American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS)

ISSN (Print) 2313-4410, ISSN (Online) 2313-4402

© Global Society of Scientific Research and Researchers

http://asrjetsjournal.org/

Fuzzy Logic Based Segmentation for Myanmar Continuous Speech Recognition System

Yin Win Chit[a*], Dr. Renu[b]

[a]Ph.D Researcher, University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Myanmar
[b]Professor, University of Technology (Yatanarpon Cyber City), Pyin Oo Lwin, Myanmar
[a]Email: yinwin.chit@gmail.com
[b]Email: renushi@gmail.com

Abstract

Speech recognition is one of the next generation technologies for human-computer interaction. Automatic Speech Recognition (ASR) is a technology that allows a computer to recognize the words spoken by a person through telephone, microphone or other devices. The various stages of the speech recognition system are pre-processing, segmentation of speech signal, feature extraction of speech and recognition of word. Among many speech recognition systems, continuous speech recognition system is very important and most popular system. This paper proposes the time-domain features and frequency-domain features based on fuzzy knowledge for continuous speech segmentation task via a nonlinear speech analysis. Short-time Energy and Zero-crossing Rate are time-domain features, and Spectral Centroid is frequency-domain feature that the system will calculate in each point of speech signal in order to exploit relevant information for generating the significant segments. Fuzzy Logic technique will be used not only to fuzzify the calculated features into three complementary sets namely: low, middle, high but also to perform a matching phase using a set of fuzzy rules. The output of the Fuzzy Logic are phonemes, syllables and disyllables of Myanmar Language. The result of the system will recognize the continuous words of input speech.

Keywords: Time-domain Features; Frequency-domain Features; Fuzzy Logic; Mel Frequency Cepstral Coefficient; Correlation Coefficient.

----

* Corresponding author.

1. Introduction

Automatic Speech Recognition (ASR) can be defined as the independent, computer-driven transcription of spoken language into readable text in real time. In ASR is technology that allows a computer to identify the words that a person speaks into a microphone or telephone and convert it to readable text. The goal of an ASR system is to accurately and efficiently convert a speech signal into a text of the spoken words of the speaker. There have been many literatures in Automatic Speech Recognition (ASR) system for the major languages of the world such as English, Japanese, Chinese, etc. However, only a few of works have been done in ASR for Myanmar Language. The major difficult in the research process of Myanmar Language ASR is the lack of Myanmar Speech Corpus. Generally, it is not easy to build the speech corpus because it requires a huge amount of speech data and it is very difficult for segmentation of Myanmar Continuous Speech. Speech segmentation is an important step of speech recognition. The aim of this paper is to implement the Myanmar Continuous Speech Recognizer that is capable of recognizing Myanmar Continuous Speech and responding the continuous words. This system will easily recognized the word by using the Fuzzy based speech segmentation. MFCC will be used to extract the significant feature from the segmented speech. Finally the system will recognize the word by using correlation coefficient method.

The remaining part of this paper is organized as follows: section 2 describes about the overview of Myanmar Language, section 3 shows the proposed system design (Training and Testing), section 4 describes the methodological steps of the proposed system, section 5 presents the expected result and the section 6 concludes this paper.

2. Nature of Myanmar Language

Myanmar Language (formerly known as Burmese) is the official Language of Myanmar. The Myanmar script was adapted from the Mon script, and descends from Brahmi script of South India [1]. It is syllabic script and one of the languages which have complex structures and unique. Myanmar words are formed by collection of syllables and each syllable consists of up to seven different sub syllabic elements. Myanmar is written from left to right without any spaces between the syllables or between the words. Nowadays, modern writing sometimes contains space after a sentence in order to enhance readability. 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. 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 [2].

3. Methodological Steps of Proposed System

The continuous speech recognition system will implement in windows environment and it will use MATLAB Tool Kit for developing this application. The proposed speech recognition system has seven major step as follows:

- Speech Acquisition
- Signal Preprocessing
- Feature Extraction for Speech Segmentation
- Speech Segmentation
- Fuzzy Logic for Speech Classification
- Extract MFCC feature from Segmented Speech
- Recognition with Correlation Coefficient

3.1. Speech Acquisition

Speech Acquisition is acquiring of continuous Myanmar Continuous 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 40 kHz and mono channels are used.

