the recorded interval.

The organization of this paper is as follows. In Section II we discuss the major difficulties in locating the beginning and end of an utterance and propose various measurements for distinguishing between speech and no speech in these cases. In Section III we describe the algorithm to locate the endpoints of the utterance. In Section IV we give examples of the use of the algorithm, and give the results of both formal and informal tests on its ability to find endpoints of a corpus of words from several speakers. Finally, in Section V we discuss the general characteristics of the endpoint-location problem and propose alternative methods of solving the problem.

II. EXAMPLES OF SPEECH ENDPOINT-LOCATION PROBLEMS

To arrive at a reasonable algorithm for separating speech from nonspeech, it is necessary first to define the acoustic environment in which the recordings are made. In this paper, we consider two specific modes of recording. In the first mode, the speaker makes recordings on analog tape using a high-quality microphone in a soundproof room. This mode of recording is useful for obtaining reasonably high-quality speech. In the second mode of recording, the speaker records directly into computer memory in a noisy environment (e.g., a computer room) using a noise-reducing, close-talking microphone. This mode of recording is a reasonable approximation to a real-world environment for most man-machine interaction problems. To eliminate 60-Hz hum, as well as any dc level in the speech, it is assumed that the speech is high-pass filtered above 100 Hz; similarly, to keep the processing simple, the speech is low-pass filtered at 4 kHz, thereby allowing a 10-kHz sampling frequency.

Figure 1 shows a comparison of the waveform* of the background silence (on a greatly amplified scale) for these two modes of recording. The top two lines of this figure show the waveform for tape-recorded

* In this and subsequent illustrations, each line shows 25.6 ms of the waveform. Successive lines show successive 25.6-ms segments of the waveform.

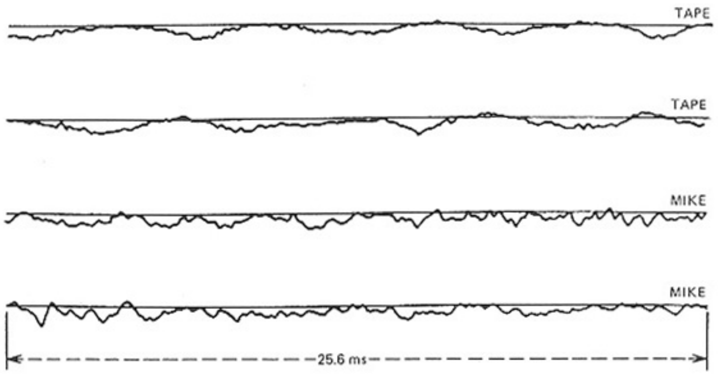

Fig. 1—Acoustic waveforms for the silences from tape and microphone.

silence from a soundproof booth, whereas the lower two lines show the waveform for the silence from the close-talking microphone. It is seen from this figure that the tape-recorded silence has a strong low-frequency component (period ≈ 8 ms) due to the recording process. The waveforms from both the close-talking microphone and the recording process appear to be quite broadband, as one would expect. Figure 2 shows typical frequency spectra of these background silences. The spectra are plotted on a log magnitude scale and are for 512-point Hamming window weighted sections. Except for the strong low-frequency-hum components for the recorded silence, the spectra of these silences are quite similar.

The problem of locating the endpoints of an utterance in these backgrounds of silence essentially is one of pattern recognition. The way one would attack the problem by eye would be to acclimate the eye (and brain) to the "typical" silence waveform and then try to spot some radical change in the pattern. In many cases this is easy to do. Figure 3 shows an example (a waveform of the word "eight") in which the silence pattern (on a reduced amplitude scale) is easily distinguished from the speech which begins just past the beginning of the third line on this figure. What one is observing in this case is a radical change in the waveform energy between the silence and the beginning of the speech.

Figure 4 shows another example (a waveform of the word "six") in which the eye can do an excellent job in locating the beginning of the speech. In this case, the frequency content of the speech is radically different from the frequency content of the background noise as manifested by the sharp increase in the zero crossing (or level crossing) rate of the waveform. For this example, the speech energy at the beginning of the utterance is not radically higher than the silence energy; however, other characteristics of the waveform signal the beginning of the speech.

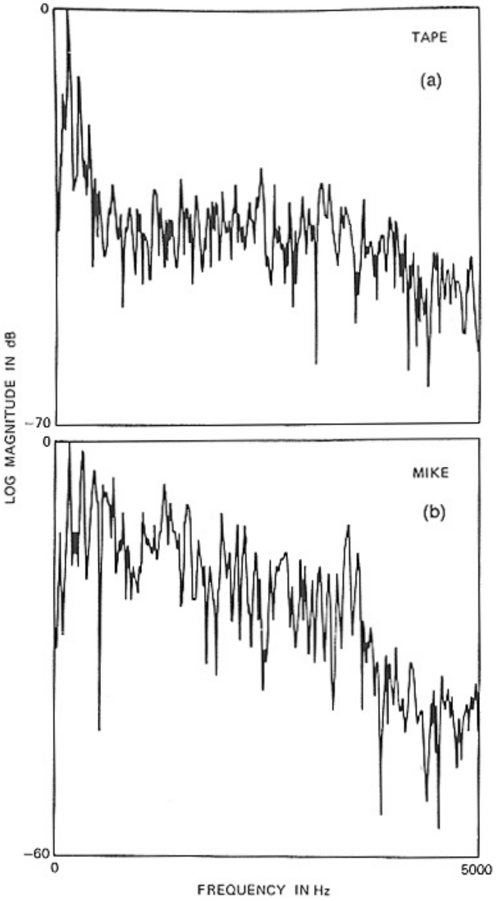

Fig. 2—Log magnitude spectra for the silences from tape and microphone.

