International Journal of Scientific Engineering and Technology Research

ISSN 2319-8885
Vol.03,Issue.24
September-2014,
Pages:4726-4729

www.semargroup.org,
www.ijsetr.com

Automatic Speech Segmentation for Myanmar Language

INGYIN KHAING[1], K ZIN LIN[2]

[1]University of Technology, Yatanapon, Cyber City, Pyin Oo Lwin, Myanmar, E-mail: ingyinkhaing@gmail.com.
[2]University of Computer Studies, Yangon, Bahan Campus, Myanmar, E-mail:kzinlin78@gmail.com.

Abstract: This paper presents simple and novel feature extraction method for segmenting Myanmar continuous speech sentences into words/sub-words. These methods are based on two simple speech features, namely time domain features and frequency features. After the features sequences are extracted, a simple dynamic thresholding criterion is applied in order to detect word boundaries and label the entire speech sentences into a sequence of words/sub-words. All the algorithms used in this work are implemented in Matlab.

Keywords: Automatic Speech Segmentation; Short Time Energy; Spectral Centroid.

I. INTRODUCTION

Segmentation of acoustics signal into words/sub-words is an important stage in ASR system. When building a Myanmar speech database, speech signals should be segmented and labeled. Manual segmentation and labeling, however, is extremely time consuming and tiresome. The process is both laborious and tedious, in that it requires extensive listening and spectrogram interpretation. Therefore automatic procedures for segmenting speech into syllables are implemented in this paper. Speech recognition systems have been developed for isolated words in Myanmar Language. Isolated word recognition, in each word is surrounded by some sort continuous of pause, is much easier than recognizing continuous speech, in which words run speech into each other and have to be segmented. Continuous speech tasks themselves vary greatly in difficulty. This paper is organized as the follows. SectionI is the introduction, SectionII explains about the overview of the standard Myanmar language. SectionIII details background theory. Proposed system design is described in SectionIV. Experimental study is presented in SectionV and the conclusion is summarized in SectionVI.

II. OVERVIEW OF STANDARD MYANMAR LANGUAGE

A. Tones

The most important feature of the Myanmar language is its use of tone to convey the lexical meaning of the syllables. Myanmar tones can be divided into two groups: static and dynamic. The static group consists of three tones (mid, low, and high) whereas the dynamic group consists of two tones (falling and rising).

B. Stress

The syllable in a word produced with a higher degree of respiratory effort is referred to as “stress.” The stressed syllables are usually louder, longer, and higher in pitch than unstressed syllables. The placement of stress on a word in Myanmar is linguistically significant and governed by rules including the monosyllabic word rule and the polysyllabic word rule. For monosyllabic words, all content words are stressed, whereas all grammatical words are unstressed. However, monosyllabic unstressed words when spoken in isolation or emphasized can be stressed as well. For polysyllabic words, stress placements are determined by the number of syllables as well as the structure of the component syllables in the word. The primary stress falls on the final syllable of a word. The secondary stress is determined by the position of the remaining syllables and whether or not they are linker or non-linker syllables.

TABLE I: IPA Code for Myanmar Words

IPA	Words in Myanmar	IPA format
ð		[ʔ əɲ àðá]
h		[hoʊ ʔ]
k		[kòʊ ɴ]
k^h		[k^hòʊ ɴ]
l		[loʊ ʔ]
l̥		[l̥oʊ ʔ]
ə		[k^həloʊ ʔ]
ˋ		[ŋà]
˷		[ŋa̰]

C. Vowels and Consonants

The Myanmar alphabet consists of 33 letters and 12 vowels, and is written from left to right. It requires no spaces between words, although modern writing usually contains spaces after each clause to enhance readability. The latest spelling authority, named the Myanma Salonpaung That pon Kyan (□ □ □ □ □ □ □ □ □ □ □ □ □ □ □

□ □ □ □ □ □ □ □ □ □ □ □),, was compiled in 1978 at the request of the Burmese government. Some of the spoken words and their International Phonetic Alphabet (IPA) format are shown in table (1).

II. BACKGROUND THEORY

The segmentation method described here is a purely bottom-up blind speech segmentation algorithm. The general principle of algorithm is to track the amplitude or spectral changes in the signal by using short-time energy or spectral features and to detect the segment boundaries at the locations where amplitude or spectral changes exceed a minimum threshold level. Two types of features are used for segmenting speech signal: time-domain signal features and frequency-domain signal features.