3.2. Signal Preprocessing

This step includes eliminating of background noise. Background noise is removed from the data so that only speech samples are the input to the further processing of speech recognition. In this paper, the system uses a noise reduction method based on the least mean square LMS adaptive filter of speech signal. It restores the desired signal by passing the noisy speech through on FIR filter whose coefficients are estimated by minimizing the mean square error (MSE) between the clean signals. In many applications, LMS adaptive filtering algorithm is widely used, partly because they require less calculation and are simple to implement [3].

3.3. Feature Extraction for Speech Segmentation

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 feature.

(i) 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. This system will be used short-time energy and zero-crossing rate as time domain features [4].

- Short-time Energy

Short-time energy is the principle and most natural feature that has been used. Physically, energy is a measure of how much signal are there at any one time. Energy is used to discover voiced sounds, which have higher energy than silence/ unvoiced, in a continuous speech. The energy of a signal is typically calculated on a short-time basis, by windowing the signal at a particular time, squaring the samples and taking the average [5].

- Zero-crossing Rate

The average zero-crossing rate refers to the number of times speech samples change algebraic sign in a given frame. The rate at which zero crossings occur is a simple measure of the frequency content of a signal. It is a measure of number of times in a given time interval/ frame that the amplitude of the speech signals passes through a value of zero. Unvoiced speech components normally have much higher ZCR values than voiced one [6].

(ii) Frequency-Domain Signal Features

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. Widely used frequency-domain features is spectral centroid that used Discrete Fourier transform.

- Spectra 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 [7].

3.4. Speech Segmentation

Speech segmentation is a process of decomposing the speech signal into smaller units. It involves segment identification in continuous speech and processes then to generate distinguishable features. It is used to detect the proper start and end point of segment boundaries. It is very important for various automated speech processing algorithm. There are two types of speech segmentation: manual and automatic segmentation. This system uses the automatic segmentation. After computing speech feature sequences, a simple dynamic threshold-based algorithm is applied in order to detect the speech word segments.

3.5. Fuzzy Logic for Speech Classification

Fuzzy logic system: the matching phase step of our Myanmar speech segmentation is performed by using the fuzzy logic model which consist of a number of conditional “if-then” rules. The system will apply the obtained fuzzy rules with the fuzzy inference system through the four parts which constitute a fuzzy logic system.

- Fuzzification

The calculated features are converted to the fuzzy sets with a corresponding membership degree. The membership function is the most important describe the relationship of the input and the fuzzy sets in the input domain.

- Inference

At this stage, the rules are applied. Each rule, which is composed of several antecedents in the IF statement and one or several consequents in the THEN statement, establishes a relationship between the linguistic values through an IF-THEN statement.

- Aggregation

The complete knowledge about a fuzzy system is contained in its rule-base (set of rules used to describe a system) and the membership functions. The inference methods are used to derive the strength of each rule and the aggregation method is used to determine an overall fuzzy region.

- Defuzzification

The process which produces numerical values after converting the degrees of membership of output linguistic variables.

The input to the matching phase are Spectral Centroid (SC), short-term energy (STE) and zero-crossing rate (ZRC), and the output is the membership degree of phoneme, syllables and disyllable. The input variables are fuzzified into three complementary sets: low, medium, high, and the output variable is fuzzified into three sets: phoneme, syllables and disyllable. Thus, the system obtained for the different coefficients considering the features values:

- STE: low–medium–high

- ZCR: low–medium–high

- SC: low–medium–high

Automatic fuzzy sets and fuzzy rules generation: The fuzzy systems are considered as knowledge-based systems. Through fuzzy inference system and fuzzy membership functions, human knowledge is incorporated into their knowledge. These fuzzy inference system and fuzzy membership functions are generally built by subjective decisions, having a great influence over the performance of the system. In most existing applications, the fuzzy rules are generated by experts in the area, especially for control problems with only a few inputs.

