3

WORD BOUNDARY DETECTION IN BODO LANGUAGE

CONTENTS

OBJECTIVE OF THE CHAPTER

The objective of this chapter work is to analysis different speech features in Bodo language for word boundary detection in normal mode. We propose one method to automatically segment the speech signal into voiced and unvoiced and silence regions. Our proposed method is based on Zero Crossing Rate, Short Time Energy and Fundamental Frequency of speech signal. We analyze energy and pitch features to find some clue for word boundary detection. The pitch contour of a sentence indicates the sentence type and its expressive style. Similarly pitch variation also indicates the boundaries between major syntactic units in a sentence..

3.1 INTRODUCTION

The identification of start and end of each word in a continuous spoken utterance is called Word Boundary Detection (WBD). Keyword spotting, speech recognition system uses WBD for increasing the performance of the system. Speech community researchers use two types of speech for their analysis: constrained speech and unconstrained speech. In the constrained speech, there are well-defined boundaries and well-defined pauses between words. But in unconstrained speech there are not any well-defined boundaries, grammar or pauses. So WBD is challenging in unconstrained speech and without a good detection algorithm, word boundary detection can result in many false alarms and misses. WBD can also be classified as speaker dependent and speaker independent [24]. Speaker dependent systems are easy to develop as compared to speaker independent systems because this type of systems contains the characteristics of speech signal like pitch, intensity of single speaker only. But speaker independent systems consist of utterances of words by several speakers. So it makes the process challenging, due to different prosody, pitch, intensity etc. of the different speakers. Word boundary detection plays a vital role for the speech based application development like keyword spotting, speech recognition etc.

Researchers proved that 50% of the failure of this type of system occurs due to the errors in the word boundary detector [25]. So designing of efficient word boundary detection algorithm can reduce the recognition errors and it can improve the performance of keyword spotting of the speech recognition systems. WBD can also

help in reducing the search space in the keyword spotting algorithm. Speech is non-stationary signal. Therefore, most of the time, no utterance of the same word will be same as another utterance of same word. This makes it challenging to develop any speech processing algorithm. The researchers in the past have used energy, acoustic features, pitch; lexical cues etc. are used to design the WBD [25].

The detection of word boundary is one of the most challenging works in the area of speech processing. The accurate detection of word boundaries is challenging in the sense that the speaker does not pause consciously between the words. Also the method developed for a particular language may not work for other language as different languages possess different features. It is more challenging in case of tonal language. Tonal languages are unique in the sense that they have two tones for certain words. In tonal language, lexical items are distinguished with the help of tone. In this type stress pattern is not used. Tones are associated with unambiguous speech units called syllables and each tone is associated with isolated utterances, which has a unique fundamental frequency (f0) curve. In continuous speech f0 curve is affected, but they can be easily recognized in articulated speech or slow speaking rate. The detection and sensitivity of tones is greatly influenced by intensity, fundamental frequency and timing of the speech signal. The fundamental frequency (f0) curve is considered as the most prominent acoustic correlate of tone in tone languages.

3.2 LITERATURE SURVEY

The emotional state, intonation pattern behavior is used supra-segmental based word boundary detection algorithm. These patterns give the clues to the linguistic structure of the speaker's message. Intonation, stress, rhythm etc. can help to signal the syntactic structure of utterances of the language. Word Boundary Detection based on Cepstral matrices is an efficient method of detecting word boundaries in continuous speech in noisy environment [26]. In pitch variation based WBD method the pitch frequency (F0) is found to rise in a word and fall to the next word. The presence of this fall is proposed as a means of detecting word boundaries. Pitch frequency is the rate of vibration of the vocal chords during production of speech. Acoustic features like short time energy, short time pitch frequency and short time zero crossing rate are used along with the Boundary Confidence Coefficient (BCC)

parameter which is based on the lexical information of the frame. These lexical and acoustic features are used to train the Hidden Markov Model (HMM) based classifier, to detect word boundaries.

Shyamal Kumar Das Mandal [27] proposed a method of detecting word boundaries by using supra-segmental features for Bengali language. Bengali is a stress-bound language. While detecting word boundaries for Bengali language, the speaker's Intonation, Stress and rhythm of the language were considered. After their research, they found that 87.8 % of word boundaries have been correctly detected. The drawback of this method is that, it can be used only for stress bound language.

JurajKacur and GregorRozinaj [28] discussed a method for Word Boundary detection using cepstral matrices. They proposed that the method is efficient for detecting word boundaries, in noisy environments. In this method, speech signals are divided into consecutive segments, which are overlapped by certain parts of their length to keep the feature vectors behavior smooth in time and moreover not losing any speech information.

Ramana Rao and Srichand [29] proposed a method for word boundary detection using pitch variation, which determines the highness and lowness of a tone in a spoken utterance for Hindi, Bengali, Marathi, Telugu and German. They used the frequency of pitch, for each word, and found that more than 85 % of word boundaries were correctly determined for the Indian languages, and about 65 % word boundaries were correctly determined for German language. The authors attributed the low detection rate in German to the fact that the misses occurred more often between words with only one vowel in that particular word and the next word.

Anurag Jain [30] developed an algorithm for Word Boundary Detection in Hindi Language which was based on intensity and pitch. Their algorithm was designed to work in a noisy environment by considering three prosodic parameters which are "defined pitch contour", "undefined pitch on silence zone" and "intensity contour".

3.3 ISSUES IN WORD BOUNDARY DETECTION

The most important issue of word boundary detection is to correctly identify the boundaries of words in a continuous spoken sentence. Once the word boundaries are identified, the sentences can be broken up into a series of words and these words

can be used as input for speech recognition. The word boundary detection problem is non-trivial. For detecting the word boundaries, the listener has to correctly place the boundaries, so that the perception and understanding of the sentence is correct [25]. The different issues in WBD are given below:

a. As word boundary detection is a non-trivial task and it is language dependent. So we cannot apply a common method for all languages. The developed method must use some universal characteristics of the languages rather than features that are language specific.

b. The spoken utterance is a combination of some connected words, a phrase or a sentence. We cannot impose any constraint on the length of the speech input. So a method used should be effective over all continuous speech input.

c. Continuous speech does not offer any explicit clues for placing the word boundaries. During the conversation mode the speaker does not make any pause between words to signify the word boundaries. So speech signal related clues must be used for remarking the word boundaries.

The segmentation of isolated words of a continuous speech for a listener, listening to mother language, seems to be easy, but the difficulty arises when one listens to a language not known to him/her. This difficulty in segmenting words occurs because in speech, no prominent marking is present between words. There are a number of potential sources of information that could be used as indicator of word boundaries which influence speech. This includes: "*metrical stress cues, Phonotactics cues, context-sensitive allophones, Co-articulation cues, and Statistical / distributional properties*" [31].

3.4 ISSUES IN ESTIMATION OF Fo

The general problem of fundamental frequency estimation is to take a portion of signal and to find the dominant frequency of repetition. Thus the different difficulties that arise in the estimation of fundamental frequency are:

a. All signals are not periodic.

b. Those that are periodic may be changing in fundamental frequency over the time of interest.

c. Signals may be contaminated with noise, even with periodic signals of other fundamental frequencies.

d. Signals which are periodic with interval T are also periodic with interval 2T, 3T etc., So we need to find the smallest periodic interval or the highest fundamental frequency and even signals of constant fundamental frequency may be changing in other ways over the interval of interest

3.5 VOICED,UNVOICED & SILENCE CLASSFICATION

Voice, unvoiced and silence detection plays a vital role in word boundary detection. So before proceed to word boundary detection classification of the signal is very important. As different language has different ways to start and end, so techniques like pitch, energy and intensity etc. helps to detect the boundary in stress bound language. For voiced, unvoiced, silence classification segmentation, ZCR, STE, hamming windows are prominent parameter [32]. These techniques are not widely addressed in Bodo language for classification.

For analysis of these speech features we have collected a number of phonetically balanced sentences in Bodo language. All the sentences spoken were declarative sentences with lengths varying from five to eight words. The average length of the sentences was about four to five words. The speech data was digitized at 16bits/sample at a sampling rate of 44 kHz and used channel Mono. In order to achieve good quality recording, recorded speech corresponding to a syllable was examined for the following defects:

- Distortion arising from clippings,
- Aliasing via spectral analysis,
- Noise effects arising from low signal amplitude-resulting in evidence of quantization noise or poor signal-to-noise ratio (SNR),
- Large amplitude variation,
- Transient contamination (due to extraneous noise).

Recorded speech samples that have one or more of the above listed defects are discarded and the recording repeated until the resulting speech signal is satisfactory. To prepare the recorded waveform for further processing, we replayed each speech

sound to ensure that they are correctly pronounced. Artifacts introduced at the end and beginning of each recording was edited out.

