International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE)
Volume 3, Issue 11, November 2014

Review of Speech Segmentation Algorithms for Speech Recognition

M.Kalamani[1], Dr.S.Valarmathy[2], S.Anitha [3] , R.Mohan [4]

Assistant Professor (Sr.G), Electronics and Communication Engineering, Bannari Amman Institute of Technology, Sathyamangalam, India [1]

Professor and Head, Electronics and Communication Engineering, Bannari Amman Institute of Technology, Sathyamangalam, India [2]

PG Students, Electronics and Communication Engineering, Bannari Amman Institute of Technology, Sathyamangalam, India[3,4]

***Abstract* – This paper presents a comparison of segmentation methods. The segmentation methods are time domain features and frequency domain features. The time domain features are short time energy, short time zero crossing rate. The frequency domain features are spectral centroid and spectral flux. After features are extracted, a simple thresholding method is used to detect the word boundaries. The segmentation is used to divide the entire speech sequence into a sequence of words or sub words. Among these methods, spectral centroid has high segmentation accuracy.**

***Index Terms* – Features Extraction, Speech segmentation, Short time zero crossing rate, Spectral flux.**

I. INTRODUCTION

Speech is the most important manner of communication for humans to exchange the information. Speech Recognition is also called as voice recognition which deals with analysis of the linguistic content of a speech signal and its conversion into a computer readable format. Speech recognition can be classified into speaker dependent or independent, isolated or continuous and can be for large vocabulary or small vocabulary. Speech signal can be segmented into words, sub words, syllables and phonemes. Segmentation is the process of decomposing the speech signal into a set of basic phonetic units. Speech segmentation was done using wavelet [1], fuzzy methods [2], artificial neural network [3], hidden markov model [4]. But it was found that results still do not meet expectations. This paper is Continuation of feature extraction for speech segmentation research.

This paper is organized as follows: Section 2 describes techniques for segmentation of the speech signal. In section 3 we will describe different short time speech features.

II. SPEECH SEGMENTATION

Speech segmentation is used to segments continuous speech into uniquely identifiable or phonemes, syllables, words or sub words and processes them to generate distinguishable features. Segmentation is an important role in speech recognition to reduce memory size and computational complexity for large vocabulary systems. Segmentation is used to detect the proper start and end point of speech events.

Automatic speech segmentation can be classified into two types: Blind segmentation and Aided segmentation algorithms. In blind segmentation there is no use of pre existing or external knowledge of linguistic properties. In aided segmentation use some sort of external linguistic knowledge of the speech. Generally there are two kinds of segmentation: Phonemic segmentation and syllable like unit segmentation. Phonemic segmentation segments speech sequence into small phonemes and aided segmentation segments speech into small syllables [5, 6].

ISSN: 2278 – 909X

III. FEATURE EXTRACTION

In feature extraction there is two types of signals are used. One is time domain features and another is frequency domain features. The time domain features are zero crossing rate and short time energy. The frequency domain features are spectral flux and spectral centroid. We will discuss about these two approaches in the following sub sections.

A. Short Time Energy

The energy associated with speech is time varying in nature. Hence the interest for any automatic processing of speech is to know how the energy is varying with time and to be more specific. By the nature of production, the speech signal consist of voiced, unvoiced and silence regions. Further the energy associated with voiced region is large compared to unvoiced region and silence region will not have least or negligible energy.

Thus short term energy can be used for voiced, unvoiced and silence classification of 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 [7].

The short time energy is defined as

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

Where x (m) - discrete-time audio signal
W (m) -Rectangle window

The RMS energy equation is given by

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

The equation of rectangle window is

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

B. Zero Crossing Rate

Zero Crossing Rate gives information about the number of zero-crossings present in a given signal [8]. If the numbers of zero crossings are more in a given signal, then the signal is changing rapidly and accordingly the signal may contain high frequency information. The numbers of zero crossing are less, hence the signal is changing slowly and accordingly the signal may contain low frequency information. Thus ZCR gives indirect information about the frequency content of the signal.