The next set of figures illustrates some of the cases in which the eye can be greatly deceived, even with the use of expanded amplitude scales to aid in the examination of the frequency content of the speech. Figure 5 shows the waveform for the beginning of the utterance "four."

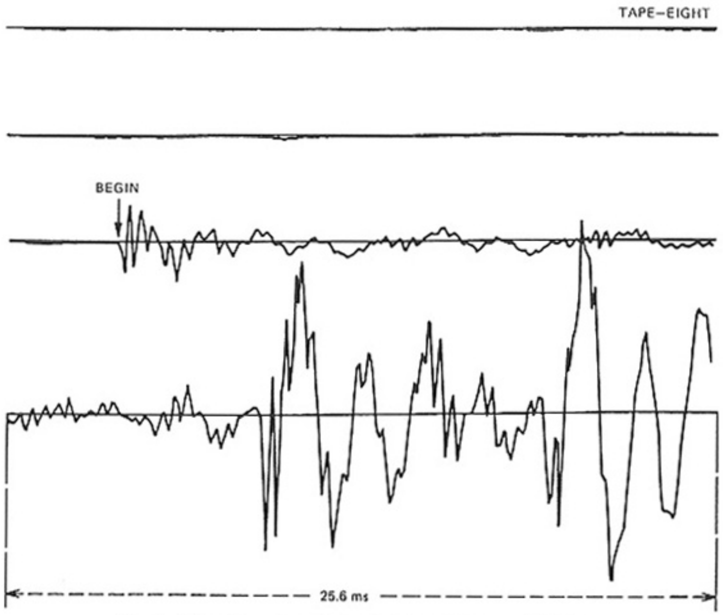

Fig. 3—Waveform for the beginning of the word "eight."

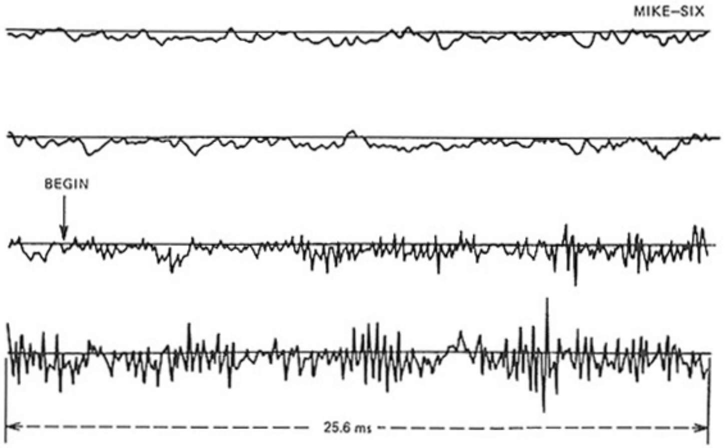

Fig. 4—Waveform for the beginning of the word "six."

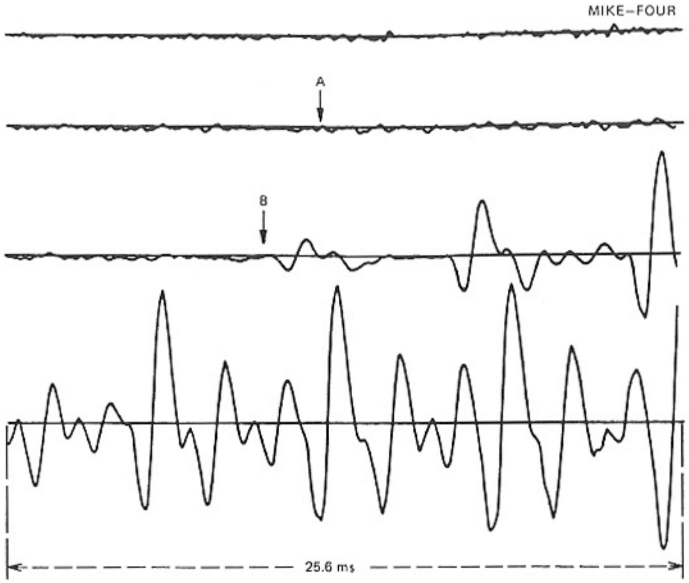

Fig. 5—Waveform for the beginning of the word "four."

This utterance begins with the weak fricative /f/. Without any *a priori* information about the utterance, the eye would select point B as the beginning of the utterance. This is incorrect, however, in that it completely misses the weak fricative /f/ at the beginning. For this example, point A is a better indication of the beginning of the speech.[*] Thus, one problem to be concerned with is weak fricatives at the beginning (or end) of the utterance.

