Speech End Point Detection using short term zerocrossing rate and energy features

Joel Jeffrey
Electrical Engineering Department
Indian Institute of Technology, Palakkad
Palakkad, Kerala

***Abstract*—Sound signals or speech is one of the most widely used mode to convey one's thoughts and intentions. Even in this digital era, the use of sounds to communicate is omnipresent. One may even communicate with robots or artificial intelligence-based assistants such as Siri, Bixby, Alexa and Google Assistant. Speech end point detection comes into the picture in such applications. The success of such assistants is directly impacted by the clarity of speech and quality of speech recognition. This report aims to model of a speech end point recognition system which uses zero-crossing to detect the start and end of words. This report also discusses the limitations and advantages of zero-crossing based speech end point recognition systems and sheds light into how such a system can either be modified or put into use in the future. Chiefly, this report aims to provide an all-encompassing picture of speech end point recognition using short term zero-crossing, its limitations and advantages and how it can prove to be beneficial in our daily life.**

Keywords—Speech End point, Zero-crossing rate, threshold, zero-axis, zero-crossing

I. INTRODUCTION

Speech End Point Detection or Speech End Point Recognition is one of the basic steps in speech processing. It involves detecting the two end points (start and end) of a speech signal. Speech End Point Detection plays a crucial role in speech processing, identity verification and other such applications.

Through speech end point detection, one aims to identify the start and end points of speech segments in an audio signal. It then involves segregation of the start and end of speech segments and differentiation of them from non-speech segments such as background sound, silences and noise.

The identification of start and end points of signals in turn help us improve the quality of speech and avoid abrupt changes in speech. The success of Speech End Point Detection and accuracy with which the segments are identified determine the success of automatic speech recognition systems.

The start and end points of speech segments can be detected by a number of ways. Few methods to detect the speech segments include using zero-crossing rate, statistical methods like Hidden Markov Model (HMM) and neural network-based methods. In this paper, however, we focus on identifying the start and end points with the help of short term zero-crossing rate.

A. Organization of the Paper

1. Literature Review.
2. Theory
3. Method (includes source code)
4. Results and Observation
5. Conclusion
6. References

II. LITERATURE REVIEW

Significant research has been done to study speech processing and speech end point detection. The sheer number of research papers indicate how vital speech end point detection is for speech processing.

The urge and necessity to drive acoustic noise out, coupled with the rise in Machine Learning algorithms have enticed researchers to use support vector machines for speech end point detection. As in [5], researchers have found SVMs to be proficient in overcoming the barrier of acoustic noise. SVMs function by classifying the problem as a two-class classification problem with +1 for noise and -1 for speech or vice-versa. However, SVMs can be computationally expensive. Training a SVM on a large dataset can take a long time and can be memory-intensive. The accuracy of SVM based systems often rely on how perfectly the parameters are tuned to. Tuning these parameters are highly time consuming and is an iterative process.

Researchers have also ventured into finding speech segments using entropy. In [3] and [4], researchers find the speech segments using spectral entropy. After obtaining the Probability Distribution Function, the spectral entropy is calculated. Sum of spectral entropy is obtained and it is filtered using the median filter. Thresholds are then placed to obtain start and end points of speech segments. Spectral entropy-based end point detection requires a huge amount of computation and hence can be time consuming.
The zero-crossing based detection relies on whether or not the signal crosses a threshold value. This in turn classifies the audio into speech segments or ignores it.

III. THEORY

A. Zero crossing detection

Zero-crossing detection is a popular technique used in digital signal processing where one estimates whether or not a signal has crossed zero-axis. Zero-axis refers to the axis which has zero voltage or zero amplitude. For zero-crossing detection, the signal is divided into overlapping frames. The number of zero crossing is then calculated. This can be implemented by checking the signs of

previous and current value or by just checking if the current value is zero.

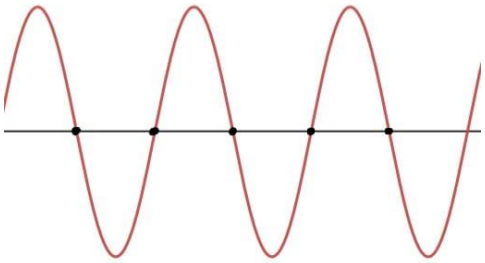

Figure 1: A sinusoidal continuous signal (The points marked in black depict the zero crossings)

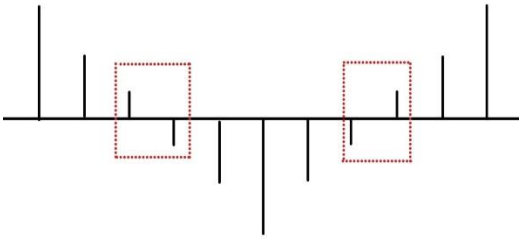

Figure 2: A discrete signal (The red boxes depict the zero crossings)

