Proc. of the 6th WSEAS Int. Conf. on Signal Processing, Computational Geometry & Artificial Vision, Elounda, Greece, August 21-23, 2006 (pp217-221)

Automatic Segmentation and Labeling for Malay Speech Recognition

S.A.R. AL-HADDAD, SALINA ABDUL SAMAD, AINI HUSSEIN, K. A. ISHAK, A.A. AZID, R. GHAFFAR, D. RAMLI, M.R. ZAINAL, *M.K.A. ABDULLAH
Lab Signal Processing
Dept. Electrical, Electronic and System Engineering,
Faculty of Engineering,
National University of Malaysia,
43600 Bangi, Selangor
Malaysia.

*Department Computer and Communication System Engineering
Faculty of Engineering,
Putra University of Malaysia
43400 UPM SERDANG, Selangor
Malaysia.

Abstract: This study is focused on Malay speech recognition with the intention to distinguish speech and non-speech segments. This study proposes an algorithm for automatic segmentation of Malay voiced speech. The calculations of log energy and zero rate crossing are used to process speech samples to accomplish the segmentation. The algorithms are written and compiled using Matlab. The algorithm is tested on speech samples that are recorded in different environment at three different places in Faculty of Engineering, National University Malaysia. As a result almost nearly 90% of the Malay Speech can be segmented.

Key-Words: Speech Recognition, Segmentation, Labeling, Signal Processing, Malay, Endpoint Detection

1 Introduction

Sentence segmentation in speech is very useful especially in speech document indexing, video summarization and speech summarization [1]. But for any system it requires a significant amount of human effort to label the phonetic boundaries [2][3][4]. In many speech segmentation, endpoint detection plays a main role to detect the presence of speech in a background of noise. The beginning and end of a word should be detected by the system that processes the word. The problem of detecting the endpoints would seem to be easily distinguished by human, but it has been found complicated for machine to recognize. Instead in the last three decades, a number of endpoint detection methods have been developed to improve the speed and accuracy of a speech recognition system.

This study uses Malay language which is a branch of the Austronesian (Malayo-Polynesian) language family, spoken as a native language by more than 33,000,000 persons distributed over the Malay Peninsula, Sumatra, Borneo, and the numerous smaller islands of the area, and widely used in Malaysia and Indonesia as a second language [5].

Meanwhile speech recognition (SR) is a technique aimed at converting a speaker's spoken utterance into a text string. SR is still far from a solved problem. It is quoted that in 2003, the best reported word-error rates on English broadcast news and conversational telephone speech were 10% and 20%, respectively [6]. Meanwhile error rates on conversational meeting

speech are about 50% higher, and much more under noisy condition [7].

This study is focused on Malay speech recognition with the intention on endpoint detection to distinguish speech and non-speech segments. This study proposes an algorithm for automatic segmentation of Malay voiced speech. The calculations of log energy and zero rate crossing are used to process speech samples to accomplish the segmentation. The algorithms are written and compiled using Matlab. The algorithm is tested on speech samples that are recorded in different environment at three different places in Faculty of Engineering, National University Malaysia. This paper is segmented in 4 sections: Introduction, material and methods, results and discussion, and conclusions.

2 Material and Methods

In speech signal processing, two basic parameters are Zero Crossing Rate (ZCR) and short time energy. The energy parameter has been used in endpoint detection since the 1970's [8]. By combining with the ZCR, the detection process can be made very accurate [9]. The beginning and ending for each utterance can be detected. Traditional endpoint detection scheme is shown in Fig.1 [10].

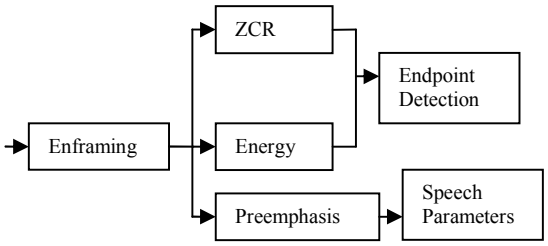

Fig. 1: Traditional Endpoint Detection Scheme.