Table 3.1: Parameters and Specification of Corpus

Sl No	Parameter	Specification
1	Sampling rate	44kHz
2	Frame size	10ms
3	Window type	Hamming
4	Window length	32 ms (512 samples)
5	Window overlap	22 ms (352 samples)
6	Analyses	Short term spectrum
7	Set no of channel	1
8	Wave form format	mswav

Table 3.2: Example of Some Typical Bodo Sentences

Sl No	Bodo sentence
1	Baishaguwa asomi harini furbow
2	Mushwa nokhari fichijagra junta
3	Afaya railao sakari mawo.
4	Rama sache gnani gotha
5	Kajirangaya gandani thakhai mungdangkha.
6	Mumbaia munche falangini nwgwr.
7	Bharatni rajdhania gwdan Delhi.
8	Ang dinoi collegeao thanga.
9	Asoma Prakritik sampatjung bungfabnai gongshe raijw.
10	Moidera mache gidir junta.

In the following sections discussed some methods in terms of Bodo language for voiced, unvoiced and silence classification. This analysis is also required for word boundary detection.

3.5.1 SEGMENTATION

Speech segmentation is the process of identifying the boundaries between words, syllables, or phonemes in spoken natural languages. The term applies both to the mental processes used by humans, and to artificial processes of natural language processing. The knowledge of acoustic features in particular voiced and unvoiced segment plays an important role in many speech analysis systems. Segmentation of speech into voiced, unvoiced and silence could help in marking the candidate word boundaries in the speech. Voiced or Unvoiced speech, followed by silence could be a candidate of word boundary and in the similar way; silence followed by the voiced or unvoiced speech could be the candidate of word boundary [33]. Based on the characteristics of the voiced and unvoiced speech listed in **Table 3.3**, many algorithms were developed to detect voiced and unvoiced speech.

Table 3.3: Differences in Voiced and Unvoiced Speech

Voiced Speech	Unvoiced Speech
Energy per frame is much higher	Energy per frame is less
Less zero crossings	More Zero crossings
Quasi periodic	Noise like

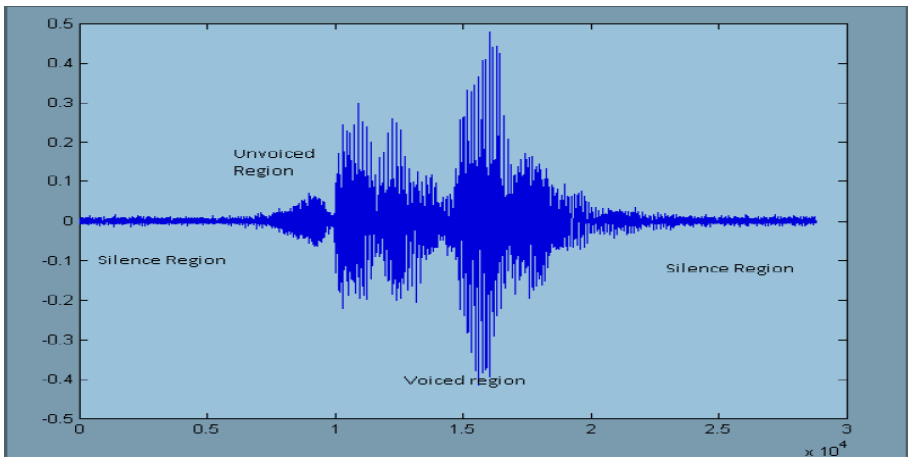

Figure 3.1: Voiced, Unvoiced and Silence Detection using Segmentation

But there are some problems associated with segmentation method. These are:

a. Due to the overlapping of unvoiced and voiced speech, it is very difficult to identify their exact end points.

b. There is an ambiguity with low energy voiced speech, like whisper. It can be easily mistaken as unvoiced speech.

To overcome this problem, Zero-Crossing-Rate (ZCR), Short-Time-Energy (STE), hamming window and fundamental frequency is used to detect the voiced, unvoiced and silence region.

3.5.2 ZERO CROSSING RATE

ZCR is a basic acoustic feature that can be computed easily. It is equal to the number of zero-crossing of the waveform within a given frame. ZCR is an important phenomenon to classify silence, voiced and unvoiced parts of a signal. With the help of zero crossing count we can indicate the frequency at which the energy is concentrated in the signal spectrum. Due to the excitation of vocal tract by the periodic flow of air at the glottis, voiced signal is produced. It usually shows a low zero-crossing count [34]. The unvoiced speech is produced by the constriction of the vocal tract narrow enough, to cause turbulent airflow, which results in noise and shows high zero-crossing count. There is high energy in voiced part of the speech signal because of its periodicity and low energy in unvoiced part. The mathematical model of ZCR is [35]

$$zcr = \frac{1}{T-1}\sum_{t=1}^{T-1} 1_{\mathbb{R}_{<0}}(s_t s_{t-1}) \qquad 3.1$$

Where S is the signal of length T and $1_{R<0}$ is an indicator function.

The mathematical model for short term ZCR defined by [35]

$$Z_n = \sum_{m=-\infty}^{\infty} 0.5|sgn\{x[m]\} - sgn\{x[m-1]\}|w[\hat{n} - m] \qquad 3.2$$

Where $sgn\{x\} = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases}$

A typical single with its ZCR is given in **Figure 3.2.**

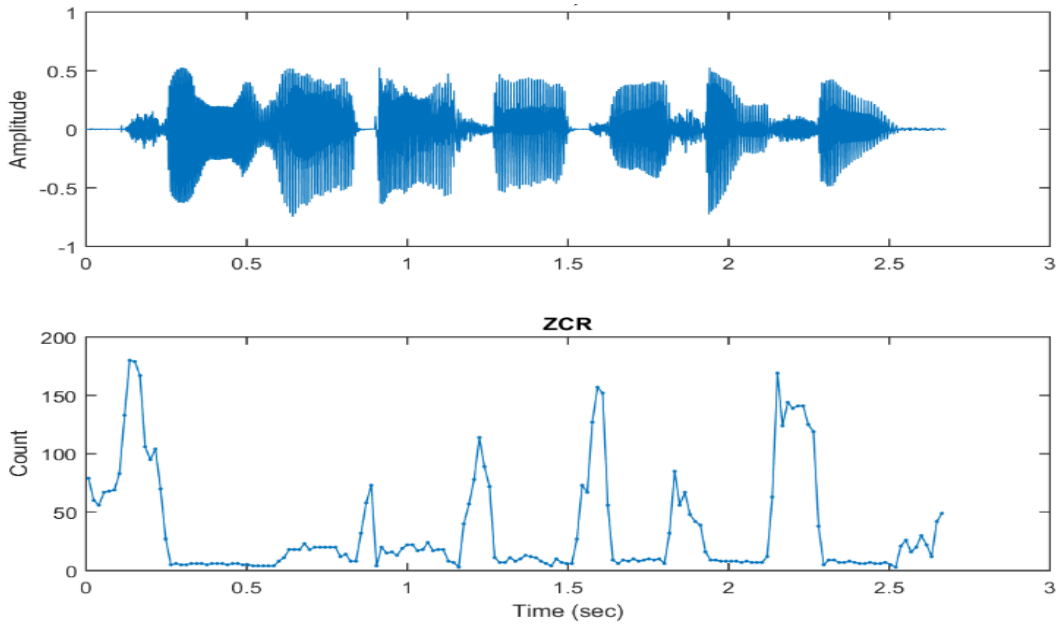

Figure 3.2: A Typical Signal with ZCR

ZCR of unvoiced sounds and environmental noise are usually larger than voiced sounds, which has observable fundamental periods. It is hard to distinguish unvoiced sounds from environmental noise by using ZCR alone since they have similar ZCR values. There for shifting of waveform is required before computing the ZCR. ZCR is often used in conjunction with energy for end-point detection. But in realistic environment is not highly unreliable unless a further refined procedure is taken for post-processing.

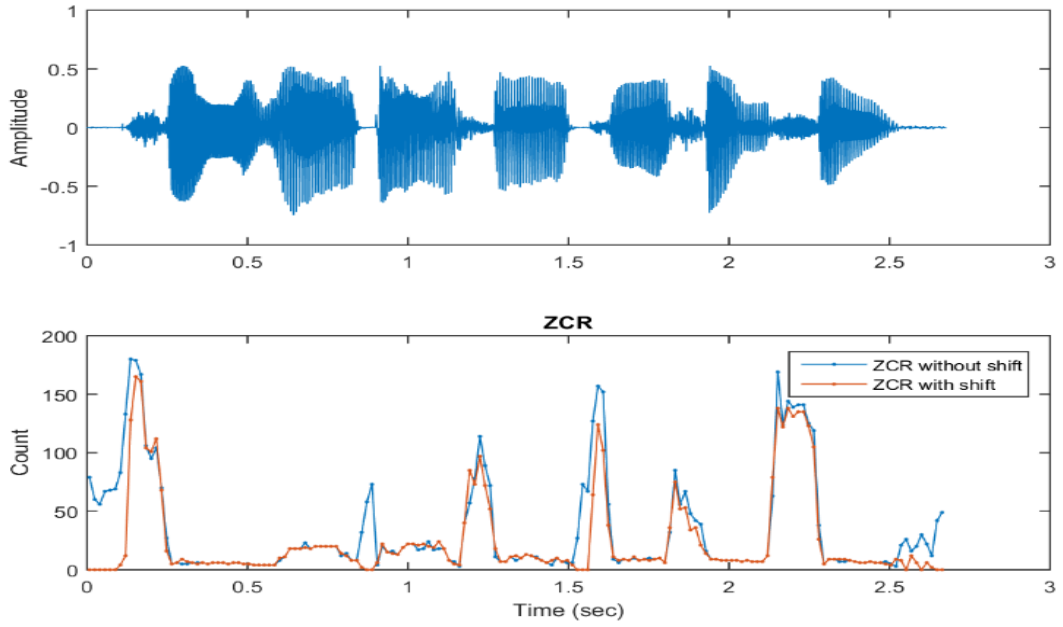

Figure 3.3: ZCR with shift and without shift.

