Open access Journal International Journal of Emerging Trends in Science and Technology
Impact Factor: 2.838 DOI: http://dx.doi.org/10.18535/ijetst/v3i06.14

Implementation of ZCR and STE techniques for the detection of the voiced and unvoiced signals in Continuous Punjabi Speech

Author
Anu Priya Sharma
Chandigarh University, Gharuan
Email: *er.anupriya33@gmail.com*

ABSTRACT

During the analysis of speech signals the evaluation of the basic characteristics of the speech is an important stage. The basic characteristics of speech are voiced, unvoiced and silence. Such characteristics are evaluated by calculating zero crossing rate (ZCR) and short term energy (STE). The speech segmentation system needs a speech signal to be segmented into some basic units like words, phonemes, syllables. In this paper some of the

steps required for the feature extraction in case of automatic segmentation of speech are discussed. There are various characteristics of speech such as: voiced, unvoiced and silence. Such basic characteristics of the speech can be evaluated by the computation of zero crossing rate and short term energy.

Keywords: *Zero crossing rate, Short Term Energy (STE), ASR, Segmentation, Syllables, Words, Automatic segmentation, Manual Segmentation.*

1. Introduction

Though computer is the most popular and effective means to access information and helps to make our work easier. But still it requires a skill to operate the computer. It's a great challenge for physically handicapped or blind people to operate computers. Thus, speech synthesis and speech recognition systems play a vital role in such scenarios. This paper discusses some of the steps required for the automatic segmentation of speech. There are various characteristics of speech such as: voiced, unvoiced and silence. Such basic characteristics of the speech can be evaluated by the computation of zero crossing rate and short term energy. The acoustic signal is segmented into some basic units. The syllables are one of the most important units of segmentation. The function STE contains useful information about the peaks and valleys thus, helps to define the segment boundaries. The peak having maximum value called the nucleus is represented as vowel is represented as consonants.

2. Segmentation of speech into its basic Units

The basic units of speech into which a signal can be segmented are words, phonemes, or syllables. The units to be chosen mainly depend on the vocabulary size. Though word is the most natural unit of speech still it is not appropriate for segmentation due to lack of generalization and more memory consumption [7]. The higher level units of speech are phonemes and are the smallest segmental units employed to form meaning. There are different realizations of same phoneme in different words. The phonemes are found inappropriate for segmentation due to its overgeneralization. To overcome this problem the combination of phone and words gives rise to nest level basic units of speech known as syllables. The syllables are composed of vowels and consonants and are defined by rules. The presence of vowel is mandatory where as the presence of consonants are optional.

3. General Characteristics of Speech

In a continuous speech signal there are two main parts: one carries the speech signal in the form of information, and the other includes silence or noise sections that are without any verbal information between the utterances. Thus, a speech can be divided into numerous voiced and unvoiced regions.

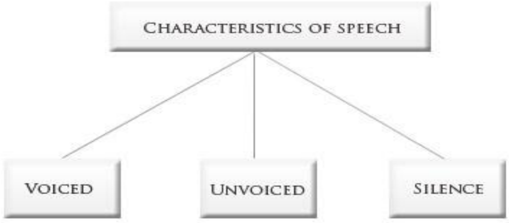

Fig 1: Block diagram of characteristic features of Voice

3.1 Voiced speech

The Voiced sound is produced when the air from the lungs passes through the larynx. The pitch is the fundamental frequency and differs among people as different people have different larynx's anatomy. The pitch of Men's commonly ranges between 50 to 250 Hz while women's lies between 120 and 500 Hz [3].

3.2 Unvoiced Speech

With the passage of air directly through the vocal tract formations the *unvoiced speech sounds* are produced. Unvoiced speech does not exhibit periodicity has it is characterized by a noise-like signal. On the other hand voiced speech shows periodicity [3].

3.3 Silence region

The speech production process is incomplete without the detection of voiced and unvoiced speech that is separated by a silence region. In case of silence region no excitation is supplied to the vocal tract and thus, no speech is produced. A regular speech is incomplete/ inaccurate without silence region. It helps to make the speech understandable [3].

5. Characteristic features for voice and its detection criteria

The two main characteristics features of voice are Zero Crossing Rate also know are ZCR and Short Term Energy also known as STE.

5.1 Zero Crossing Rate

The rate at which the signal crosses zero provides the information regarding its (source of creation) i.e. zero crossing rate. In case of unvoiced speech the signal crosses zero more number of times that means the unvoiced speech has higher zero crossing rate. Whereas in case of voiced speech the zero crossing rate is low that means the signal crosses zero less number of time. Thus, the amplitude of unvoiced segments is lower than that of the voiced segments.

