A COMPATIVE STUDY OF SILENCE AND NON SILENCE REGIONS OF SPEECH SIGNAL USING PROSODY FEATURES FOR EMOTION RECOGNITION

J. Naga Padmaja
Assistant Professor of CSE
KITS, KHAMMAM
srija26@gmail.com

R. Rajeswara Rao
Professor of CSE, JNTUK- UCEV.
VIZIANAGARAM
rajaraob4u@gmail.com

Abstract

The objective of this work is the comparative study of the speech signal between the silence and non- silence regions of the speech signal. In this work our main goal is to observe the pitch contour, energy and duration are time-varying and also study how these changes play an important role in emotion recognition. An important step in emotion recognition from speech is to select significant features which carry large emotional information about the speech signal. It was given that emotion recognition from speech has different types of features, among them is prosody, spectral and acoustic features. Sometimes prosody features are called supra-segmental features. It deals with the auditory qualities of the sound and it can also reflect aspects of meaning, intention and emotional state of the characters [1] [2].Prosody Feature consists of more pitch information which is used in identifying the emotion such as Pitch, Energy, and Duration. In this work we also explored the importance of the speech signal which doesn't have silence regions and how the signal varies due to of pitch contour, energy, duration values, to analyze their contribution towards the recognition of emotion. The main intention of this work was to utilize the speech properly by means of actual speech content, i.e., with no other silence parts or unnecessary parts of the signal.

Keywords- Pitch, Energy, Duration, Prosody Features, silence regions.

I. INTRODUCTION

An emotion is a human state of a mental behavior which expresses the feeling by physical movements or by words. As the physical movements are the facial expressions and the body language, words are the way they speak and the way they pronounce the words. Humans are very good at hiding the feelings from others at that particular situation human couldn't judge at what state the person is really in, but in speech, humans can, by the pronunciations of a word and stress on a particular letter. Even if they try to hide they couldn't do as longer as they want, because at some time they could be caught, so selecting the system to do on speech have to be given the highest priority.

Speech is the most natural form of human communication. Speech is one of the most information-laid signals; speech sounds have a rich and multi-layered temporal-spectral variation that convey words, intention, expression, intonation, accent, speaker identity, gender, age, style of speaking, state of health of the speaker and emotion. Speech sounds are produced by air pressure vibrations generated by pushing inhaled air from the lungs through the vibrating vocal cords and vocal tract and out from the lips and nose airways. The air is modulated and shaped by the vibrations of the glottal cords, the resonance of the vocal tract and nasal cavities, the position of the tongue and the openings and closings of the mouth. Speech signals contain information like

intended message, speaker identity and emotional state of speaker. An important issue in speech emotion recognition is to determine a set of important emotions to be classified by an automatic emotion recognizer.

Speech is an immensely information-rich signal exploiting frequency-modulated, amplitude-modulated and time-modulated carriers (e.g. resonance movements, harmonics and noise, pitch intonation, power, duration).

Speech is a sequence of elementary acoustic sounds or symbols known as phonemes that convey the spoken form of a language. Speech signals convey much more than spoken words. The information conveyed by speech is multi-layered and includes time frequency modulation of such carriers of information as formants and pitch intonation. Formants are the resonances of vocal tract and pitch is the sensation of the fundamental frequency of the opening and closings of the glottal folds.

Normally human beings use dynamics of long term speech features like energy profile, intonation pattern, duration variations and formant tracks, to perceive and process the emotional content from the speech utterances. In this work we proposed the method that removes the silence from the speech signal, while recording the speech there may have the noise and other unrelated information like silence regions in the speech signal due to this reason we cannot get results accurately. Generally the speech signal is classified as three parts: that is, voiced regions. In this region the vocal cords are vibrated then a speech sound is produced. Actually this is very useful for emotion recognition. Another classification is unvoiced region. In this region, vocal cords are not vibrated so the resulting speech is random in nature like sounds of whisper or aspiration [1]. The final classification is the silence. In this region, energy and the amplitude of the signal is very low so we can remove this silence parts [ ]. Here in this work we believed that removing the silence parts from the speech signal will give better results to identify the emotion and also break the signal into three parts like initial, middle and final of the words for observing the pitch contour, energy and duration values are changes over the time. Through this observation, we made a decision on how prosody features play an important role in emotion recognition based on the changes over in the speech signal which will be studied independently for the speech with silence regions and speech without silence regions. If we observe the pitch contour, energy and duration of the each individual emotional speech signal might vary through the signal processing .For example, if we considered the anger wav file and sadness wav file, in that pitch contour, energy values are different compared to that in general. Anger has high pitch and energy values compared to sadness. The duration also tells us the difference of both speech files. Anger will be identified in short duration but whereas sadness will identified after the length of the speech signal utterance.