A. Time-Domain Signal Features

Time-domain features are widely used for speech segment extraction. These features are useful when it is needed to have algorithm with simple implementation and efficient calculation.

1. Short-Time Signal Energy: Short-time energy is the principle and most natural feature that have been used. Physically, energy is a measure of how much signal there is at any one time. Energy is used to discover voiced sound, which have higher energy than silence/unvoiced, in a continuous speech. The energy of a signal is typically calculated on a short-time basic, by windowing the signal at a particular time, squaring the samples and taking the average [1]. The square root of this result is the engineering quantity, known as the root-mean square (RMS) value, also used. The short-time energy function of a speech frame with length N is defined as

$$E_n = \frac{1}{N}\sum_{m=1}^{N}[x(m)w(n-m)]^2 \quad (1)$$

The short-term root mean square (RMS) energy of this frame is given by:

$$E_{n(RMS)} = \sqrt{\frac{1}{N}\sum_{m=1}^{N}[x(m)-w(n-m)]^2} \quad (2)$$

Where x(m) is the discrete-time audio signal and w(m) is rectangle window function.

$$w(m) = \begin{cases} 1 & 0 \le n \le N-1 \\ 0 & Otherwise \end{cases} \quad (3)$$

B. Frequency-Domain Signal Features

The most information of speech is concentrated in 250Hz-6800Hz frequency range [2]. In order to extract frequency-domain features, discrete Fourier transform (that provides information about how much of each frequency is present in a signal) can be used. The Fourier representation of a signal shows the spectral composition of the signal. In this system spectral centroid feature is used.

Spectral Centroid: The spectral centroid is a measure used in digital signal processing to characterize a spectrum. It indicates where the "center of gravity" of the spectrum is. This feature is a measure of the spectral position, with high values corresponding to "brighter" sound [3]. The spectral centroid, SC_i , of the i^{th} frame is defined as the center of "gravity" of its spectrum and it is given by the following equation:

$$SC_i = \frac{\sum_{m=0}^{N-1} f(m)X_i(m)}{\sum_{m=0}^{N-1} X_i} \quad (4)$$

Here, f(m) represents the center frequency of i-th bin with length N and $X_i(m)$ is the amplitude corresponding to that bin in DFT spectrum. The DFT is given by the following equation and can be computed efficiently using a fast Fourier transform (FFT) algorithm [4].

$$X_k = \sum_{n=0}^{N-1} x(n)e^{-j2\pi k\frac{n}{N}}; k = 0,\ldots,N-1 \quad (5)$$

Here, X_k is the DFT coefficients of i^{th} short-term frame with length N, given in equation (4).

III. PROPOSED SYSTEM

The proposed speech segmentation system has five major steps, as shown in fig.1.

A. Speech Acquisition

Speech acquisition is acquiring of continuous Myanmar speech sentences through the microphone. Speech capturing or speech recording is the first step of implementation. Recording has been done by native female speaker of Myanmar. The sampling frequency is 44100 Hz; window length is 512, and both mono and stereo channels are used.

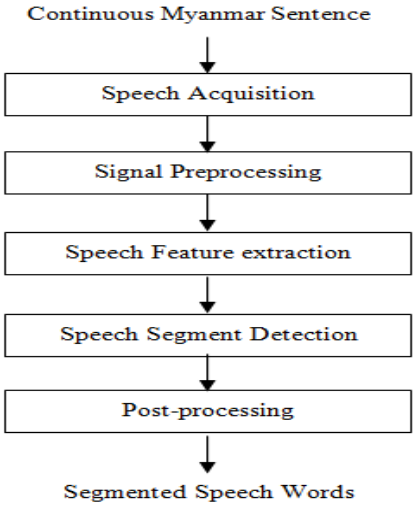

Fig.1. The Proposed Automatic Speech Segmentation System.

B. Signal Pre-pocessing

This step includes elimination of background noise, framing and windowing. Background noise is removed from the data so that only speech samples are the input to the further processing. Continuous speech signal has been separated into a number of segments called frames, also known as framing. After the pre-emphasis, filtered samples have been converted into frames, having frame size of 50

msec. Each frame overlap by half. To reduce the edge effect of each frame segment windowing is done.

