Detection of Silence, Voiced and Unvoiced segments in Speech

Agnel Waghela1, Rohan Reddy2, Shivangi Rai3, Aditya Pawar4

1-4 BE Student, Electronics and Telecommunication Department,

K.J. Somaiya Institute of Engineering and Information Technology, Sion, 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 [x], 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 expected 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. 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 [x].

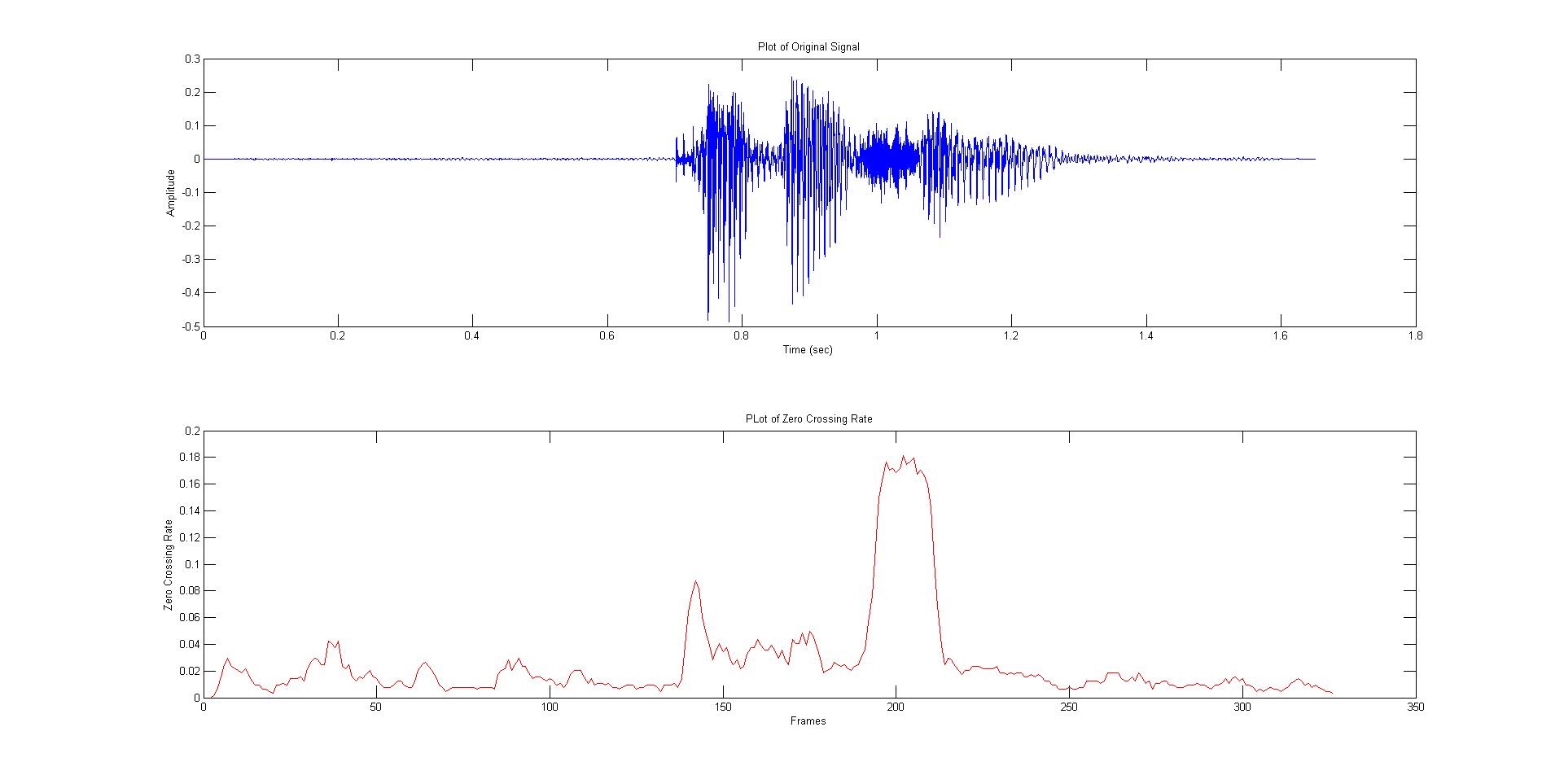
## Zero Crossing Rate

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 [bachu]. 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 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’



