© 2015 IJSRSET | Volume 1 | Issue 2 | Print ISSN: 2395-1990 | Online ISSN: 2394-4099
Themed Section: Engineering and Technology

Myanmar Continuous Speech to Isolated Word Segmentation

Taryar Myo Tun, Khin Thida Lynn

University of Computer Studies Mandalay, Myanmar

ABSTRACT

This paper proposes a word segmentation method for Myanmar continuous speech. This system consists of speech processing inclusive of segment boundary detection for isolated words which used zero crossing, duration and energy techniques. Inaccurate segment boundaries are a major cause of errors in automatic speech recognition and a pre-processing stage that segments the speech signal into periods of speech and non-speech is invaluable in improving the recognition accuracy. We propose a combination of three audio features that is energy based voice activity detection, zero crossing rate (ZCR) and duration length for the speech/non-speech detection. Each feature has unique properties to differentiate speech and non-speech segments. We evaluate the results by dividing the speech sample into some segments and used the zero crossing rate, energy based VAD and Myanmar tone length to separate the parts of speech. The algorithm is tested on speech samples that are recorded as sentences of Myanmar speech. The results show that the algorithm managed to segment almost 98.5% of the Myanmar words for all recorded sentences.

Keywords: Myanmar language, Energy based VAD, ZCR, Myanmar Tone Length

I. INTRODUCTION

Language Technology has a potential to play a major role in the process of learning a language. Myanmar language is like Chinese, Japanese, India and Thailand and so on. Myanmar is a kind of tonal languages. This means that all syllables in Myanmar have prosodic features that are an integral part of their pronunciation. Myanmar (Burmese) is official language in Myanmar. Standard Myanmar is based on the dialect spoken in the lower valleys of the Irrawaddy and Chindwin rivers. It is spoken in most of the country with slight regional variations. Typically, Myanmar word consist of a root or stem and zero or more affixes. Words can be combined to form phrases, clauses and sentences. A word consisting of two or more stems joined together is known as a compound word.

Audio signals are generally referred to as signals that are audible to humans. Audio signals usually come from a sound source which vibrates in the audible frequency range. There are many ways to classify audio signals. An audio stream can be segmented into many categories such as silence, environmental sound, music, and speech. Acoustics is a branch of physics that studies sound.

Speech/non-speech detection aims to distinguish the speech segments and non-speech segments of spoken speeches. It is one of the pre-processing steps in sentence boundary detection and is crucial as it affects the accuracy of boundary detection. Speech/non-speech detection is simply the task of discriminating noise-only frames of a signal from its noisy speech frames. In the literature, this process is usually known as voice activity detection (VAD) and it becomes an important problem in many areas of speech processing such as real-time noise reduction for speech enhancement, speech recognition, digital hearing aids, and modern telecommunication.

The organization of this document is as follows. Section 2 explains the Myanmar spoken speeches as the data set and details of methods and algorithms are also presented. Section 3 discusses the results of the experiment and finally, Section 4 concludes the result and limitation with recommendations for further work.

II. METHODS AND MATERIAL

Myanmar language is said to have basically 33 consonants, 12 vowels, other medial and consonant diphthongs. Syllables or words are formed by consonants combining with vowels. However, some syllables can be formed by just consonants, without any vowel. Other characters in the Myanmar script include special characters. Myanmar language has four tones and a simple syllable structure that consists of an initial consonant followed by a vowel with an associate tone. Different tone makes different meanings for syllables with the same structure of phonemes. Each of the processes is explained in the next subsections.

A. Frame Blocking and Windowing

Our ear cannot response to very fast change of speech data content, we normally cut the speech data into sampling and frame segment before analysis. Human frequency range is between 20Hz and 20K Hz. Frames can be overlapped, normally the overlapping region ranges 50% of the frame size. For this speech signal is divided into frames of small duration typically 20 ms with overlap of 10 ms for short-time spectral analysis. Then a short piece of signal is cut out of the whole speech signal. This is done by multiplying the speech samples with hanning window function to cut out a short segment of the speech signal.

