International Journal on Applications of Information and Communication Engineering
Volume 1: Issue 1: January 2015 , Pages:39-46

Hybrid Speech Segmentation Algorithm for Continuous Speech Recognition

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

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 Student, Electronics and Communication Engineering, Bannari Amman Institute of Technology, Sathyamangalam, India[3]

Abstract – **In this a hybrid segmentation methods are proposed for Automatic Speech Recognition. The segmentation methods are based on 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 continuous speech into a sequence of words or sub words. In this project, short time energy and spectral centroid are combined. Among these methods, hybrid of short time energy and spectral centroid has high segmentation accuracy. The results shown to be computationally efficient for real time applications and it perform better than conventional methods.**

Keywords – Short time energy, 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 [1]. Segmentation is the process of decomposing the speech signal into a set of basic phonetic units. Speech segmentation was done using wavelet, fuzzy methods, artificial neural network, hidden markov model. But it was found that results still do not meet expectations. This paper is Continuation of feature extraction for speech segmentation research.

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 [2]. 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 [4]. 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].

This paper is organized as follows: Section 2 describes techniques for segmentation of the speech signal. Section 3 describes segments detection of speech signal. Section 4 describes the proposed method. In section 5 describes the performance measures. Section 6and 7 describes the results and conclusion.

II. SPEECH SEGMENTATION

In speech segmentation two types of methods 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 [3].

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

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

Where, x (m) is a discrete-time audio signal and w (m) is a rectangle window

The RMS energy equation is given by

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

The equation of rectangle window is

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

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

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

..(4)

Where,

$$\mathrm{sgn}[x(m)] = \begin{cases} 1 & x(m) \ge 0 \\ -1 & x(m) < 0 \end{cases} \quad \ldots\ldots (5)$$

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

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

f(m) is a Center frequency, X_i (m) is a amplitude of the signal

The DFT is given by the following equation

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

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

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

Here, X_i (k) is the DFT coefficients of i^{th} frame.

III. SEGMENTS DETECTION

A simple dynamic based threshold method is used to detect the speech segments. The following steps are present in these thresholding methods.

1. Get the feature sequence from the previous feature extraction module.
2. Apply median filtering to smooth the feature sequences.
3. Compute the Mean or average values of these sequences.
4. Find the threshold value.

Threshold

$$T = \frac{Mean}{2} \quad \ldots\ldots\ldots (9)$$

where,

W- weight parameter.

Here, the both short time energy and spectral centroid the above steps are applied to find the threshold value.

The two threshold values are T1 and T2. T1 is threshold value for energy and T2 is threshold value for spectral centroid. Based on these two threshold values speech segment is detected.

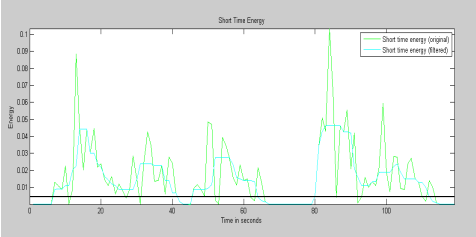


Figure1. Original and filtered signal of short time energy.

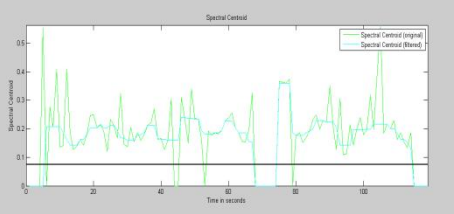


Figure2. Original and filtered signal of spectral
Centroid

In figures 2 & 3 shows the details of the original speech signal of short time energy and spectral centroid and how it will be after the preprocessed signal. This preprocessed stage will makes the signal in standard format which leads at increasing the segmentation accuracy rate. In filtered output DC component is removed and it gives standardized signal.

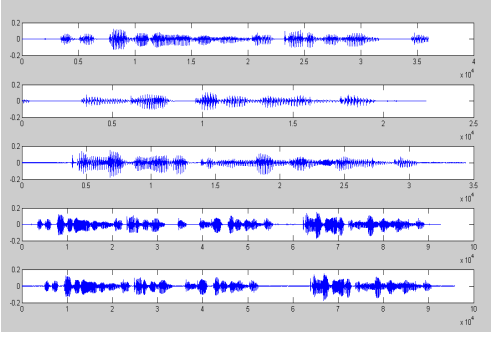


Figure3. Time Domain results for short time energy and spectral centroid.

In this figure3. Indicates the time domain results of short time energy and spectral centroid. It shows the segmented output of the input signal

IV. PROPOSED METHOD

In this paper we have experimented to design the speech segmentation using the combination of time domain feature and frequency domain feature and segmented into syllables or sub words. The proposed method has five major steps [11].