B. Zero crossing rate

Zero crossing rate is the number of times a given signal crosses the zero axis in a specified amount of time. The zero-crossing rate is an indicator of speech and the rate of zero-crossing can be used to infer if a person is speaking.

$$\text{Zero Crossing Rate (ZCR)} = \frac{\text{Number of Zero Crossings}}{\text{Time}} \quad (1)$$

For a discrete time signal x[n],

$$\text{ZCR} = \left(\frac{1}{2*N}\right)*\sum mod(sign(x[n] - sign(x[n-1])) \quad (2)$$

For a continuous time signal x(t),

$$\text{ZCR} = \left(\frac{1}{2*T}\right)*\int mod\left(sign(x(t)) - sign(x(t-dt))\right) dt \quad (3)$$

C. Threshold

The threshold is a key parameter in speech end point detection. It is a fixed multiple of the mean absolute deviation. Here,

$$\text{Threshold} = 7* \text{ Mean Absolute Deviation (MAD)} \quad (4)$$

The basic principle behind speech end point detection using zero-crossing detection is the fact that, when one speaks, the zero-crossing rate would be lower compared to that when there is noise. Hence, by comparing the zero-crossing rate to a threshold, one can detect if a person is speaking and obtain speech segments.

The success of this speech end point detection algorithm heavily depends on how the threshold is set. A low value of threshold may result in missed speech segments and a high value for threshold can result in non-speech segments such as silences and noise being included.

IV. METHOD

A. Proposed Method

A summary of the steps followed can be obtained from the flowchart presented below

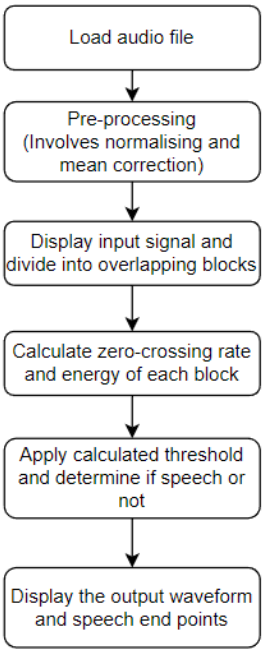

Figure 3: Flowchart depicting the proposed method

The signal is first normalized by dividing the signal with the highest amplitude. This makes the signal have values between 1 and -1.
Then the mean correction is done by shifting the components below. This helps in accurate zero crossing detection.
The signal is then divided into overlapping blocks. Zero-crossing rate and energy is determined for each block.

V. RESULTS AND SECTION

A sample signal was given as input. The signal can be found in this link. The sound input was a beat produced by the drum. The input was fed into the GUI after selecting the file.