Following figures gives the ZCR of voiced and unvoiced Bodo signal

Figure 3.4: ZCR of a Voiced Signal

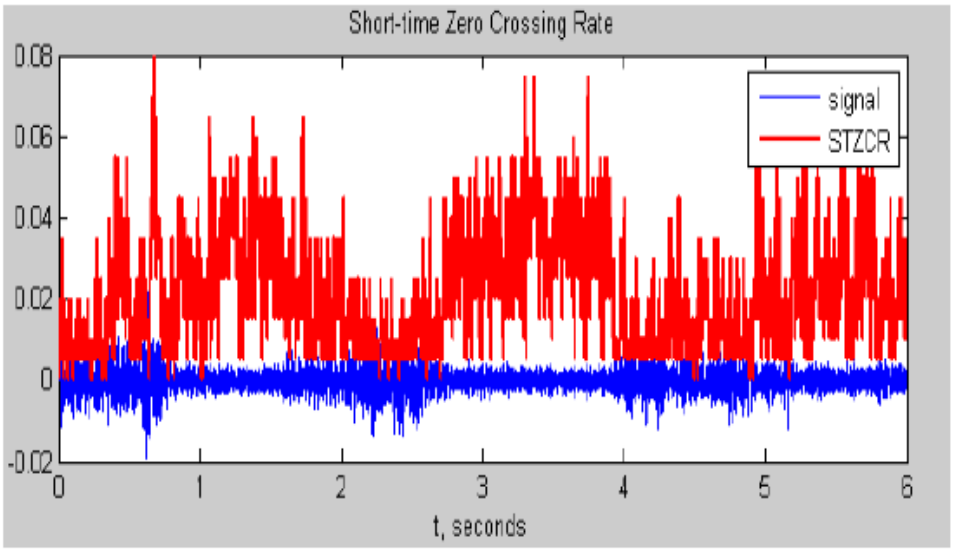

Figure 3.5: ZCR of an Unvoiced Signal

3.5.3 SHORT TIME ENERGY

The amplitude of the speech signal varies with time. Generally, the amplitude of an unvoiced speech segment is much lower than the amplitude of voiced segments. STE can compute the number of voiced segments but it cannot compute the phonetic content of the speech. STE can be defined as [32]

$$E_n = \sum_{m=-\infty}^{\infty} [x(m)w(n-m)]^2 \qquad 3.3$$

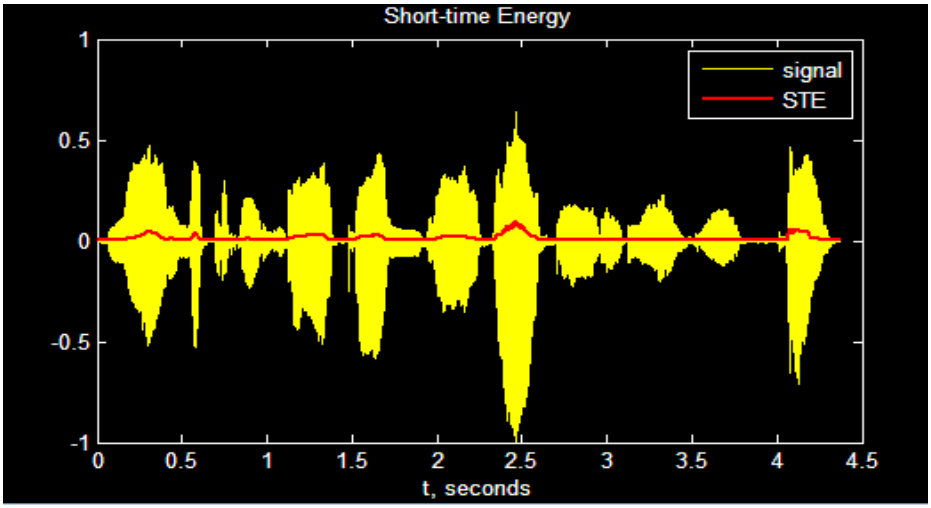

Figure 3.6: STE of a Typical Signal

3.5.4 HAMMING WINDOW

The main aim of windowing in spectral analysis is the ability of zooming into the finer details of the signal rather than looking at the whole signal as such. Short Time Fourier Transforms (STFT) are of prime importance, in case of speech signal processing, where the information like pitch or the formant frequencies are extracted by analyzing the signals through a window of specific duration. The width of the windowing function relates to how the signal is represented, i.e., it determines whether there is a good frequency resolution or a good time resolution [33]. A wide window gives better frequency resolution but poor time resolution. A narrower window gives good time resolution but poor frequency resolution. This is the exact reason as to why a wavelet transform was developed, where a wavelet transform is capable of giving good time resolution for high frequency events and good frequency resolution for low frequency events. This type of analysis is well suited for real signals.

3.6 PROPOSED METHOD

After the observation of basic speech properties following methods are proposed to classify the voice/ unvoiced and silence region of a speech signal. **Figure.3.7** demonstrates the flow chart of voiced/unvoiced classification [30].

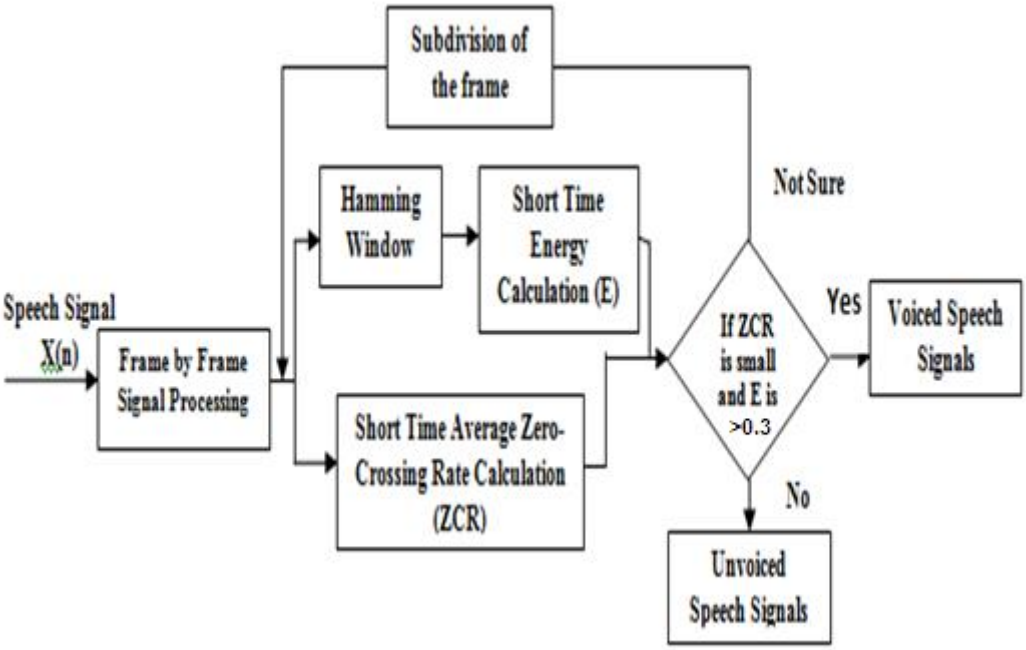

Figure 3.7: Voiced -Unvoiced Classification

Above mentioned voiced/unvoiced classification algorithm uses short-time zero-crossing rate and energy of the speech signal. In the frame-by-frame processing stage, the speech signal is segmented into a non-overlapping frame of samples. It is processed into frame by frame until the entire speech signal is covered. It has 3600 samples with 44 KHz sampling rate. In the beginning, we set the frame size as 160 samples. At the end of the algorithm, if the decision is not clear, energy and zero-crossing rate is recalculated by dividing the related frame size into two frames. It has been also observed that for classification, the number of times the speech signal crosses the zero mark is considered; more times the crossing, it is a voiced speech of higher frequency content and vice versa for unvoiced.