1. Speech Acquisition
2. Signal Preprocessing
3. Speech Segmentation
4. Dynamic Thresholding.
5. Speech Segments Detection

1) Speech Acquisition

Speech acquisition is acquiring of continuous speech sentence s through the microphone. Speech recording is the first step of implementation.

2) Signal Preprocessing

Preprocessing is elimination of back ground noise, framing and windowing. Back ground noise is removed from the data. Continuous speech has been separated into frames. That method is known as framing. Windowing is used to determine the portion of the speech signal.

3) Speech Segmentation

In this section we have been computed the hybrid of short time energy and spectral centroid of each frame of the speech signal.

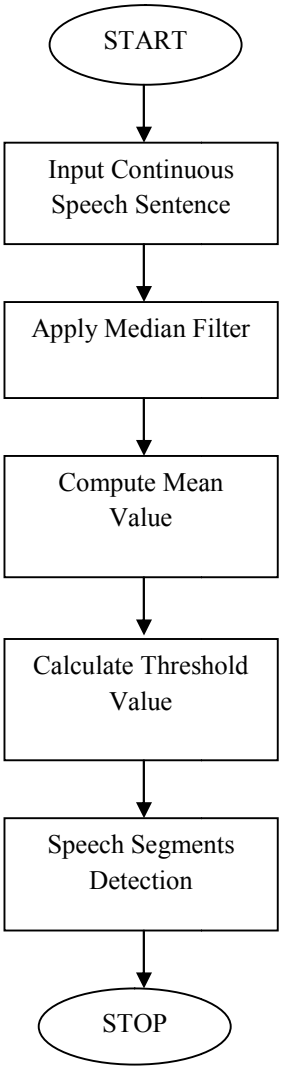
4) Dynamic Thresholding

This method is used to find the threshold values. The two threshold values are T1 and T2. After computing two thresholds, the speech word segments are formed by successive frames for which the respective feature values are larger than the computed threshold values.

5) Speech Segments Detection

A simple dynamic based threshold method is used to detect the speech segments.

FLOW CHART OF THE PROPOSED METHOD



V. PERFORMANCE MEASURES

Percentage of correctness regarding extraction of voiced sample from a speech signal is defined as follows:

Hit rate is defined as number of correctly recognized words. It is the ratio of number of correctly identified words to the total number of words in the sentence.

$$\text{Hit Rate} = \frac{\text{No. of correctly identified words}}{\text{Total no. of words}}$$

False Alarm rate is defined as number of words incorrectly recognized. It is the ratio of number of incorrectly recognized words to the total number of words in the sentence.

$$\text{False Alarm Rate} = \frac{\text{No. of erroneous word identified}}{\text{Total no. of words}}$$

VI. RESULTS AND DISCUSSION

The proposed technique has been implemented in Matlab 2013a.Various speaker speech in Tamil language have been recorded and segmented. Proposed method has been implemented and analyzed. Table 1 summarizes the results showing percentage of correctness in detection of word boundaries languages.

Table .1 Comparison of no. of identified words for existing and proposed methods

Si.no	Sentences	Total no. of words	Total no. of identified words			
			STE	SC	SF	STE&SC
1.	Tamil 1	3	3	2	1	3
2.	Tamil 2	5	4	5	1	4
3.	Tamil 3	5	5	5	1	5
4.	Tamil 4	6	5	5	1	5
5.	Tamil 5	3	2	3	1	3
6.	Tamil 6	5	5	1	1	5
7.	Tamil 7	5	5	5	1	5
8.	Tamil 8	3	2	3	3	3
9.	Tamil 9	5	5	5	1	5
10.	Tamil 10	4	3	3	1	4
	Total	**44**	**39**	**37**	**12**	**43**

In table.1 shows the details no. of identified words results for existing and proposed method. The existing method is SF, SC and STE. The proposed method is hybrid of STE and SC.

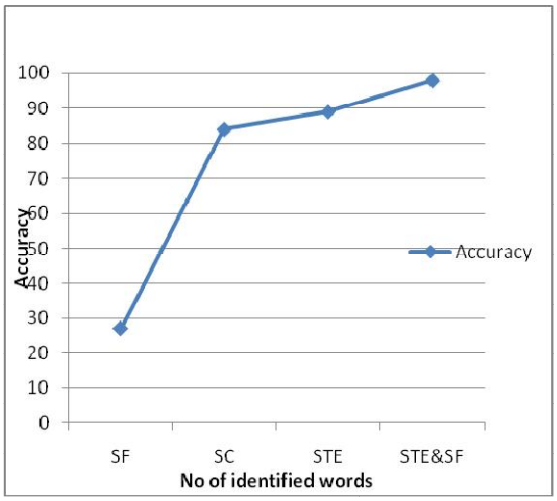


Figure4. Comparisons of Segmentation Algorithms

The above line chart gives the comparison of various segmentation methods. Accuracy of four segmentation methods is compared and it shows that the hybrid of short time energy and spectral centroid has high segmentation accuracy. Spectral flux has low segmentation accuracy. Accuracy is exponentially increased.