C. Speech Feature Extraction

After windowing, compute the short-time energy features and spectral centroid features of each frame of the speech signal. These features have been discussed in detail in Section 3. In this step, median filtering of these feature sequences also computed.

D. Speech Segment Detection

After computing speech feature sequences, a simple dynamic threshold-based algorithm is applied in order to detect the speech word segments.

- Compute the Mean or average values of smoothed feature sequences.
- Find the local maxima of histogram.
- If at least two maxima M_1 and M_2 have been found, then:

Threshold,

$$T = \frac{W * M_1 + M_2}{W + 1} \quad \textbf{(6)}$$

Otherwise,
Threshold,

$$T = \frac{Mean}{2} \quad \textbf{(7)}$$

Where W is a user-defined weight parameter [5], Large values of W obviously lead to threshold values closer to M_1, Here, W=10.

The above process is applied for both feature sequences and finding two thresholds: T_1 based on the energy sequences and T_2 based on the spectral centroid sequences. After computing two thresholds, the speech word segments are formed by successive frames for which the respective feature values are larger than the computed thresholds (for both feature sequences).

E. Post Processing

As a post-processing step, the detected speech segments are lengthened by 5 short term window. Finally, successive segments are merged.

IV. EXPERIMENTAL STUDY

All the techniques and algorithms discussed in this paper have been implemented in Matlab. In this implement, various speech sentences in Myanmar Language have been recorded, analyzed and segmented by using time-domain and frequency-domain features with dynamic threshold technique. Fig.2 shows the filtered short-time energy and spectral centroid features of the Myanmar sentence. Fig.3 shows the segmented sub-word. The percentage of accuracy rate and failure rate of segmentation had been calculated using the following eq:

$$\text{segmentation accuracy} = \frac{\text{No.of correct words segmented by the system}}{\text{No.of words in the sentence}} * 100\% \quad \textbf{(8)}$$

In this experiment, example of 10 spoken sentences was allowed to segments by the system. The system output was word and sub-word. An example of Myanmar continuous sentences are given in Table 2 and the detailed segmentation results were shown in Table 3.

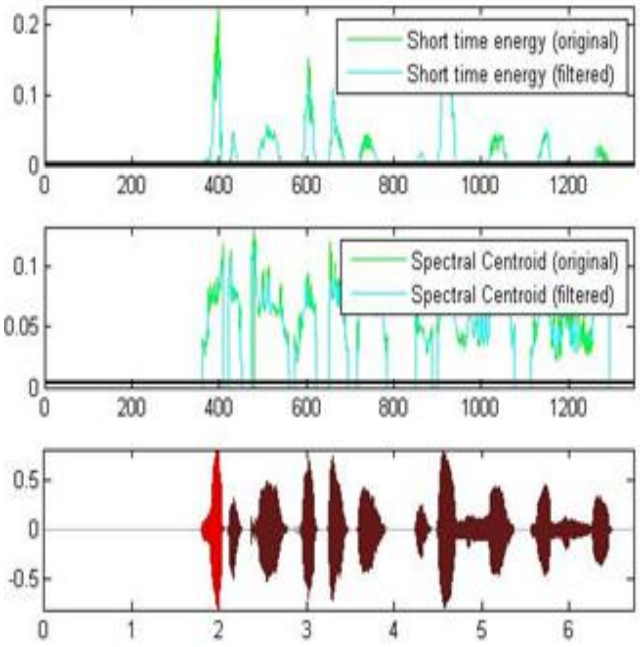

Fig.2. The first sub figure shows the sequence of the signal's energy. In the second sub figure the spectral centroid sequence is presented. In both cases, the respective thresholds are also shown. The third figure presents the whole audio signal. Red color represents the detected speech segments.