The problem with above algorithm is that, classification of silence speech is not possible. To overcome the problem and to classify the voiced, unvoiced and silence part we have taken the help of fundamental frequency to classify the speech sample. The procedure of the method is as follows:

Procedure

1. Read a speech single of length =n and create speech vector SV(n)
2. Calculate the ZCR vector(ZC)) taking hamming window size=20 ms
3. Calculate Fundamental frequency (Fo) of the single taking hamming window size=20 ms

4. Map Fo vector to the vector of length similar to that of speech signal.
5. calculate STE taking hamming window= 50 ms
6. Map STE vector to vector of length similar to that of speech signal.
7. calculate threshold value(T_value) for STE
8. Make output vector (OUT) of length n and initialize all its values to zero
9. Repeat for i=1 to n

    If ZC (i) and STE (i)>T_value the set OUT (i) =0.1 for voiced

        Else if ZC (i) between 0.1 and 0.3 and STE (i) <T_value

        If Fo (i) =0 then set OUT (i) =0.2 (for silence)

        End if

    Elseif ZC (i)>0.3

        If Fo (i) =0 then set OUT (i) =0.3 (for voiced)

        End if

    End if

10. Plot speech vector and OUT.

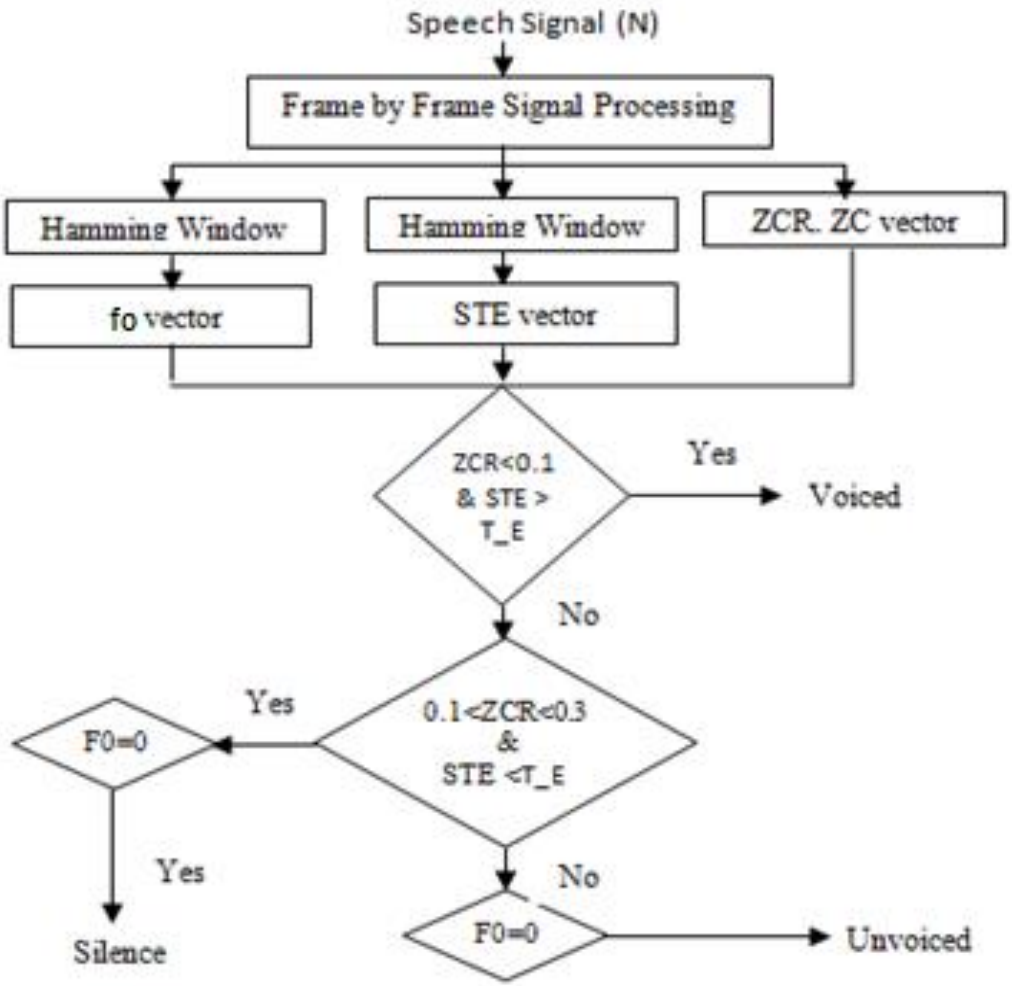

Figure 3.8: Voiced-Unvoiced-Silence Classification

The algorithm was applied to all the words present in the sentence spoken in Bodo by 10 persons (5 male and 5 female). Accuracy of the algorithm was calculated by checking how many samples in the spoken word were identified correctly compared to the manual classification of the voiced, unvoiced and silence region in the word and then dividing them by total no of samples. The accuracy of the algorithm for four different speakers taking all the 5 words spoken3 times is shown in **Table 3.4, 3.5, 3.6 and 3.7**

Table 3.4: Accuracy of first speaker (male)

Word Spoken	Acc. in 1st Time utterance	Acc. in 2nd Time utterance	Acc. in 3rd Time utterance	Avg. Accuracy

Asoma	98.4	98.4	98.4	98.4
Prakritik	92.6	92.6	92.6	92.6
sampatjung	98.9	98.9	98.9	98.9
bungfabnaigongsha	95.4	95.4	95.4	95.4
raijw	96.4	96.4	96.4	96.4
Average accuracy of second speaker				96.34

Table 3.5 Accuracy of second speaker (male)

Word Spoken	Acc. in 1st Time utterance	Acc. in 2nd Time utterance	Acc. in 3rd Time utterance	Avg. Accuracy
Asoma	98.2	98.56	96.7	97.82
Prakritik	94.42	88.2	92.3	92.3
sampatjung	98.4	89.9	92.3	93.53
bungfabnaigongsha	96.7	95.62	96.9	96.41
Raijw	98.7	98.34	99.18	98.74
Average accuracy of second speaker				95.76

Table 3.6 Accuracy of third speaker (male)

Word Spoken	Acc. in 1st Time utterance	Acc.in 2nd Time utterance	Acc. in 3rd Time utterance	Avg. Accuracy
Asoma	96.5	98.18	97.7	97.7
Prakritik	97.5	97.73	98.18	97.80
Sampatjung	99.47	99.47	97.67	98.79
Bungfabnaigongsha	98.7	95.67	97.3	97.3
Raijw	98.3	98.57	98.18	98.35
Average accuracy of second speaker				97.98

Table 3.7 Accuracy of fourth speaker (female)

Word Spoken	Acc. in 1st Time utterance	Acc. in 2nd Time utterance	Acc. in 3rd Time utterance	Avg. Accuracy
Asoma	95.5	95.5	95.5	95.5
Prakritik	96.85	96.85	96.85	96.85
Sampatjung	95.36	95.36	95.36	95.36
Bungfabnaigongsha	94.74	94.74	94.74	94.74
Raijw	94.64	94.64	94.64	94.64
Average accuracy of second speaker				95.41

This algorithm is efficient in solving the problem of identifying the unvoiced, voiced and silence chunks in speech. Three fundamental features namely: ZCR, STE and F0 are used in the algorithm for the classification purpose and an accuracy of 95.41 % is achieved. The errors in the system are mainly in the starting and the ending of the word due to little noise or lower energy during the starting and ending of the word. But the main drawback of the algorithm is that the performance is degraded under noisy conditions.

3.6.1 ANALYSIS OF THE PROPOSED METHOD

ZCR rate is computing by finding the number of times the sound signal crosses the level zero within the analysis frame. We assume threshold value is equal to 0.06 ms. The duration between two letters less than 0.06 ms, we consider it to be one letter. When we get the duration less than 0.035 then it is considered as conjunct consonant. Steps involved in this process:

- Take input: We take the recorded files and separate the data into some blocks. Each block size is considered as N.
- Block data: Each block data consists of 600 samples and the length of the block data is 30 ms. then next block is considered by leaving 200 samples from the first block and it is incremented by 200 samples. Similarly the whole data is divided into blocks.
- Strip data: The block data are now separated into strips by strips by removing leading blocks.

- Zero cross data: It shows the number of times the input data crosses the zero level.

We tested our recorded utterances of the ten speakers, which have shown variation in their outcomes. This is because each speaker's accent is different. In the experimental studies, we have analyzed that different speakers have different correction rates and rejection rate due to the pronunciation tone of Bodo. In Bodo language there are so many words which end with a letter of long vowels. Because of which the duration between two words reduces and the algorithm may not detect the boundary. The reason of rejection rate of ZCR is given below:

a. In case of zero crossing we found that when one word ends the intensity value is decreased and crosses the zero line then we consider it as one word. But in Bodo language there are many points where intensity value is decreased but does not touch the zero line because of the tone of pronunciation. If the end tone level of ending point and beginning point is high then it does not crosses the zero line. Because of what the algorithm cannot detect it as boundary.

b. There are many points where the intensity value is decreased but every such point is not word boundary. This can be happened when in one letter is low and the tone of next letter is high.

c. Again there are many points where intensity value almost cross the zero level but every such point is not a word boundary. This is happened when the ending point is ended with long vowels.