B. Prosodic Feature Extraction

Within each frame, we can observe the three most distinct acoustic features. Three acoustic features that are, energy and zero-crossing rate (ZCR) and duration length of Myanmar tones are extracted.

1) Energy: Energy of the speech dataset is one of the features used to classify it to speech/non-speech segments. The energy preceding and succeeding the non-speech segments are also used to detect sentence boundary detection. Energy is very much related to the amplitude. It is a way of representing the amplitude changes in speech signal. . The most common way to calculate the full-band energy of a speech signal is

$$E_j = \frac{1}{N} \cdot \sum_{i=(j-1).N+1}^{j.N} x^2(i)$$

Where, E_j – energy of the j-th frame and x_j is the j-th frame is under consideration.

2) Zero-Crossing Rate: Zero crossing rate (ZCR) is measured based on the number of times the audio signal crosses the zero amplitude line by transition from a positive to negative or vice versa. High frequencies imply high zero crossing rates, and low frequencies imply low zero-crossing rates. If the zero-crossing rate is high, the speech signal is unvoiced, while if the zero-crossing rate is low, the speech signal is voiced.

$$Z_k = \sum_{n=1}^{N-1} |sgn[x_i(n)] - sgn[x_i(n-1)]|$$

where

$$sgn[x_i(n)] = \begin{cases} +1, x_i(n) \geq 0 \\ -1, x_i(n) < 0 \end{cases}$$

The zero crossing value (Z_k) for the k-th segment is computed using Eq.(1), where sgn(x_i(n)) can be three possible value that is +1, 0, -1 depending on whether the sample is positive, zero or negative.

3) Duration length of tones: Myanmar toneme is described with the variety of rate or duration. Length of the tone is defined as rate or duration. Myanmar language has four tones. The lengths of tones are:

- Tone 1 has 18.50 Cs
- Tone 2 has 21.03 Cs
- Tone 3 has 15.44 Cs and
- Tone 4 has 10.35 Cs

Tone 2 is defined as a longest rate and tone 4 is defined as a shortest rate in these four tones.

C. Classification of End Point Detection (VAD)

VAD is a classification problem in which features of the audio signal are used to separate the input speech and non-speech. The end points accuracy is one of the major factors in recognition performance. Voice activity (VAD=1) is declared if the measured values exceed or lower the thresholds. Frame segment speech signal is calculated the short term energy and zero crossing rate for end point detection by VAD. If zero crossing rate is small and energy is high, we define speech signal is voiced. Otherwise Speech signal is unvoiced. If it is not sure for detection, we also calculate the duration length between start point and end point of the signal. Otherwise, no

speech activity or noise, silence (VAD=0) is present. VAD design involves selecting the features, and the way the thresholds are updated for next frame. Both of E_{max} and E_{min} are considered simultaneously with parameter value p for E_{thr}. There is another parameter, mean of energy and standard deviation of energy for each frame, should be also calculated to considering E_{thr}. A ZCR threshold value (ZCR_{thr}) is defined as 0.1 to determine the speech/non-speech segments. The threshold of duration length is also defined as the length of the longest Myanmar's tone.

III. RESULTS AND DISCUSSION

This section provides the experimental results and further discussions of our proposed speech/non-speech detection. Fig. 1 illustrates a flowchart of the speech/non-speech detection processes.