Figure 6 shows another example of the difficulty in locating the endpoint of an utterance. This figure shows the waveform for the end of the word "five." Without any *a priori* information, point A might be chosen by eye as the endpoint of the utterance. However, the actual endpoint occurs approximately at point B. In this example, the final /v/ in "five" becomes devoiced and turns into an /f/, a weak fricative. Such weak fricatives are difficult to locate by eye (and sometimes even by ear).

[*] The criterion for deciding the actual beginning and ending points of the utterances was to use a combination of careful listening combined with precise visual examination of the waveform.

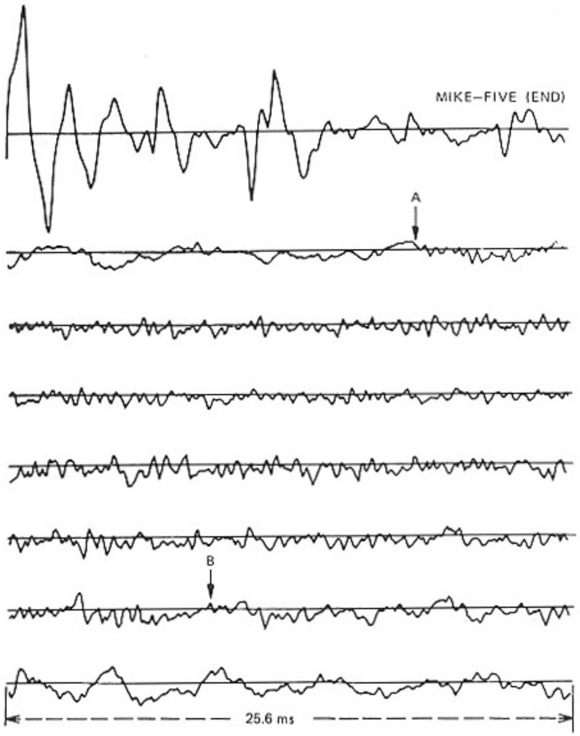

Fig. 6—Waveform for the end of the word "five."

As a final example, Fig. 7 shows the waveform for the end of the word "nine." It is quite difficult to say where the final nasal ends and where the silence begins. A reasonable location for the endpoint is the point marked END in this figure, although it is not clear how accurate this choice actually is.

Rather than give several more examples of situations in which it is difficult to locate either the beginning or the end of an utterance, we list below the broad categories of problems encountered. These include:

(*i*) Weak fricatives (/f, th, h/) at the beginning or end of an utterance.
(*ii*) Weak plosive bursts (/p, t, k/).
(*iii*) Final nasals.

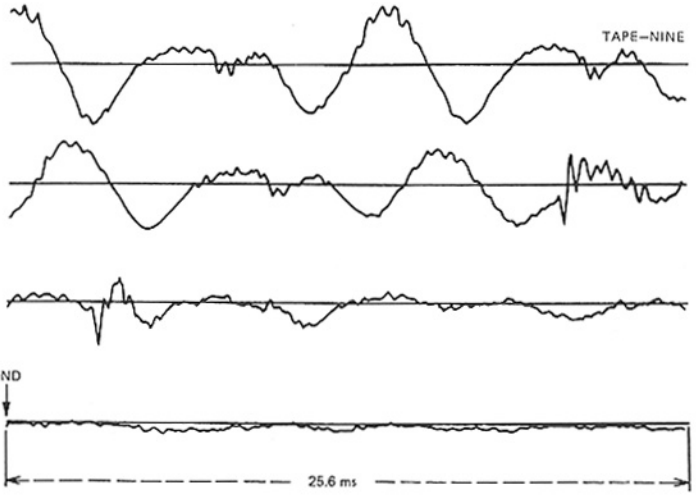

Fig. 7—Waveform for the end of the word "nine."

(*iv*) Voiced fricatives at the ends of words which become devoiced.
(*v*) Trailing off of certain voiced sounds—e.g., the final /i/ becomes unvoiced sometimes in the words "three" (/th-r-i/) or "binary" (/b-aI-n-e-r-i/).

The approach we have taken to solve these problems in an automatic endpoint-location algorithm is a pragmatic one. Our goal is to isolate enough of the word (utterance) so that a reasonable acoustic analysis of what is obtained is sufficient for accurate recognition of the word. Thus, it is not necessary to locate *exactly* the point where the word begins or ends, but instead it is important to include all significant acoustic events within the utterance. For a word like "binary," it is of little consequence if the trailing off unvoiced energy is omitted (in fact, it is probably quite helpful for a "phonetic" word-recognition strategy); however, for a word like "four" it is important to be able to reliably locate and include the initial weak fricative /f/. For this last example, the word "four," it is not necessary to include the entire initial unvoiced interval; in fact, experience has shown that 30 to 50 ms of unvoiced energy is sufficient for most word-recognition purposes. This type of knowledge is of great importance in an endpoint-finding algorithm because it enables you to set conservative values on all decision thresholds (thereby guaranteeing a very low false-alarm rate) and, for the word-recognition application, the concomitant high miss rate will be of little practical significance. In Section III, we give the details of one practical implementation of an endpoint-location algorithm.

III. THE ENDPOINT-LOCATION ALGORITHM

Based on the preceding discussion, the goals of the endpoint algorithm are:

(*i*) Simple, efficient processing.
(*ii*) Reliable location of significant acoustic events.
(*iii*) Capability of being applied to varying background silences.