In the literature survey. Steepness of F0 contour during rise and falls, articulation rate number and duration of pauses are explored in Cahn (1990) Murray and Arnott (1995), for characterizing the emotion troughs in the profile of fundamental frequency and intensity, duration of pauses and bursts are prepared for identifying four emotion namely fear, anger, sadness and joy (McGilloway et al.2000). In Iida et al. (2003) explained a complex relation between the pitch, energy and duration parameters are exploited. Poonam Sharma et al [1]. Speech signal can be automatically divided into voiced, unvoiced and silence regions which are very beneficial in increasing the accuracy and performance of system recognition. In this they considered four different speakers of data for performance of the proposed method with the three features and overall accuracy of 96.61% is achieved. Tushar Ranjan Sahoo et al [2]. Here they proposed a composite silence removal technique comprising of short time energy and statistical method. The performance of the algorithm is highly appreciable in the presence of low SNR. The algorithm resulted in increase of accuracy by 20% when the silence is removed from the speech signal. J. Meribah Jasmine et al [3]. In the research work they done on the removal of noise from original speech signal and analyze various parameters. They used different methods like framing, windowing threshold values, and also for the removing of the silence by using the maximum envelope level. G.Saha et al [4]. Pre-processing of speech signal serves various purposes in any speech processing application. Apart from all silence/unvoiced portion removal along with end point detection is the fundamental step for application like speech and speaker recognition. This work shows better end point detection as well as silence removal than conventional Zero Crossing Rate (ZCR) and Short Time energy (STE) function methods.

This paper is organized as follows: Section-2 consists of Introduction and Database.Section-3 discusses about Prosody Features.Section-4 comprise information about of Signal analysis with results.Section-5 derives the conclusion about this paper .The last section list out the references to various papers.

II. DATABASE

Berlin emotional database comprises of ten utterances recorded by 7 emotions (anger, fear, neutral, Happiness, sadness, disgust, boredom). Ten professional native German actors (5 female and 5 male) simulated these emotions, producing 10 utterances (5 short and 5 longer sentences), which could be used in every-day

communication and are interpretable in all applied emotions [6]. The database considered of a total of 535 files. Out of which some files are considered for testing and remaining files for training the data models.

III. PROSODIC FEATURES

One of the most important parts of emotion recognition from speech a system is the feature extraction process, selecting the right features is crucial for successful classification. Speech signal is composed of large number of features which indicate its emotion contents. Changes in these features indicate changes in the emotions. Therefore proper choice of feature vectors is one of the most important tasks. There are many approaches towards automatic recognition of emotion in speech by using different feature vectors. Feature vectors can be classified as long-time and short-time feature vectors. The long time ones are estimated over the length of the utterance, while the short time feature vectors are determined over window size of usually less than 100ms. In this it has taken window size of 20ms and window shift 10ms.

Based on the acoustic correlates described in the previous section in the literature relating to automatic emotion detection from speech, we selected features namely the pitch, energy and duration. The pitch, energy, and duration are represented as contours.

In total, we selected 4 features which are used as a starting point for describing the variation between angry and neutral speech. For the extraction of the pitch contour, energy, duration.

a) Pitch:-

The pitch signal, also known as the glottal waveform, has information about emotion, because it depends on the tension of the vocal folds and the sub-glottal air pressure. The pitch signal is produced from the vibration of the vocal folds. Two features related to the pitch signal are widely used, namely the pitch frequency and the glottal air velocity at the vocal fold opening time instant. The time elapsed between two successive vocal fold openings is called pitch period T, while the vibration rate of the vocal folds is the fundamental frequency of the phonation F0 or pitch frequency.