3.7 ENERGY

Energy is one of the basic acoustic features in speech. It is mainly used in the Voice Activity Detection (VAD). It is easy to mark the word boundaries using other techniques if one can detect the voice activity. In 2008 Li et al used the speech energy edge information which is the change in energy of the incoming speech. It has a capability of detecting voicing activity in low SNR environments. There are two steps, the first step is to employ an optimized edge filter to identify the sudden rise or fall in the energy is defined in the equation below [36]:

$f(x) = eAx[K1\ Sin(Ax) + K2cos(Ax)]+ e-Ax\ [K3\ sin(Ax)+K4cos(Ax)] +$

K5 + K6eπ (3.4)

Edge (g) = $\sum_{i=-w}^{w} f(i)g(t+i)$ (3.5)

Where A and Ki are parameters of the edge filter, s is a positive constant, g is the speech energy, t is the current frame number and W is the window size. This filter is used to compute speech energy and obtain the edge information. In the next stage a four state finite state is applied with silence, starting speech, during speech and ending speech as its four states. Either silence or starting speech can be the starting point of the incoming speech data and then each frame will be classified into one of the four states of the transitions state diagram as shown as shown in **Figure. 3.9**, where Gap is the minimum frame number between the detected end points. Emax is the average energy value of the adjacent frames; F is the feature value of the current frame. The starting and ending boundaries of the speech are detected using this algorithm, which also identifies the background noise and silence.

Energy can be an important parameter in identifying the boundaries, and voicing activity. However, for automatic speech segmentation, it is difficult to segment the speech reliably on energy changes as these cues often are unreliable due to co-articulation. However it is an important acoustic feature and will produce good results when combined with other techniques like pitch detection in identifying the boundaries [32].

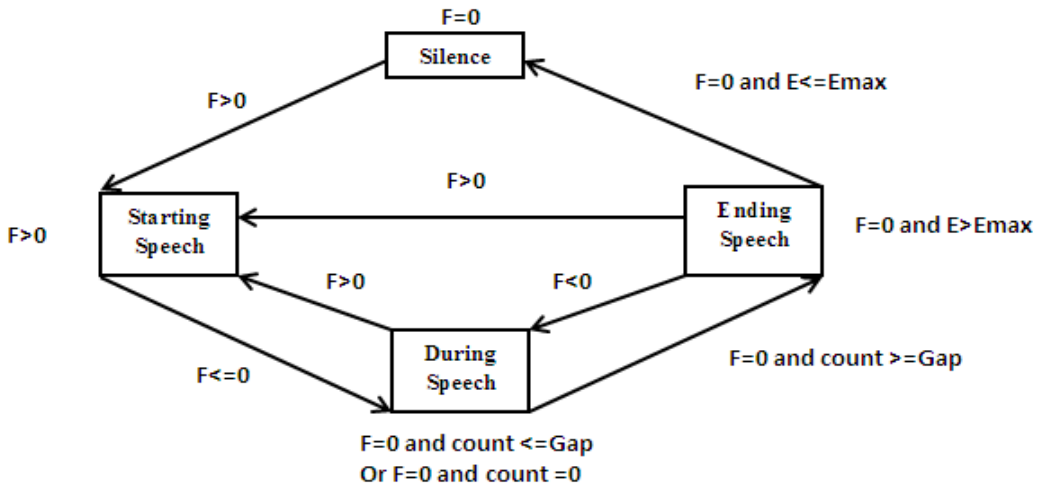

Figure 3.9: Four Sate of Finite State Diagram

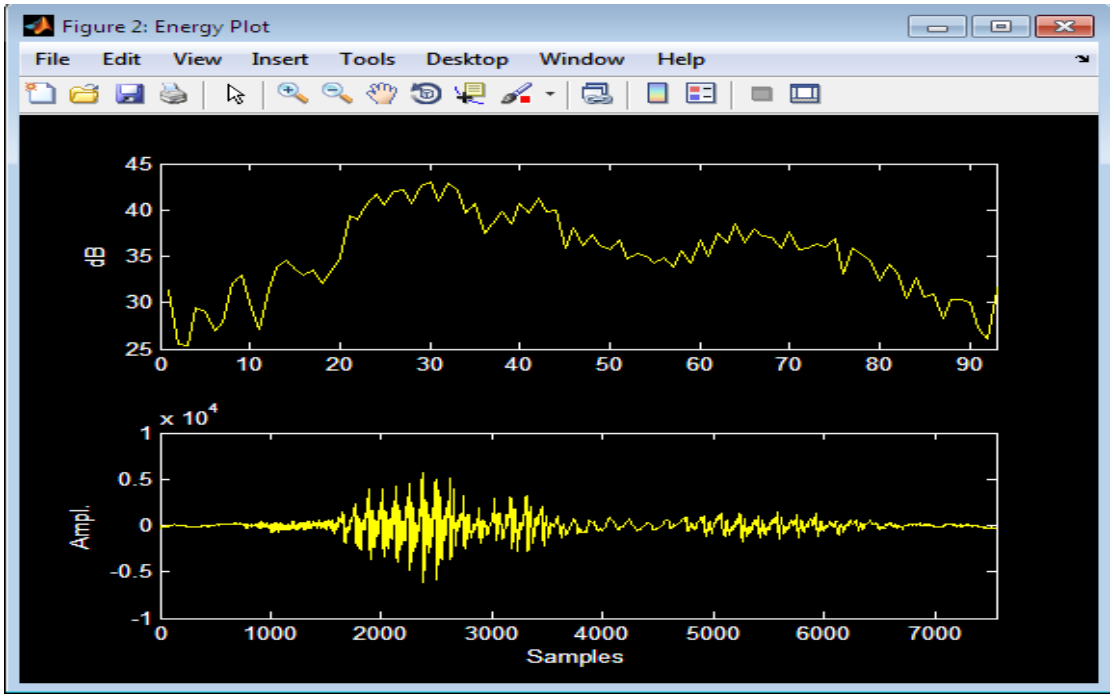

Figure 3.10: Energy plot of a "furbow"

3.7.1 ANALYSIS USING ENERGY

The changes in energy at the word boundaries were observed to obtain the cues to identify them. From **Figure 3.11** it can be observed that, at the word boundaries, amplitude has sudden rises or falls. The increase in amplitude from the previous frame to the current frame is taken as the cue for word boundary; also a decrease in energy from the previous frame to the current frame is taken as cue to identify the word boundary. A threshold is identified for each utterance, which is mean of energy of the speech file, was used to determine the word boundaries.

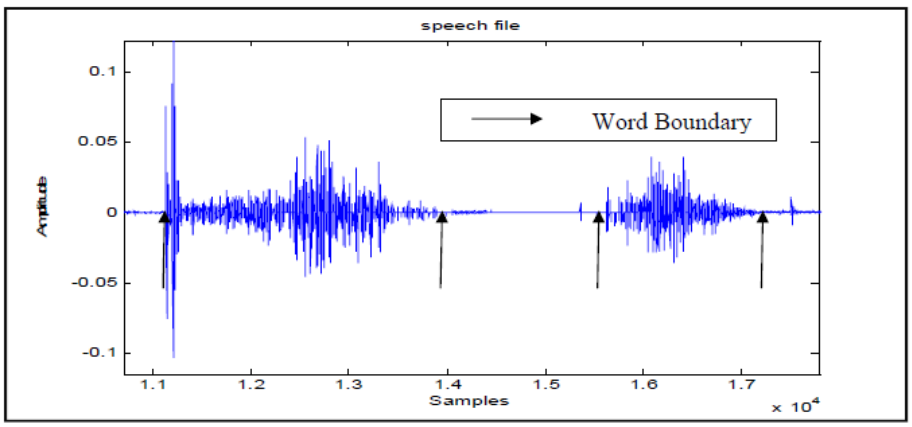

Figure 3.11: Word boundaries identified using transcriptions

Procedure:

a. Each speech file is divided into frame with frame size 20ms

b. Calculate energy [33] of the frame with the equation

$$E(n) = \sum_{i=1}^{k} s(i)^2 \quad \textbf{3.6}$$

Where

- E(n) denotes the energy of frame
- n, s denotes the speech signal
- k denotes the frame length.

c. Calculate the threshold value which is the mean of the energy of speech file.

d. Now find the word boundary based on the following criteria and equations:

- Word boundary has a rise in energy from the previous frame which can be the beginning of the word:

if (**E(i-2)<= t** and **E(i-1)<t** and **E(i)>=t** and **E(i+1)>t** and **E(i+2)>t**) **3.7**

Where **i** is the possible place of the word boundary which is marked as a candidate word boundary, where **E(i)** is the energy of frame **i** and **t** is the threshold of energy

- Word boundary has fall in energy from the previous frame which can possibly be the end of the word: if (**E(i-2)> t** and **E(i-1)>t** and **E(i)<=t** and **E(i+1)<t** and **E(i+2)<t**) **3.8**

