Improvement of Endpoint Detection for Thai Isolated Word Recognition

Narongrit Sae-ngan and Narong Buabthong

Department of Electrical Engineering
Thammasat University Rangsit Campus
Klong Luang, Pathumthani 12120, Thailand
pao_kmutt@hotmail.com, narongbt@engr.tu.ac.th

***Abstract*— This paper proposes an approach for speech detection with Mel-Frequency Cepstral Coefficients (MFCC) for Thai isolated word recognition by increasing the difference between pre-frames and post-frames (Delta) to signal and distinguishing speech level in terms of speech energy, adaptive zero-crossing rate, and Euclidean distance of cepstrum-domain coefficients. Experimental result shows that the accuracy of speech recognition system gains the highest rate at 98.49% according to back-propagation neural network, from the database of 10 speakers by testing MFCC comparison of 3 different endpoint detection methods.**

I. INTRODUCTION

Speech recognition is a process to find the most suitable interpretation for a signal which represents human speech. Then, a computer recognizes all data including a meaning of word. If the computer output matches with speech, users can control machine by voiced command through a microphone. Isolated word recognition is suitable for any voiced command with few words. It is because the database and training period are shorter than continuous speech recognition.

According to the isolated word recognition, a speech signal naturally contains both a speech segment (i.e. voiced, unvoiced) and a background noise including noises from speaker (e.g. breath, lip smacks) or environmental noises (e.g. air-conditioners, car engines). Thus, an endpoint detection process is used to distinguish the beginning and the ending of actual speech as well as to remove non-speech signal including silence and noise segments. This process is essential for the determination of the correct isolated word boundary and the rejection of the background noise. If the speech signal including noisy environments is detected, the received features will be different from the learning system and inefficient recognition will be processed. The most commonly used method of endpoint detection for isolated word recognition is the use of short-time energy and zero-crossing rate. This method can work efficiently in a clean environment. In the case of low signal to noise ratio (SNR), the proper estimation of the start and end of speech degrades due to their undistinguished signal.

There is an increasing interest in the study of endpoint detection under various real-life circumstances. In the year of 2000, for instance, Liang-sheng Huang and Chung-ho Yang [1] proposed the features of combining energy (time domain) and entropy (frequency domain) for endpoint detection in a noisy in-car environment. Moreover, in the year of 2002, Sahar E. Bou-Ghazale and Khaled Assaleh [2] applied Mel-Frequency Cepstral Coefficients (MFCC) and Euclidean distance in order to improve endpoint detection in various noisy environments. Their improvement of endpoint detection incisively indicates the difference of the background noise and the speech segment. Nevertheless, it is difficult to distinguish the endpoints accurately in various environments, with the result that threshold can be adapted to different types of noise.

This paper presents a new method of endpoint detection for Thai isolated word recognition. This proposed method is modified in three aspects. First, there is an increase in the difference between pre-frames and post-frames (Delta) and a computation of the speech level, determined by short-time energy, adaptive zero-crossing rate, and Euclidean distance of MFCC. Next, the threshold is obtained from the silent signal of the first three frames and the last three frames. Finally, the frame of the start and finale of endpoint is determined to detect MFCC which will subsequently match the speech.

In this introductory part, we briefly review some concepts, methods, and relevant studies, which proved helpful in highlighting the speech endpoint detection. Section II reviews the speech recognition background. Section III describes the proposed method of Delta and endpoint detection algorithm. Section IV sets out the experimental evaluation and discussion of speech detection with MFCC. Finally, Section V summarizes major conclusions.

II. BACKGROUND

The speech recognition process begins with the digital sampling of the speech signal when the discrete speech is generated.

A. Pre-processed Signal

The original incoming speech data, *S(n)*, is first pre-processed using a pre-emphasis filter [3]. The function of this pre-emphasis is to take the focus to the spectral characteristics for compatible signal. The pre-processed signal is then computed as follows:

$$\tilde{S}(n) = S(n) - aS(n-1) \quad (1)$$

Where a is the filter coefficient (Experiment: $a = 0.95$).

The signal using a pre-emphasis filter is divided into short frames, each of which contains approximately 20-40 milliseconds (ms). The short signal in the last frame is discarded.

