Computer Information and Systems Engineering (Batch 2018) 20-31 January, 2022, Karachi, Pakistan

Speech end point detection using Zero Crossing Rate and Short Term Energy

Fasiha Shakil, Maham Noor, Firdous Riaz, Hoor Soomro

CS-18144, CS-18137, CS-18141, CS-18110

Ned university of Engineering and Technology, Karachi, Pakistan

[fasihashakil98@gmail.com](mailto:fasihashakil98@gmail.com), [mahamnoor200@gmail.com](mailto:mahamnoor200@gmail.com), [firdousriaz4@gmail.com](mailto:firdousriaz4@gmail.com)

**Abstract**

*Speech processing is very crucial concept in this highly sophisticated world where, signals are harnessed everywhere. This paper focuses the two algorithms for detecting speech begin and end points of a recorded signal. Zero crossing rate and short-term energy. Our findings suggest that short-term energy is much more efficient at detecting end points than zero crossing rate. Short-term energy as the name suggests uses energy of the signal for detection. On the contrary, zero crossing rate checks the signal where it changes the sign. In other words, when the signal crosses the zero or x-axis to analyze end-points.*

**Index terms:** zero crossing rate, short term energy, framing, hamming window.

1. INTRODUCTION

To begin with, Speech endpoint detection is an important part of modern speech information processing technology. Identified as the starting point and endpoint of speech signal accurately. Plays an important role in speech signal processing. Endpoint detection directly improves the performance and quality of speech coding, speech recognition, speech synthesis, speech enhancement and echo cancellation. A new, simple speech detection algorithm is implemented and applied to isolated Korean utterances with encouraging results. The algorithm makes use of decision parameters and threshold values based on three features: the logarithmic energy, the zero-crossing rate, and the modified zero crossing rate. All threshold values are fixed through the training procedure by applying a discrete optimization technique for the prepared training data set. The vocabulary for the experiment includes Korean digits and some other control commands designed for an automatic dialing system. Tested on 384 utterances from three male and three female speakers, the algorithm produces 4.6 ms average error while the well-known Rabiner and Sambur's endpoint detection algorithm gives 18.0 ms average error for the same speaker-

independent data set. It is also shown that the new algorithm can be improved by training for male and female separately.[1] Both zero crossing rate and shot time energy are used to separate voiced and unvoiced regions. They both give efficient result, increasing processing speed and are reliable.

1. ZERO CROSSING RATE

A zero-crossing is a point where the sign of a mathematical function changes. Used in speech recognition & music information retrieval. Accurate endpoint detection is important for improving the speech recognition capability. This paper proposes an effective endpoint detection algorithm based on the acoustic frequency feature for Uyghur. The spectrum of each speech frame is divided into several sub-bands, and the maximum average spectral density of these sub-bands is used as the detection criteria to distinguish the speech and noise. At the same time, a dynamically updated threshold and a smoothing window are used to improve the performance of the algorithm. It is characterized by higher accuracy or flexibility, faster processing speed and less computation. Experimental results show that the proposed algorithm achieves better performance compared to energy based and zero-crossing rate-based algorithms.[2]

1. SHORT TERM ENERGY

By the nature of production, the speech signal consists of voiced, unvoiced and silence regions. It can be used for voiced, unvoiced and silence classification of speech. Converting the voice signals (uttered in Maithili language) to text has lots of applications including speaker identification, voice mode interaction, bio-metric identification etc. Most of the research papers reported in literature, till date, concentrated mainly in English. Some works are also available for the languages Hindi, Bengali and Tamil. But working with regional languages like

Maithili, Bhojpuri, and Magahi etc remain untouched. So, in this research work one of the most popular regional languages of Bihar state under India territory, Maithili language is selected for our case study. In order to segment a voice signal firstly the word boundary is detected from the voice signals of sentences using Short Term Energy (STE) and Zero Crossing Rate (ZCR). Since there exist a large number of words for the Maithili language; converting the words directly to a text is a complex and tedious job. Hence words are further dived into syllables, which are mostly unique in terms of signal and in manageable in size. Detection of rough syllable boundary from the word boundary has been performed by group delay algorithm. To make these syllable boundaries more accurate group delay algorithm has been modified by considering the differences of consecutive peaks in the negative derivative phase. Accuracy is quantified through error rate measured by taking the differences between the ground truth boundary of the signals (determine manually) and system detected boundary of the syllable signals. Experiment using the data set of 25 different sentences from 10 different speakers have been performed and an accuracy rate of 85.62% is computed using above algorithm.[3]