Fundamental frequency is often processed on a logarithmic scale, rather than a linear scale, to match the resolution of the human auditory system. Normally, to that correlation function [22] for the centre clipped section is computed over a range of frequency from 50 to500 Hz for voiced speech (the normal range of human pitch frequency).

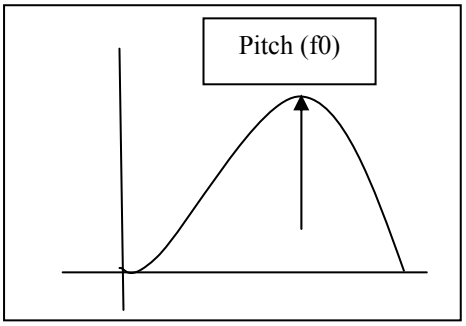

Fig. 3.1: Pitch on the Speech Signal

b) Energy:-

The amplitude of unvoiced segments is noticeably lower than that of voiced segments. The short time energy of speech signals reflects the amplitude variation and is defined [5] in equation 1:

$$1).\ E_n = \sum_{m=-\infty}^{\infty} x^2(m) \cdot h(n-m).$$

In order for E_n to reflect the amplitude variation in time(for this a short window is necessary), and considering the need for a low pass filter to provide smoothing, h(n) was chosen to be a Hamming window powered by 2. It has been shown to give good results in terms of reflecting amplitude variation. [5]

c) Duration:-

Duration was taken to be the length of each wav file. The speaking rate is good indicator of the emotional content in the speech.

IV. SIGNAL ANALYSIS

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.

Here in this work the main goal is to study the signal contours how the signals are varying due to the time and with respective of emotion and also observe that difference between the speech signal before the silence removal and after the silence removal .For this analysis we considered three basic prosodic features through this feature we can make a decision about the silence regions parts in the speech signal from this analysis we understand particular emotion will have the specific energy and pitch values and based upon the duration will decide the emotion. In this analysis considered Berlin database in 7 emotions (anger, fear, happy, neutral, sadness, disgust, boredom).Here we have shown the result of each emotion before and after silence was removed from the speech signals are represented by using the wavsurfer-1.8.5 speech recorded tool. These results are shown through the fig1 to fig13.