B. Windowing

The signal in the short-time frame multiplied with a window function, Hamming window [3] is used.

$$h(l+1) = 0.54 - 0.46\cos\left(2\pi\frac{l}{L-1}\right), \quad 0 \le l \le L-1 \qquad (2)$$

Where L is the window length or the total number of samples in a window. As for this experiment, L equals 360.

Each frame is partitioned into overlapping 1/3 of window length for maintaining all significant speech constantly and continuously. The result should be:

$$\hat{S}_j(l) = S'_j(l).h(l), \quad 1 \le l \le L \qquad (3)$$

Speech energy [4] for frame j is obtained by the sum of squares:

$$E_j = \sum_{l=1}^{L} \hat{S}_j(l)^2 \qquad (4)$$

Zero-crossing rate [4] can be defined below.

$$Z_j = \frac{1}{2}\sum_{l=1}^{L} \left| \operatorname{sgn}\left(\hat{S}_j(l+1)\right) - \operatorname{sgn}\left(\hat{S}_j(l)\right) \right| \qquad (5)$$

Where sgn(x) is the sign function,

$$\operatorname{sgn}(x) = \begin{cases} 1 & \text{if} \quad x > 0 \\ 0 & \text{if} \quad x = 0 \\ -1 & \text{if} \quad x < 0 \end{cases} \qquad (6)$$

C. Mel-Frequency Cepstral Coefficients

MFCCs are features which represent the speech signal [5]. The signal sample of each frame uses Fast Fourier Transform (FFT) for calculating spectral vector and offering power spectrums. The power spectrums multiplied with filter bank, which is on Mel-scale, and then obtains Mel power spectrums. Next, Logarithm is computed before operating Discrete Cosine Transform (DCT). Finally, MFCC duly appears as in Figure 1.

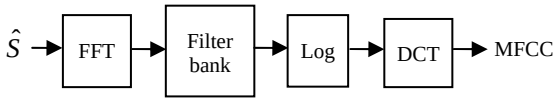


Fig. 1 Block diagram of extracting MFCC

D. Euclidean Distance of Cepstral Coefficients

When assuming that the silent segment is the first 3 frames (approximate length equals 90-100 ms) and the last 3 frames, the average MFCC of silent frame coefficients (C_{mean}) define as follows:

$$C_{mean}(i) = \frac{1}{F}\sum_{s=1}^{F} MFCC_s(i), \quad 1 \le i \le I \qquad (7)$$

Where I is the number of MFCCs within a frame (Experiment: I = 15), F is the amount of silent frames, defining F = 6, and s is the silent frames (the first 3 frames and the last 3 frames).

Thus, Euclidean distance between current frame coefficients and C_{mean} are determined as:

$$U_j = \sum_{i=1}^{I}\left(MFCC_j(i) - C_{mean}(i)\right)^2 \qquad (8)$$

Euclidean distance in [2] presents the difference between each frame coefficients and the average of silent coefficients within cepstrum domain which clearly distinguishes low-energy sounds such as fricatives and silent segment.

III. Proposed Method

A. Difference of Pre-frames and Post-frames

After pre-emphasis process, the signal is divided into frames (approximate length equals 20 ms) and computing difference between pre-frames and post-frames of signal (Delta), D, as below:

$$D_j(l) = \mu\sum_{k=-T}^{T} k\tilde{S}_{j+k}(l), \quad 1 \le l \le L \qquad (9)$$

Where $\tilde{S}$ is the signal after pre-emphasis in frame j, k is the Delta coefficients, T is the number of pre-frames and post-frames (Experiment: T = 2), and the gain of Delta is $\mu = T^{-2}$.

Combining all pre-emphasis signal with Delta, it is defined as follows:

$$S'(m) = \tilde{S}(m) + D(m), \quad m < n \qquad (10)$$

New signal contains the different value between pre-frames and post-frames. Such difference contributes significant speech features and higher energy signal in the beginning and ending endpoint detection.