The first goal implies that only simple measurements can be made on the speech waveform as a basis for the decision. If speed and simplicity were not major issues, far more sophisticated processing could be used to give a better, more accurate result.

With the above considerations in mind, the endpoint location algorithm that was implemented is based on two simple measurements, energy and zero crossing rate, and uses simple logic in the final decision algorithm. Both energy and zero crossing rate are simple and fast to compute, and, as seen in Section II, can give fairly accurate (although conservative) indications as to the presence or absence of speech. Before proceeding to a description of the algorithm, we first define how the energy and zero crossing rate are measured. The speech "energy," $E(n)$, is defined as the sum of the magnitudes of 10 ms of speech centered on the measurement interval,[2] i.e.,

$$E(n) = \sum_{i=-50}^{50} |s(n+i)|, \qquad (1)$$

where $s(n)$ are the speech samples and it is assumed that the sampling frequency is 10 kHz. The choice of a 10-ms window for computing the energy and the use of a magnitude function rather than a squared-magnitude function were dictated by the desire to perform the computations in integer arithmetic and, thus, to increase speed of computation. Further, the use of a magnitude de-emphasizes large-amplitude speech variations and produces a smoother energy function. By way of example, Fig. 8 shows typical energy functions for the words "directive" and "multiply." (The beginning and end of these words is noted on these energy plots.) For this example, the energy function is computed once every 10 ms, or 100 times per second.

The zero (level) crossing rate of the speech, $z(n)$, is defined as the number of zero (level) crossings per 10-ms interval. Although the zero crossing rate is highly susceptible to 60-Hz hum, dc offset, etc., in most cases it is a reasonably good measure of the presence or absence of unvoiced speech.

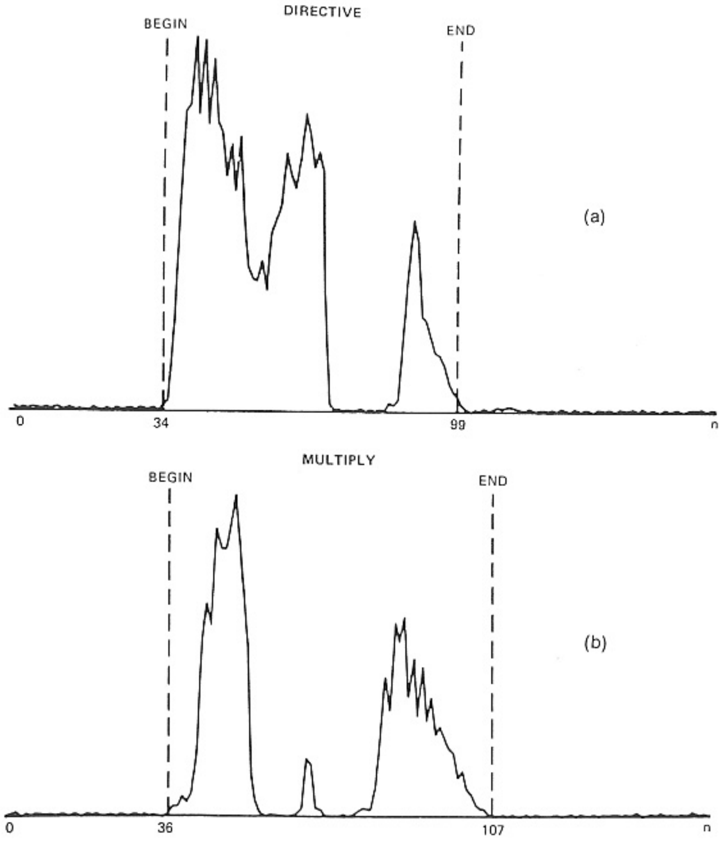

Fig. 8—Typical energy plots for the words "directive" and "multiply" with markers indicating the beginning and end of the utterance.

Figure 9 shows a flowchart of the endpoint-location algorithm. The speech waveform is filtered prior to sampling at 10 kHz by a bandpass filter with a 100-Hz low-frequency cutoff and a 4000-Hz high-frequency cutoff and having 48 dB per octave skirts. It is assumed that during the first 100 ms of the recording interval there is no speech present. Thus, during this interval, the statistics of the background silence are measured. These measurements include the average and

