Detection of Silence, Voiced and Unvoiced segments in Speech

Agnel Waghela1, Rohan Reddy2, Shivangi Rai3, Aditya Pawar4, Namrata Gharat5

1-4 BE Student, Electronics and Telecommunication Department

5 Faculty of Electronics and Telecommunication Department

K.J. Somaiya Institute of Engineering and Information Technology, Mumbai, India

1agnel.w@somiya.edu, 2reddyrohan93@gmail.com, 3raishivangi33@gmail.com, 4b4.aditya.bpawar@gmail.com

*Abstract*—One of the most influential device in human lives is the `Computer`. The usage of computers is increasing exponentially in every other field existing on this earth. No doubt it simplifies the way we handle data. It also helps in easily exchanging information among humans, with everyday new technologies coming up for making the sharing of data more and more easy. But it does require some input when an interface is utilized to transfer data. Currently most common form of input provided is touch or by mouse and keyboard. With pioneering research in `Speech Recognition` from many years, it has become one of the most prominent part of modern technological systems. Thus, `Speech` is also considered as an input to computer, making it a natural and faster way of providing input to computers. A high degree of accuracy is required for good speech recognition system. The predominant problems faced are silence detection and removal, voiced unvoiced distinguishing, detection of word boundary, noise removal. This paper discusses the implementation of an algorithm which automatically detects the silence, voiced and unvoiced parts of a speech signal, which can drastically improve the accuracy of a speech recognition system. The algorithm is based on three important characteristics of a speech Signal – Zero Crossing Rate, Short Time Energy and Fundamental Frequency.

*Index Terms*—Short Time Energy, Zero Crossing Rate, Fundamental Frequency, SUV detection.

# Introduction

The NEED for deciding whether a given speech signal should be classified or segmented as voiced part, unvoiced part or silence (considered as absence of speech) arises in many speech analysis systems. A broad variety of approaches have been described in speech literature to make the approximate decision [1] – [6]. Most of the times the decision of voiced-unvoiced (V-U) usually made in conjunction with analysis of pitch of the speech segment. For example, in the well-known cepstrum based pitch detection algorithms [7], the V-U decision is taken based on the amplitude of largest peak in the computed cepstrum.

In actual practice, some additional features are required which must be included in the designed decision procedure. Because no one parameter obtained from the speech signal can accurately contribute in making the approximate but close to ideal decision for the segmentation of the speech signal.

In this paper, we are describing a method which uses number of speech derived measurements for classifying a given speech segment into three classes: silence, voiced segment and unvoiced segment (SUV). The SUV detection approach provides an effective method of combining the individual contributions of each measurements—which individually may not be enough to distinguish between the three classes—into a single output capable of providing reliable separation between each of the three classes. Then, based on the measurements and standard facts for each of the classes the discrimination is done.

A SUV detection algorithm (SUVDA) is then designed to automatically generate an output which contains individual flag values for each of the three classes. For example, the output signal generated by SUVDA will contain 0.1 corresponding to a sample value which is classified as silence and 0.2 for unvoiced, while 0.3 for voiced sample. Thus, the final output shall contain a constant signal which has the above amplitudes based on the class of the speech segment.

The SUVDA takes into account three features namely – Zero Crossing Rate, Short Time Energy, Fundamental Frequency Contour. Each of these three features are computed individually by using the windowing technique. Before the features are calculated a pre-processing step is done to remove silence parts because in most signal processing systems silence parts of any speech signal are useless, hence are discarded.

# Preprocessing Step

Preprocessing of speech signal serves various purposes in any speech processing application. It includes noise removal, pre-emphasis, windowing, framing, etc. Only windowing, framing and silence removal was needed for SUVDA. For the removal of silence parts, a feature called Spectral Centroid is computed [13]. The speech signal is divided into frames using windowing technique considering the overlap of frames as well.

Spectral Centroid: The spectral centroid, *C­i*, of the *i*-th frame is defined as the center of “gravity” of its spectrum, i.e., *Xi(k)*, k=1…,N, is the Discrete Fourier Transform (DFT) coefficients of *i*-th short-term frame, where N is the frame length. This feature is a measure of the spectral position, with high values corresponding to “brighter” sounds. If unvoiced segments simply contain environmental sounds, then the spectral centroid for the voiced segments is again larger, since these noisy sounds tend to have lower frequencies and therefore the spectral centroid values are lower.