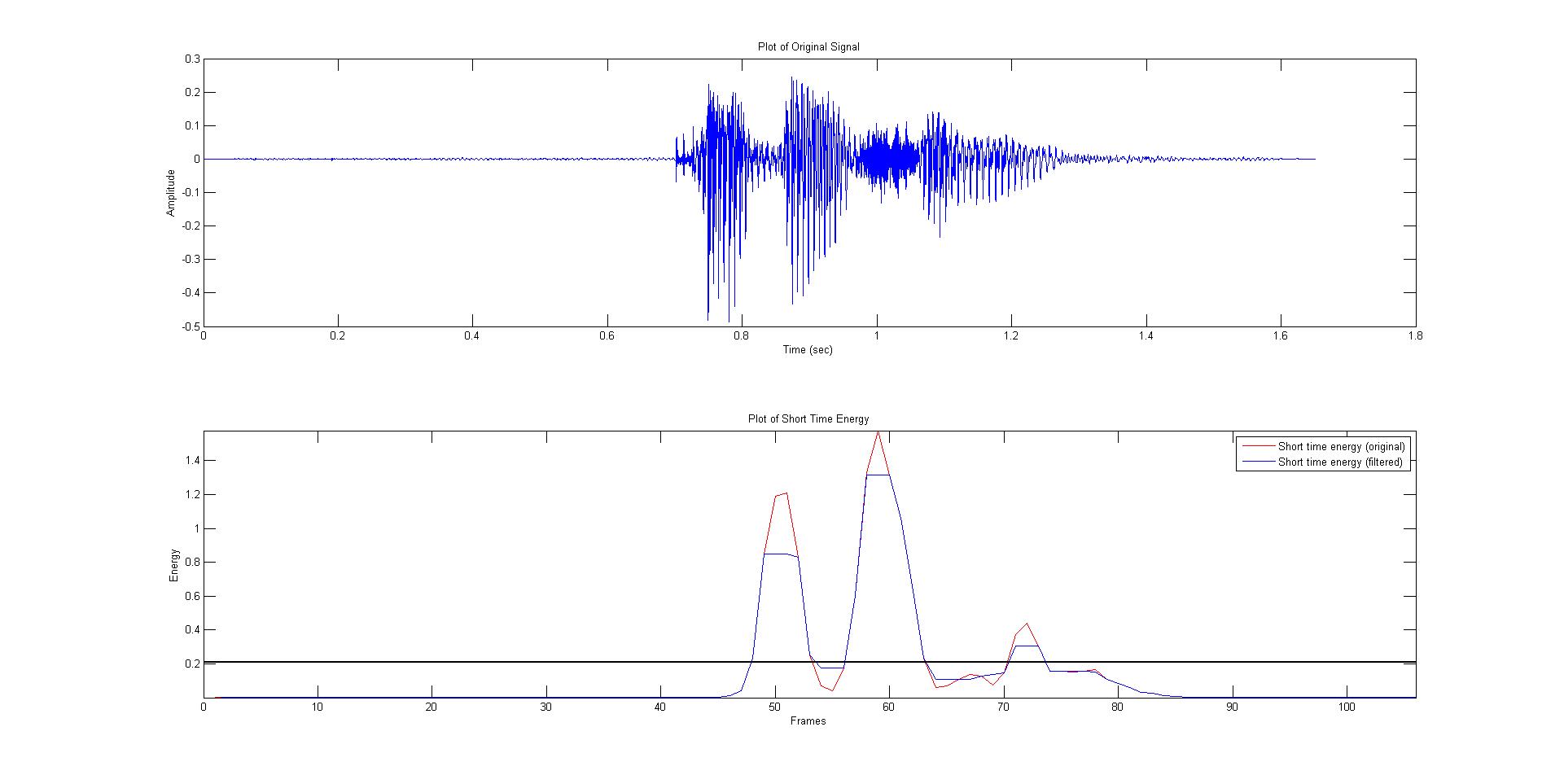
## Short Time Energy

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’