Where **i** is the possible place of the word boundary which is also called candidate word boundary, and **E(i)** is the energy of frame **i** and **t** is the threshold of energy. The energy of each frame along with the speech file is shown in **Figure. 3.12**

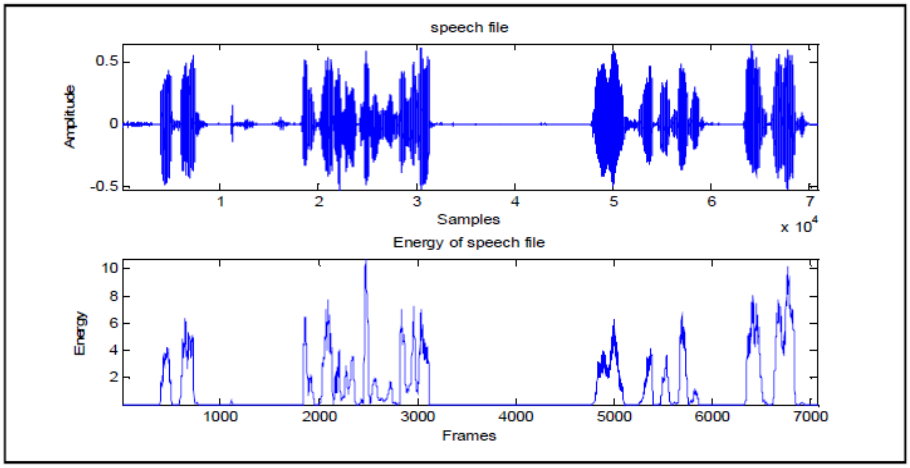

Figure 3.12: Speech file (top panel) and energy of speech file (bottom panel)

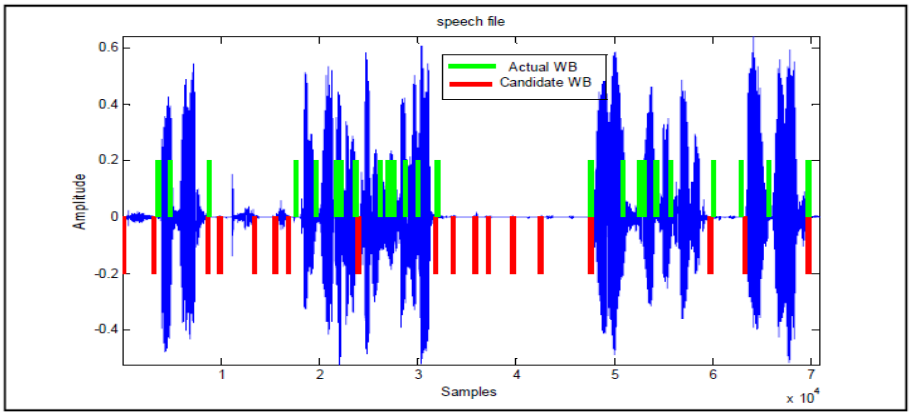

Figure 3.13: Shows the actual word boundaries (green) and candidate word boundaries (red) which were determined using energy

Strengths of energy in identifying the word boundaries are [34]:

- It is very good at identifying the word boundaries with well-defined word boundaries.
- It helps in confirming the predictions of the word boundaries using other techniques.
- The hits of energy are more closely identified to the actual word boundaries than Pitch.

Drawbacks of energy in identifying the word boundaries are:

- The word boundaries are identified only if there is a well-defined boundary. If there is no well-defined boundary between two words, this technique results in misses.

- It also identifies some of the boundaries of the noise as word boundaries which results in false alarms.

3.8 PITCH

The opening and closing of the vocal folds that occur during speaking break the air steam into chains of pulses. The rate of repetition of these pulses is the pitch and it defines the fundamental frequency of the speech signal .In other words, the rate of vibrations of the vocal folds is the fundamental frequency of the voice. Relative differences in the fundamental frequency of the voice are utilized in all languages to study the various aspects of linguistic information conveyed by it. The estimation of pitch and formant frequencies finds extensive use in speech encoding, synthesis and recognition. In adult, generally the length of vocal folds in male is more than that of female counterpart [36]. The more is the vocal fold length, less is the pitch frequency. Thus the pitch differs in male and female informants. Figure 3.14 shows the pitch counter of Bodo vowel.

