S10.12

FAST ENDPOINT DETECTION ALGORITHM FOR ISOLATED WORD RECOGNITION IN OFFICE ENVIRONMENT

Evangelos S. Dermatas, Nikos D. Fakotakis, George K. Kokkinakis

Department of Electrical Engineering
Wire Communications Lab.
University of Patras, 261 10 Patras, Greece

ABSTRACT

This paper presents a simple and fast algorithm for accurately locating the endpoints of isolated words spoken in office environment. This algorithm is based on energy measures and threshold logic and is adaptive to almost any low noise acoustic environment. Experimental results from the algorithm's application were both acoustically checked and compared to hand-edited results by skilled personnel. It was shown that the accuracy of the algorithm is quite acceptable.

1. INTRODUCTION

For the detection of the endpoints of speech signals in speech recognition systems several algorithms have been proposed during the last years [1-5]. Most of these algorithms are based on simple parameters such as energy contours, zero crossings or LPCs and have a low accuracy which decreases significantly the overall accuracy of the speech recognition system [1-2]. On the contrary, algorithms with high accuracy are based on acoustic parameters, the calculation of which is rather time consuming making the response time of the system slow [5].

In this paper we present a simple, fast and accurate algorithm for the detection of the endpoints of isolated words spoken in office environment. This algorithm is based on energy measures and threshold logic, and is adaptive to almost any low background noise.

2. DESCRIPTION OF THE ALGORITHM

The proposed algorithm consists of four steps. In the first, the speech signal corresponding to a single word is preprocessed and the background noise is estimated which is used to adapt the decision threshold values of the following steps. In the second step, the starting-point of the first and the ending-point of the last voiced sound are located to be used as reference points for searching the endpoints. In the third step, a low energy level area at the beginning and the ending of the input signal are located, where the endpoints are supposed to lie. Finally in the fourth step, the endpoints of the utterance are located.

2.1. Background Noise Estimation

The input signal is pre-emphasized, to eliminate the d-c component and to emphasize the higher frequency components, using an one-zero filter, i.e.

$$s'_n = s_n - s_{n-1} \tag{1}$$

where, s_n is the quantized input speech signal, for $n = 1,2,...N$ samples.

From samples taken at the beginning and the ending of the input signal, the background noise (acoustic environment) is estimated. To this end, the energy levels of two wide-length non-overlapping frames at the beginning and two at the ending of the signal, are calculated, using the relation

$$E_k = \sum_{n=(k-1)W_L+1}^{kW_L} (s'_n)^2 , \tag{2}$$

where, $k=1,2,K-1,K$ the two frames at the beginning and the ending of the signal respectively, and W_L an 80 msec frame (W_L = 800 for 10 kHz sampling frequency).

The noise level at the front-end of the signal (E_F) is estimated using the first two energy frames. If the difference between the energy levels of the two frames is equal or less than twice the value of one frame, the noise level is taken to be equal to the mean value of the two frames, otherwise to be the minimum of the energy levels of the two frames, i.e.

$$E_F = \begin{cases} \dfrac{E_1 + E_2}{2}, & \text{if } 0.5 \le E_1 / E_2 \le 2 \\ \min(E_1, E_2), & \text{elsewhere.} \end{cases} \tag{3}$$

The noise level at the back-end of the signal (E_B) is estimated in the same way, using the last two energy frames, i.e.

$$E_B = \begin{cases} \dfrac{E_{K-1} + E_K}{2}, & \text{if } 0.5 \le E_{K-1} / E_K \le 2 \\ \min(E_{K-1}, E_K), & \text{elsewhere.} \end{cases} \tag{4}$$

Finally, the background noise of the input

CH2977-7/91/0000-0733 $1.00 © 1991 IEEE

signal (E_N) is estimated using the noise levels at the front and back ends. If the difference between the two values is equal to or less than twice the value of one level, the background noise is taken to be equal to the mean value of the two noise levels, otherwise the background noise cannot be estimated. In that case, the speech signal is assumed to be pruned, therefore it is rejected, i.e.

$$E_N = \begin{cases} \dfrac{E_F + E_B}{2}, & \text{if } 0.5 \le E_F / E_B \le 2 \\ \text{unknown}, & \text{elsewhere.} \end{cases} \quad (5)$$

However, the background noise obtained, should lie within two limits, $T_N < E_N < T_S$, otherwise the speech signal is not acceptable as being either to noisy or underamplified. T_N is the maximum noise threshold for an acceptable noisy environment and T_S is the minimum silence threshold for an indication of possible misconnection in the input circuit or underamplification of the input signal. These two thresholds, depend on the type of microphone and the quantization error. In our tests, they have been experimentally derived once as $T_N = 3.3 W_L$ and $T_S = 0.6 W_L$.