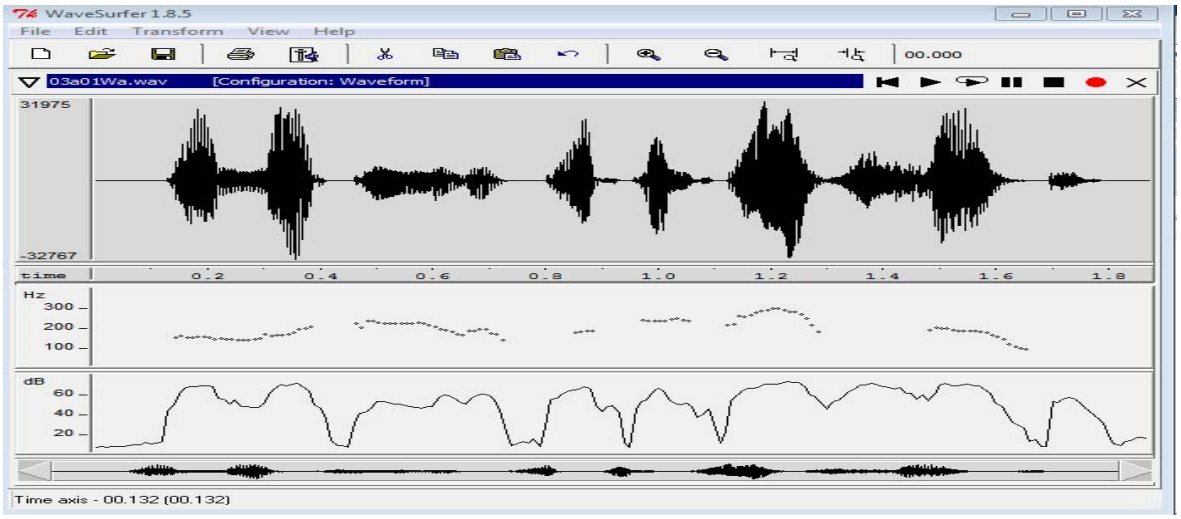

Fig. 1: Anger speech signal

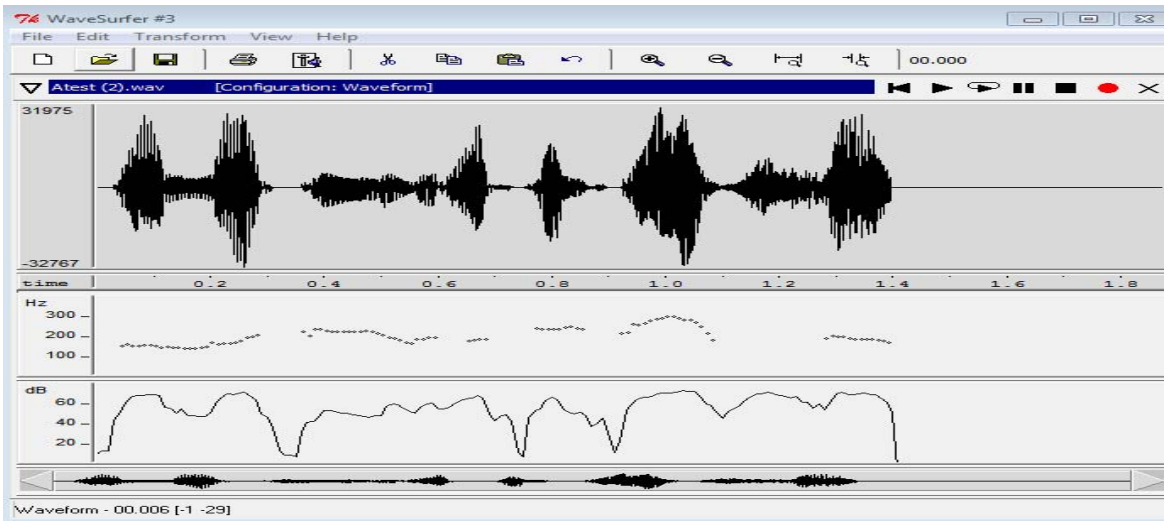

Fig. 2: Silence of Anger signal

