A SILENCE REMOVAL AND ENDPOINT DETECTION APPROACH FOR SPEECH PROCESSING

Muhammad Asadullah & Shibli Nisar, National University of Computer and Emerging Sciences, Peshawar

Abstract

In this paper a brief overview of silence removal and voice activity detection is discussed and a new method for silence removal is suggested. The objective of suggested method is to delete the silence and unvoiced segments from the speech signal which are very useful to increase the performance and accuracy of the system. Endpoint detection is used to remove the DC offset value from the signal after silence removal process. Silence removal and Endpoint detection are main part of many applications such as speaker and speech recognition. The proposed method uses Root Mean Square (RMS) to delete the unvoiced segments from the speech signal. This work showed better results for silence removal and endpoint detection than existing methods. The performance of this research work is evaluated using MATLAB tool and accuracy of 97.2% is achieved.

Key words: Digital Signal Processing; Root Mean Square; Noise Removal; Voice Activity Detection

Introduction

Speech processing is study of human speech signals and its processing methods. Speech signals are normally processed in digital form, so speech signal processing is exceptional case of digital signal processing (DSP). Characteristic of speech processing includes storage, acquisition, manipulation, transfer and output of the speech signal. It contain a lot of information and its classification into voiced, unvoiced and silence regions helps to increase the performance of system. In silence region of speech signal no data is being transferred so it is very necessary to identify and delete the silence region from the speech signal. Once it deleted then it will get ignored from the further processing. For that purpose many algorithms are used such as Voice activity detection (VAD). VAD is used to detect the presence and absence of human voice. VAD mainly used in speech recognition and speech coding, it also used for noise estimation using pitch of speech signal [1]. It deactivate the process of system during the silence region of audio signal. It also avoid the unwanted transmission of silence frames and saves the processing time and bandwidth of system. The bandwidth is amount of data which can be transmitted in particular time duration. Researcher have developed the different type of silence removal algorithm according to computation cost and accuracy of the system [2-4]. Two broadly accepted techniques Zero Crossing Rate and Short Time Energy have been used for silence removal [5], however they have their own pros and cons regarding setting of threshold. While Endpoint detection techniques are mostly used in speaker and speech recognition system in order to increase the performance of system it detect the start and stop point of speech from a noisy signal. It is used to remove DC offset value from the speech signal. In endpoint detection algorithm, the start point is where the signal magnitude start to increase and exceeds the threshold value and stop point is where the magnitude of signal drops below the threshold value [6].

This research work aims to detect and delete the silence and unvoiced frames from the speech signal using Root Mean Square (RMS). After the deletion of silence frames the system consumes less bandwidth and processing times reduces. That's how this technique helps to increase overall performance and accuracy of system. In proposed method new algorithm and equations are designed to get better results than existing methods.

Related Work

Several silence removal techniques that used to remove silence region from the speech signal have been studies. In one of them fundamental frequency, zero crossing rate and short time energy is used for the identification of silence and unvoiced frames. This research work achieved 96.61% accuracy [7].

Similar research work was also studied in which probability density function (PDF) and Z-Score was used for end point detection and silence removal. This algorithms was designed for speaker and speech recognition system it achieved better results than short time energy and zero crossing rate function method [8].

A voiced detection method was also studied in which silence features such as normalization and speech energy maximization were used. According to this method the performance of speech recognition was improved [9].

Similarly, in another research work a composite silence region deletion method was proposed and compared with statistical and short time energy method by increase signal to noise ratio (SNR). It was observed this method increased the performance of speaker recognition by 20% [10].

In A. KInghorn and M. Greenwood research work zero crossing rate and short time energy were used together and it achieved 65 % accuracy [11].

Methodology

The suggested method is consist of three important parts: Noise Removal, Silence Removal and Endpoint detector. System gets input signal from the microphone for specific time duration at the particular sampling frequency. The total length of input signal is equal to product of time duration and sampling frequency of input signal.

$$N = Input\ signal_{duration} \times Fs \qquad (1)$$

Where N represents the total length of input signal and Fs is sampling frequency. Flow chart of suggested technique is shown in Figure 1.