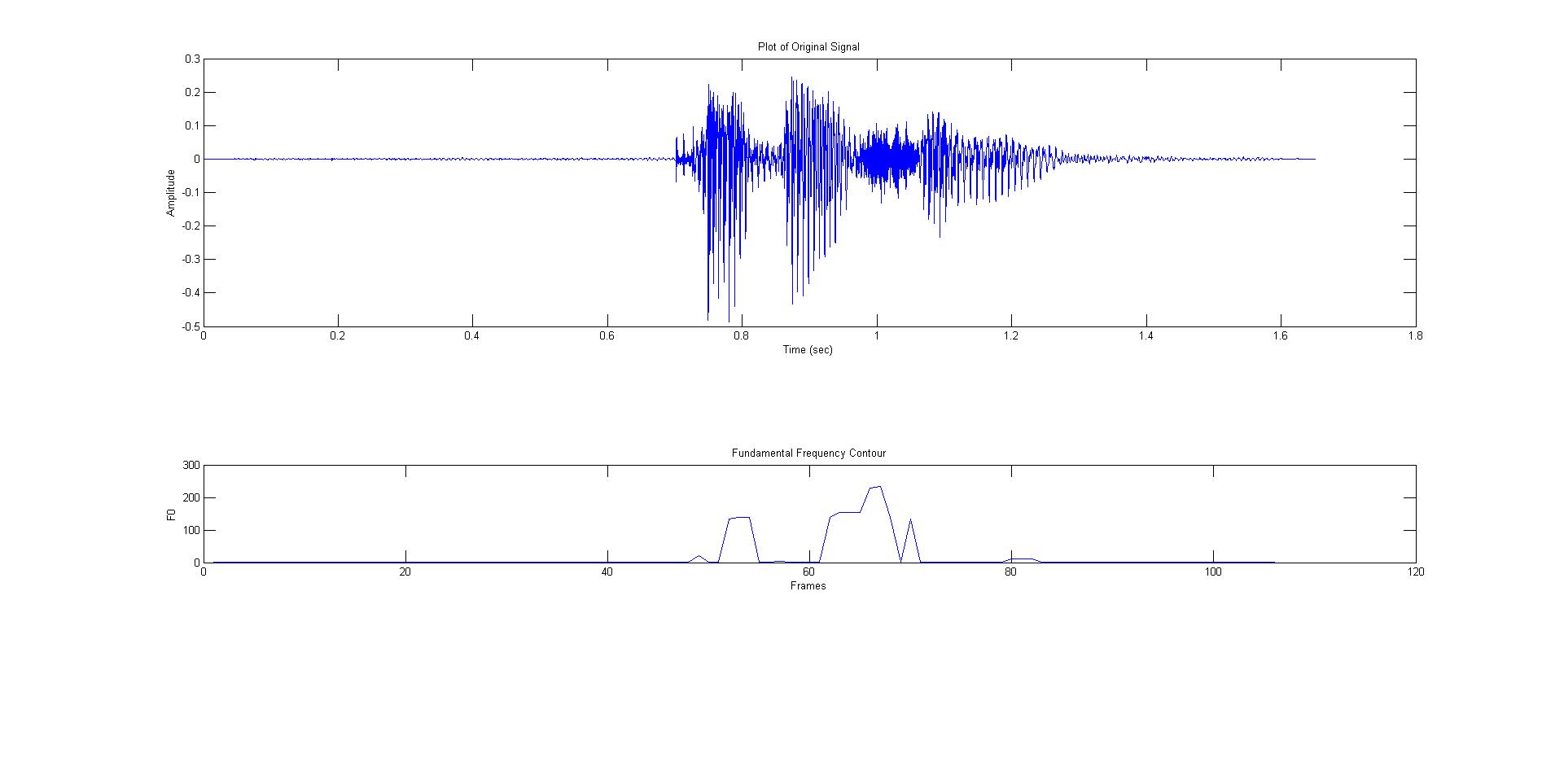


## Fundamental Frequency (F0)

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. [Xufang Zhao].

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.

# 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 []. 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.

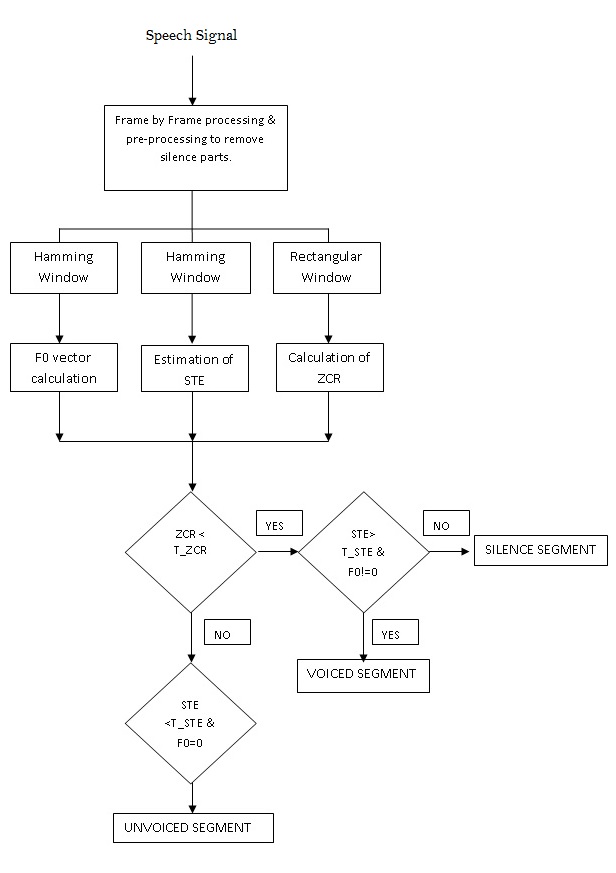
The energy of voiced region is always higher than the silence and unvoiced region, whereas the energy of the unvoiced region is less than voiced region but greater than silence region. Hence possibility of having errors due to the zero crossing rate is solved using short time energy for the purpose of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

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

# Expected Results

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

# Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads—the template will do that for you.