The ZCR is defined as

$$Zn = \frac{1}{2}\sum_{m=1}^{N}\left|\text{sgn}[x(m)] - \text{sgn}[x(m-1)]\right| w(n-m) \quad \text{..(4)}$$

Where $\text{sgn}[x(m)] = \begin{cases} 1 & x(m) \ge 0 \\ -1 & x(m) < 0 \end{cases}$ …… (5)

C. Spectral Centroid

Spectral centroid 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” sounds [9].

The Spectral centroid is given by

$$SC_i = \frac{\sum_{m=0}^{N-1} f(m)X_i(m)}{\sum_{m=0}^{N-1} X_i(m)} \quad \text{…….. (6)}$$

F(m)- Center frequency
Xi (m)-Amplitude of the signal
The DFT is given by the following equation

$$X_k = \sum_{n=0}^{N-1} x(n)e^{-j2\Pi k\frac{n}{N}} \quad \text{; k=0… N-1 …. (7)}$$

D. Spectral Flux

Spectral flux refers to a measure of how quickly the power spectrum of a signal is changing, calculated by comparing the power spectrum for one frame against the power spectrum from the previous frame. The spectral flux can be used to determine the timbre of an audio signal [10].

The Spectral Flux is given by

$$SF_i = \sum_{k=1}^{N/2}\left(\left|X_i(k)\right| - \left|X_i(k-1)\right|\right)^2 \quad \text{……. (8).}$$

Here, X i (k) is the DFT coefficients of i-th frame.

IV. CONCLUSION

In this paper, a review on segmenting the continuous speech into words/sub-words is proposed. In addition, comparisons of the segmentation methods are presented. The short-term speech features have been selected for several reasons. First, it provides a basis for distinguishing voiced speech components from the unvoiced speech components, i.e., if the level of background noise is not very high, the energy of the voiced segments is larger than the energy of the silent or unvoiced segments. Second, if unvoiced segment

ISSN: 2278 – 909X

Simply contain environmental sounds, and then the spectral centroid for the voiced segments is again larger. Third, its change pattern over the time may reveal the rhythm and the periodicity nature of the underlying sound.

This reduces the memory requirement and computational time in any speech recognition system. Among these methods, spectral centroid provides the high segmentation accuracy.

ACKNOWLEDGEMENT

The authors would like to thank friends, reviewers and Editorial staff for their help during preparation of this paper.

REFERENCES

[1] Hioka Y and Namada N, "Voice Activity Detection with array signal processing in the wavelet domain " ,IEICE TRANSACTIONS on Fundamentals of Electronics, Communication and computer sciences,86(11):2802-2811, 2003.

[2] Beritelli F and Casale S, "Robust Voiced/unvoiced classification using fuzzy rules," 1997 IEEE workshop on speech coding for telecommunications proceeding, pages 5-6,1997.

[3] Qi Y and Hunt B," Voiced-unvoiced-silence classification of speech using hybrid features and network classifier ," IEEE Transaction on Speech and Audio Processing,I(2):250-255,1993

.

[4] Basu S ," A Linked-HMM model for robust voicing and speech detection," IEEE international conference on acoustics, speech and signal processing (ICAASSP'03),2003

[5] C. T. Hsieh and J. T. Chien, "Segmentation of continuous speech into phonemic units", Proceedings of International Conference on Information and Systems, 1991, pp. 420-424

[6] R. G. Chen, "Autoatic segmentation techniques for Mandarin speech recognition," M. S. Thesis, Department of Electrical Engineering, National Taiwan University, 1978.

[7] Tong Zhang and Jay C CKuo, "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.

[8] L R Rabiner and M R Sambur, "An Algorithm for determining the endpoints ofIsolated Utterances", The Bell System Technical Journal,February 1975, pp 298-315.

[9] 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.

[10] Bello J P, Daudet L, Abdallah S, Duxbury C, Davies M, and Sandler MB, "A Tutorial on Onset Detection in Music Signals", IEEE Transactions on Speech and Audio Processing 13(5), pp 1035–1047,2005.

ISSN: 2278 – 909X