A 3-Level Endpoint Detection Algorithm for Isolated Speech Using Time and Frequency-based Features

Goh Kia Eng*, and Abdul Manan Ahmad**

* Faculty of Computer Science and Information System, Universiti Teknologi Malaysia, Johor, Malaysia.
(Tel.: +60 (07)-5532070; E-mail: isaac604@hotmail.com)
** Faculty of Computer Science and Information System, Universiti Teknologi Malaysia, Johor, Malaysia.
(Tel.: +60 (07)-5532070; E-mail: manan@fsksm.utm.my)

Abstract: This paper proposed a new approach for endpoint detection of isolated speech, which proves to significantly improve the endpoint detection performance. The proposed algorithm relies on the root mean square energy (rms energy), zero crossing rate and spectral characteristics of the speech signal where the Euclidean distance measure is adopted using cepstral coefficients to accurately detect the endpoint of isolated speech. The algorithm offers better performance than traditional energy-based algorithm. The vocabulary for the experiment includes English digit from one to nine. These experimental results were conducted by 360 utterances from a male speaker. Experimental results show that the accuracy of the algorithm is quite acceptable. Moreover, the computation overload of this algorithm is low since the cepstral coefficients parameters will be used in feature extraction later of speech recognition procedure.

Keywords: Endpoint detection, speech recognition, rms energy, zero crossing rate, cepstral coefficients.

1. INTRODUCTION

A major cause in isolated speech recognition system is the inaccurate detection of the beginning and ending boundaries of test and reference patterns. Thus, it is essential for automatic speech recognition algorithms that speech segments be reliably separated from silence [1]. Recently, a real-world evaluation of a discourse system using isolated speech recognition showed that more than half of the recognition errors were due to the incorrect detection of speech endpoint detector [2]. According to Savoji, the required characteristics of an ideal speech endpoint detector are reliability, robustness, accuracy, adaptation, simplicity, real-time processing and no a priori knowledge of the noise [5].

For the detection of the endpoints of speech signals in speech recognition systems several algorithms have been proposed during the last few years. Most of these algorithms are based on simple parameters and have a low accuracy which decreases significantly the overall accuracy of the speech recognition system. 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.

The followings are typical speech endpoint detection algorithms [3]:
1) Energy-based algorithms with automatic threshold adjustment (EPD-ATA): They are intuitive approaches based on energy levels and durations of silence and speech. 5 energy thresholds are adapted automatically according to the voicing peak and ambient noise estimated from the first few frames.
2) Use of pitch information (EPD-PCH): This kind of algorithms relies on pitch extraction and energy variations.
3) Noise adaptive algorithms (EPD-NAA): This kind of algorithms use the log of the rms signal energy, the zero crossing rate, duration information and a set of heuristics. The thresholds used for the energy and the zero crossings are adapted automatically from a few frames provided by the signal environment.
4) Voice activation algorithms (EPD-VAA): This algorithm is based on energy and zero crossing parameters and a set of decision rules and threshold setting.

The proposed algorithm for endpoint detection in this paper is based on time and frequency features. Time features are rms energy and zero crossing. Frequency feature is cepstral coefficients (cepstrum). The proposed algorithm has the advantage of low computation overhead.

2. DESCRIPTION OF PROPOSED ALGORITHM

The algorithm consists of four basic steps: the pre-processing step, background noise estimation, initial endpoint detection, and actual endpoint detection.

2.1 Pre-processing

The speech signal was sampled at 16kHz and quantized to 16 bits per sample. In this paper, 400 samples per frame (25 ms) and the overlapping of 80 samples (5ms) were assumed. This overlapping is essential for a smooth transition feature from one frame to another frame. The speech data is normalized with respect to the maximum of the speech data and then preemphasized with first order low-pass filter to eliminate the d-c component and to emphasize the higher frequency component as show in equation (1). The factor of preemphasis *a*, was 0.95.

$$\tilde{S}[i] = S[i] - aS[i-1] \tag{1}$$

in which *i* is the number of samples included in the current utterance.

Before calculating the time and frequency features, the speech data in the frame will be smoothed by Hamming window to reduce the amplitude at the edges.

2.2 Background Noise Estimation

The background noise is estimated which is used to decide the threshold values of the following steps. From the samples taken at the beginning and the ending of the input signal, the background noise is estimated. rms energy is computed as (2).

$$E_n = \left[\frac{1}{W} \sum_{i=1}^{W} \tilde{S}_n^2[i] \right]^{\frac{1}{2}} \tag{2}$$

in which, i=1, 2, …, W-1, W. W is the length of a frame (we use W=256), n is the number of frame 1, 2, … N-1, N (N = Total Frame).