2.2. Location of the first and the last voiced sound

The starting-point of the first voiced sound of the input utterance and the ending-point of the last one are located to be used as reference points for the location of the actual endpoints of the speech signal.

Searching the amplitude-time function from left to right with an 80 msec frame, in a 0.1 msec step (one sample), the first frame with V peaks above an amplitude threshold T_A is assumed to lie at the beginning of the first voiced sound of the utterance (V is experimentaly derived). Thus, the starting-point (t_{F3}) of the front voiced sound is obtained by

$$t_{F3} = \underset{i}{\text{argmin}} \left(\sum_{n=i-W_L+1}^{i} u(s'_n - T_A) > V \right), \; W_L \le i < N \quad (6)$$

where, $u(x)$ is a step function defined by

$$u(x) = \begin{cases} 1, & x \ge 0 \\ 0, & \text{elsewhere.} \end{cases} \quad (7)$$

The amplitude threshold is experimentally derived from the background noise E_N, using the relation

$$T_A = C\sqrt{E_N / W_L} + A \quad (8)$$

where $A = \sqrt{E_N / W_L}$ is a d-c level equivalent to the background noise of the signal and C is an experimentally derived constant ($C = 7$).

In the same way, searching the amplitude-time function backwards from its far end (from right to left), the ending-point (t_{B3}) of the back voiced sound is obtained by

$$t_{B3} = \underset{i}{\text{argmax}} \left(\sum_{n=i+1}^{i+W_L} u(s'_n - T_A) > V \right), \; 0 \le i < N - W_L \quad (9)$$

If the relations (6) and (9) can not be satisfied or if the distance between the points t_{F3} and t_{B3} is below a certain threshold, i.e. $t_{B3} - t_{F3} < t_{min}$, where t_{min}=20 msec is a minimum duration of vowel phonemes, the algorithm recognizes absence of speech in the input signal and the procedure terminates. Figure 1 shows the starting-point of the first voiced sound and the ending-point of the last one in the Greek word /COr'is/ (without).

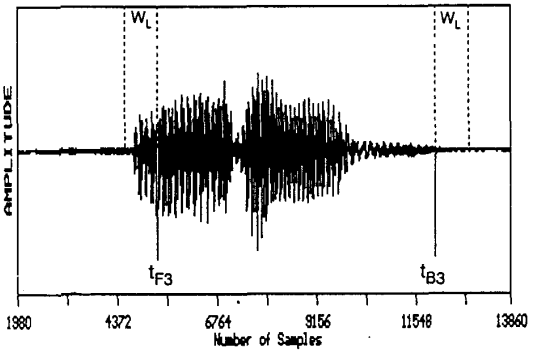

Figure 1: The starting-point of the first voiced sound t_{F3} and the ending-point of the last voiced sound t_{B3}, in the Greek word /COr'is/, as estimated by the proposed algorithm.

2.3. Location of Low Energy Areas

At the beginning of the input signal a low energy area is located, where it is assumed that the front endpoint lies. Also, at the ending of the signal another low energy area is located, where the back endpoint is assumed to lie. With these areas the search effort necessary for the location of the end-points in the final step, is significantly reduced.

An 80 msec frame is moved backwards, from the front voiced sound starting-point t_{F3}, in a 0.1 msec step, to calculate the energy contour of the signal. This energy function is compared to two thresholds in order to locate the boundaries of the front low energy area. The analytical expression for the estimation of the boundaries is given by,

$$t_{Fq} = \underset{i}{\text{argmax}} \left(\sum_{n=i-W_L+1}^{i} (s'_n)^2 < T_{Fq} \right), \; W_L < i < t_{F3} \quad (10)$$

where q=1,2 correspond to the two boundaries. T_{F2} and T_{F1}, are two experimentally derived energy thresholds ($T_{F1} = 1.1 E_N$, $T_{F2} = 2.2 E_N$). Figure 2 shows the energy thresholds T_{F2}, T_{F1} and the corresponding time boundaries, t_{F2}, t_{F1}, of the front low energy area in the Greek word /COr'is/.