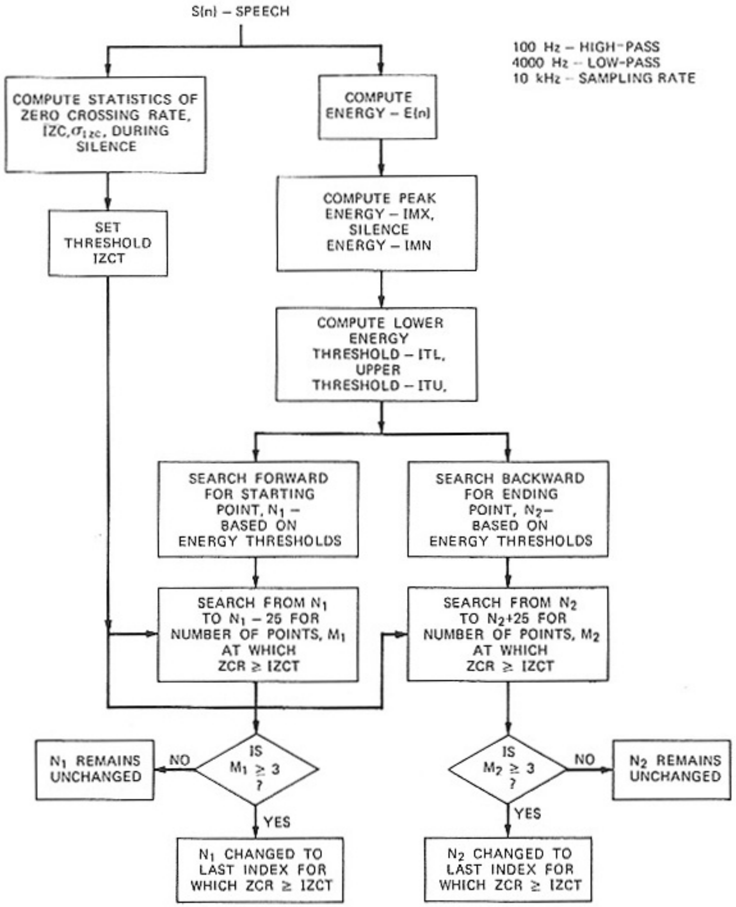
S(n) – SPEECH
100 Hz – HIGH-PASS
4000 Hz – LOW-PASS
10 kHz – SAMPLING RATE
COMPUTE STATISTICS OF ZERO CROSSING RATE, $\overline{IZC}$, σ_{IZC}, DURING SILENCE
COMPUTE ENERGY – E(n)
SET THRESHOLD IZCT
COMPUTE PEAK ENERGY – IMX, SILENCE ENERGY – IMN
COMPUTE LOWER ENERGY THRESHOLD – ITL, UPPER THRESHOLD – ITU.
SEARCH FORWARD FOR STARTING POINT, N_1 – BASED ON ENERGY THRESHOLDS
SEARCH BACKWARD FOR ENDING POINT, N_2– BASED ON ENERGY THRESHOLDS
SEARCH FROM N_1 TO $N_1 - 25$ FOR NUMBER OF POINTS, M_1 AT WHICH ZCR ≥ IZCT
SEARCH FROM N_2 TO N_2+25 FOR NUMBER OF POINTS, M_2 AT WHICH ZCR ≥ IZCT
N_1 REMAINS UNCHANGED
NO
IS $M_1 \geq 3$?
IS $M_2 \geq 3$?
NO
N_2 REMAINS UNCHANGED
YES
YES
N_1 CHANGED TO LAST INDEX FOR WHICH ZCR ≥ IZCT
N_2 CHANGED TO LAST INDEX FOR WHICH ZCR ≥ IZCT

Fig. 9—Flowchart for the endpoint algorithm.

standard deviation of the zero crossing rate and the average energy. If any of these measurements are excessive, the algorithm halts and warns the user. Otherwise, a zero crossing threshold, $IZCT$, for unvoiced speech is chosen as the minimum of a fixed threshold, IF (25 crossings per 10 ms), and the sum of the mean zero crossing rate during silence, $\overline{IZC}$, plus twice the standard deviation of the zero crossing rate during silence, i.e.,

$$IZCT = \mathrm{MIN}(IF, \overline{IZC} + 2\sigma_{IZC}). \tag{2}$$

The energy function for the entire interval, $E(n)$, is then computed. The peak energy, IMX, and the silence energy, IMN, are used to set two thresholds, ITL and ITU, according to the rule

$$I1 = 0.03*(IMX - IMN) + IMN \tag{3}$$

$$I2 = 4*IMN \tag{4}$$

$$ITL = \mathrm{MIN}(I1, I2) \tag{5}$$

$$ITU = 5*ITL. \tag{6}$$

Equation (3) shows $I1$ to be a level which is 3 percent of the peak energy (adjusted for the silence energy), whereas (4) shows $I2$ to be a level set at four times the silence energy. The lower threshold, ITL, is the minimum of these two conservative energy thresholds, and the upper threshold, ITU, is five times the lower threshold.

The algorithm for a first guess at the endpoint locations is shown in Fig. 10. The algorithm begins by searching from the beginning of

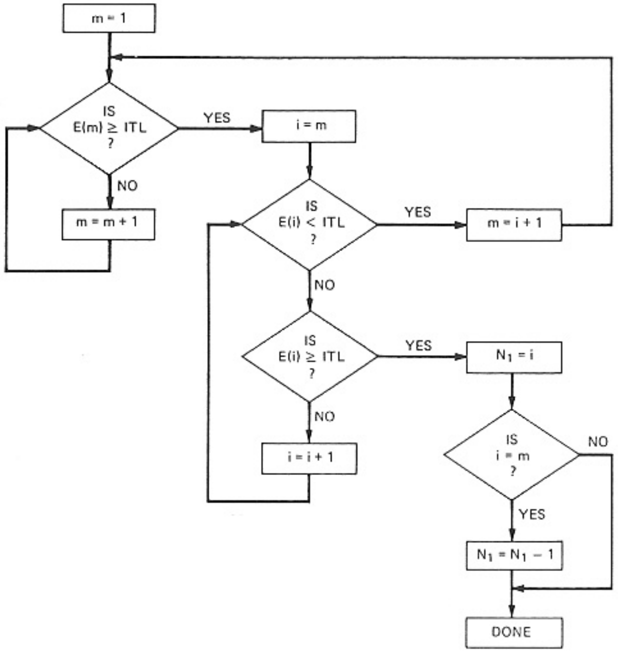

Fig. 10—Flowchart for the beginning point initial estimate based on energy considerations.