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar.

## Abbreviations and Acronyms (Heading 2)

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE and SI do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

## Units

* Use either SI or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
* Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
* Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter”, not “webers/m2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.
* Use a zero before decimal points: “0.25”, not “.25”. Use “cm3”, not “cc”. (*bullet list*)

## Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in Eq. 1, using a right tab stop. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in

 

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “Eq. 1” or “Equation 1”, not “(1)”, especially at the beginning of a sentence: “Equation 1 is . . .”

## Some Common Mistakes

* The word “data” is plural, not singular.
* The subscript for the permeability of vacuum **0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
* In American English, commas, semi-/colons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
* A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
* Do not use the word “essentially” to mean “approximately” or “effectively”.
* In your paper title, if the words “that uses” can accurately replace the word “using”, capitalize the “u”; if not, keep using lower-cased.
* Be aware of the different meanings of the homophones “affect” and “effect”, “complement” and “compliment”, “discreet” and “discrete”, “principal” and “principle”.
* Do not confuse “imply” and “infer”.
* The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
* There is no period after the “et” in the Latin abbreviation “et al.”.
* The abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

An excellent style manual for science writers is given by Young [7].

# Using the Template

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Authors and Affiliations

The template is designed so that author affiliations are not repeated each time for multiple authors of the same affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization). This template was designed for two affiliations.

### For Author/s of Only One Affiliation (Heading 3): To change the default, adjust the template as follows.

#### Selection (Heading 4): Highlight all author and affiliation lines.

#### Change Number of Columns: Select Format > Columns >Presets > One Column.

#### Deletion: Delete the author and affiliation lines for the second affiliation.

### For Authors of More than Two Affiliations: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change Number of Columns: Select Format > Columns > Presets > One Column.

#### Highlight Author and Affiliation Lines of Affiliation 1 and Copy this Selection.

#### Formatting: Insert one hard return immediately after the last character of the last affiliation line. Then paste down the copy of affiliation 1. Repeat as necessary for each additional affiliation.

#### Reassign Number of Columns: Place your cursor to the right of the last character of the last affiliation line of an even numbered affiliation (e.g., if there are five affiliations, place your cursor at end of fourth affiliation). Drag the cursor up to highlight all of the above author and affiliation lines. Go to Format > Columns and select “2 Columns”. If you have an odd number of affiliations, the final affiliation will be centered on the page; all previous will be in two columns.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is *“Heading 5”*. Use *“figure caption”* for your Figure captions, and *“table head”* for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named *“Heading 1”*, “*Heading 2”*, *“Heading 3”*, and *“Heading 4”* are prescribed.

## Figures and Tables

Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table captions should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1” in the text, and “Figure 1” at the beginning of a sentence.

Use 8 point Times New Roman for figure labels. Use words rather than symbols or abbreviations when writing figure-axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”.

If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

## Footnotes

Use footnotes sparingly (or not at all) and place them at the bottom of the column on the page on which they are referenced. Use Times 8-point type, single-spaced.

To help your readers, avoid using footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence).

Number footnotes separately from reference numbers, and in superscripts. Do not put footnotes in the reference list. Use letters for table footnotes.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
|  | Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

a. Sample of a table footnote. *(table footnote)*

1. Example of a figure caption. *(figure caption)*

We suggest that you use a text box to insert a graphic (ideally 300 dpi, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on Example of a figure caption. *(figure caption)* your frame, use the MSWord pull-down menu, select Format > Borders and Shading > Select ”None”.

# Copyright Forms

You must submit the IEEE Electronic Copyright Form (ECF) as described in your author-kit message. THIS FORM MUST BE SUBMITTED IN ORDER TO PUBLISH YOUR PAPER.

# Acknowledgment

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression, “One of us (R. B. G.) thanks . . .” Instead, try   
“R. B. G. thanks”. Put applicable sponsor acknowledgments here; DO NOT place them on the first page of your paper or as a footnote.

# References

List and number all bibliographical references in 9-point Times, single-spaced, at the end of your paper. When referenced in the text, enclose the citation number in square brackets, for example: [1]. Where appropriate, include the name(s) of editors of referenced books. The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in “[3]”—do not use “Ref. [3]” or “reference [3]”. Do not use reference citations as nouns of a sentence (e.g., not: “as the writer explains in [1]”).

Unless there are six authors or more give all authors’ names and do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.