ZCR can be defined as:

$$Z_n = \sum_{m=-\infty}^{\infty} \left|\operatorname{sgn}[x(m)] - \operatorname{sgn}[x(m-1)]\right| w(n-m)$$

Where

$$\operatorname{sgn}[x(n)] = \begin{cases} 1, x(n) \geq 0 \\ -1, x(n) < 0 \end{cases}$$

Matlab code for Zero Crossing Rate:

zc = zerocross1(x,wintype, winamp(1), winlen)

The ZCR of unvoiced speech is much higher than the ZCR of voiced speech. The ZCR of Voiced and Unvoiced speech is shown in the following diagrams.

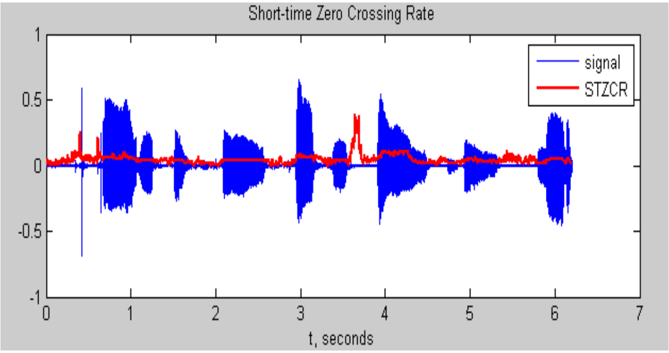

Fig 2: Representation of ZCR for voiced signal

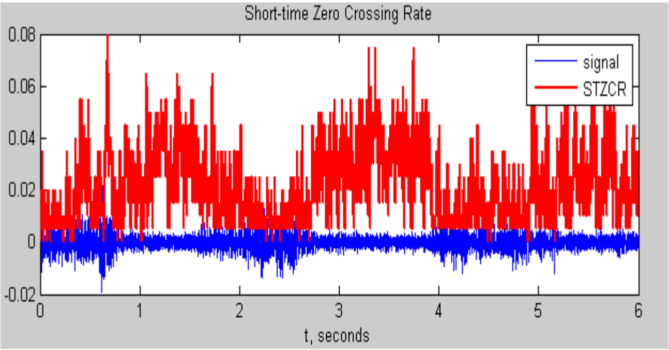

Fig 3: Representation of ZCR for unvoiced signal

5.2 Short Term Energy (STE)

Short-time energy of speech signals reflects the amplitude variation. By processing STE function the speech can be segmented. STE shows the voiced content of the signal. The number of voiced segments can be computed by calculating STE but it cannot compute the phonetic content of the speech. Due to some local energy fluctuations this method cannot directly perform segmentation thus; this approach is used for language independent segmentation of multilingual speech.

The STE can be defined as follows:

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2$$

Matlab energy function:

E = energy(x,wintype, winamp(2), winlen);

The STE of voiced signal is always much greater than that of unvoiced signals. In a speech signal where there are voiced signal its STE will be high, the peaks in the signal represents nucleus that is denoted as vowel where as the valleys at both the ends represents the coda. STE for unvoiced signal shown in the following diagram:

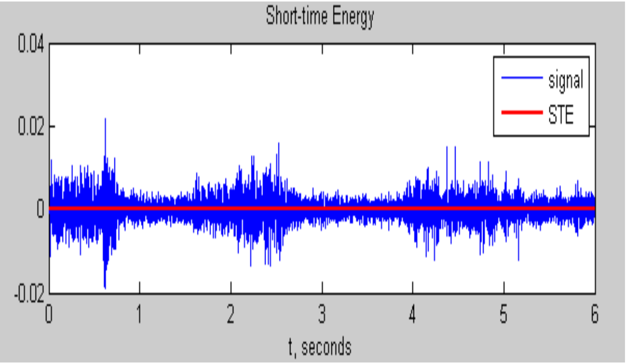

Fig 4: Representation of STE for unvoiced signal

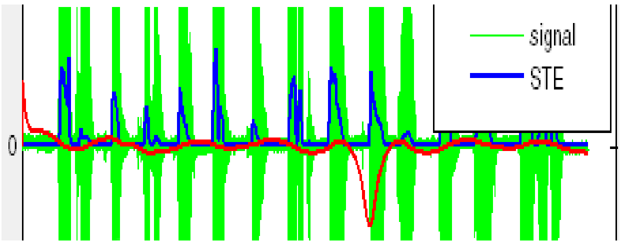

Fig 5: Representation of STE for voiced signal

The syllable centers are the high energy regions in the STE functions and the valleys at both ends of the syllable nuclei are the syllable boundaries. For continuous speech STE functions are not reliable [2].