the interval until the lower threshold is exceeded. This point is preliminarily labeled the beginning of the utterance unless the energy falls below ITL before it rises above ITU. Should this occur, a new beginning point is obtained by finding the first point at which the energy exceeds ITL, and then exceeds ITU before falling below ITL; eventually such a beginning point must exist. A similar algorithm (shown in Fig. 11) is used to define a preliminary estimate of the endpoint of the utterance. We call these beginning and ending points N_1 and N_2, respectively.

Until now, we have only used energy measurements to find the endpoint locations; and these endpoint locations are conservative in that fairly tight thresholds are used to obtain these estimates. Thus, at this point, it is fairly safe to assume that, although part of the utterance may be outside the (N_1, N_2) interval, the actual endpoints are not within this interval. In relation to this, the algorithm proceeds to examine the interval from N_1 to $N_1 - 25$, i.e., a 250-ms interval preceding the initial beginning point, and counts the number of intervals where the zero crossing rate exceeds the threshold $IZCT$. If the number of times the threshold was exceeded was three or more, the starting point is set back to the first point (in time) at which the threshold was exceeded. Otherwise, the beginning point is kept at N_1. The rationale behind this strategy is that for all cases of interest, exceeding a tight threshold on zero crossing rate is a strong reliable indication of unvoiced energy. Of course, it is still possible that a weak fricative will not pass this test, and will be missed. However, in these cases there is no simple, *reliable* method of distinguishing such a weak fricative from background silence.

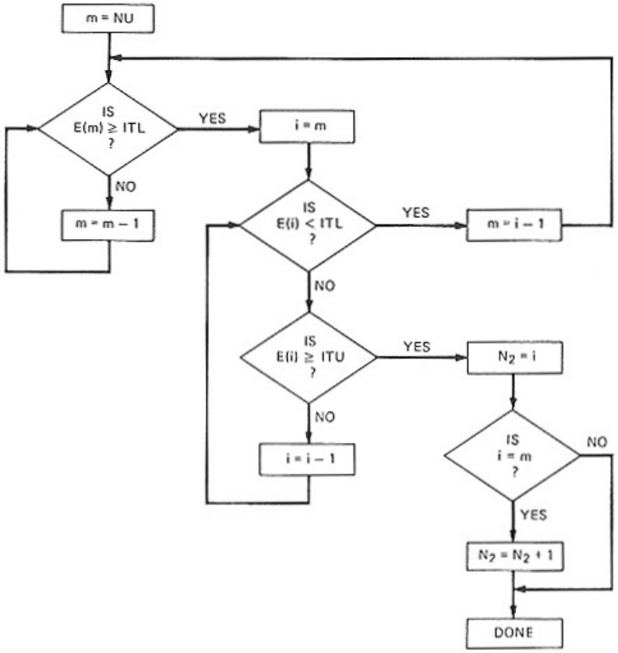

Fig. 11—Flowchart for the ending point initial estimate based on energy considerations.

A similar search procedure is used on the endpoint of the utterance to determine if there is unvoiced energy in the interval from N_2 to $N_2 + 25$. The endpoint is readjusted based on the zero crossing test results in this interval.

To illustrate the use of the endpoint algorithm, Fig. 12 shows representative contours of the energy and zero crossings for an utterance. Using the energy criterion alone, the algorithm chooses the point N_1 as the beginning of the utterance and N_2 as the end of the utterance. By searching the interval from N_1 to $N_1 - 25$, the algorithm finds a large number of intervals with zero crossing rates exceeding the threshold; thus, the beginning point is moved to $\hat{N}_1$, the first point (in time) that exceeded the zero crossing threshold. Similar examination of the interval from N_2 to $N_2 + 25$ shows no significant number of

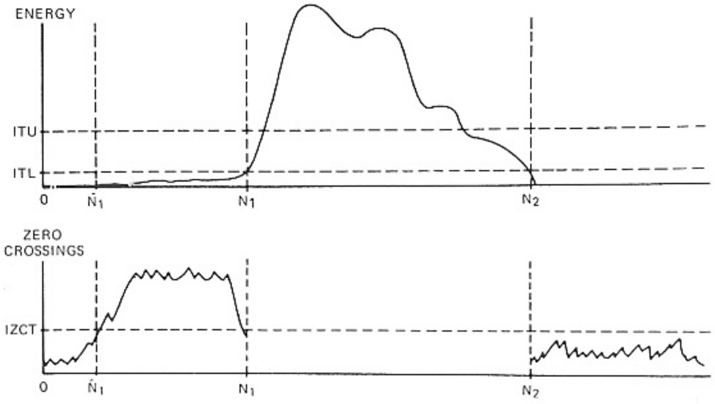

Fig. 12—Typical example of energy and zero crossings data for a word beginning with a strong fricative.

intervals with high zero crossings; thus, the point N_2 is retained as the endpoint of the utterance.