# Feature Extraction

In feature extraction, the speech signal is converted in feature vectors containing only the information that is needed for the classification purpose. There are two types of features. First, the temporal features, which are simple to extract with physical interpretation like the energy of signal, zero crossing rate, maximum amplitude, minimum energy, *etc*. Second, the spectral features, are the ones obtained by converting into frequency domain like fundamental frequency or pitch detection and estimation using cepstrum [9].

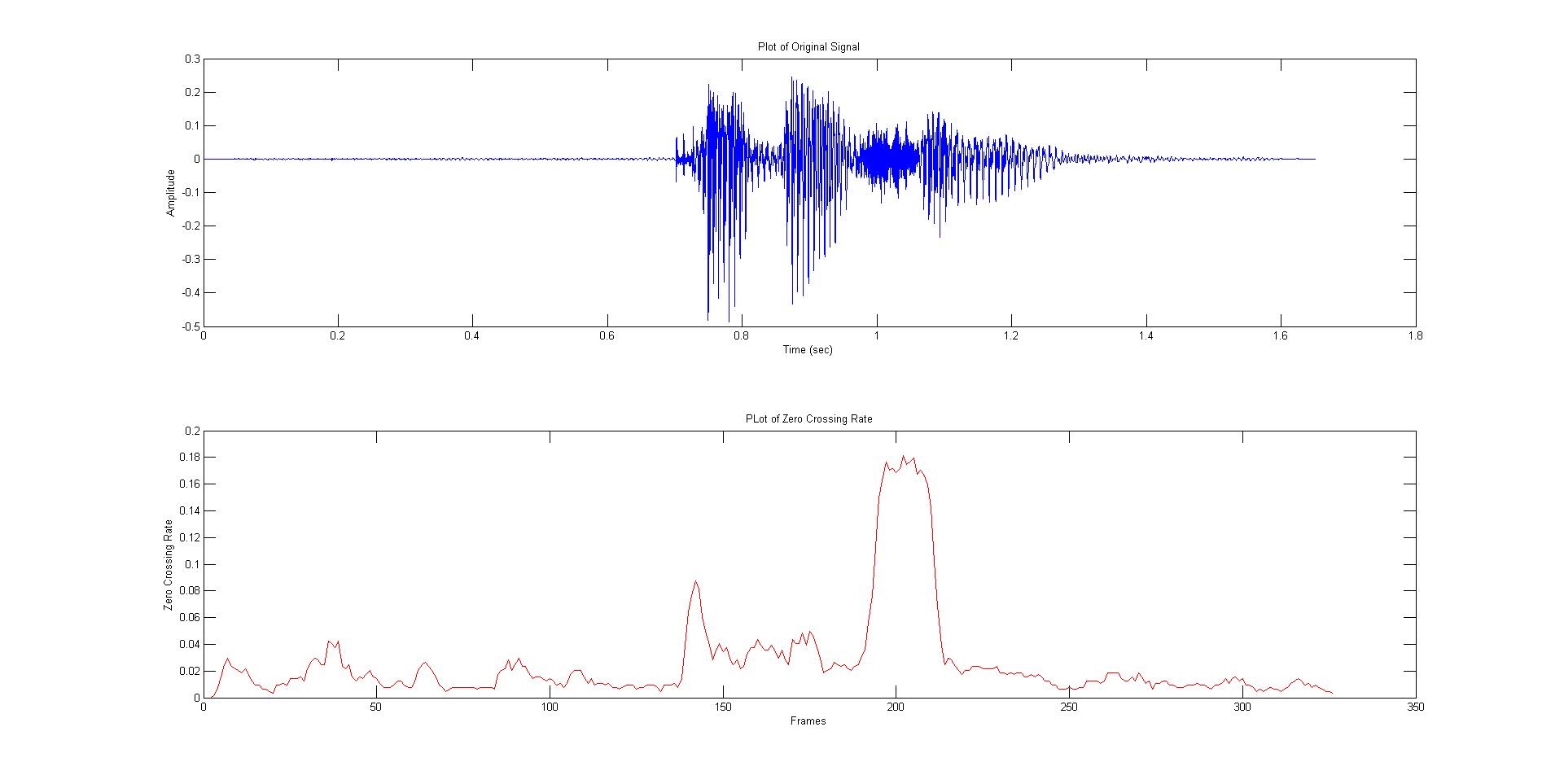
## Zero Crossing Rate

The zero crossing rate is the measure of the number of times in given time interval or frame that the amplitude of speech signal passes through a value of zero. In the context of discrete-time signals, a zero crossing is said to occur if successive samples have different algebraic signs [11]. In mathematical terms short time zero crossing rate can be defined as the weighted average of the number of times the speech signal changes sign within the time window [Rabiner and Schafer 2007].

Where

Since 0.5|*sgn*{*x*[*m*]} – *sgn*{*x*[*m*]}| is equal to 1 if *x*[*m*] and *x*[*m*–1] have different algebraic signs and 0 if they have the same sign and, it follows that Z*n* is a weighted sum of all the instances of altering sign that fall within the support region of the shifted window w[–*m*]. Figure 1 shows the plot of the original speech signal of word ‘Edition’ spoken by Male Speaker M1 and corresponding plot of the Short Time Zero Crossing Rate.