At the back-end, an 80 msec frame is moved forwards from the ending-point of the last voiced sound (t_{B3}), in 0.1 msec steps, to calculate the energy contour of the signal. This energy function is compared to two energy thresholds to locate the boundaries of the back low energy area:

$$t_{Bq} = \underset{i}{\text{argmin}} \left(\sum_{n=i-W_L+1}^{i+W_L} (s'_n)^2 < T_{Fq} \right), t_{B3} < i < N - W_L \quad (11)$$

where q=1,2 correspond to the two boundaries. T_{B2} and T_{B1} are two experimentally derived energy thresholds (in our tests they have the values T_{B2}=3.33E_N, T_{B1}=3.0E_N). It has to be noted that the back thresholds are higher than the front ones. This is because the back-end of the speech signal contains the breath noise.

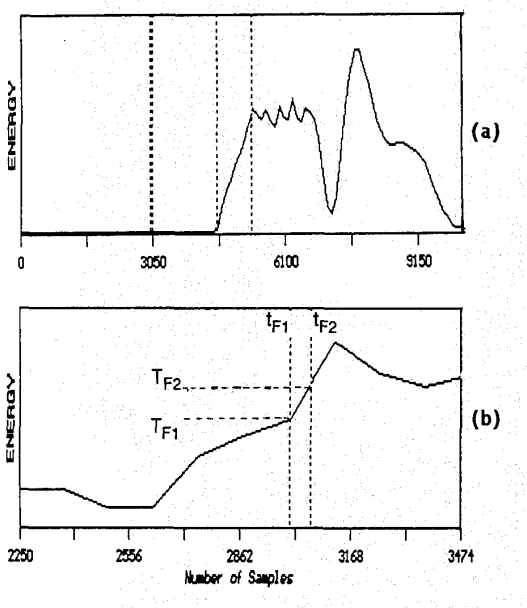

Figure 2: The two boundaries t_{F2}, t_{F1}, of the front low energy area, and the corresponding thresholds T_{F2}, T_{F1}, in the Greek word /COr'is/. Figure 2(a) gives the whole function, Figure 2(b) gives a magnification of the function's front-end.

2.4. Endpoint Detection

Within the two low energy areas estimated in the previous step, the location of the actual endpoints of the speech signal is searched. At the front area, from point t_{F1} to t_{F2} the contour of the energy ratio of two successive non-overlapping 30 msec frames is estimated, in a 0.1 msec step. The actual beginning of the speech signal corresponds to the maximum value of this energy ratio contour. The analytical expression for the estimation of the front endpoint is given by

$$t_F = \underset{i}{\text{argmax}} \left(\frac{\sum_{n=i+1}^{i+W_S} (s'_n)^2}{\sum_{n=i-W_S+1}^{i} (s'_n)^2} \right), \quad i \in [t_{F2}, t_{F1}] \quad (12)$$

where W_S is a 30 msec frame (i.e. W_S=300 for 10 KHz sampling rate).

In the same way, at the back low energy area, from point t_{B2} to t_{B1}, the contour of the energy ratio is estimated. The actual ending of the speech signal corresponds to the maximum value of this contour:

$$t_B = \underset{i}{\text{argmax}} \left(\frac{\sum_{n=i-W_S+1}^{i} (s'_n)^2}{\sum_{n=1-i-W_S}^{i+W_S} (s'_n)^2} \right), i \in [t_{B1}, t_{B2}] \quad (13)$$

3. EXPERIMENTAL RESULTS

The described algorithm has been evaluated with several tests using as vocabulary the alphadigits of the Greek language (34 words). The recordings were made in three different acoustic environments:

- Anechoic chamber, (S/N≈60dB)
- Quiet office environment (S/N≈40dB)
- Noisy environment (S/N≈25dB)

A number of 25 speakers (18 males and 7 females) has repeated the vocabulary 68 times in anechoic environment, 29 times in quiet office environment and 33 times in noisy environment, to establish the test database of the system.

The tests have been performed in the following way. The vocabulary was divided in seven categories according to the phoneme at the begininig or at the end of each word. The algorithm was then applied to the words of these categories and the result, endpoint definition or rejection of the speech signal, was measured. Table I shows the rejections of the speech signal measured, which made impossible the endpoint detection. As a whole, the algorithm works in almost all cases.

Environment	Rejections/tests	Accuracy (%)
Anechoic champ.	0/2312	100
Low-noise env.	3/986	99.6
Noisy env.	16/1122	98.3

Table I: Rejections during the steps of background noise estimation and low energy areas estimation, for the three acoustic environments.