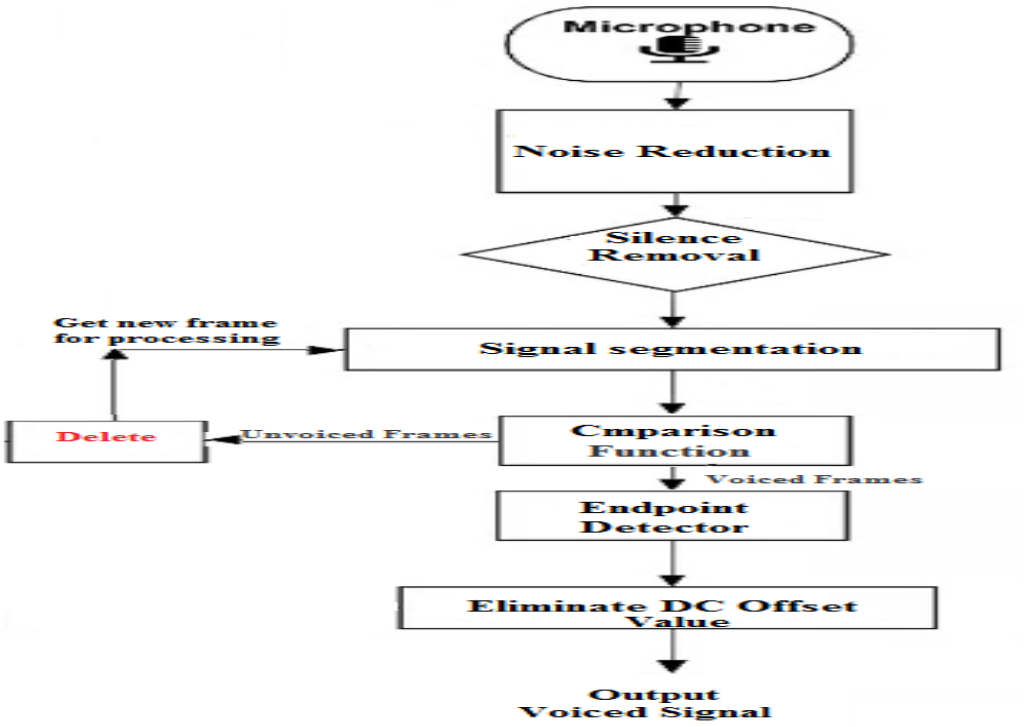

Fig. 1. Flow Chart of proposed methodology

Noise reduction

This block aims to remove the background noise from the speech signal. Different types of algorithms and filters are used to remove the noise from the signals. In Digital Signal Processing (DSP) filtering is a process which is use to eliminate the unwanted features and components from the signal. Fundamental frequency of speech signal is lies between 85 to 255 Hz. A usable ranges for voice frequency band is lies between 300 Hz to 3400 Hz. In proposed method a band pass filter is designed to suppress all the frequency contents below 300 Hz and above 3400 Hz without disturbing the voiced portion of signal.

Silence Removal

Silence removal block is used to eliminate the unvoiced and silent portion of the speech signal. For this purpose input signal is divided into small segments (frames) and root mean square (RMS) of each individual segment is calculated and compared with a specific threshold value. The total length of each individual segment is equal to product of time duration and sampling frequency of segment.

$$Segment_{lengt\ h} = Segment_{duration} \times Fs \quad (2)$$

Accuracy and performance of silence removal block depend on total number of segments. The total number of segments can be calculated from dividing total length of input signal by length of individual segment. Equation to find total number of segments is expressed as:

$$Total_{Segme\ ts} = \frac{N}{Segment_{\ lengt\ h}} \quad (3)$$

RMS value of each segment is calculated and compared with threshold value. RMS value of each individual segment can be calculated from equation 4.

$$RMS_{Segment} = \sqrt{mean(Segment)^2} \quad (4)$$

Threshold value for this block is computed from equation (5).