1. Short Time Zero Crossing Rate of word ‘Edition’



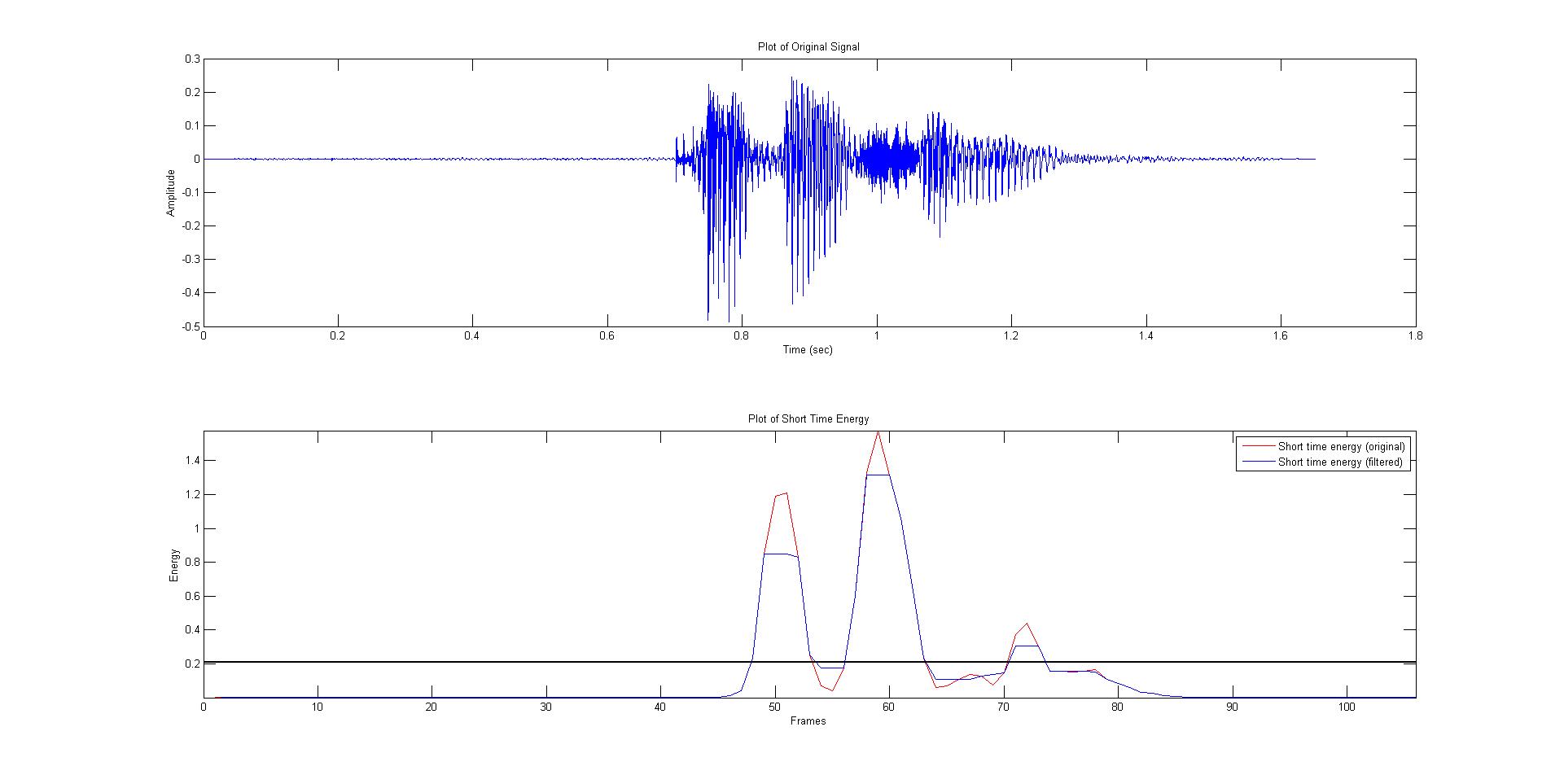
## Short Time Energy

The amplitude of the Speech signal varies over time. The energy of speech signal provides us a representation which reflects these amplitude variations. For a discrete-time signal *x*[*n*], the short time energy measure at a sample n is defined as

Since Speech signal is highly variable in nature and is assumed to have stationary properties only within short time frame so short time energy is calculated after windowing. So the definition modifies as

Here n lies between zero and N-1 where N is length of the window. Figure 2 shows the plot of the original speech signal of the word ‘Edition’ spoken by Male Speaker M1, and the corresponding plot of the Short Time Energy.