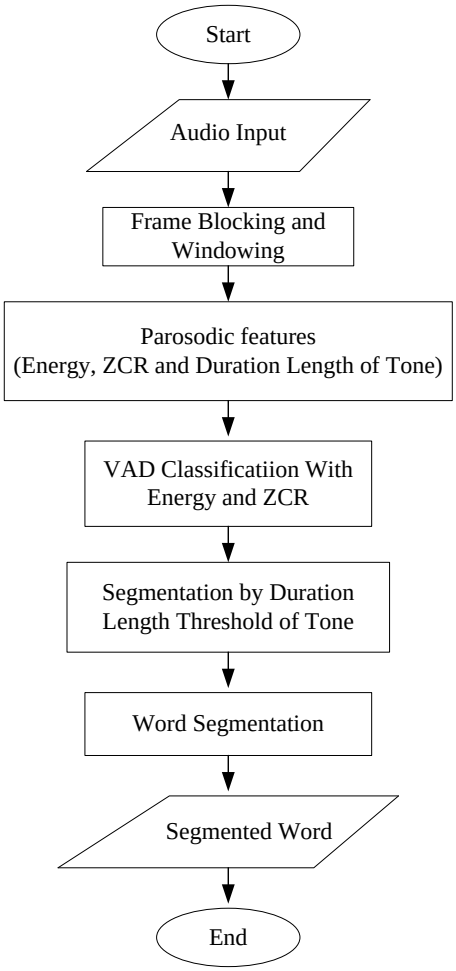

Figure 1: Myanmar word segmentation flowchart

We can see the boundaries of words and the boundaries between continuous speech and even the boundaries between speech and non-speech endings are determined correctly. Two female and one male speaker were recorded reading Myanmar sentences. The signals were ranging from 3-9s in length, and organized into 100 speech files. Each sentence contains five words to eight words. The speech signals were digitized at a sample rate of 44.1k Hz using 16bits and save in wav file. We also recorded 25 kinds of isolated word signals for ten times in clean environment for training database. Every token has sufficient non-speech period at the start and end of utterance. In Fig. 2, the prosodic features of a sample audio data is illustrated. The first plot shows the signal's waveform, followed by silence removal, energy and segmentation boundary.

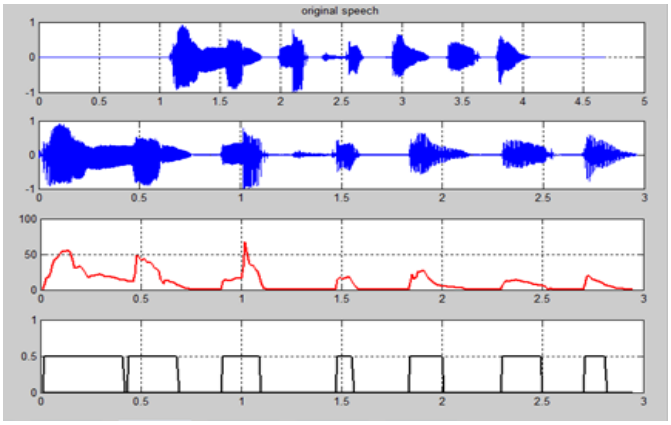

Figure 2 : Myanmar word segmentation result

We evaluate word boundary segmentation performance of the proposed method. For the comparison, we also evaluate the segmentation performance using each feature separately. The accuracy of the syllable segmentation obtained by the method was evaluated using word error measure, WER, which represents the the total number of erroneous boundaries in each sentence to overall number of boundaries in each sentence. Table. 1 shows the WER of the proposed method and other energy threshold methods for word boundary segmentation.

As performance criteria, E_{max} and E_{min} with the various parameter value, mean and standard deviation of energy with various parameter value, are employed. When we evaluate the E_{thr} by using proposed method, mean, standard deviation and duration length, with $0.1<p<0.9$, $p=0.4$ is better segmentation rate than the other methods. Using E_{max} and E_{min} with different parameter values,

TABLE I

TO COMPARISON OF WORD SEGMENTATION RATES OF THE PROPOSED METHOD OTHER METHODS

Energy Threshold	$E_{thr}= (1-p) E_{max} + pE_{min}$			$E_{thr}= E_{min} + p*(E_{max} - E_{min})$		$E_{thr}=$ mean- p*std				
Sentence ID	p=0.45	p=0.6	p=0.9	p=0.1	P=0.05	p=0.1	p=0.2	p=0.4	p=0.6	p=0.9
1	0.83	0.67	0	0	0	0	0	0	0	0.67
2	0.71	0.57	0	0	0	0.4	0.4	0	0	0.71
3	0.67	0.5	0	0	0	0	0	0	0	0.67
4	0.71	0.57	0	0	0	0	0	0	0	0.71
5	0.6	0.6	0	0	0	0.2	0.2	0	0	0.67
6	0.83	0.67	0	0	0	0	0	0	0	0.67
7	0.33	0.57	0	0.16	0	0.17	0.17	0	0	0.67
8	0.57	0.42	0.14	0.14	0	0.14	0.14	0	0	0.71
9	0.86	0.86	0.14	0.14	0	0.14	0.14	0	0.14	0.71
10	0.89	0.67	0.11	0.11	0	0	0	0	0	0.78