$$R_{th} = \frac{\mu + v}{2} \quad (5)$$

Where v is minimum RMS value of K voiced signals and μ is mean RMS value of K unvoiced signals. Formula to compute μ is expressed as:

$$\mu = \frac{1}{K}\sum_{i=1}^{K} RMS_{Unvoiced} \quad (6)$$

If $RMS_{Segment}$ of individual segment is less than R_{th} then eliminate that segment. Similarly all the segments are compared with threshold value and system will delete all the unvoiced portion from the input speech signal. The function of silence removal block is given in equation 5.

$$f(x) = \begin{cases} RMS_{Segment} > R_{th}, \ Voiced\ signal \\ RMS_{Segment} \leq R_{th}, \ Unvoiced\ Signal \end{cases} \quad (7)$$

Where R_{th} indicates the RMS threshold value. Silence removal is very helpful portion of proposed technique to reduce processing time and increase the performance of system by eliminating unvoiced segments from the input signal. A novel idea is used to set the threshold value for silence removal it eliminates 97.2% of unvoiced segments from speech signal.

Endpoint Detector

After the elimination of silent segments the new (remaining) signal entered into Endpoint detector block. Length of new signal is always less than the length of original signal. Endpoint detector is used to compute the stop point of signal where the magnitudes of signal drops to zero. After the deletion of silence segments the endpoint of new signal is equal to its length.

$$T = \frac{End_{point}}{Fs} \quad (8)$$

Where T represents the time period of new signal. Endpoint detection is important feature of speech processing it plays an important role in speaker and speech recognition for the identification of individuals.

Results

The suggested method was tested and analyzed by using MATLAB. For results 50 voiced and unvoiced signals were recorded for 10 seconds at Fs = 11025 Hz. Initially background noise was eliminated by using noise reduction block then signal was entered into silence removal block where $\mu = 0.00097$ was calculated from equation (6) and v=0.0013 was computed from minimum RMS of fifty voiced signal. Threshold value was calculated by following equation 5.

$$R_{th} = \frac{\mu + v}{2} = \frac{0.00097 + 0.0013}{2} = 0.001135$$

Input Signal was divided in 100 segments. According to equation (7) RMS value of each individual segment is calculated and compared with R_{th}. All the segments with RMS less

than R_{th} were eliminated from the speech signal and only voiced segments left. Graphical representation of silence removal process is shown in Figure 2.

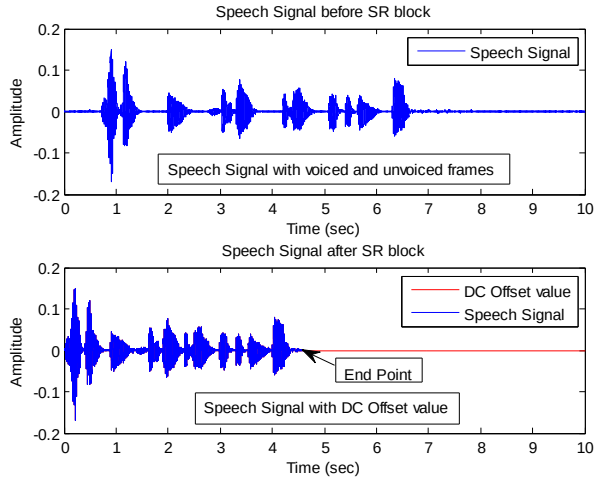

Fig. 2. Graphical representation of silence removal process

From the above figure it can observe that silence removal process eliminate all the unvoiced segments from the speech signal. DC Offset value is also shown at origin from 4.5 seconds to 10 seconds. Endpoint detector was used to remove the DC offset value and find the total time duration of speech signal. Endpoint detector measure the start point from where magnitude start to increase and stop (end) point from where magnitude start to decrease and it deleted all the DC offset value from speech signal. Graphical representation of endpoint detector process is shown in Figure 3.