1. Short Time Energy plot of word ‘Edition’

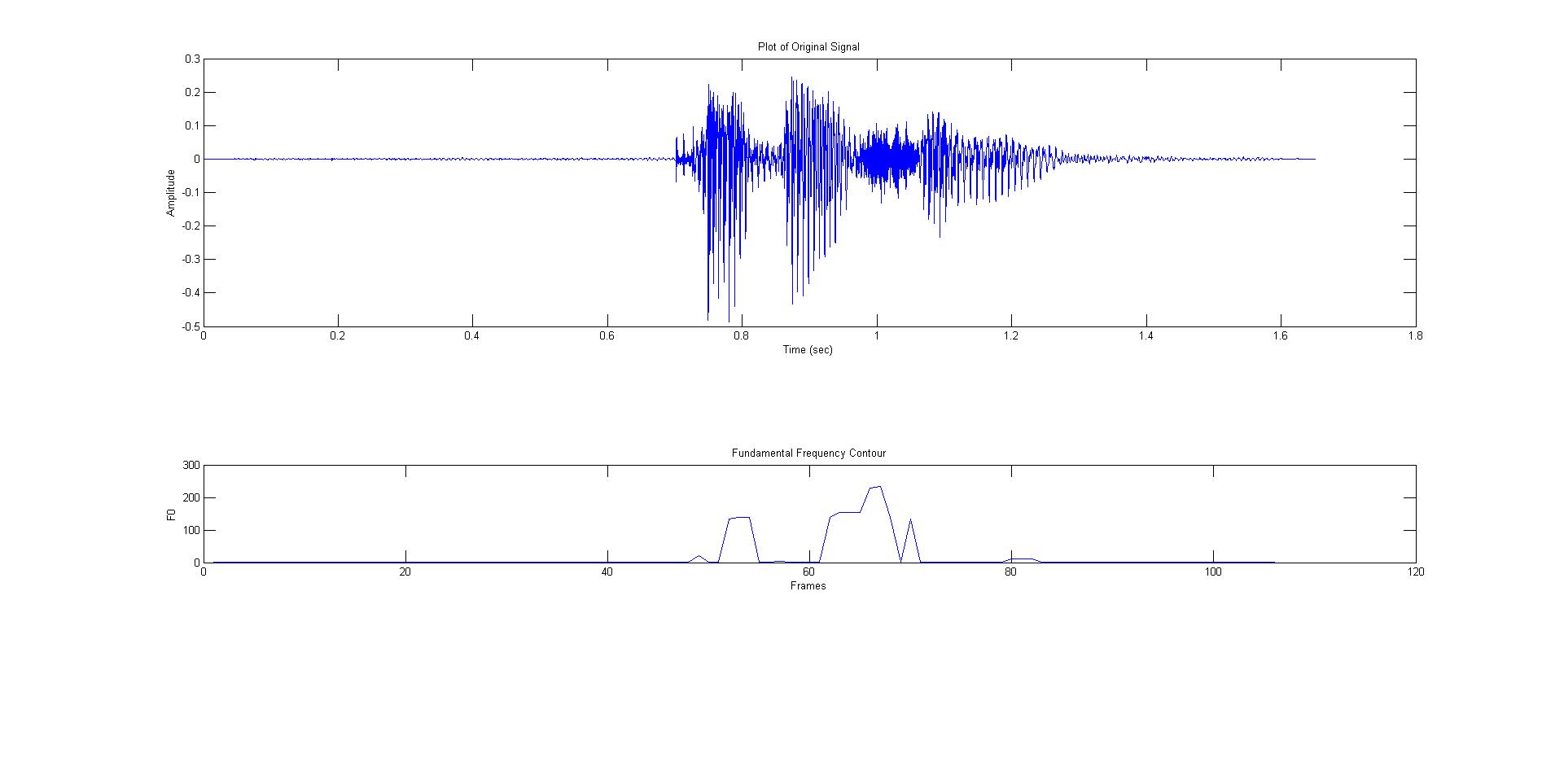


## Fundamental Frequency (F0)

A reliable way of obtaining an estimate of dominant fundamental frequency for long, clean, stationary speech signal is to use the cepstrum. The cepstrum is a Fourier analysis of the logarithmic amplitude spectrum of a signal. If the log amplitude spectrum contains many regularly harmonics, the Fourier analysis of the spectrum will show a peak corresponding to the spacing between harmonics i.e. the fundamental frequency. [14].

The Cepstrum has the units of quefrency and the peaks are called rahmonics. It is defined as

1. Plot of the F0 Contour for the word ‘Edition’.



For estimation of F0 from the obtained cepstrum, we look for the peak in the quefrency region corresponding to typical speech fundamental frequencies.

Figure 3 shows the plot of original speech signal of word ‘Edition’ spoken by Male Speaker M1 and corresponding plot of the F0 Contour. The pitch contour obtained is then used for the classification purpose. A new method of pitch tracking makes use of nonlinear processing to partially restore the missing fundamental frequency. This method utilizes a modified autocorrelation function to obtain best pitch track [8]. This method is a modified version of previously developed one [10].

# Dynamic Threshold Computation

For the classification purpose some threshold needs to be set. If the threshold value was set to some static value then it was observed that the energy of the voiced region was well above it, but in some cases the energy of the unvoiced region was also above it. Hence the threshold value was calculated dynamically to the speech data [13]. For calculating the threshold value after getting the feature vector of the energy following process was carried out:

* Compute the histogram of the feature sequence’s values
* Detect the histogram’s local maxima.
* Let M1 and M2 be the position of the 1st and 2nd local maxima respectively. The threshold value is computed using the following equation:

Where W is a user-defined parameter. Large values of W obviously lead to threshold values closer to . The computed threshold is named as ‘T\_STE’ for representing the threshold for the feature vector of Short Time Energy.

# Algorithm Description

For most of the signals it was observed that the zero crossing rate for the silence region was around 0.05 and for unvoiced region it much higher, hence this was considered to be threshold ‘T\_ZCR’ corresponding to the Zero Crossing Rate feature vector for classification purpose. In some cases, the zero crossing rate of the unvoiced region was less due to constant background noise. Hence, zero crossing alone could not be used for the classification purpose.