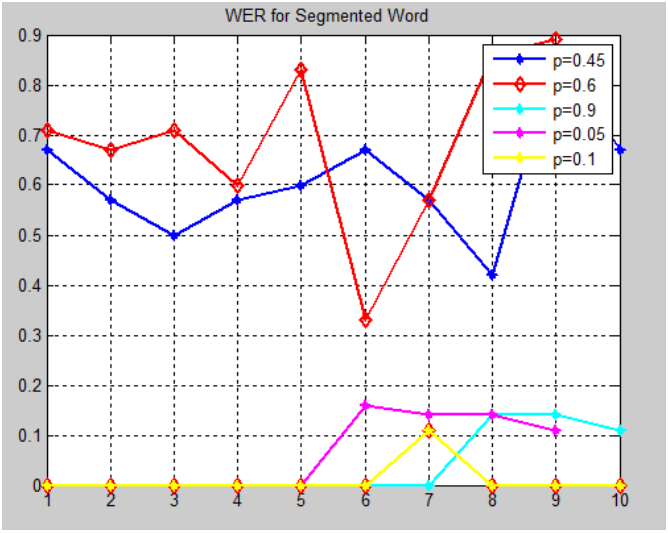

Figure 3 : Different threshold results using E_{max} and E_{min} with different parameter value

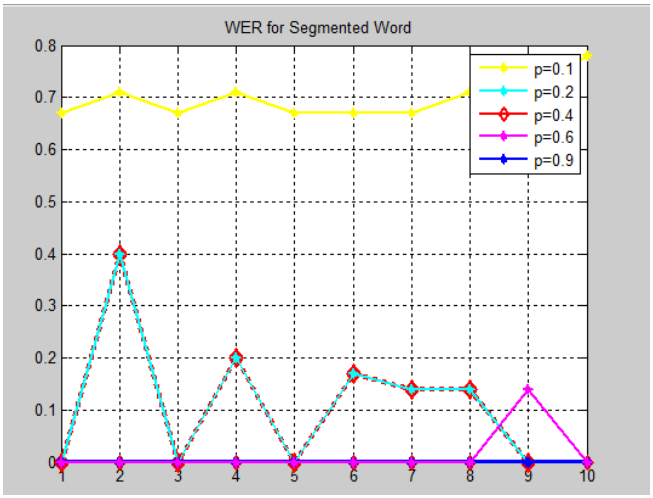

Figure 4 : Different threshold results using mean and standard deviation with different parameter value

speech segmentation accuracy decreases, the parameter value increases. When we evaluate the E_{thr} by using E_{max} and E_{min} with p=0.9, the segmentation accuracy is the better performance than the lower parameter value. Although we evaluate the E thr with E_{max} and E_{min}, we use the other equation for segmentation boundary using various parameters. Word error rate of the proposed method is nearly constant for varying p of 0.1, 0.2, 0.4, 0.6 and 0.9 with their respective values of 0.18%, 0.18%, 0%, 0.09% and 2.81%. If we evaluate the proposed method with p=0.9, all examples speech sentence are divided into two segments only. When the parameter value is increase, the word error rate of segmentation is also increase. Although the parameter value is decrease, the word error rate of segmentation is also increase but it is not the best result. The proposed method achieves the lowest error rates with p=0.4 is the best performance. Word error rate with E_{max} and E_{min} is nearly constant for varying p of 0.45, 0.6 and 0.9 with their respective values of 0.62%, 0.62%, and 0.45%. Word error rate for other equation is nearly constant for varying p of 0.05 and 0.1 with their respective values of 0.015% and 0.06%. Moreover, the proposed method can detect both speech and non-speech frames with least error probabilities for all levels of parameter in clean environment for Myanmar language. If a detected speech segment is shorter than an actual speech segment, phoneme information at the beginning or the end of the utterance is missing, resulting in a speech recognition error. On the other hand, even if a detected speech segment has additional non-speech ranges before and after the speech utterance, it would not cause significant damage to speech recognition performance. We assume