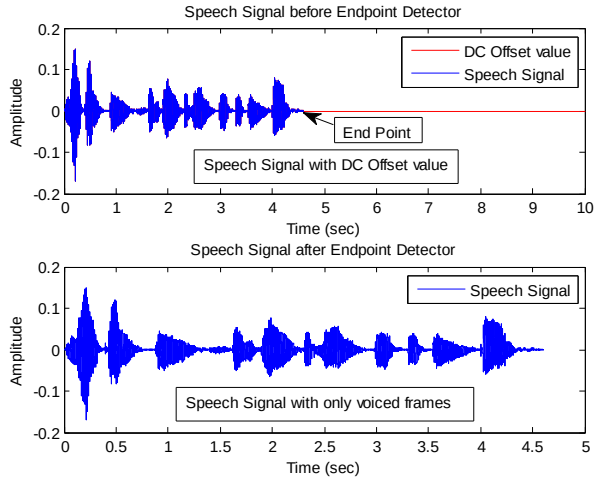

Fig. 3. Graphical representation of Endpoint detection process

Speech signal was recording for 10 seconds after all the process of proposed method the time duration of speech signal is 4.6 seconds left as shown in Figure 3. It can clearly analyzed that the proposed method has deleted all the unvoiced and silent segments from the speech signal and remaining signal contain only voiced data.

Conclusion

In this paper a latest idea for silence removal and endpoint detection is suggested. Silence removal and endpoint detection plays an important role in speaker and speech recognition to increase the performance and reduce the processing time of the system. In proposed work a new formula is suggested to set the threshold value for silence removal. It is concluded from the results proposed method eliminates 97.2% unvoiced segments from the speech signal without corrupting the voiced segments. Accuracy and performance of proposed method can be increase by adding more features.

References

[1] A. M. Cordovilla, N.Ma, V. Sánchez, J. L. Carmona, A. M. Peinado, J. Barker, "A Pitch Based Noise Estimation Technique for Robust Speech Recognition with Missing Data", IEEE ICASSP, 2011 , pp. 4808 – 4811.

[2] N. Soo Kim, W. Sung, "A statistical model-based voice activity detection", IEEE Signal Processing Letters, 1999, vol. 6, pp. 1 – 3.

[3] D. G. Childers, J. M. Larar, M. Hand "Silent and Voiced/Unvoied/ Mixed Excitation (Four-Way), Classification of Speech", IEEE Transaction on ASSP, IEEE, 1989, Vol.37, pp. 1771-1774

[4] H. Dou, Z. Wu, Y. Feng, Y. Qian, "Voice Activity Detection Based on the Bi-spectrum", IEEE 10th International conference on Signal Processing, IEEE, 2010, pp. 502-505.

[5] D. Enqing, L. Guizhong, Z. Yatong, C. Yu "Voice activity detection based on short-time energy and noise spectrum adaptation" IEEE, 6th international conference on signal processing, 2002, vol. 1, pp. 464 – 467.

[6] E. A. E-Sotelo, E. E-Hernandez, E. G-Rios, H. M. P-Meana "Endpoint Detector Algorithm for Speech Recognition Application", 2012 22nd International Conference on ELECOMP, IEEE, 2012, pp. 252 - 256

[7] Poonam Sharma, Abha Kiran, "Automatic Identification of Silence, Unvoiced and Voiced Chunks in Speech" Academy & Industry Research Collaboration Center (AIRCC), Computer Science & Information Technology, 2013, 3 (5), pp. 87-96.

[8] G. Saha, S. Chakroborty, S. Senapati, "A New Silence Removal and Endpoint Detection Algorithm for Speech and Speaker Recognition Applications", IJIGSP, December 2014, pp. 1-5.

[9] In-Sung Han, Chan-Shik Ahn, "Voice Detection using Speech Energy Maximization and Silence Feature Normalization", Advanced Science and Technology Letters, Vol.49 (ICSS 2014), pp.25-29.

[10] T.R Sahoo, S. Patra, "Silence Removal and Endpoint Detection of Speech Signal for Text Independent Speaker Identification", I.J. Image, Graphics and Signal Processing, 2014, vol. 6, pp. 27-35.

[11] Andrew KInghorn and Mark Greenwood, "SUVing: Automatic Silence/Unvoiced/Voiced Classification of Speech", Presented at the university of Sheffield.