1. METHODOLOGY

With the knowledge related to short-time energy and zero-crossing rate, an algorithm was created which had taken some built-in functions. First recording e voice which goes through framing and hamming window so be chopped into sections and then those sections are filtered off with the of hamming. Assigning a frame length and frameshift value we move on to calculating the signal length we find out the scale of the signal with these values we further move on to calculate the number of frames and corresponding time of those frames, all these values will help in the plotting of our output signal. After performing these necessary calculations, we will finally apply the two algorithms to the voice signal to short-time energy and zero-crossing rate to which we have defined a threshold to assign boundaries to the beginning and endpoint to the signal. Finally, the signal is plotted where the original signal, short-time energy, and zero-crossing rate are plotted to show the difference between them and thus proving which of the two algorithms is better.

The methodology is summarized in the flow diagram shown in fig. 1.

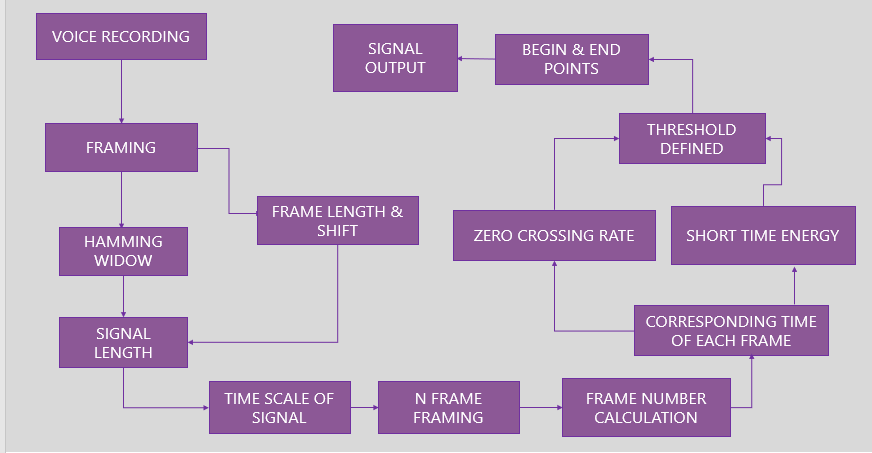


Fig. 1.

1. EXPERIMENTATION

The results were analyzed using the plots in MATLAB. The following diagram (fig. 2.) shows the result with scenario one of normal voice with almost zero noise. And the subsequent fig. 3. shows the number of frames for the given recorded sample. Total number of frames used were 399 and the number of frames for each algorithm can not exceed this number. For short-term energy the frames are 212 and for zero crossing the frames are 117. In the recorded voice only two words were said with a pause in between. Short-term energy has efficiently found the begin and the end points of the speech of the voiced parts. Whereas, in zero-crossing rate has only found one. As it could not distinguish between the tiny waves of noise and second voiced part, as it was a low pitch word said slowly. Hence, it was closely related to the noise. In fact, more noise is also taken into account for zero-crossing rate. On the contrary, short-term energy almost completely eliminates the noise. And clearly identifies the two words using their energy content. Any amount of noise in this signal is purely environmental noise, which was not generated intentionally to analyze the signal. Moreover, the initial begin point is calculated erroneously in the zero-crossing rate. Which, is slightly deviated from the actual word or voiced part.