The noise level at the front-end of the signal , E_f is estimated using the first 5 energy frames where the energy values in these 5 frames are consistent between each others.

$$E_f = \frac{1}{5}\sum_{i=1}^{5} E_i \tag{3}$$

The noise level at the back-end of the signal, E_b is estimated in the same way, using the last 5 frames where their energy values are consistent between each others.

$$E_b = \frac{1}{5}\sum_{i=N-4}^{N} E_i \tag{4}$$

Finally, the background noise level of the input signal, E_N is estimated using the noise levels at the front and back ends at the following:

$$E_N = \frac{E_f + E_b}{2} \tag{5}$$

However, the rms energy of background noise obtained should lie within two limit thresholds, otherwise the speech signal is not acceptable as being either too noisy or under-amplified.

Another parameter zero crossing of the background noise is also been estimated in the similar way as that of the parameter rms energy. The following equations (6) – (9) are the estimation of background noise of zero crossing:

$$Z_n = \frac{1}{2}\sum_{i=1}^{W-1}\left\|\mathrm{sgn}\left(\tilde{S}_n\right) - \mathrm{sgn}\left(\tilde{S}_{n+1}\right)\right\| \tag{6}$$

which

$$\mathrm{sgn}(x) = \begin{cases} 1, & if \quad x \geq 0 \\ -1, & if \quad x < 0 \end{cases}$$

$$Z_f = \frac{1}{5}\sum_{i=1}^{5} Z_i \tag{7}$$

$$Z_b = \frac{1}{5}\sum_{i=N-4}^{N} Z_i \tag{8}$$

$$Z_N = \frac{Z_f + Z_b}{2} \tag{9}$$

However, the zero crossing of background noise obtained should lie within two limit thresholds, otherwise the speech signal is not acceptable as being either too noisy or under-amplified.

2.3 Initial Endpoint Detection

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

This part begins with the searching the rms energy function from frame with highest rmse to left with a frame in shifting step. The first frame whose rms energy is below an energy threshold T_e is assumed to lie at the beginning of the first voiced speech. Thus, the starting point, P_{F1} of the front voiced speech is obtained by

$$P_{F1} = \arg_n \max\left\{E_n < T_e, \quad n = m, m-1, ..., 0\right\} \tag{10}$$

in which E_n is defined by equation (1) and m is the index for frame with highest rms energy.

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

$$T_e = C_e \times E_N \tag{11}$$

which C_e is an experimentally derived constant.

In the same way, the ending point, P_{B1} of the last voiced speech is obtained by searching the energy function backwards from right to the left.

$$P_{B1} = \arg_n \min\left\{E_n < T_e, \quad n = m, m+1, ..., N\right\} \tag{12}$$

If the equations (10) and (12) cannot be satisfied or if the distance between the points P_{F1} and P_{B1} is below a certain threshold, the algorithm recognizes absence of speech in the input signal and the procedure terminated. Otherwise, the speech signal between these two reference points is assumed to be voiced speech segment.

Next, we utilize the parameter zero crossing to relax the endpoints. It begins with searching the zero crossing function from point P_{F1} backwards. The reference starting point, P_{F2} is obtained by

$$P_{F2} = \arg_n \max\left\{Z_n > T_{ZF}, \quad n = P_{F1}, P_{F1}-1, ...1\right\} \tag{13}$$

in which Z_n is defined by equation (6), T_{ZF} is the zero crossing threshold for front-end defined by

$$T_{ZF} = C_{ZF} \times Z_N \tag{14}$$

in which C_{ZF} is obtained experimentally.

In the same way, the reference ending point, P_{B2} is obtained by searching the zero crossing function from P_{B1} forwards:

$$P_{B2} = \arg_n \min\{Z_n > T_{ZB}, \quad n = P_{B1}, P_{B1}+1, ..., N\} \quad (15)$$

where T_{ZB} is the zero crossing threshold for back-end defined by

$$T_{ZB} = C_{ZB} \times Z_N \quad (16)$$

which C_{ZB} is obtained experimentally.

Due to the different characteristic of starting and ending phonemes of an isolated speech, different zero crossing thresholds are utilized for determining the starting-point and ending-point respectively.

2.4 Actual Endpoint Detection

In this part, the implementation is based on the discrimination between current frame and the last retained frame *j* and compare this distance with a distance threshold. The simplest discrimination measure we used which was also proved to be successful is the weighted Euclidean distance.