The measurements of the short time energy can be defined as follows [9]:

a. logarithm energy:

$$E= \sum_{i=1}^{N} \log x(i)^2 \qquad (1)$$

b. sum of square energy:

$$E= \sum_{i=1}^{N} x(i)^2 \qquad (2)$$

c. sum of absolute energy:

$$E= \sum_{i=1}^{N} | x(i)^2 | \qquad (3)$$

As mentioned in the definition above, we write the algorithm E as energy, N is samples in a frame, the frame size is 256, sample rate is 8K, the upper level energy is -10db and lower level energy is -20db.

The flowchart of the experiment is shown on Fig.2. The system begins with recording a WAV file using Logitech USB headset 250 and Sony Ericsson hand phone P800. The wav file is recorded from three different places at Faculty of Engineering, National University of Malaysia. The places are at Signal Processing Laboratory; near an air-condition compressor, and inside the Cafeteria. The word of spoken speech stored in the wav file is "KELMARIN KAMI PERGI KE RUMAH KAWAN KAMI SEPULUH BATU DARI SINI. DI SANA ADA PESTA HARI ULANG TAHUN".

The wav file is played to hear the sound before it is processed. Then ZCR is adjusted to the number of times in a sound sample the amplitude of the sound wave changes sign by getting their mean (y= y-mean(y)). A tolerance of threshold is included in the function that calculates zero crossing which is 10% of maximum ZCR. Next log energy allows us to calculate the amount of energy in a sound at specific instance. For specific window size there are no standard values of energy. Log energy depends on the energy in the signal, which changes depending on how the sound was recorded. In a clean recording of speech the log energy is higher for voiced and zero or close to zero for silent.

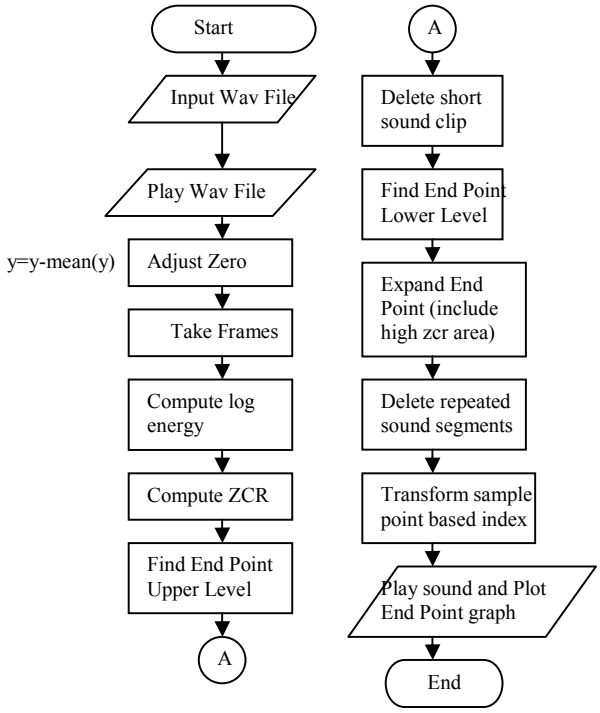

Fig. 2: Flowchart of Automatic Segmentation

3 Result and Discussion

From the methodology discussed above the system managed to show the voiced speech and unvoiced speech (included silence). For voiced speech, energy is high and zero crossing rates are low. On the other hand, for unvoiced speech the energy is low and zero crossing rates are high.

For labeling the segmented speech frame the zero crossing and energy were applied to the frame. Unfortunately it contains some level of background noise and it is difficult also to detect the silent of speech samples at air-condition compressor and inside restaurant due to the fact that energy for breath and surround can quite easily be confused with the energy of a fricative sound [11]. As shown in Fig.3, the voiced speech can be distinguished from unvoiced speech as it has much greater amplitude displacement when the speech is viewed as a wave form.