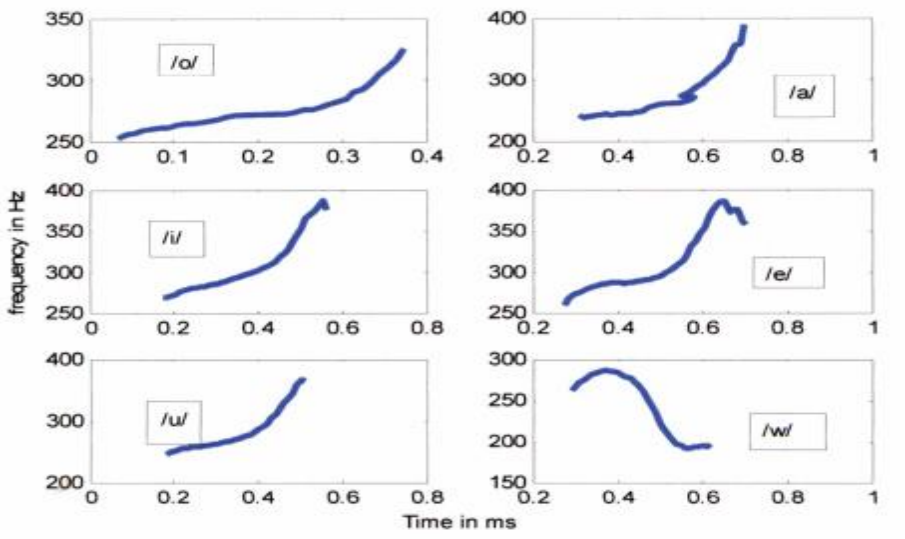

Figure 3.14: Pitch Contour of Bodo Vowel

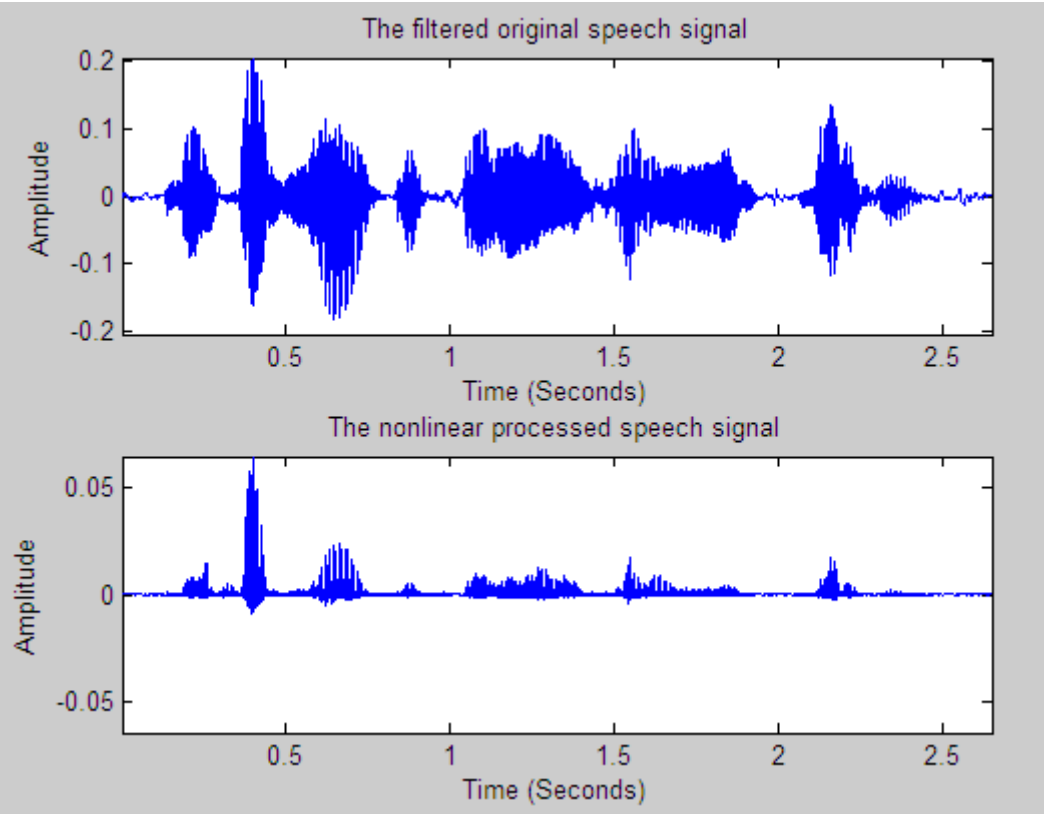

Figure 3.15: Nonlinear Processed Bodo Speech Signal

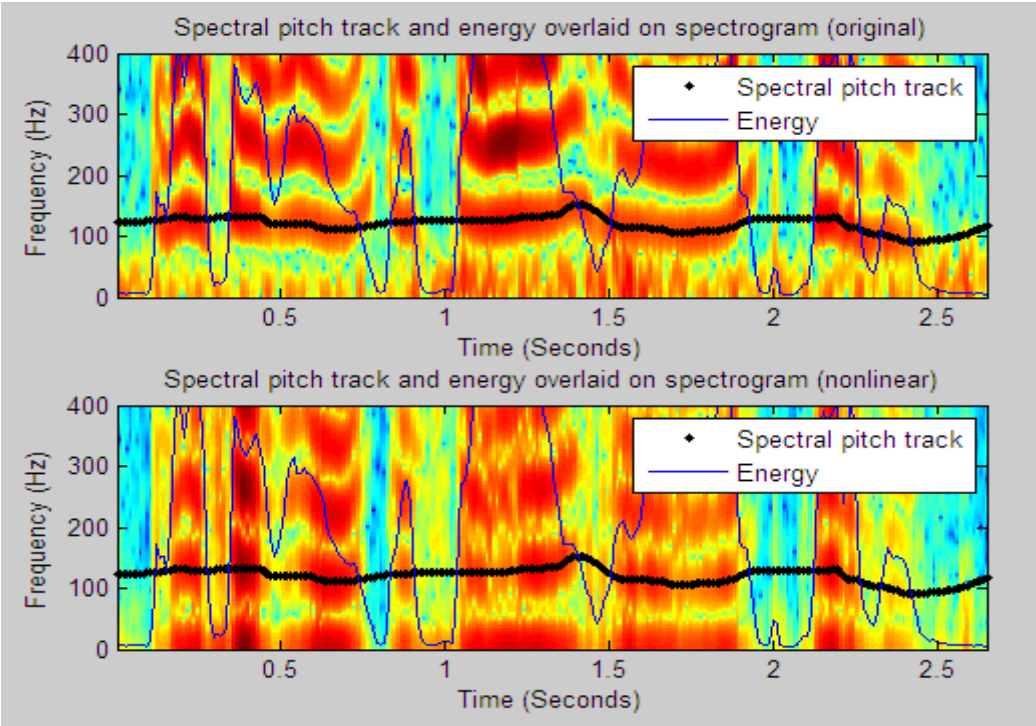

Figure 3.16: Spectral Pitch Track and Energy Overlaid on Spectrogram

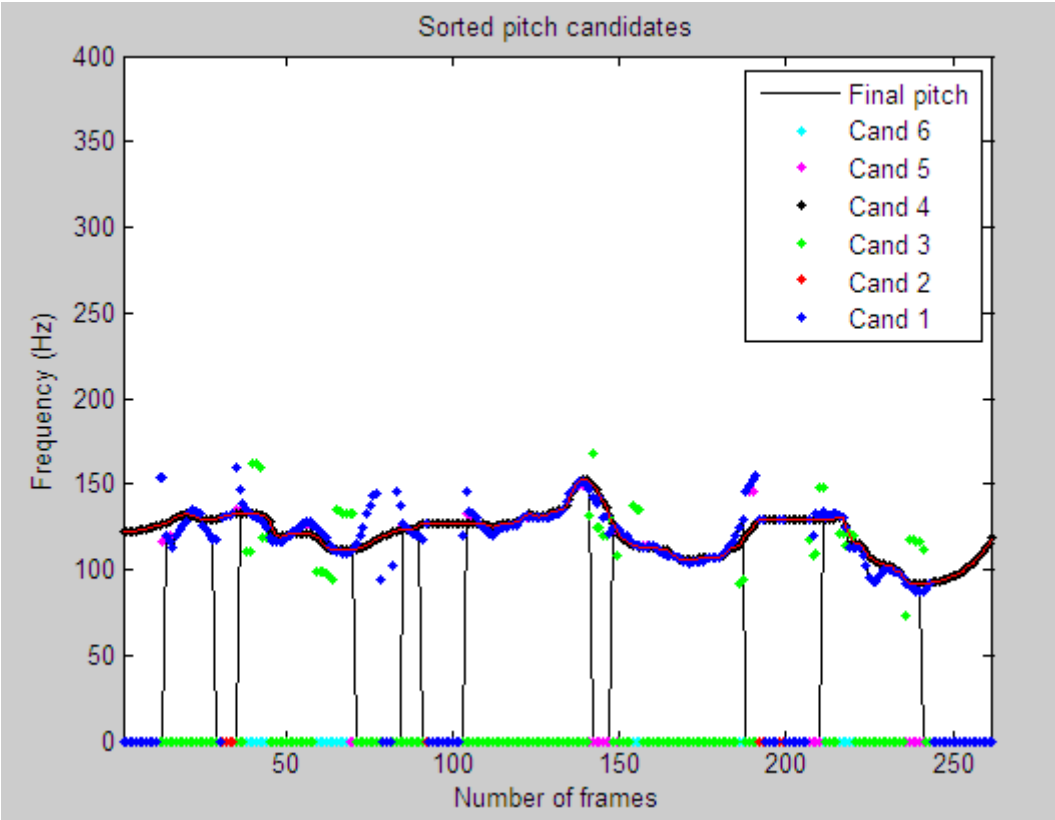

Figure 3.17: Pitch Candidates of a Bodo sentence

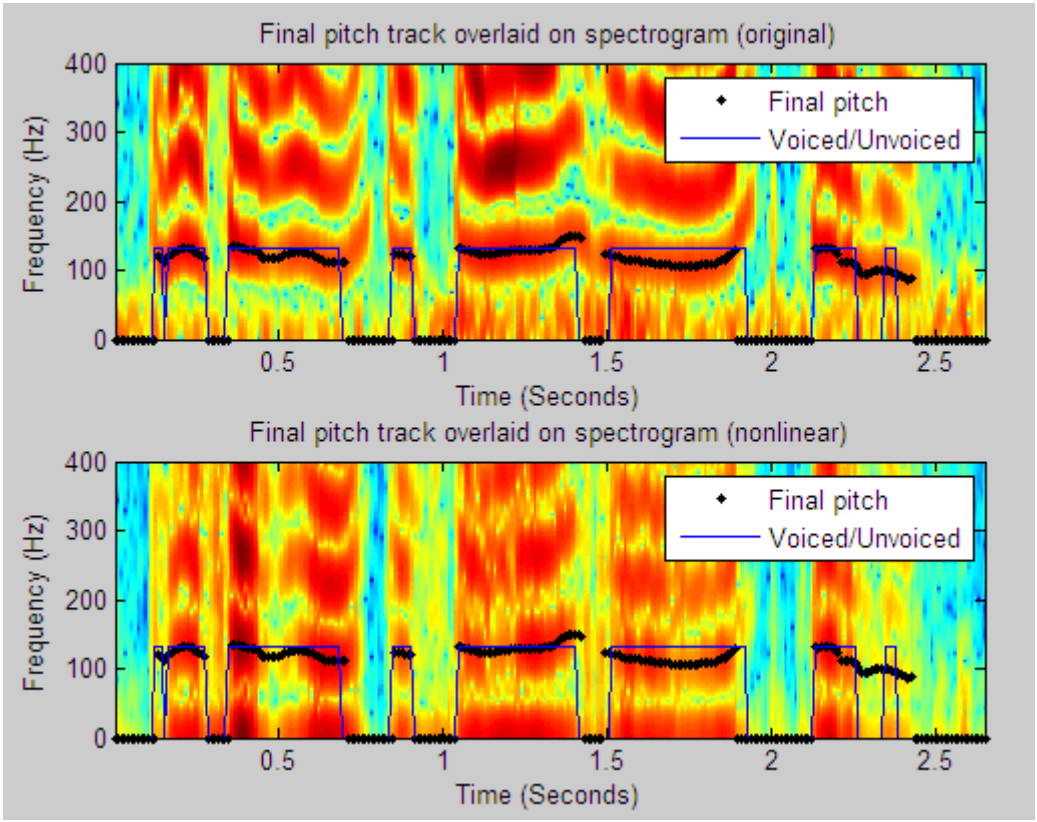

Figure 3.18: Voiced –Unvoiced separation of a Bodo sentence

Cestrum is the inverse Fourier transform of the logarithmic spectrum of the signal. Cestrum of voiced speech contains strong peaks corresponding to pitch periods. The typical graphical representation of Fo with autocorrelation and Capstrum for a Bodo word is shown at **Figure 3.19** and **Figure 3.20** respectively.