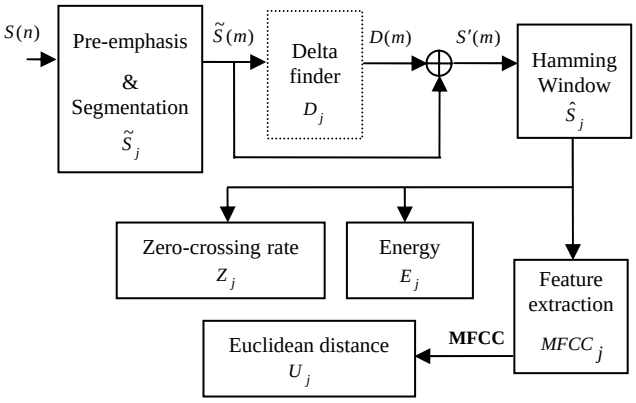


Fig. 2 Block diagram of increasing Delta to signal

B. Adaptive Zero-Crossing Rate

In order to modify the high frequency segment of unvoiced speech equalizing the frequency of voiced speech, the all-frame of zero-crossing rate in (5) are adapted:

$$Z_j = \begin{cases} Z_j & \text{if} \quad Z_j < \lambda \times \text{Median}(\mathbf{Z}) \\ \text{Median}(\mathbf{Z})/\lambda & \text{if} \quad Z_j \ge \lambda \times \text{Median}(\mathbf{Z}) \end{cases} \qquad (11)$$

In (11) $\mathbf{Z}$ is the all-frame of zero-crossing rate and λ = 1.9 by experience.

C. Endpoint Detection

Endpoint detection is to compare the energy of each frame with a threshold. The beginning speech is a frame of higher energy than the threshold whereas a frame of less energy than the threshold is the ending speech. This paper proposes the speech level (**X**) for detecting the endpoints computed from speech energy (**E**), adaptive zero-crossing rate (**Z**), and Euclidean distance (**U**), as follows:

$$X_j = \sqrt{\frac{E_j}{Z_j} \cdot U_j} \tag{12}$$

The threshold, *t*, is defined by:

$$t = w + \alpha\,, \tag{13}$$

$$w = \min\left(\max(\mathbf{X_{begin}}), \max(\mathbf{X_{end}})\right) \tag{14}$$

Where $\mathbf{X_{begin}}$ is **X** of the first 3 frames, $\mathbf{X_{end}}$ is **X** of the last 3 frames and α = 0.1 by experiment.

Obtaining:

$$\mathbf{Y} = \mathbf{X} - t \tag{15}$$

Any frame of **Y**, higher than 0, states the speech segment. Then, both of beginning and ending endpoint detection are formulated:

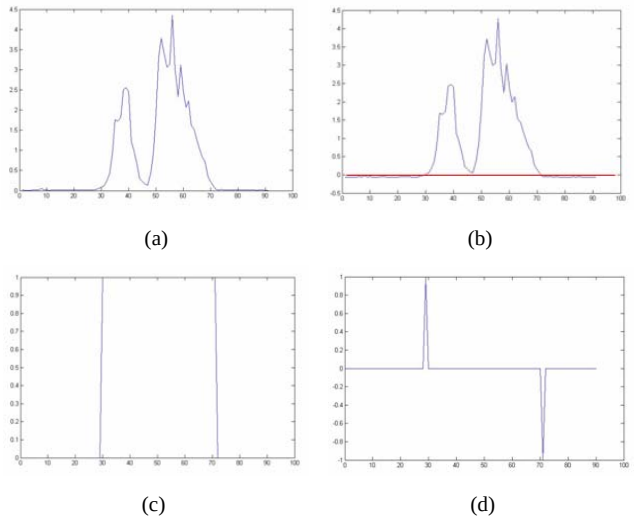
$$P_j = \frac{1}{2}\left(\frac{Y_j}{|Y_j|} + 1\right) \tag{16}$$

$$Q_j = P_{j+1} - P_j \tag{17}$$

Q = 1 (the frame of the beginning speech (Q_{begin}))

Q = -1 (the frame of the ending speech (Q_{end}))

If there is more than one syllable, there might be more than a pair of frame of the beginning and ending endpoints.