Link to the code - Source Code

A. *Figures and Tables*

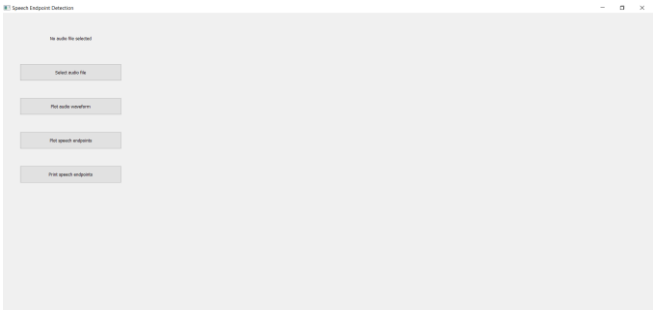

Figure 5: GUI Developed

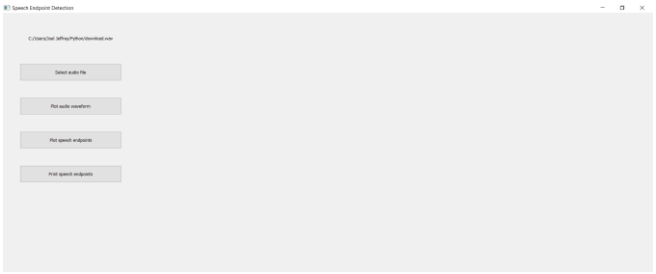

Figure 6: GUI after selecting the file

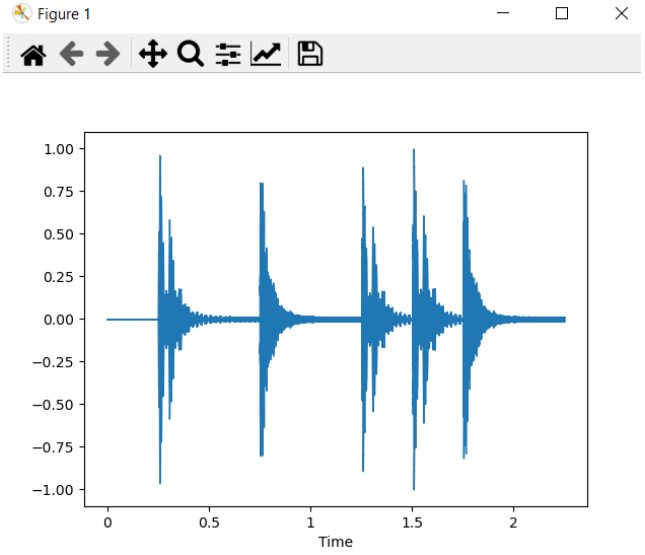

Figure 7: The input signal obtained from GUI

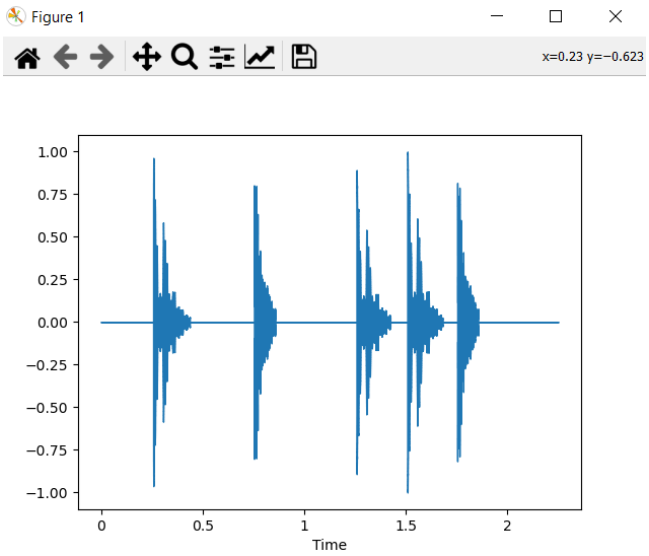

Figure 8: The output signal obtained from GUI

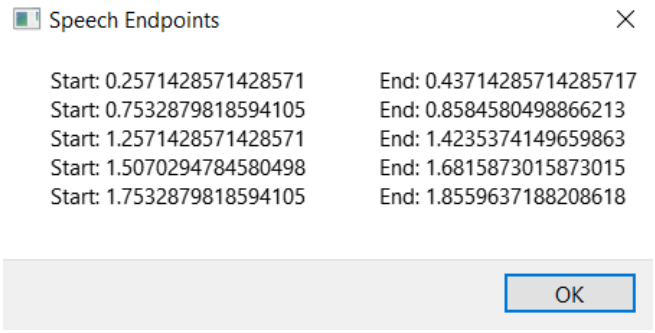

Figure 9: Speech Endpoints Printing using GUI

B. *Description of Results*

For the above-mentioned input signal as shown in Fig 7., the signal processing was done, zero-crossing rates obtained and compared for each block and the output signal was plotted. Fig. 6 depicts the output signal where the noise elements have been neutralized/nullified and speech components are present.
For more clarity, the speech end points were obtained and was printed in time domain. The printed output is displayed in Fig. 7. Comparing Fig.6 and Fig.7, it can be observed that the right speech segments have been obtained.

VI. CONCLUSIONS

The method presented in this paper has distinguished between speech segments and non-speech segments such as noise and silence. In order to distinguish, the method made use of short term zero-crossing rates and energy thresholds. Following the distinguishing, using flag variables, the method estimated the speech end points (both the start and end points).

The developed model was tested by an input sound generated from a drum. The observations and plots obtained justify the fact that the method is functioning as intended.

VII. REFERENCES

[1] Zhang, T., Shao, Y., Wu, Y., Geng, Y., & Fan, L. (2020). An overview of speech endpoint detection algorithms. Applied Acoustics, 160, 107133.

[2] Aye, Y. Y. (2009, February). Speech recognition using Zero-crossing features. In 2009 International Conference on Electronic Computer Technology (pp. 689-692). IEEE.

[3] Shen, J. L., Hung, J. W., & Lee, L. S. (1998). Robust entropy-based endpoint detection for speech recognition in noisy environments. In Fifth international conference on spoken language processing.

[4] Jia, C., & Xu, B. (2002). An improved entropy-based endpoint detection algorithm. In International symposium on chinese spoken language processing.

[5] Ramırez, J., Yélamos, P., Górriz, J. M., & Segura, J. C. (2006). SVM-based speech endpoint detection using contextual speech features. Electronics letters, 42(7), 877-879.