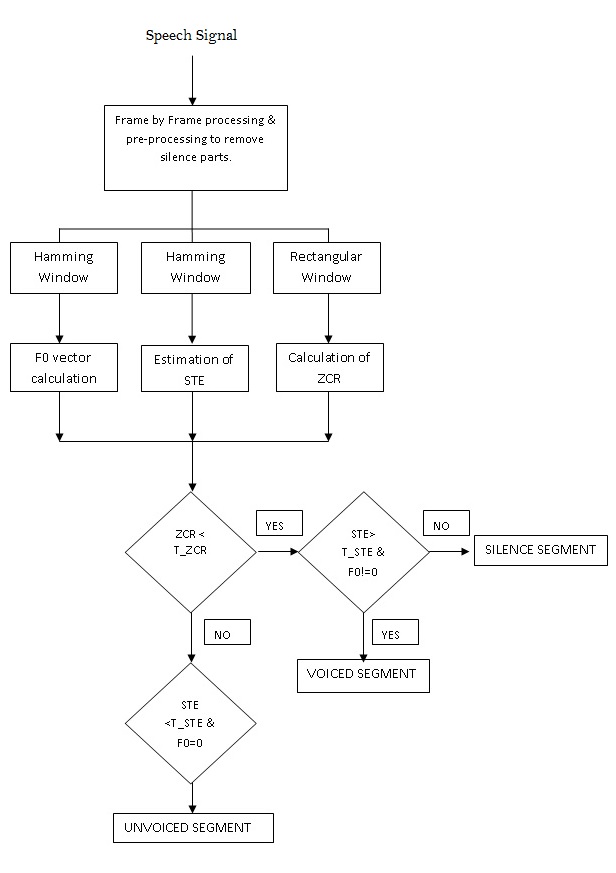
The energy of voiced region is always higher than the silence and unvoiced region, whereas the energy of the unvoiced region is less than voiced region but greater than silence region. Hence possibility of having errors due to the zero crossing rate is solved using short time energy for the purpose of classification.

There are some methods which only used zero crossing rate and short time energy for classification [12]. Whenever something is spoken voiced, the pitch of the speech signal rises and then falls when the voiced region is absent. Thus fundamental frequency is found to be present only the voiced region. This quality of speech signal is used for the classification purpose.

After analyzing the features obtained from the speech signal, the SUVDA was designed as follows:

1. Store the sound file in the vector y(n), where n is the length of the speech signal.
2. Compute the feature vectors — Zero Crossing Rate store as ZCR(n), Short Time Energy as STE(n), Fundamental Frequency as F0(n).
3. Compute the Dynamic threshold T\_STE for the feature vector Short Time Energy.
4. Mapping of the feature vectors to the length of the original signal.
5. Now if the ZCR at *n*-th sample was greater than T\_ZCR and F0 was equal to zero then that sample is declared as ‘unvoiced’ sample.
6. If the ZCR is less than T\_ZCR and F0 was found to be nonzero then it was declared as ‘voiced’ sample.
7. If either of the above two conditions fail then the sample is ‘silence’ sample.
8. Finally, an output vector is created which contains three flag values corresponding to each of the three classes.

Figure 4 shows the flowchart of the proposed SUVDA.



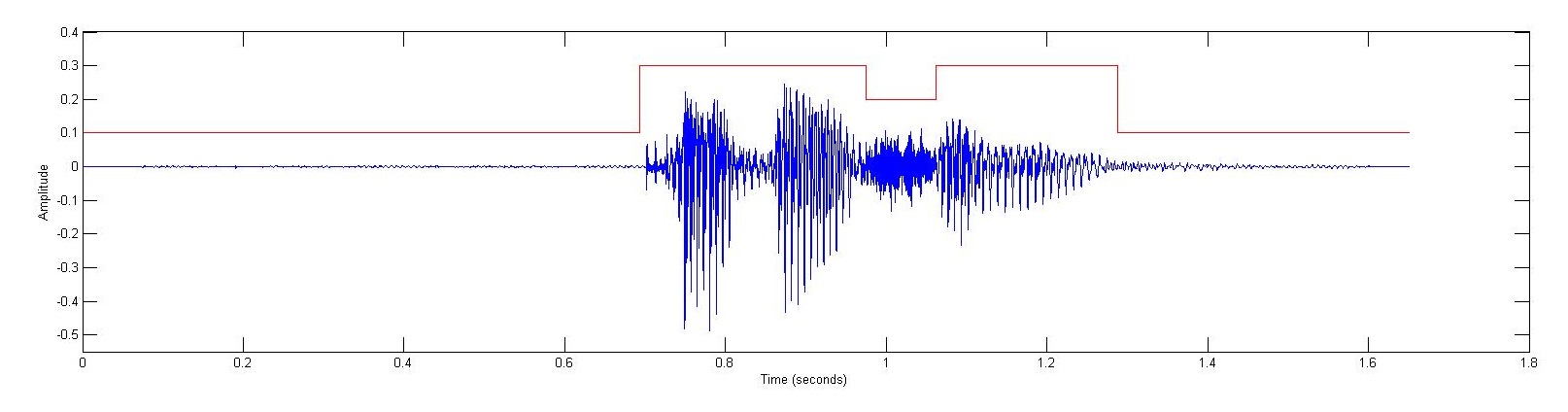
1. Flowchart of SUVDA

# Results

The final output will contain a step signal which exhibits three different amplitudes corresponding to the three classes.

