Comparative Study of Endpoint Detection Algorithms Suitable for Isolated Word Recognition

A. Akila[1] and **E. Chandra**[2]

Submitted in December, 2013; Accepted in September, 2014

Abstract - Voice Activity Detections (VAD) are used all over the speech processing applications such as speech recognition, speech enhancement etc. In Isolated word speech recognition, the end point detection reduces the computational process. In this paper, a comparative study of three VAD algorithms and the algorithms were analyzed using performance evaluation criteria. The algorithm suitable for the dataset used in the proposed research work is found using the performance criteria like misdetection, speech quality and compression.

Index Terms – Frequency domain, Short Term Energy (STE), Voice Activity Detection (VAD), Zero Crossing Rate (ZCR).

NOMENCLATURE
VAD – Voice Activity Detection
STE – Short Term Energy
ZCR – Zero Crossing Rate
ASR – Automatic Speech Recognition

1.0 INTRODUCTION

Automatic Speech Recognition (ASR) is a wide area in signal processing where the recognition of the utterance is done. The accuracy of recognition will improve if the input utterance contains only of speech after removing silence from the speech (i.e.) the accuracy will increase by the accurate end point detection. The speech can be classified as silence, voiced and unvoiced. The process of separating the speech segments of an utterance from the background noise is called the End point detection [1]. The end point detection is used for segmenting the input utterance into its subunits also. There are many end point detection algorithms developed. In Isolated word ASR, the detection of endpoints in a speech is done to separate the speech signal from unwanted background noise [2]. This process is called Voice Activity Detection. In isolated word automatic speech recognition, the detection of endpoints in a speech has been done to separate the speech signal from unwanted background noise. The main use of endpoint detection is in speech coding and speech recognition. It is an important enabling technology for a variety of speech based applications. The proposed research work is a Tamil speech recognition system which performs segmentation of the given speech data into syllables and recognize the syllable. The speech data has to undergo the preprocessing step of endpoint detection to improve the performance of the speech recognition system.

[1]*Department of Computer Science, D.J Academy for Managerial Excellence, Coimbatore, India,*
[2]*Department of Computer Science, Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore-32, India.*
E-mail: [1]*akila.ganesh.a@gmail.com and*
[2]*crcspeech@gmail.com*

2.0 VOICE ACTIVITY DETECTION

VAD is a technique used in speech processing in which the presence or absence of human speech is detected. It is also known as speech activity detection or speech detection. The main uses of VAD are in speech coding and speech recognition [3]. It is usually language independent. VAD algorithm is used as first step in speech recognition system.

2.1 Characteristics of VAD

- Reliability – the endpoints computed should be correct taking into consideration the weak fricatives.
- Robustness – Suitable for any type of application
- Computation of end points should be accurate
- Adapting to non stationary background noise should be good.
- Simplicity- easy to compute
- Real Time Processing
- No prior knowledge of noise

The essential characteristics are simplicity and robustness [4].

2.2. Features used in VAD Algorithm

- Short term energy
- Zero Crossing Rate
- Autocorrelation function based Features
- Spectrum based Features
- Power in band limited region
- Mel Frequency Cepstral Coefficients
- Delta Line Spectral Frequencies
- Features based on higher order statistics

The use of multiple features leads to more robustness against different environment. Most of the VAD Algorithms use Short Term Energy and Zero Crossing Rate features because of their simplicity.

2.2.1 Short Term Energy

The amplitude of the speech signal varies with time. As in Fig 1, the amplitude of unvoiced segments is generally much lower than that of voiced segments. The amount of energy carried by a wave is related to the amplitude of the wave.

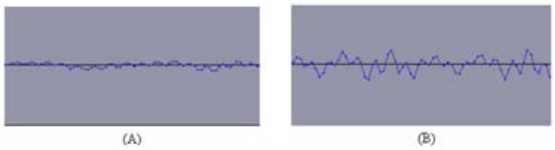

Figure1: A sample wave signal representing (A) the amplitudes of unvoiced segment (B) the amplitude of voiced segment

A high energy wave is characterized by high amplitude; a low energy wave is characterized by low amplitude. The energy of a segment indicates the presence of voice data. The energy (E) transported by wave is directly proportional to the square of the amplitude (A) of the wave which is specified in (1).

$$E \propto A^2 \quad (1)$$

The short term energy can be calculated using the formula[1] as specified in the (2)

$$E = \sum_{n=1}^{N} |S(n)^2| \quad (2)$$

where s(n) is the amplitude of each frame and n is the current frame of the N frames in the signal.