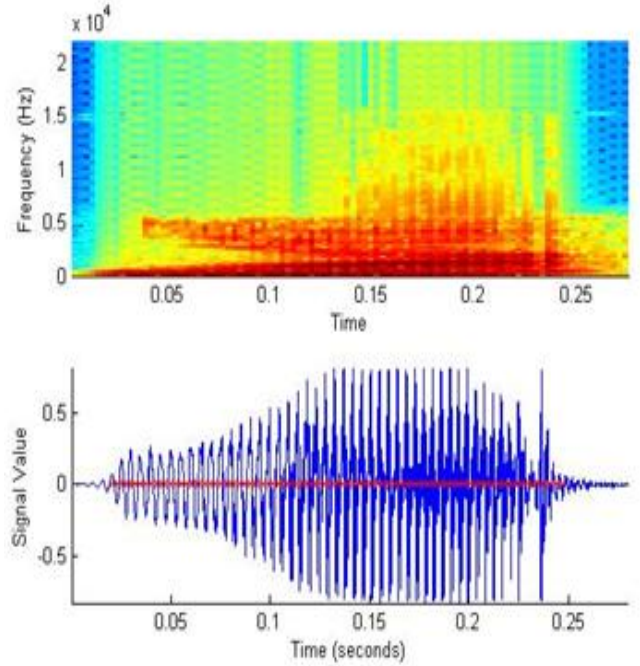

Fig.3. The first sub figure shows the spectrogram of the segment "□" and the second sub figure shows the segmented sub-word "□".

TABLE II: Example of Myanmar Continuous Sentences

Sentence ID	Myanmar continuous sentence
S1	
S2	
S3	
S4	
S5	
S6	
S7	
S8	
S9	
S10	

TABLE III: Segmentation Result

Sentence ID	No. of expected words in the sentence	No. of correct words segmented by the system	Accuracy rate
S1	13	11	84.6%
S2	19	17	89.4%
S3	14	12	85.7%
S4	25	22	88%
S5	14	13	92%
S6	18	16	88.9%
S7	17	16	94.1%
S8	24	20	83.3%
S9	22	20	90.9%
S10	18	15	83.3%
Total	184	162	88%

V. CONCLUSION

It has presented a simple speech features extraction approach for segmenting continuous speech into word/sub-words in a simple and efficient way. From the experiments, it was observed that some of the words were not segmented properly. This is due to different causes: (i) the utterance of words and sub-words differs depending on their position in the sentence, (ii) the pauses between the words or sub-words are not identical in all cases because of the variability of the speech signal and (iii) the non-uniform articulation of speech. Also, the speech signal is very much sensitive to the speaker's properties such as age, sex, and emotion. This reduces the memory requirement and computational time in any automatic speech recognition system.

VI. ACKNOWLEDGMENT

I am very grateful to Dr. K Zin Lin for fruitful discussion during the preparation of this paper and also specially thank to Rector, Professors and colleagues from Technology University (Yatanarpon Cyber City), Myanmar.

VII. REFERENCES

[1] C. T. Zhang and C. J. Kuo, "Hierarchical classification of audio data for acrhiving and retrieving", In International Conference on Acoustics, Speech and Signal Processing, volume VI, pages 3001–3004. IEEE, 1999.

[2] R. Niederjohn and J. Groteluesche n , "The enhancement of speech intelligibility in high noise level by high-pass filtering followed by rapid amplitude compression", IEEE Transactions on Acoustics, Speech and Signal Processing, 24(4), pp277-282, 1976.

[3] T. Giannakopoulos, "Study and application of acoustic information for the detection of harmful content and fusion with visual information" Ph.D. dissertation, Dept.of Informatics and Telecommunications, University of Athens, Greece, 2009.

[4] C.W. James and T.W. John. "An algorithm for the machine calculation of complex Fourier series", Mathematics of Computation: Journal Review, 19: 297–301, 1965.

[5] T. Giannakopoulos, A. Pikrakis and S. Theodoridis "A Novel Efficient Approach for Audio Segmentation", Proceedings of the 19th International Conference on Pattern Recognition (ICPR2008), December 8-11 2008, Tampa, Florida, USA.

Author's Profile:

Ingyin Khaing received Bachelor of Computer Technology from University of Computer Studies, Yangon in Myanmar. She completed the Master Course from University of Computer Studies since 2008 and especially studied and finished the thesis by Digital Signal Processing. She is working now in University of Computer Studies, Maubin as at tutor under Hardware Department. Now, she is a Ph.D student in University of Technology (Yatanarpon Cyber City) near PyinOoLwin, Upper Myanmar. Her fields of interest are Digital Signal Processing, speech recognition.

.