In Section IV, we give examples of the use of the endpoint algorithm for a large number of words with different speakers and different acoustic environments.

IV. EXAMPLES OF THE USE OF THE ENDPOINT ALGORITHM

The endpoint algorithm described in Section III was implemented on the DDP-516 computer facility of the Bell Laboratories Acoustics Research Department. The algorithm was tested using the two modes of recording described in Section II: high-quality tape recordings from a soundproof booth and on-line recordings using a close-talking microphone.

Figures 13 and 14 show examples of how the algorithm worked on typical isolated words. In Fig. 13 there are eight plots of the energy function for eight different words (of two different speakers). Some of the words were recorded on-line (marked MIKE) and others were recorded on tape (marked TAPE) from the soundproof booth. The markers on each plot show the beginning point and ending point of each word, as determined by the automatic algorithm. For the example in Fig. 13a (the word "nine"), the energy thresholds were sufficient to locate the endpoints. For the example in Fig. 13b (the word "replace"), the zero crossing algorithm was used to determine the ending point due to the final fricative /s/. It should be noted that even though the final /s/ has fairly large energy, since the energy thresholds were set conservatively, the energy criterion was not able to find the actual endpoint of the word. Instead, the zero crossing algorithm was relied upon in this case. In Fig. 13c, the final /t/ in the word "delete" was correctly located because of the significant zero crossing rate over the 70-ms burst when the /t/ was released. Thus, even though there was little energy or zero crossing activity for about 50 ms in the stop gap, the algorithm was able to correctly identify the endpoint because of the strength of the burst. On the other hand, if the burst had been weak, the ending point would have been located at the beginning of the stop gap.

Figure 13d is an example in which the energy during the silence was significant in a couple of places prior to the beginning of the word "subtract," yet the algorithm successfully eliminated these places from consideration because of the low zero crossing rates. In this example, a relatively weak burst in the final /t/ was correctly labeled as the endpoint.

Figures 13e through 13h show examples of words with fricatives at either the beginning or end of the word. In all cases, the algorithm was

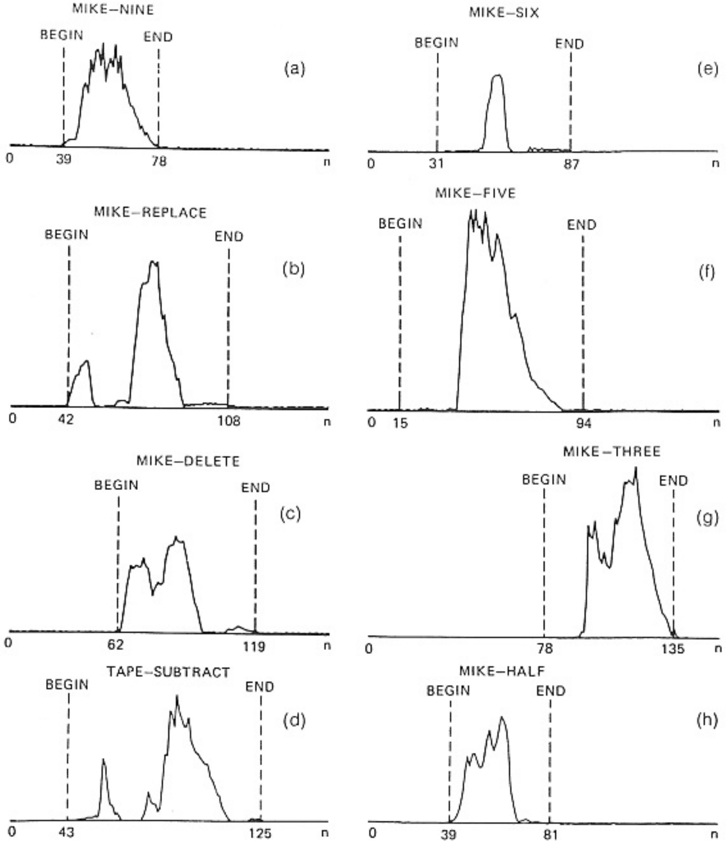

Fig. 13—Sequence of energy plots showing how the endpoint algorithm performed over a variety of words.

able to correctly place the appropriate endpoint so that a reasonable amount of unvoiced duration was included within the boundaries of the word.

Figure 14 shows three examples of how the algorithm performed for the word "four." It can be seen from the location of the beginning point that, although the level of the initial /f/ varied from strong to weak, the zero crossing indicator was able to find positive indications of the frication noise in all three cases. As discussed earlier, there are many examples where initial or final fricatives (mainly /f/ and /th/)