This method emphasizes the transient regions, which are more relevant for speech recognition. The boundary between voiced speech signal and silence can be determined by adopting the principle of this method. The decision criterion then becomes the following: leave the current frame out if $D(i,j) < T_D$.

Since the changes of speech signal can be better embodied in the frequency domain and cepstrum can be measured by Euclidean distance, cepstrum is adopted to determine the actual endpoints [4, 6, 7].

Let $D(i,j)$ be the Euclidean distance between the cepstrum vectors of frame *i* and *j*. If $D(i,j)$ is great than threshold T_D in the searching procedure, the transient of voiced and unvoiced speech segment is assumed to occur. In order to avoid the sudden high-energy noise, three frames are detected.

Searching from frame P_{F2} forwards until frame P_{B2}, the actual starting point P_{F3} is determined by

$$P_{F3} = \arg_n \min \left\{ \begin{matrix} D(n,n+1) > T_D \ \&\& \\ D(n,n+2) > T_D \ \&\& \\ D(n,n+3) > T_D \end{matrix} \right\} + 1 \quad (17)$$

in which $P_{F2} \le n \le P_{B2}$

Searching from frame P_{B2} backwards until frame P_{F2}, the actual ending point P_{B3} is determined by

$$P_{B3} = \arg_n \max \left\{ \begin{matrix} D(n,n-1) > T_D \ \&\& \\ D(n,n-2) > T_D \ \&\& \\ D(n,n-3) > T_D \end{matrix} \right\} - 1 \quad (18)$$

in which $P_{B2} \ge n \ge P_{F2}$

The final result or actual endpoint for the proposed algorithm are the starting point , P_{F3} and the ending point, P_{B3}.

3. EXPERIMENTAL RESULTS

The proposed algorithm is evaluated with isolated English digit utterances ranging from 1 to 9. The database consists of 360 utterances where each digit words is repeated for 40 times in laboratory environment.

The accuracy of the results obtained from the implementation of the proposed algorithm was evaluated both acoustically and optically by skill personal and speech analysis software. The acoustic evaluation is based on listening to locate the boundary of speech. The optical evaluation is based on the inspection to manually locate the boundary point. Both tests showed that the accuracy achieved by the proposed algorithm is quite acceptable. The results of these tests are shown in Table 1 and 2.

Table 1 Results of the detection for beginning boundary.

Digit word	Errors/Tests	Accuracy(%)
One	0/40	100.0
Two	0/40	100.0
Three	1/40	97.5
Four	2/40	95.0
Five	1/40	97.5
Six	4/40	90.0
Seven	2/40	95.0
Eight	0/40	100.0
Nine	0/40	100.0

Table 2 Results of the detection for ending boundary.

Digit word	Errors/Tests	Accuracy(%)
One	2/40	95.0
Two	0/40	100.0
Three	4/40	90.0
Four	8/40	80.0
Five	6/40	85.0
Six	10/40	75.0
Seven	9/40	77.5
Eight	2/40	95.0
Nine	1/40	97.5

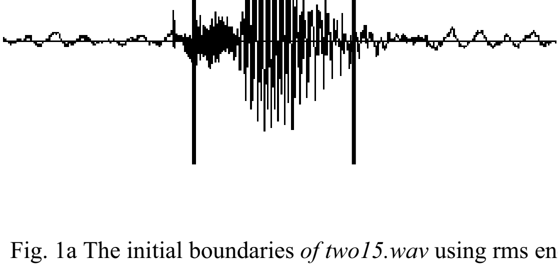

Fig. 1a The initial boundaries *of two15.wav* using rms energy in initial endpoint detection.

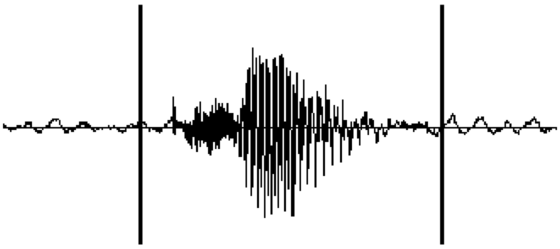

Fig. 1b The initial boundaries of *two15.wav* using zero crossing in initial endpoint detection.