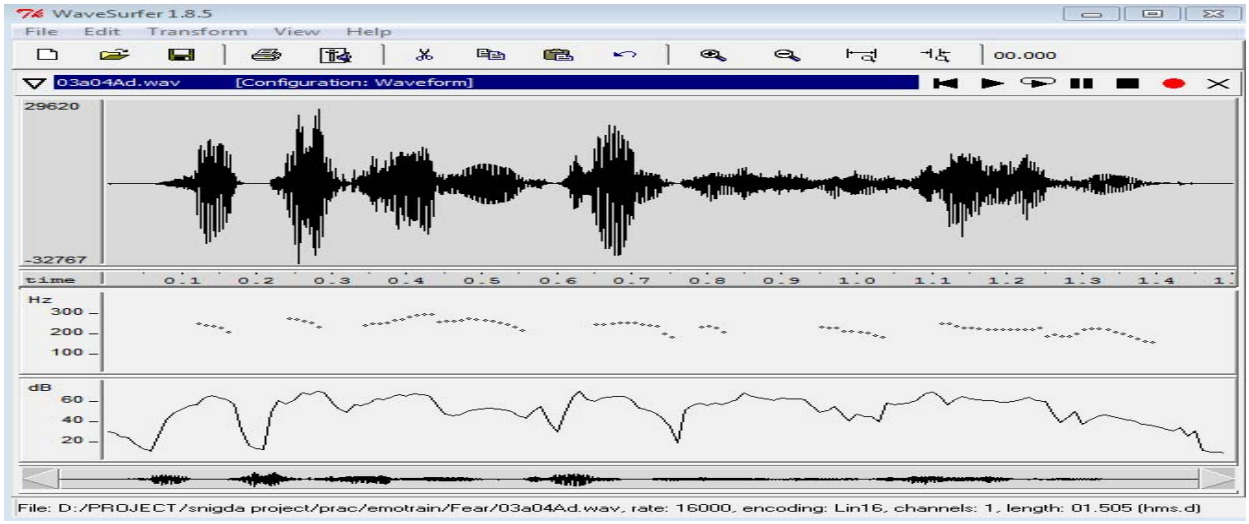

Fig. 3: Fear signal

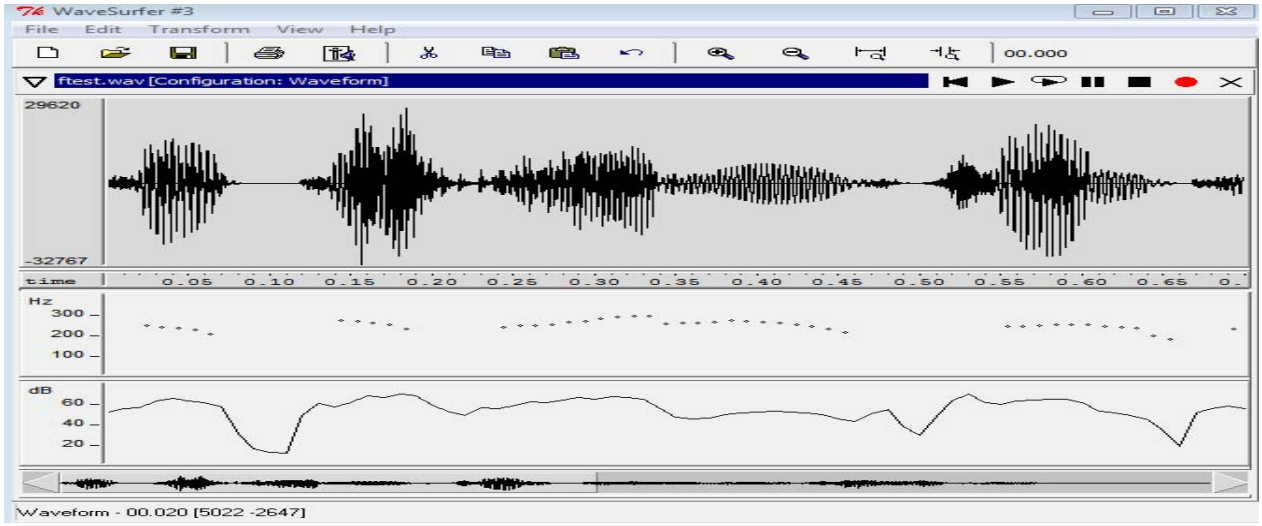

Fig. 4: Silence of Fear signal

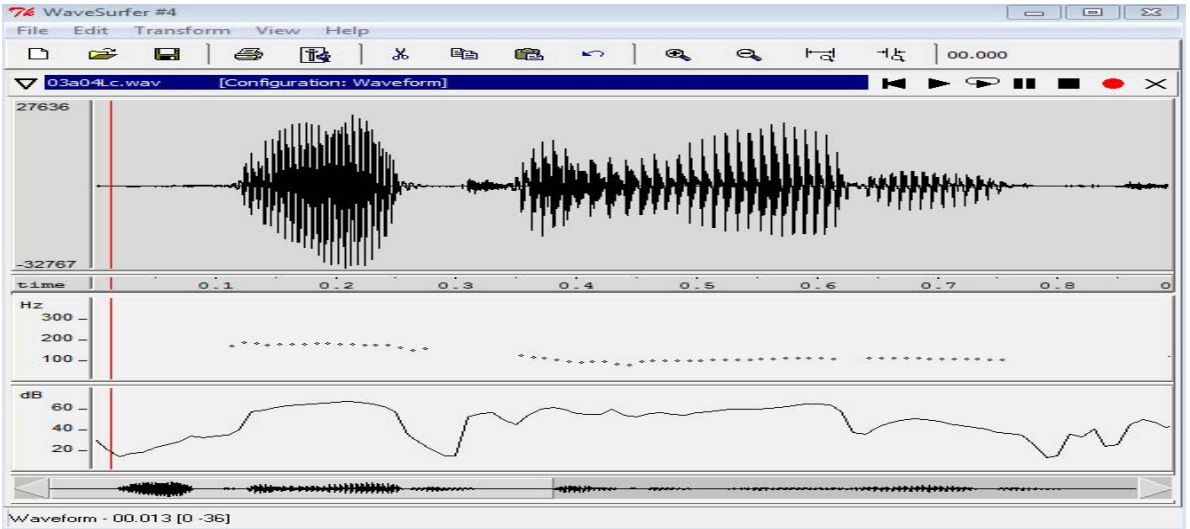