The final output will contain a constant value for a particular region of the speech signal. Hence, the region of the one class can be segmented out and stored. Then the segment can be further used for processing by any recognition system as and when required at ease. Figure 5 shows the final output mapped on to the original signal with clear distinction of the three classes using the user–defined flag value. When the preprocessing step is done the following output will not contain the silence regions as it would have been removed.

1. The Final Output of the Algorithm



# References

1. K. Abdullah-Al-Mamun, F. Sarker and G. Muhammad, “A High Resolution Pitch detection algorithm based on AMDF and ACF,” J. Sci. Res. 1(3), 508-515 August 2009.
2. Mojtaba Radmard, Mahdi Hadavi and Mohammad Mahdi Nayebi, “A new method of Voiced/Unvoiced Classification based on clustering,” in JSIP, vol. 2, 336-347, October 2011.
3. Bishnu S. Atal and Lawerence R. Rabiner, “A Pattern recognition approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition,” in IEEE Trans., Acoustics, Speech, Signal Processing, vol. ASSP-24, no. 3, June 1976.
4. M. M. Sondhi, “New methods of pitch extraction,” IEEE Trans. Audio Electroacoust., vol. AU-16, pp. 262-22, June 1968.
5. J. D. Markel, “The SIFT algorithm for fundamental frequency estimation,” IEEE Trans. Audio Electroacoust., vol. AU-20, pp. 367-377, Dec. 1972.
6. Etan Fisher, Joseph Tabrikian and Shlomo Dubnov, “Generalized likelihood Ratio Test for Voiced-Unvoiced decision in noisy speech using the harmonic model,” in IEEE Trans. Audio, Speech and Lang. Processing, vol. 14, no. 2, Mar. 2006.
7. A. M. Noll, “Cepstrum pitch determination,” Journal of Acoustic Society Amer., vol. 41, pp. 293-309, Feb 1967.
8. Stephen A. Zahorian, Princy Dikshit and Hongbing Hu, “A Spectral-Temporal method for Pitch Tracking,” in INTERSPEECH, ICSLP, pp. 1710-1713, 2006.
9. R. W. Schafer and L. R. Rabiner, “System for Automatic Formant Analysis of Voiced Speech,” in Journal of Acoustic Society Amer., vol. 47, pp. 634-648, 1970.
10. Kavita Kasi, Stephen A. Zahorian, “Yet another algorithm for pitch tracking,” in ICASSP 2002, pp. 361-364, Orlando.
11. Bachu R. G., Kopparthi S., Adapa B., Barkana B. D., “Separation of Voiced and Unvoiced using zero crossing rate and energy of the speech signal,” in American Society for Engineering Education (ASEE) Zone Conference Proceedings, 2008, pp. 1-7.
12. D. S. Shete, S. B. Patil and S.B. Patil, “Zero crossing rate and Energy of the Speech Signal of Devanagari Script,” in IOSR Journal of VLSI and Signal Processing, vol. 4, issue 1, ver. I, pp. 01-05, e-ISSN: 2319-4200, p-ISSN: 2319-4197, Jan. 2014.
13. Theodoros Giannakopoulos, “A method for silence removal and segmentation of speech signals, implemented in Matlab,” Department of Informatics and Telecommunications University of Athens, Greece, 2009. Software available at https://www. mathworks.com/matlabcentral/fileexchange/28826-silence-removal-in-speech-signals.
14. X. Zhao, D. O’Shaughnessy and M. Nguyen, “A Processing Method for Pitch Smoothing Based on Autocorrelation and Cepstral F0 Detection Approaches, ISSSE, (2007), pp. 59-62.