2.2.2 Zero Crossing Rate

Zero crossing is a commonly used term in electronics, mathematics, and signal processing which refers to a point where the sign of a signal changes by crossing of the axis. Fig 2 shows some of the zero crossing point of a sample wave signal.

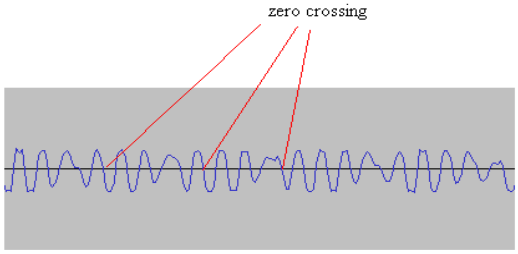

Figure 2: Zero Crossing points of a sample wave signal

The rate of sign changes along a signal is called Zero Crossing Rate. It is simple measure of the frequency content of a signal. It is a measure of the number of times in a given signal, the amplitude passes through a value zero. The zero crossing rate [1] is calculated using the (3)

$$ZCR = \frac{1}{len-1} \sum_{n=1}^{len-1} \left| \frac{sgn(s(n)) - sgn(s(n-1))}{2} \right| \quad (3)$$

where sgn(s(n))=+1 if s(n) >0 and sgn(s(n))=-1 if s(n)<0, s(n) is the amplitude of current frame, n is the current frame and len is the number of frames in the signal.

3.0 REVIEW OF END POINT DETECTION ALGORITHMS

There are many algorithms to find the end point of the input signal. The three algorithms discussed below are suitable for end point detection of isolated word.

3.1 Rabiner's Endpoint Detection Algorithm

The algorithm proposed in [5] uses the two basic features zero crossing rate and short term energy. The uttered speech signal is divided into n frames each of 80ms length. The first 10 frames are considered for calculating the threshold value. The threshold value is used to find the initial point and the endpoint of a speech signal. The algorithm can be used in any environment with a signal to noise ratio of at least 30dB.

3.2 Qiang He Algorithm

The VAD Algorithm was developed by Qiang He in the year 2001. It was implemented using MATLAB. The code of the algorithm is available as a free software. The signal is split into overlapping frames of length 80ms. The Short Term energy (STE) and the Zero Crossing Rate (ZCR) are calculated for each frame using equation 2 and 3 respectively. Then ZCR and STE are compared with the threshold values which are chosen between 2 and 10. The comparison result specifies whether the segment or frame is a silent, Voiced or noise signal. If the STE and ZCR are within the threshold values, the frame is a voiced signal or noise signal. If they are below the initial threshold values then the frame contains silent signal. To differentiate between noise and voiced frames, a count of frames which has voice or noise is manipulated. If the total frame length of the counted frames is below 150 ms, then the signal is noise else it is a voiced signal.

3.3 VAD with Frequency Domain Approach

This algorithm takes its decisions based on energy comparisons of the signal frame with a reference energy threshold in the frequency domain. The frequency domain (F) of the frame is obtained by (4) where FFT is Fast Fourier Transform function. The signal is divided into frames of length 80ms.

$$F(f_j) = FFT\{f_j\} \quad (4)$$

The Frequency domain is used to find whether the frame is Active or Inactive by comparing with the threshold value. The initial point is the frame where the current frame is Active and the predecessor is Inactive. Similarly the endpoint is frame where current frame is Inactive and the predecessor is Active [6].

4.0 EXPERIMENTAL RESULTS

4.1 Dataset

The signals that are used for comparison of the three different End point detection Algorithms were 6 simple words of Tamil language which are equivalent to yes and no in English language. The words were recorded for 3 seconds with a frequency of 8 KHz.

The utterances tested were drawn from a single female speaker. For recording the speech, Audacity was used. The sounds were recorded in a normal room where environment may not be quiet. Reliability of endpoint detection Algorithms will be more when environment condition is quiet. But it is not always practical.

4.2 Computation

MATLAB environment was used to test the algorithms on the 6 signals. The Initial and the endpoint of voiced segments for each signal are computed. Table 1 list the initial and end points of each signal computed using the three algorithms.

4.3 Criteria for assessing the end point detection algorithms

Performance of the algorithms was analyzed based upon the following criteria[7].