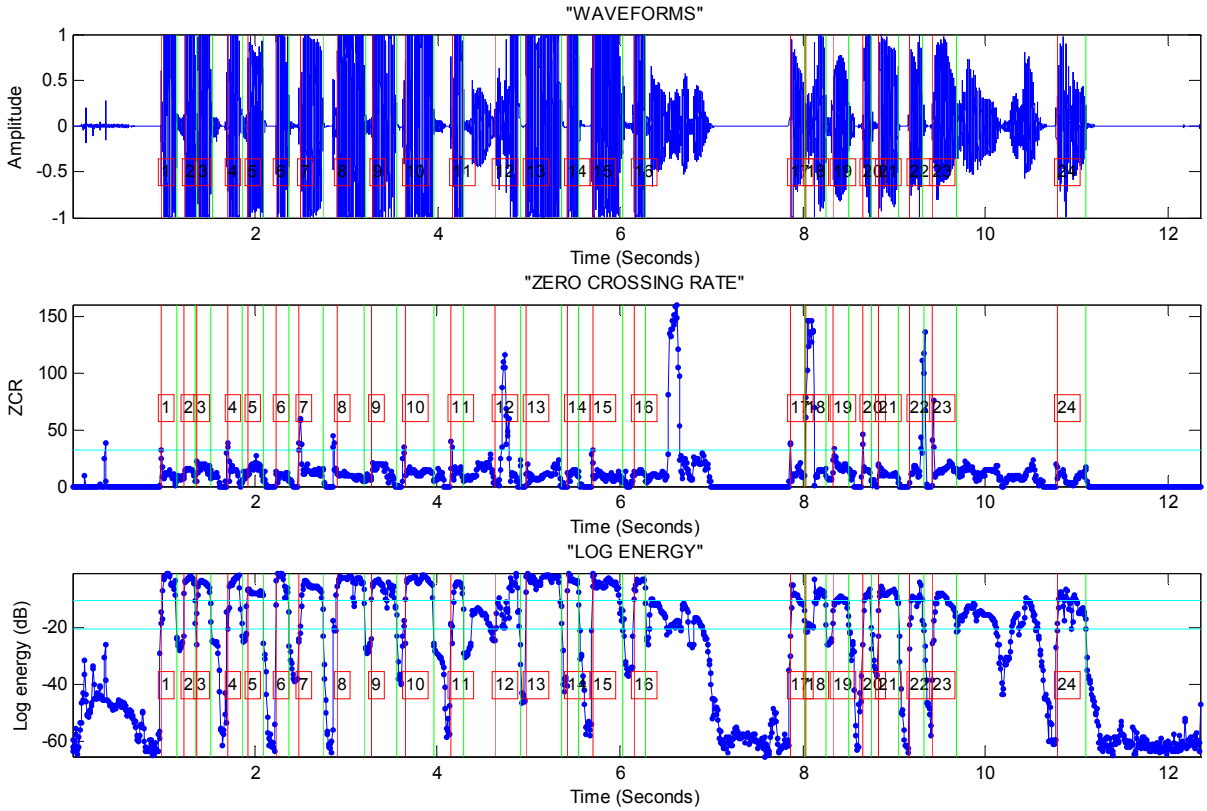

Fig 3: The waveform, zero crossing rate and energy for words spoken in wav file.

As a result, some of the words are read twice as shown in Table 1. This happens for voice recorded near the air-condition compressor and inside the restaurant. But this algorithm performs almost perfect segmentation for voice recoded inside laboratory.

Table 1: Segmentation of wav file recorded at three different places by using Logitech headset 250 and hand phone Sony Ericsson P800.