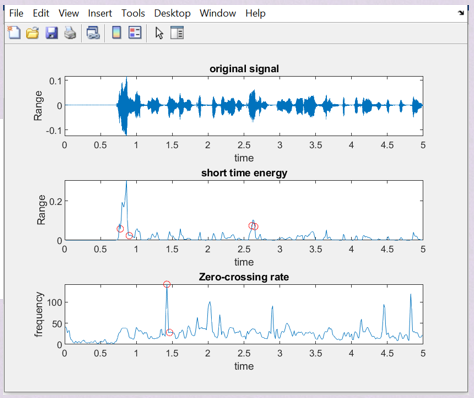


Fig. 2.

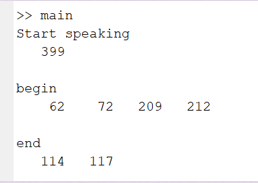


Fig. 3.

The second set of analyzed results of voice combined with great magnitude of noise is shown in fig. 4. And its subsequent number of frames used in fig. 5. As in previous set, the total number of frames is 399. Zero-crossing rate was concluded at 85th frame. Whereas, short-term energy was concluded at 369th frame, which seems is way much higher and efficient at detecting this type of speech. Where, a sentence of multiple words was said with quite unnoticeable pauses. And purposely, some continuous noise was recorded along with this signal. The noise has been shown in red line it can be noticed that noise is way much smaller after the implementation of short-term energy algorithm. Opposingly, the noise is still retained with high proportions after the zero-crossing rate. This also indicates the inefficiency of zero-crossing rate.

All the eight words of the sentence have been detected in short-term energy algorithm. On the other hand, only one word is detected in zero-crossing rate. Other words of medium intensity have been considered as noise; thus, those words have not been detected by the respective algorithm.

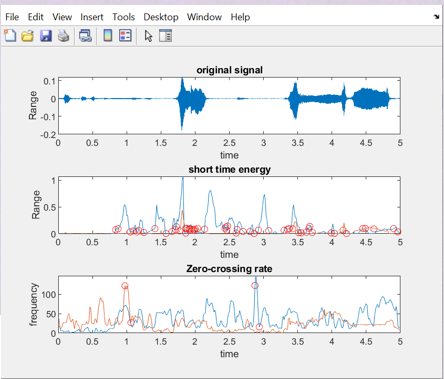


Fig. 4.



Fig. 5.

1. CONCLUSION

The digital signal processing calculates the zero crossing of the spoken words and accordingly generates different analog signals at it output. So does that Short-term energy algorithm is performed by digital processing systems. These analog signals are further processed so as to operate the appliances. The results obtained in order to conduct this research paper are summarized in the graph shown in fig. 6. The analysis of the frames through graph shows that short-term energy is more efficient at detecting begin and end points of a signal. As short-term energy is using maximum number of frames after the original signal. And zero-crossing rate uses the least number of frames for end point detection.

Hence, it can be concluded that end points of the entire signal are not being detected. Whereas, short-term energy is able to detect them with more efficiency.

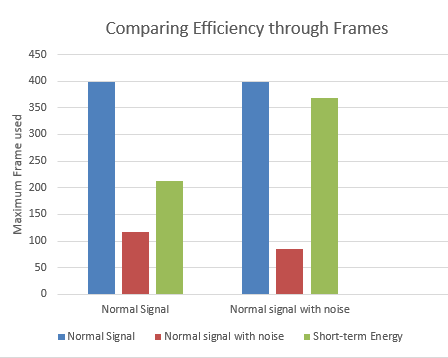


Fig. 6.

**REFERENCES**

[1] M. Hahn and C. K. Park, "An improved speech detection algorithm for isolated Korean utterances," [Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992, pp. 525-528 vol.1, doi: 10.1109/ICASSP.1992.225856.

[2] Y. Yang, B. Ma, O. Turghun and X. Li, "Speech endpoint detection algorithm for Uyghur based on acoustic frequency feature," IEEE 10th INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING PROCEEDINGS, 2010, pp. 553-556, doi: 10.1109/ICOSP.2010.5655119.

[3] S. Kumar, S. Phadikar and K. Majumder, "Modified segmentation algorithm based on Short Term Energy & Zero Crossing Rate for Maithili speech signal," 2016 International Conference on Accessibility to Digital World (ICADW), 2016, pp. 169-172, doi: 10.1109/ICADW.2016.7942535.