Table.2 Results for Hit Rate

Hit Rate (%)					
Sentences	Total no of words	SF	SC	STE	STE&SC
Tamil 1	3	33.33	66.67	100	100
Tamil 2	5	20	100	80	100
Tamil 3	5	20	100	100	100
Tamil 4	6	16.6	83.33	83.33	83.33
Tamil 5	3	33.33	100	66.67	100
Tamil 6	5	20	20	100	100
Tamil 7	5	20	100	100	100
Tamil 8	3	100	100	66.67	100
Tamil 9	5	20	100	100	100
Tamil 10	4	25	75	75	100
	Average Hit Rate (%)	**30.82**	**84.50**	**87.16**	**98.33**

In table.2 showing the details Hit rate for existing and proposed method. The existing method is SF, SC and STE. The proposed method is hybrid of STE and SC. From the above tabulation is observed that the average hit rate of the proposed method is increased by 11.17 % when compared with existing method STE.

Table.3 Results for False Alarm Rate

False Alarm Rate (%)					
Sentences	Total no of words	SF	SC	STE	STE&SC
Tamil 1	3	66.67	33.33	0	0
Tamil 2	5	80	0	20	0
Tamil 3	5	80	0	0	0
Tamil 4	6	83.33	16.67	16.67	16.67
Tamil 5	3	66.67	0	33.33	0
Tamil 6	5	80	80	0	0
Tamil 7	5	80	0	0	0
Tamil 8	3	0	0	33.33	0
Tamil 9	5	80	0	0	0
Tamil 10	4	75	25	25	0
	Average False Alarm Rate (%)	**69.16**	**15.5**	**12.83**	**1.66**

In table.3 showing the details False Alarm rate for existing and proposed method. The existing method is SF, SC and STE. The proposed method is hybrid of STE and SC. From the above tabulation is observed that the average False Alarm rate of the proposed method is decreased by 11.17 % when compared with existing method STE.

VII. CONCLUSION

In this paper, a hybrid of short time energy and spectral centroid are proposed and comparisons are made between various segmentation methods. It performs the segmentation of word and sentence boundaries automatically. The proposed method results better performance of word detection. The Average Hit rate of proposed method is increased from the range of 11.17% to 67.51% compared to existing methods. The Average False alarm rate of proposed method is decreased from the range of 67.5% to 11.17% compared to existing methods. This method increases the accuracy rate and decreases the error rate. This reduces the memory requirement and computational time in any speech recognition system. Among these methods, hybrid of short time energy and spectral centroid provides the high segmentation accuracy. The hybrid of speech segmentation algorithm achieved segmentation accuracy of 98.33% and error rate is 1.67%.

ACKNOWLEDGEMENT

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

REFERENCES

[1] *Md.Mijanur Rahman and Dr. Md. Al-AminBhuiya "Continuous Bangla Speech Segmentation using Short-term Speech Features Extraction Approaches" International Journal of Advanced Computer Science and Applications, Vol. 3, No. 11, 2012*

[2] *J. Sangeetha and S. Jothilakshmi " Robust Automatic Continuous Speech Segmentation for Indian Languages to Improve Speech to Speech Translation" International Journal of Computer Applications (0975 – 8887) Volume 53– No.15, September 2012*

[3] *Hemakumar G, Punitha P " Automatic Segmentation of Kannada Speech Signal into*

Syllables and Sub-words: Noised and Noiseless Signals " International Journal of Scientific & Engineering Research, Volume 5, Issue 1, January-2014

[4] *Md. Mijanur Rahman, Md. Farukuzzaman Khan and Mohammad Ali Moni "speech recognition front-end for segmenting and clustering continuous bangla speech" daffodil international university journal of science and technology, volume 5, issue 1, January 2010*

[5] *Nipa Chowdhury, and Md. Abdus Sattar "Separating Words from Continuous Bangla Speech "Global Journal of Computer Science and Technology.*

[6] *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*

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

[8] *Tong Zhang and Jay C CKuo, "Hierarchical classification of audio data for archiving and retrieving", In IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 6, pp. 3001–3004, 1999.*

[9] *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.*

[10] *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.*

[11] *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.*