Segment	**Original**	**Recorded by Using** Logitech USB headset 250			**Recorded by Using** hand phone Sony Ericsson P800		
		Inside Signal Processing Lab	**Near Air-condition Compressor**	**Inside Restaurant**	**Inside Signal Processing Lab**	**Near Air-condition Compressor**	**Inside Restaurant**
1	KELMARIN KAMI PERGI KE RUMAH KAWAN KAMI SEPULUH BATU DARI SINI. DI SANA ADA PESTA HARI ULANG TAHUN	Kel	Kelmarin	Kelmar	Kelm	Kelmarin	Kelmar
2		mar	kam	rin	marin	kami	kelmarin
3		rin	mip	kamip	kamip	mi per	kamip
4		kam	per	per	pergi	mi pergi	per
5		mi	gi	gi	ke rumah	ke rumah	gi
6		per	keru	keru	kawan	mah kawan	ke rumah
7		gi	mah	mah	kami	kami	kawan
8		Keru	kawan	kawan	sepuluh	sepuluh	kami
9		mah	kam	ka	bat	puluh bat	sep
10		kawan	sep	mi	tu	puluh batu dari	puluh
11		kam	puluh	sep	dar	di sana ad	bat
12		misep	bat	puluh	ri	da	tu dari
13		puluh	tu	bat	sini	da pest	tu dari sini
14		bat	dar	tu	di	ta	di sana
15		tu	di sana	dar	sana	ta tah	di sana ad
16		dar	pest	di sana	ad	un ulang	da
17		ri	ta	ad	da	lang tahun	pes
18		san	ulang	da	ha		ta
19		na		pest	ri		har
20		ad		ta ha	ulang		ri ulang
21		da		ri ulang	ta		tah
22		pes		tahun	hun		hun
23		ta					
24		tahun					

For the more noisy places, segmentation problem happens because in some cases the functions produce different values caused by background noise. This causes the cut off for silence to be raised as it may not be quite zero due to noise being interpreted as speech by the functions. On the other hand under clean speech both zero

crossing rate and short term energy should be zero for silent regions.

Furthermore the way people talk, the volume and speed also cause problems to detect the endpoints. This is because zero crossing has a low value for silence and voiced speech, therefore there is more chance of an error between these values, but energy is only high when voiced speech occurs.

4 Conclusion and Suggestion

It can be concluded that algorithm in this research is reasonably accurate for speech recorded in a quiet place and use a good microphone with noise cancellation such as Logitech USB headset 250. However the Malay word segmentation accuracy problem can be improved by focusing on tweaking the cut-off values used by the algorithm to label the different parts of speech especially breathy-voice like "hari."

References:

[1] Zechner, K., 2002. "Summarization of Spoken Language- Challenges, Methods, and Prospects" Technology eZine, Issue 6.

[2] Ljolje A., Hirschberg J. and van Santen J.P.H, 1997. "Automatic Speech Segmentation for Concatenative Inventory Selection", *Progress in Speech Synthesis*, Springer.

[3] Wang H. C., Chiou R. L., Chuang S. K., and Huang Y. F., 1999 ."A phonetic labeling method for MAT database processing" , *Journal of the Chinese Institute of Engineers*, vol. 22, no. 5, 1999, pp. 529-534.

[4] Cosi P., Falavigna D. and Omologo M., 1991. "A Preliminary Statistical Evaluation of Manual and Automatic Segmentation Discrepancies", *Proceedings of European Conference on Speech Communication and Technology*, pp. 693-696.

[5] Britannica, 2006. Encyclopedia Britannica Online, http://www.britannica.com/eb/article-9050292.

[6] Le, A, 2003. "Rich Transcription 2003: Spring speech-to-text transcription evaluation results," Proc. RT03 Workshop, 2003.

[7] Le, A, Fiscus, J., Garofolo, J., Przybocki, M., Martin, A., Sanders, G., and Pallet, D., 2002. "The 2002 NIST RT evaluation speech-to-text results", in Proc. RT02 Workshop, 2002.

[8] Rabiner, L.R. and Sambur, M.R., 1975. "An Algorithm for Determining the Endpoints of Isolated Utterances", Bell Sjjsl. Tech. J., Vol. 54,pp.297-3 1 5.

[9] Rabiner, L.R.and Schafer, R.W., 1978 Digital Processing of Speech Signals, Prentice-Hall Inc.

[10] Analog Devices Inc., 1992. "Digital Signal Processing Applications" using the ADSP-2100 Family Vol. 2, Prentice Hall.

[11] Gold, B., and Morgan, N., 2000. "Speech and Audio Signal Processing", John Wiley and Sons.