(a) (b) (c) (d)

Fig. 3 Endpoint detection process: (a) speech level of **X**, (b) speech level after subtracting threshold of **Y**, (c) speech segment of **P**, (d) beginning and ending endpoint of **Q**

D. Removing Noises

The pair of endpoint detection might be wrong in the case that the noise energy is higher than the threshold. Thus, the noise repeat is removed by identifying another threshold ($\hat{t}$):

$$\hat{t} = \beta.\max(\mathbf{X}) \tag{18}$$

Where β is the threshold factor, the value is 0.2 by experience.

Next, for each pair of endpoint detection, the highest speech level is computed:

$$\hat{Y}(r) = \max\left(\hat{\mathbf{X}}(r)\right) \tag{19}$$

Where $\hat{\mathbf{X}}$ is **X** of frame Q_{begin} to Q_{end} on endpoint pair *r*.

If $\hat{Y}(r) < \hat{t}$ the pair of $Q_{begin}(r)$ and $Q_{end}(r)$ is removed, otherwise the pair of $Q_{begin}(r)$ and $Q_{end}(r)$ remains. In this case, the pair of speech frame appears for detecting MFCC.

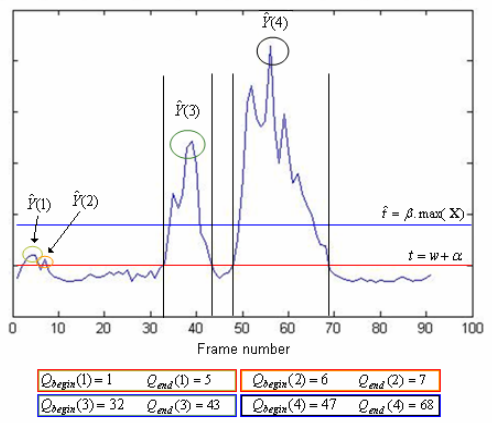


Fig. 4 Histograms of removing noise by speech level

IV. Experiments

A. Database

In this experiment, the speech signal was collected by MATLAB program. A sampling rate of 11.025 KHz 16 bit was used for all data. From the database, we chose 10 speakers, each of whom was recorded on 10 training files and 10 experimental files, making a total of 20 files. Each file includes 7 Thai isolated word of voiced command: /liaw saːj/, /liaw kʰwaː/, /dɤːn naː/, /tʰɔːj laŋ/, /rew kʰɯn/, /tɕʰaː loʔŋ/ and /jut/.

B. Experimental Process

After the speech signal in a set of 10 training files (70 words) per each speaker was detected the endpoints, MFCC matches the speech signal. MFCC of these 10 files is normalized to be an input. These data are then inserted to feed-forward back-propagation neural networks for learning. Neural networks contains 500 nodes for input layer and 3 hidden layers using sigmoid function, whereas 7 nodes of output layer equalizing the number of isolates word of voiced command. After the learning process, we obtain weights and biases which are reasonable with word group for further experiment.

C. Efficient Comparison

This efficient comparison uses the accuracy of neural networks recognition in the set of 10 experimental files (70 words) per speaker. There are three methods in detecting the

endpoint, each of which manipulates the comparison between the signal of increasing Delta and the signal of non-increasing Delta.

1) Method-1: The detection is based on speech energy (**E**).

2) Method-2: The detection is based on speech energy and Euclidean distance (**E** & **U**) [2].

3) Method-3: The detection is based on speech level (**X**).

D. Experimental Results

The recognition system with endpoint detection by using only speech energy recognizes the lowest average of 95.60% while the common speech energy and Euclidean distance in [2] recognizes a high average of 96.37%. On the other hand, the proposed speech level presents the highest average of 96.71%. All three methods will have a higher average of recognition when computing the signal of increasing Delta.

TABLE I
THE EFFICIENCY OF AVERAGE RECOGNITION

Endpoint detection methods	**Accuracy (%)**	
	The signal of non-increasing Delta	**The signal of increasing Delta**
E	95.60	97.40
E & **U**	96.37	97.51
X (proposed method)	96.71	98.49

V. CONCLUSION

The speech energy is an elementary method for endpoint detection of isolated word recognition. To sharply distinguish the speech segment and background noise, Euclidean distance computing from coefficients in cepstrum domain is applied to detect the endpoints. In addition, the signal of increasing difference between pre-frames and post-frames strengthen the high accuracy of endpoint detection (both beginning and ending points which obtain low speech energy) for greater efficient recognition.