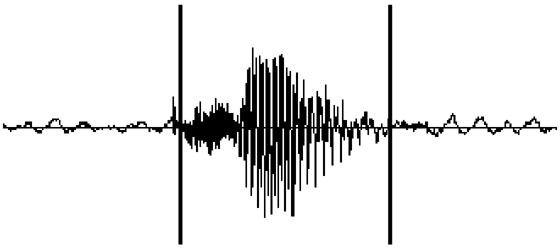

Fig. 1c The actual boundaries of *two15.wav* using cepstrum in actual endpoint detection.

Figure 1a, 1b and 1c show the endpoint results of word two15. These figures show the endpoints obtained by utilizing rms energy, zero-crossing and Euclidean distance of cepstrum, respectively. Figure 2a, 2b and 2c show the results of word *six30*. Figure 1c and 2c shows that the effectiveness of the utilization of Euclidean distance measurement to the frequency-based feature, cepstrum. So, this method is quite effective, acceptable and encouraging in a noisy environment.

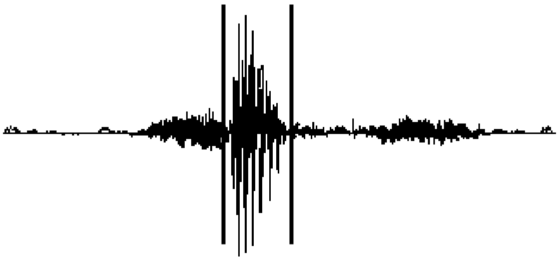

Fig. 2a The initial boundaries of *six30.wav* using rms energy in initial endpoint detection.

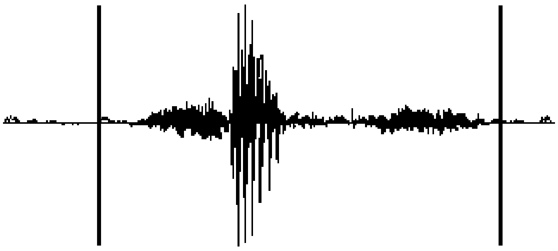

Fig. 2b The initial boundaries of *six30.wav* using zero crossing in initial endpoint detection.

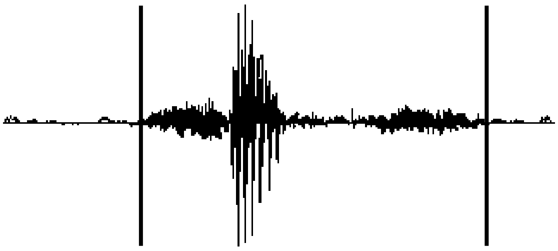

Fig. 2c The actual boundaries of *six30.wav* using cepstrum in actual endpoint detection.

4. CONCLUSION

A 3-level adaptive endpoint detection algorithm for isolated speech based on time and frequency parameters are introduced in this paper. The concept of Euclidean distance adopted in this algorithm can determine the segment boundaries between silence and voiced speech as well as unvoiced speech. The proposed algorithm was evaluated on a vocabulary of 360 isolated utterances. The results showed that this endpoint detection algorithm is effective in noisy environment.

5. REFERENCES

[1] L. R. Rabiner and M.R. Sambur, "An algorithm for determining the endpoints of isolated utterances", *The Bell System Technical Journal*, Vol.54, No.2, pp.297, February 1975.

[2] Lamel, L. F., Rabiner, L.R., Rosenberg, A. and Wilpon, J. "An improved endpoint detector for isolated word recognition", *IEEE ASSP Mag.*, Vol.29, pp. 777-785, 1981.

[3] Junqua J. C. B. Mak, and B. Reaves, "A robust algorithm for word boundary detection in the presence of noise", *IEEE Transactions on Speech and Audio Processing*, Vol.2, No.3, Jul. 1994.

[4] Hsiao-Fen Pai and Hsiao-Chuan Wang, "A two-dimensional cepstrum approach for the recognition of Mandarin syllable initials", *Pattern Recognition*, Vol.26, No.4, pp.569-577, 1993.

[5] M. H. Savoji, "A robust algorithm for accurate endpointing of speech", *Speech Commun*, Vol.8, pp 45-60, 1989.

[6] Yiying Zhang, Xiaoyan Zhu and Yu Hao, "A robust and fast endpoint detection algorithm for isolated word recognition", *IEEE International Conference on*

Intelligent Processing Systems, Vol.4, No.3, pp. 1819-1822, 1997.

[7] J. A. Haigh, and j. S. Mason, “Robust voice activity detection using cepstral features”, *Computer Communication, Control and Power Engineering. Proceedings of the IEEE Region 10 Conference TENCON*, Vol. 3, pp. 321-324, 1993