Subjective speech quality: The quality of the samples was rated on a scale of 1(poor) to 5(best). The initial signal is assumed to have the best quality of rating 5. The speech samples after end point detection were played and rated.

- Compression Ratio : The ratio of total inactive frames detected to the total number of frames formed
- Misdetection: The number of frames which have speech content, but were classified as inactive and number of frames without speech content but classified as active are counted. The ratio of this count o the total number of frames in the signal is taken as misdetection ratio [8].

Table 1: Initial and End point (in frame number) of each word without background noise

Word uttered	Rabiner's End point Detection		Qiang He VAD Algorithm		VAD with Frequency Domain Approach	
	Initial point	End point	Initial point	End point	Initial point	End point
AAM	105	169	106	169	107	186
AAMAA	106	163	105	163	108	179
AAMAAM	103	178	102	178	105	194
IILLAI	103	201	103	201	105	217
IILLA	102	155	102	154	105	171
IILLLA	199	261	197	261	200	278

The effective algorithm should have high compression and with low misdetection and should maintain speech quality [7]. The misdetection was calculated with manual end point detection using the audacity tool.

4.4 Observations

The performance of the three algorithms was analyzed and the result is graphically represented in Fig 3. We observed the following from the result.

- The algorithm that is using Frequency domain has more misdetection when compared to Rabiner and Qiang He Algorithms.
- Compression was slightly better in Qiang He when compared with Rabiner
- Speech Quality was very less in frequency domain when compared with the other two algorithms.

5.0 CONCLUSION

The study of three end point detection algorithm was presented. From the experimental result, it is observed that Qiang He Algorithm has good compression ratio and speech quality with less misdetection which is shown in Fig 3. So Qiang He algorithm suits better for end point detection of the given dataset. After performing the end point detection the speech signal can be used as an input to the speech recognition system. The performance of the speech recognition system can be enhanced by using proper end point detection algorithm. Qiang He algorithm was used in our research work of syllable based Tamil speech recognition system at the preprocessing phase.

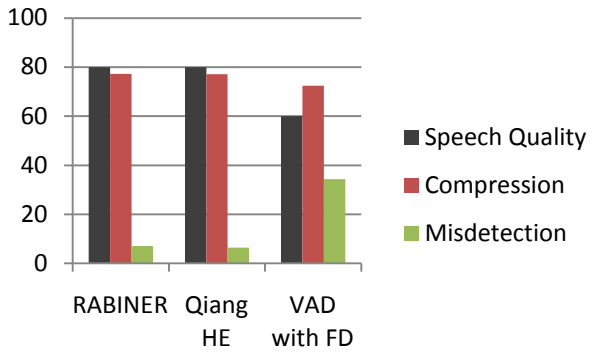

Figure 3: Graphical representation of the performance criteria of the three algorithms

6.0 REFERENCES

[1]. Miael Nilsson and Marcus EJnarsson, "Speech Recognition using Hidden Markov Model", Department of Telecommunications and speech Processing, Biekinge Institute of Technology, 2002.

[2]. Lawrence Rabiner and Biing-Hwang Juang, "Fundamentals of speech Recognition", Prentice Hall, Englewood Cliffs, N.J. 1993

[3]. Jonathan Kola, Carol Espy-Wilson and Tarun Pruthi, "Voice Activity Detection", MERIT BIEN 2011, pp 1-6

[4]. M.H.Moattar and M.M.Homayounpour, " A Simple but Efficient Real-Time Voice Activity Detection Algorithm", 17th European Signal Processing Conference, August 2009, pp 2549- 2553.

[5]. L.R.Rabiner and M.R.Sambur, "An Algorithm for Determining the endpoints of isolated utterances", The Bell System Technical Journal, Vol. 54, No. 2, February 1975, pp 297-315

[6]. Kirill Sakhnov, "Approach for Energy-Based Voice Detector with Adaptive Scaling Factor", IAENG International Journal of Computer Science, Vol. 36, No. 4, November 2009

[7]. T.Ravichandran and K.Durai Samy, "Performance Evaluation and Comparison of Voice Activity Detection Algorithms", Medwell Journals, International Journal of Soft Computing, 2007, pp: 257-261

[8]. R. Venkatesha Prasad, Abhijeet Sangwan, H.S. Jamadagni, Chiranth M.C, Rahul Sah, "Comparison of Voice Activity Detection for VoIP", Seventh International Symposium on Computers and Communications, IEEE, Tacrmina, 2002