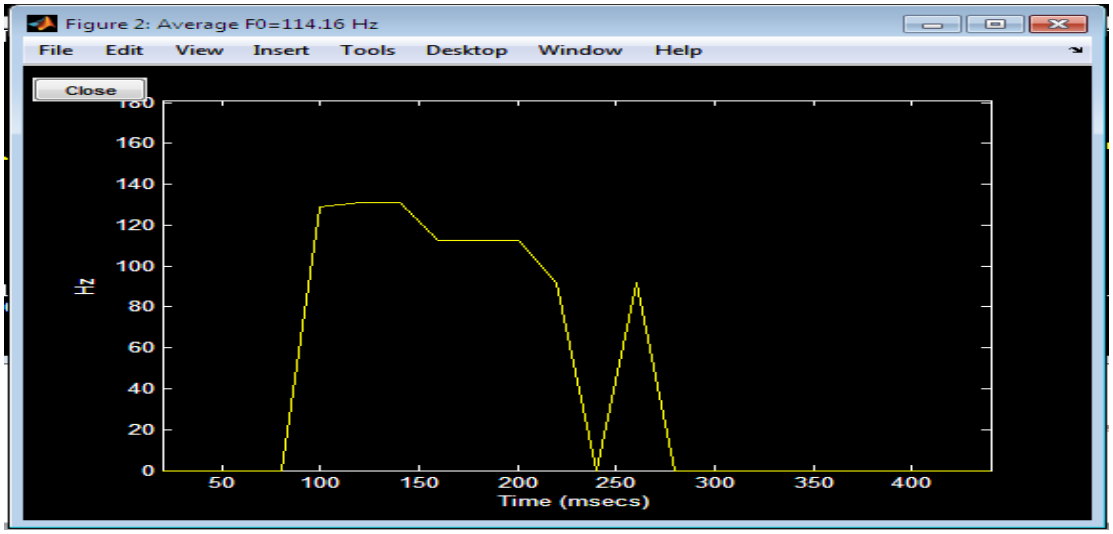

Figure 3.19: Fo Contour with Autocorrelation Method

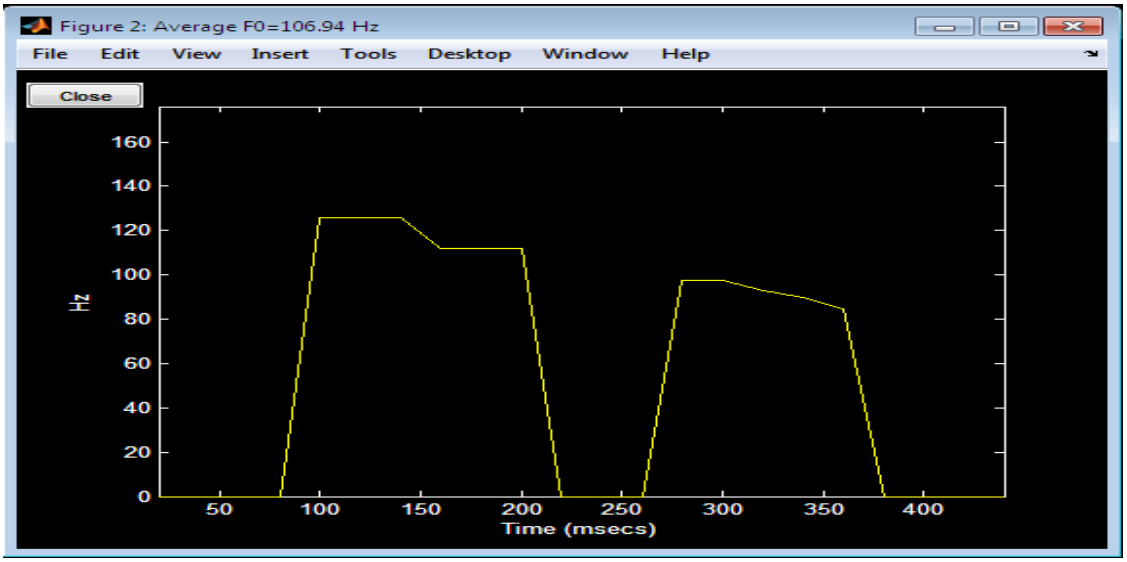

Figure 3.20: Fo contour with capstrum approach

3.8.1 ANALYSIS USING PITCH

We make the following observations based on the pitch analysis of the Bodo speech corpus. In our study, we have used Fo variations to identify word boundaries.

Referring to **Figure 3.21**, which is the pitch plot of a sentence spoken in Bodo. The sentence is "**Baishaguwa asomia harini furbow"**. It can be seen that there is a fall in the Fo value from a high of 300 Hz to a low of 175 Hz from the beginning to the end of the utterance. And within this overall fall, there are local Fo rises and falls across the words in the sentence.

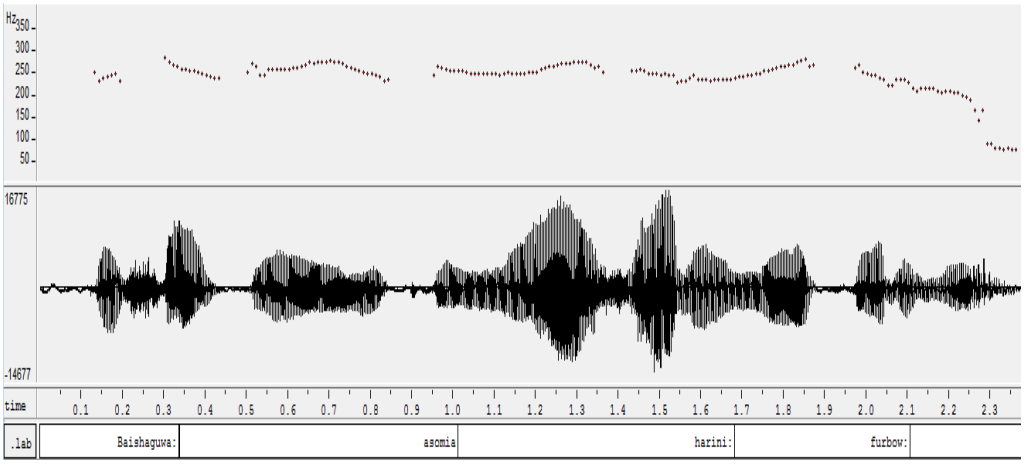

Figure 3.21: Bodo speech files (Baishaguwa asomia harini furbow)

We made the following observations based on the intensity curve of the Bodo speech file:

- **a.** Intensity value decreases at the point where one word ends and new word is started
- **b.** The decrease of intensity cannot be marked as a word boundary as intensity may decrease in case of some syllables also.
- **c.** Sometimes we have several low intensity points where pitch value is defined and there may be word boundary in this region also.

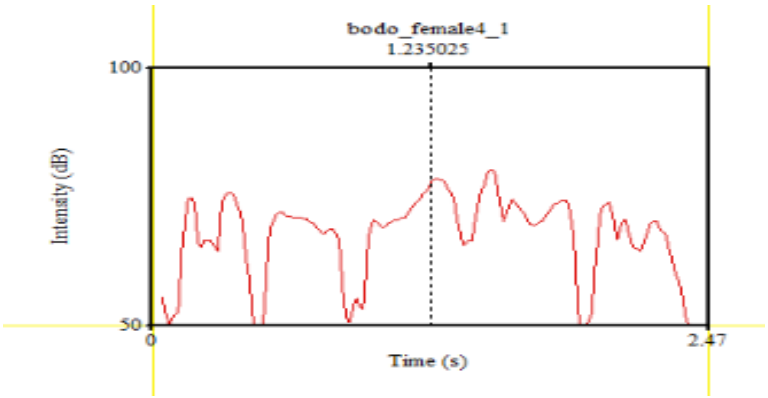

Figure 3.22: Intensity curve of the Bodo speech file

It is observed that there is local Fo rises (peaks) and falls (valleys) present. The Fo contour displays some characteristic properties across words, which we call the local properties. The Fo contour starts at a low value at the initial syllable of a word. The contour rises steadily to a peak, which is the final syllable of the word. Thus across the words, there exists a rising pattern of the Fo pattern. Syllables are formed with vowels being their nucleus, and it during the articulation of voiced speech that pitches is observed. Henceforth, we refer to vowels instead of syllables. After rising to the word final vowel in a word, the Fo contour falls to the word initial vowel of the next word in the sentence. Thus a valley occurs between the last vowel of a word, and the first vowel of the next word, i.e. at the word boundary. Some of drawbacks of pitch in identifying the WB are:

a. If the word is beginning with unvoiced speech There is some offset between actual and predicted word boundaries

b. Some false alarms may occur due to the pitch reset due to the prosodic units in the same word.

SUMMARY

In speech analysis, the voiced-unvoiced decision is usually performed in extracting the information from the speech signals. In this work, we use Fo, ZCR and STE to separate the voiced-unvoiced-silence parts of speech from a speech signal. Here, we evaluated the results by dividing the speech sample into some segments and used the zero-crossing rate and energy calculations to separate the voiced and unvoiced parts of speech. The results suggest that zero-crossing rates are low for voiced part and high for unvoiced part whereas the energy is high for voiced part and low for unvoiced part. Pitch, intensity and energy features also analyzed to detect the word boundary. The spectrographic representation of pitch track also shown to voiced-unvoiced classification of the speech signal.