REFERENCES

[1] Liang-sheng Huang and Chung-ho Yang, "A Novel Approach to Robust Speech Endpoint Detection in Car Environments", *Proceedings of IEEE ICASSP'00*, Vol. 3, pp.1751-1754, 2000

[2] Sahar E. Bou-Ghazale and Khaled Assaleh, "A Robust Endpoint Detection of Speech for Noisy Environments with Application to Automatic Speech Recognition" *Proceedings of IEEE ICASSP'02*, Vol. 4, pp.3808-3811, 2002

[3] C. Gonzalez-Concejero, V. Rodellar, A. Alvarez-Marquina, E. Martinez de Icaya and P.Gomez-Vilda, "Designing an Independent Speaker Isolated Speech Recognition System on an FPGA", *Research in Microelectronics and Electronics, Ph. D.*, pp.81-84, 2006

[4] Fan Yingle, Li Yi and Wu Chuanyan, "Speech Endpoint Detection Based on Speech Time-Frequency Enhancement and Spectral Entropy", *International Conference of the IEEE-EMBS*, pp.4682-4684, 2006

[5] Avishay Amsalem and Ilan D. Shallom, "Time Frequency Representation for Speech Recognition", *International Conference on ITRE'06*, pp.99-103, 2006

[6] Bo Yin, Eliathamby Ambikairajah and Fang Chen, "Combining Cepstral and Prosodic Features in Language Identification", *International Conference on ICPR'06*, 2006

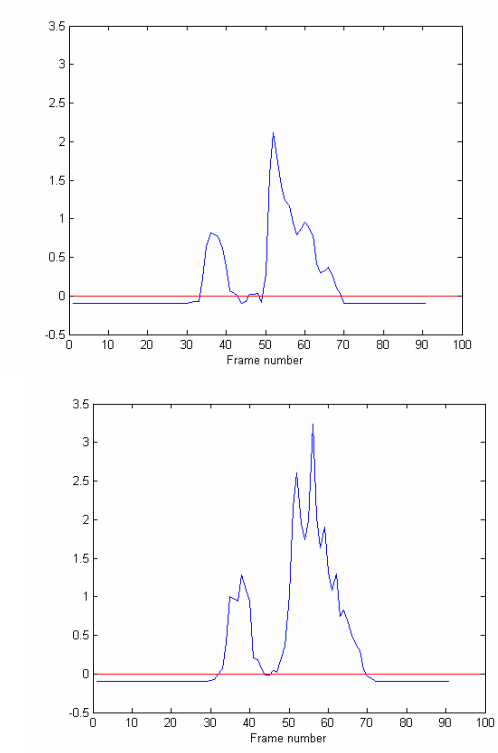


Fig. 5 The speech energy to detecting endpoint of "liaw sa:j". The upper figure shows speech energy without increasing Delta, while the lower figure shows the one with increasing Delta

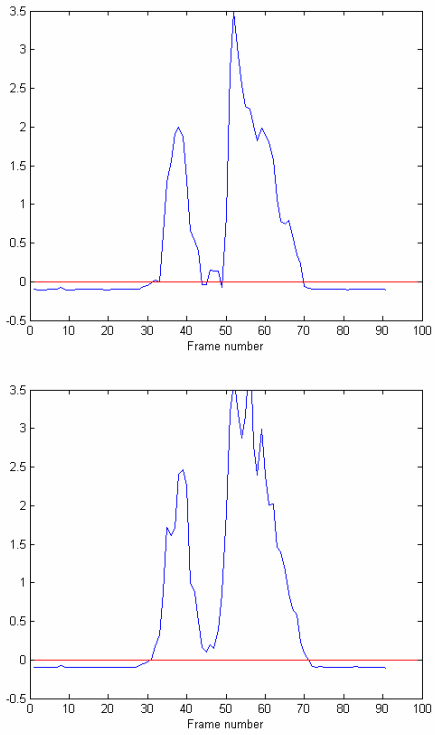


Fig. 6 The speech level to detecting endpoint of "liaw sa:j". The upper figure shows speech level without increasing Delta, while the lower figure shows the one with increasing Delta