6. Basic Methods Of Segmentation

The two basic approaches for segmenting the speech signal into various acoustic units is by: Manual segmentation (hand labeling) or automatic segmentation (ASR). According to a research the (ASR) Automatic segmentation method is found to be the better approach. The onset and offset values of the syllable boundaries for various Punjabi speech signals were computed and compared. It was concluded that the syllable boundaries marked with the automatic segmentation method were more accurate than the approach of manual segmentation [7]. Thus, the short term energy computed helps to detect the voiced segments in the speech signals. The values obtained by computing STE are very much useful for the automatic segmentation of Punjabi speech signals into syllable like units.

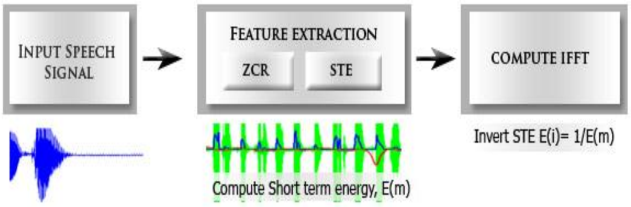

Fig 4: Feature extraction during automatic speech segmentation

Conclusion

STE is found to be the best method for speech segmentation. The higher value of short term energy refers to the voiced segments. The peaks and valleys in the speech signal refer to the nuclei and coda respectively. The valleys at both the ends represent the boundary of the syllable. On the other hand the zero crossing rate only defines the number of times the signal passes through zero. It can only categorize the signal into voiced or unvoiced where as the Short term energy can compute the energy content of the signal and also helps to mark the syllable boundaries by detecting the peaks and valleys in the speech signal. In various researches the STE method is efficiently used for marking the syllable boundaries as well as for segmentation.

References

1. T.Nagarajan et al. "*Segmentation of speech into syllable-like units*," in Eurospeech Sixth biennial conference of signal processing, Geneva, 2003.
2. T. Nagarajan and H. A. Murthy, "*Subband-Based Group Delay Segmentation of Spontaneous Speech into Syllable-Like Units*," in Eurasip Journal on Applied Signal Processing , Hindawi Publishing Corporation 2004:17, pp. 2614–2625.
3. N. Mikael, E. Marcus, "*Speech Recognition using Hidden Markov Model, Performance evaluation in noisy environment*", Degree of master of science in Electrical Engineering, Department of telecomminications and engineering, Blekinge Institute of Technology, March 2002.
4. G. Pradeep "*Text-to-Speech Synthesis for Punjabi Language*", Thesis degree of Master of Engineering in Software Engineering submitted in Computer Science and Engineering Department of Thapar Institute of Engineering and Technology (Deemed University), Patiala, May 2006.
5. V. Kamakshi Prasad, T. Nagarajan, Hema A. Murthy, "*Automatic segmentation of continuous speech, using minimum phase group delay functions*," in the proceedings of science direct, Speech Communication 42, 2004, pp. 429–446.
6. G Lakshmi Sar ada, et al. "*Automatic transcription of continuous speech into syllable-like units for Indian languages*," in Sadhana, Vol. 34, Part 2, April 2009, pp. 221–233
7. K. Amanpreet, and S. Tarandeep, "*Segmentation of Continuous Punjabi Speech Signal into Syllables*," in the Proceedings of the World Congress on Engineering and Computer Science 2010 Vol I, WCECS 2010, San Francisco, USA, October 20-22, 2010.
8. S. Parminder, L. Gurpreet, "*Corpus Based Statistical Analysis of Punjabi Syllables for Preparation of Punjabi Speech Database*," in International Journal of Intelligent Computing Research (IJICR), Volume 1, Issue 3, June 2010.
9. A.Hema, and B.Yegnanarayan, "*Group delay functions and its applications in speech technology*," in Sadhana, Vol. 36, Part 5, October 2011, pp. 745–782.
10. S. Nishi, and S. Parminder, "*Automatic Segmentation of Wave File*," in International Journal of Computer Science & Communication Vol. 1, No. 2, July-December 2010, pp. 267-270.
11. A. Hema, B. Ashwin, et al., IIT-Madras, IIT-Kharagpur, CDAC-Trivandrum, CDAC- mumbai, IIIT-Hyderabad, "*Building Unit Selection Speech Synthesis in Indian Languages*," An Initiative by an Indian Consortium, 2009.
12. Zhihong Hu, Johan Schalkwyk, Etienne Barnard, Ronald Cole, "*Speech Recognition Using Syllable-Like Units*," Center for Spoken Language Understanding, Oregon Graduate Institute of Science and Technology, September 2008, pp. 218-222.
13. R.G. Bachu, S. Kopparthi, B. Adapa, B.D. Barkana, " *Separation of Voiced and Unvoiced using Zero crossing rate and Energy of the Speech Signal*," Electrical ¬Engineering Department School of Engineering, University of Bridgeport, March 2010, volume 7340, 2012, pp. 539-546.