3.6. Extract MFCC Features from Segmented Speech

In this system, MFCC will be used to extract the significant features from each segmented speech phonemes, syllables and disyllables for the recognition stage. Most of today’s automatic speech recognition (ASR) systems are based on some type of Mel-frequency cepstral coefficients (MFCC), which have proved to be effective and robust under various conditions. To enhance the accuracy and efficiency of the extraction process, speech signals are normally pre-processed before feature are extracted. The MFCC technique is often used to create the fingerprint of the sound files. The MFCC are based on the known variation of the human ear's critical bandwidth frequencies with filters spaced linearly at low frequencies and logarithmically at high frequencies used to capture the important characteristics of speech. MFCC can be computed by using the formula [8].

$$\mathrm{Mel}\ (g) = 2595 * \log 10\ (1+g / 700) \qquad (1)$$

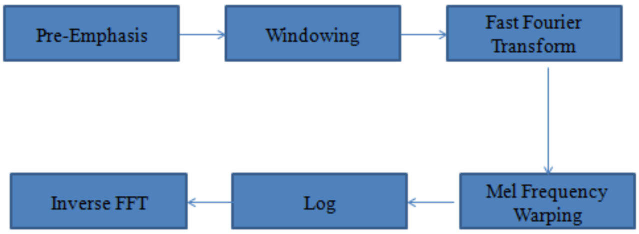

Figure 1: Block Diagram of MFCC

3.7. Recognition with Correlation Coefficient

It is a method which is used to find out the similarity between any two signals. Correlation coefficient is a measure of association between two variables, and it ranges between –1 and 1. The correlation coefficient will be used to determine the relationship between two properties [9]. It can be expressed as:

$$\boldsymbol{r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)}$$

Where X and Y are the voice features to be compared.

Sx and Sy are the standard deviations.

n is the number of data.

For better performance, the correlation coefficients are calculated on the input matrix X whose rows are the frames and columns represent the 20 MFCCs.

4. Experimental Result

In this system, various speech sentences in Myanmar Language will be recorded, analyzed and segmented by using time-domain and frequency-domain features with dynamic thresholding techniques. This system will use external expert knowledge and an automatic generation procedure of fuzzy sets and fuzzy rules then the system will provide the good performance compared to other methods. The best performance will be obtained with fuzzy logic that adding extra expert knowledge by using fuzzy rules improves the speech segmentation with MFCC features. But it requires a long time for the fuzzy rules generation which constitutes a limitation in comparison with non-fuzzy logic method. The accuracy of the syllables segmentation will be obtained by using the method that was evaluated word error measure, WER, which represents the total number of erroneous boundaries in each sentence to overall number of boundaries in each sentence.

5. Conclusion

In this system, a nonlinear approach for the speech analysis will be used for continuous speech segmentation. The system will be developed by using short-time energy, zero-crossing rate as time-domain features and spectral centroid as frequency-domain features at which the system will apply the fuzzy logic. The algorithm provides in output the segments which are phoneme, syllables and disyllables can be integrated in a speech recognition process. Experiments will be performed on Myanmar speech dataset and the results will be shown that the proposed system compare to the non-fuzzy methods. The proposed system will be achieved the better performances but this has some limitations as the high temporal cost of the fuzzy sets and fuzzy rules generation.

Acknowledgements

I am very grateful to Dr. Renu for fruitful discussion during the preparation of this paper and also specially thanks to Rector, Professor and colleagues from Technology University (Yatanarpon Cyper City), Myanmar.

References

[1] T. T. Thet , J. Na and W. K. Ko, “Word Segmentation for the Myanmar Language”, Journal of Information Science, 2008.

[2] T. M. Tun and K. T. Lynn, “Myanmar Continuous Speech to Isolated Word Segmentation”, Engineering and Technology, IJSRSET, Issue 2, Volume 1, 2015.

[3] Haykin, S (2001), Minimum mean square error adaptive filter. In Adaptive Filter Theory, 4th ed. Prentice Hall, Upper Saddle River, 183-228.

[4] M. M. Rahman and M. A. Bhuiyan, “Continuous Bangla Speech Segmentation using short time Speech Features Extraction Approaches.”, IJACSA, Volume 3, No.11, 2012.

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

[6] L R Rabiner and M R Sambur, “An Algorithm for determining the endpoints of Isolated Utterances”, The Bell System Technical Journal, February 1975, pp 298-315.

[7] 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.

[8] Vimala, C., Radha, V., " A review on speech recognition challenges and approaches", World Computer. Sci. Inf. Technol., 2012, 2, (1), pp. 1-7.

[9] Mr. Sridhar Chandramohan Iyer, Speaker Recognition System using Coefficients and Correlation Approaches in MATLAB, IJERT, Vol. 3 Issue 5, May – 2014.