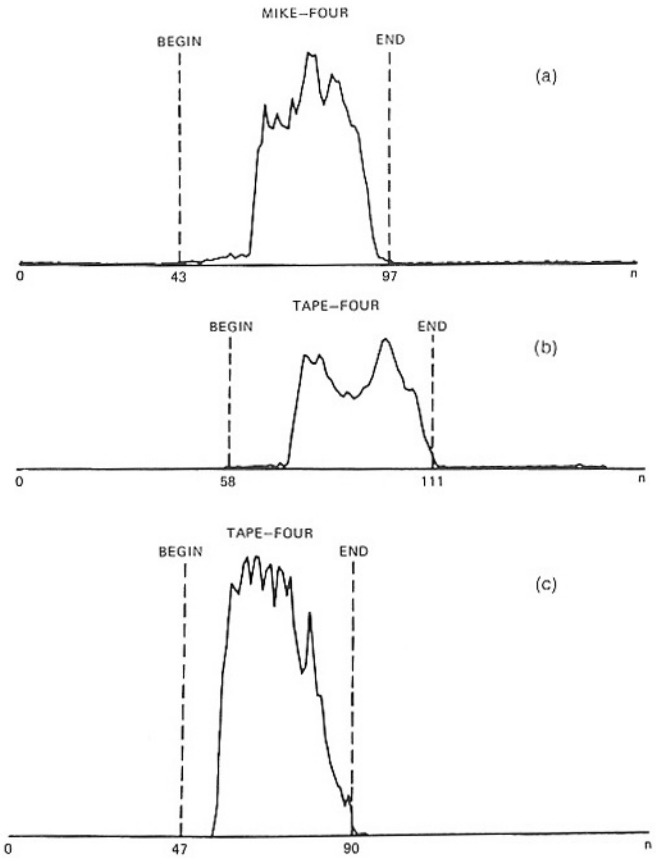

Fig. 14—Energy plots and endpoint assignments for three variations of the word "four."

were so weak they were indistinguishable from the background silence. In Section V, we discuss more sophisticated techniques for distinguishing such weak fricatives from background silence.

Two sets of formal tests were made on the algorithm. In one test, the 54-word vocabulary used by B. Gold in his word-recognition experiments[3] was read by two males and two females. For this vocabulary, the algorithm made no gross errors in locating the beginning and ending points. The algorithm did make a number of small errors of the type discussed earlier, such as losing weak fricatives or releases of

stops; however, none of these errors seriously affected the human recognition (based solely on listening) of the utterance from the portion that the algorithm did locate correctly. Thus, in some pragmatic sense, such errors can be tolerated for word recognition purposes; although for such applications as computer voice response, these small errors would probably be significant.

The second test involved 10 speakers each repeating the 10 digits from zero to nine in 10 separate sessions. (These data were actually measured for a digit-recognition experiment that used this endpoint location algorithm.) For this test, there were essentially no gross errors in locating the endpoints; in fact, it was determined that for purposes of word recognition, the algorithm was essentially error free.

V. DISCUSSION OF THE ENDPOINT-LOCATION PROBLEM

The problem of accurately locating the endpoints of an utterance is actually a specific case of the more general problem of labeling an interval of a signal as silence, unvoiced, or voiced. If one had a perfect technique for this three-level decision, the endpoint-location problem would be trivially solved. However, such an ideal algorithm does not exist as yet. Therefore, we have looked for partial solutions to this more specific problem of isolating speech from a noisy background.

The solution to this problem was based on the premise that somewhere within the given interval there was an utterance and that it would be easy to isolate the broad region in which the speech was located using energy measures alone. From this interval, we set very conservative thresholds on the speech energy (normalized to the maximum speech energy) to get a good first guess at the endpoints of the utterance. The zero crossing rate of the waveform outside these initial estimates of the endpoint was used to provide better estimates as to the existence of unvoiced speech energy in a broad region on either side of the initial endpoints.

The question now arises as to how to make the algorithm work better. One of our key goals in the original formulation was to make the algorithm fast and efficient. To this end, the readily available parameters of short-time energy and zero crossing rate were the only ones used in the decision-making process. To increase the sophistication and thereby the accuracy of the algorithm would require the inclusion of other speech parameters, such as predictor coefficients, autocorrelation coefficients, etc. The use of such additional measurements is predicated upon knowledge of how they differ for silence and for speech. Atal[4] has suggested a reasonable pattern-recognition approach for making the distinction between the three classes of silence, unvoiced speech, or voiced speech. This method, although promising, is much slower in running and, thus, cannot be relied upon in an on-line environment. It does, however, give good indications that the problems associated with this decision are not totally untractable.

VI. SUMMARY

We have presented a fast, efficient algorithm for locating the endpoints of an utterance in a background of noise. The algorithm is based on two measurements made on the speech: short-time energy and zero crossing rate. Although the algorithm does make small errors in finding the exact endpoints of the utterance, it was designed to minimize the number of gross errors (off by more than 50 ms) in the analysis. The algorithm has been found to be sufficiently reliable and accurate that it is currently being used in on-line experiments on word recognition.

REFERENCES

1. H. F. Silverman and N. R. Dixon, "A Parametrically Controlled Spectral Analysis System for Speech," IEEE Trans. on Acoustics, Speech, and Signal Processing, *ASSP-22*, No. 5 (October 1974), pp. 362–381.
2. R. W. Schafer and L. R. Rabiner, "Parametric Representations of Speech," Proc. IEEE Speech Recognition Symposium, Pittsburgh, April 1974.
3. B. Gold, "Word Recognition Computer Program," MIT Research Lab. of Electronics, Technical Report 452, Cambridge, June 1966.
4. B. Atal, personal communication.