that, if a detected speech segment includes all speech intervals of an utterance without overlapping either the preceding or succeeding speech utterance, it can be a candidate for correct detection.

IV. CONCLUSION

The end points accuracy is one of the major factors in recognition performance. Word boundary detection (SBD), also known as word breaking decides where a word begins and ends. This paper describes isolated word boundary detection using acoustic and prosodic features for Myanmar continuous speech. The proposed method was recorded in varying clean conditions. This system needs to understand which parts of signal will be segmented and how to segment them to isolated word for speech recognition. VAD determines which parts of a voice signal are actual data and which are noise based on short term energy, zero crossing rate and duration length. The recorded speech with appropriate global constraint is to set a valid segment region because the variation of the speech rate of the speaker is considered to be limited in a reasonable range, which means that it can prune the unreasonable segment space. The acoustic front-end of a speech recognizer has been trained on adults' speech to achieve a better performance when speech from children has to be recognized.

This paper can be extended the state of feeling of the person about the sad, happy, angry and other emotional. The phonological rules can be extended to get the complete speech to text system. This system can be extended by using coding method such as Linear Predictive Coding, Mel frequency coefficient coding and Sinusoidal Modelling for speech recognition.

V. REFERENCES

[1] Bachu R.G., Kopparthi S., Adapa B and Barkana B.D, "Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal", Electrical Engineering Department, University of Bridgeport

[2] Mohammad Abushariah, , Raja Ainon, Roziati Zainuddin, Moustafa Elshafei and Othman Khalifa, "Arabic Speaker Independent Continuous Automatic Speech Recognition Based on a Phonetically Rich and Balanced Speech Corpus", The International Arab Journal of Information Technology, Vol. 9, No. 1, January ,2012

[3] Ei Phyu Phyu Soe, Aye Thida, "Diphone-Concatenation Speech Synthesis for Myanmar Language", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 2, Issue 5, May 2013

[4] Moe Pwint and Farook Sattar, "Speech/Nonspeech Detection Using Minimal Walsh Basis Functions", EURASIP Journal on Audio, Speech, and Music Processing, Mark Clements, Volume 2007First Author and Second Author. 2002.

[5] Moe Pwint, Student Member, IEEE and Farook Sattar, Member, IEEE," A Segmentation method for noisy Speech Using Genetic Algorithm", School of Electrical and Electronic Engineering, Nanyang Technological University Nanyang Avenue, Singapore 639798, 0-7803-8874-7/05©2005 IEEE

[6] Won-Ho Shin, Byoung-Soo Lee, Yun-Keun Lee and Jong-Seok Lee," Speech/ Non-speech Classification Using Multiple Features For Robust Endpoint Detection", LG Corporate Institute of Technology

[7] Runshen Cai," An Automatic Syllable Segmentation Method for

[8] Mandarin Speech", Computer Science & Information Engineering College, Tianjin University of Science and Technology, Tianjin, China

[9] F. Pan, N. Ding," Speech De-noising and Syllable Segmentation Based on Fractal Dimension", International Conference on Measuring Technology and Mechatronics Automation(ICMTMA2010), pp. 433—43, IEEE Computer Society(2010)

[10] M. A. Ben Messaoud, A. Bouzid, N. Ellouze," Automatic Segmentation of the Clean Speech Signal", World Academy of Science, Engineering and Technology, International Journal of Electrical, Computer, Electronics and Communication Engineering Vol:9, No:1, 2015