The endpoint results taken from the algorithm were then checked both acoustically and optically for their accuracy by skilled personnel. Acoustic tests included gradual segmentation of the speech signal and listening to locate the endpoints. Optical tests included comparisons between automatically and manually located endpoints on the amplitude- and energy-time functions of the words. Both tests showed that the accuracy achieved by the proposed algorithm is quite acceptable, approaching that of the acoustic and manual definition of the endpoints. Results of these tests are shown on table II.

Words	Errors / Tests		
	Anechoic	Office	Noisy
Beginning with Vowel	0/275	0/275	0/275
Beginning with Voiced Consonant	0/300	0/300	1/300
Beginning with Unvoiced Consonant	0/75	1/75	3/75
Beginning with Stop	0/200	0/200	1/200
Ending with Vowel	0/725	0/725	0/725
Ending with Voiced Consonant	0/100	1/100	1/100
Ending with Unvoiced Consonant	0/25	1/25	2/25

Table II: Number of words with unacceptable endpoint detection by the algorithm vs. number of tested words.

Examples of the endpoint location in two different words belonging to different categories, are shown in figures 3 and 4.

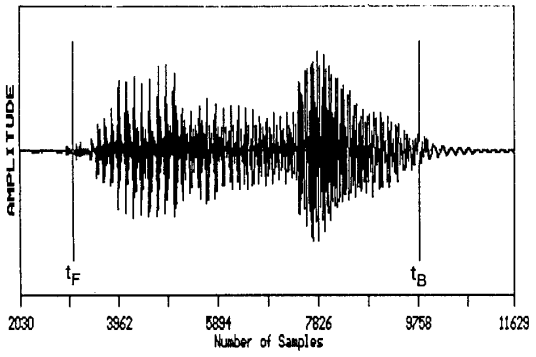

Figure 3: The endpoints t_F and t_B of the Greek word /kaL'u/, which begins with stop. Recording in quiet office environment.

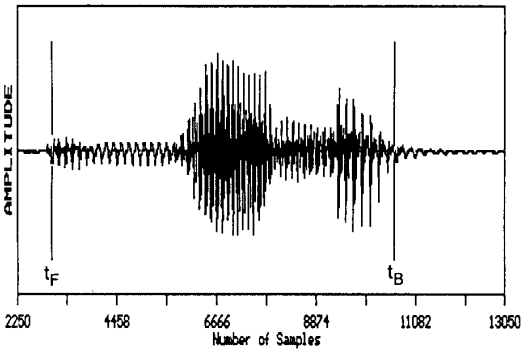

Figure 4: The endpoints t_F and t_B of the Greek word /v'ima/, which begins with unvoiced consonant. Recording in quiet office environment.

Concerning the computation time of the algorithm data of the mean multiplications (C_M), additions (C_A), incrementations (C_I) and comparisons (C_C) for recordings in the three environments, are presented on Table III.

Environment	C_M	C_A	C_I	C_C
Anechoic champ. (female & male)	4315	9172	17996	5463
Low-noise env. (male)	4042	9019	17570	5379
Office env. (female)	3875	7936	14625	4728
Noisy env. (female & male)	4744	11237	18552	7447

Table III: Average number of calculations of the endpoint detection algorithm in multiplications C_M, additions C_A, incrementations C_I and comparisons C_C.

4. CONCLUSION

A noise adaptive endpoint detection algorithm based on energy measures and rules has been described. The algorithm was tested in a vocabulary of 34 words, uttered by 25 speakers and recorded in different acoustic environments. The results were checked both acoustically and optically by skilled personnel and proved to approach those taken by acoustic tests or manually.

REFERENCES

[1] L. R. Rabiner and M. R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances", Bell Syst. Tech. J., Vol. 54, pp. 297-315, Feb. 1975.

[2] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon, "An Improved Endpoint Detector For Isolated Word Recognition", IEEE ASSP, Vol. 29 (4), pp. 777-785, Aug. 1981.

[3] D. Childers, M. Hahn, J. Larar, "Silent and Voiced/Unvoiced/Mixed Excitation (Four-Way) Classification of Speech", IEEE ASSP, Vol. 37 (11), pp. 1771-1774, Nov. 1989.

[4] M. H. Savoji, "A Robust Algorithm for Accurate Endpointing of Speech Signals", Speech Communication 8 (1989), pp. 45-60.

[5] L. R. Rabiner, C. E. Schmidt, B. S. Atal, "Evaluation of a Statistical Approach to Voiced-Unvoiced-Silence Analysis for Telephone-Quality Speech", Bell System Technical Journal, pp. 455-487, Mar. 1977.