Fig. 5: Boredom signal

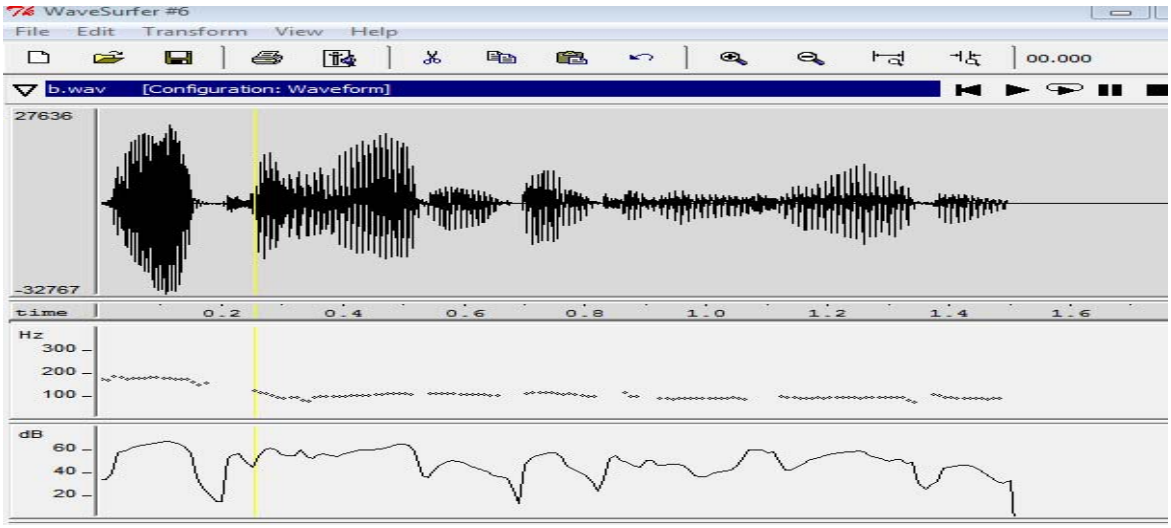

Fig. 6: Silence of Boredom signal

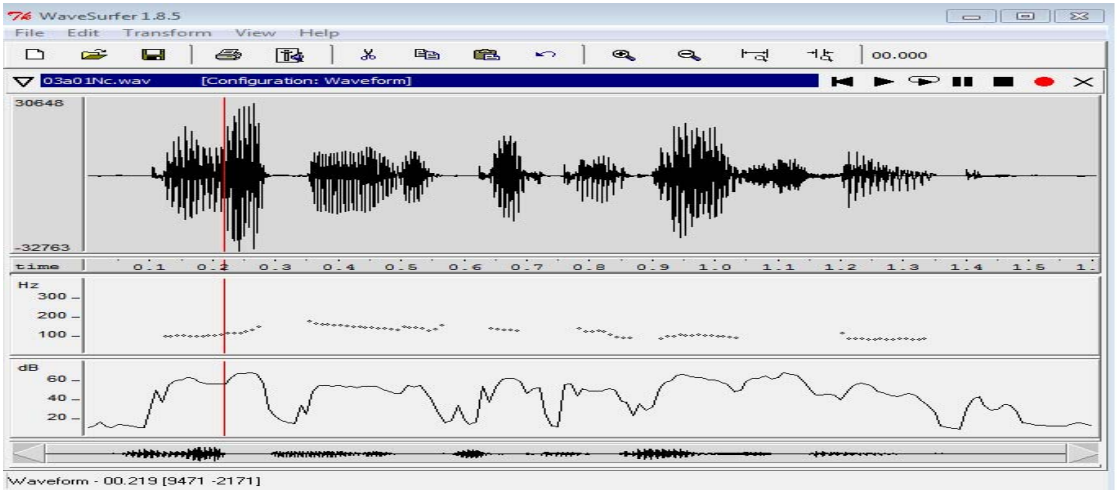

Fig. 7: Neutral signal

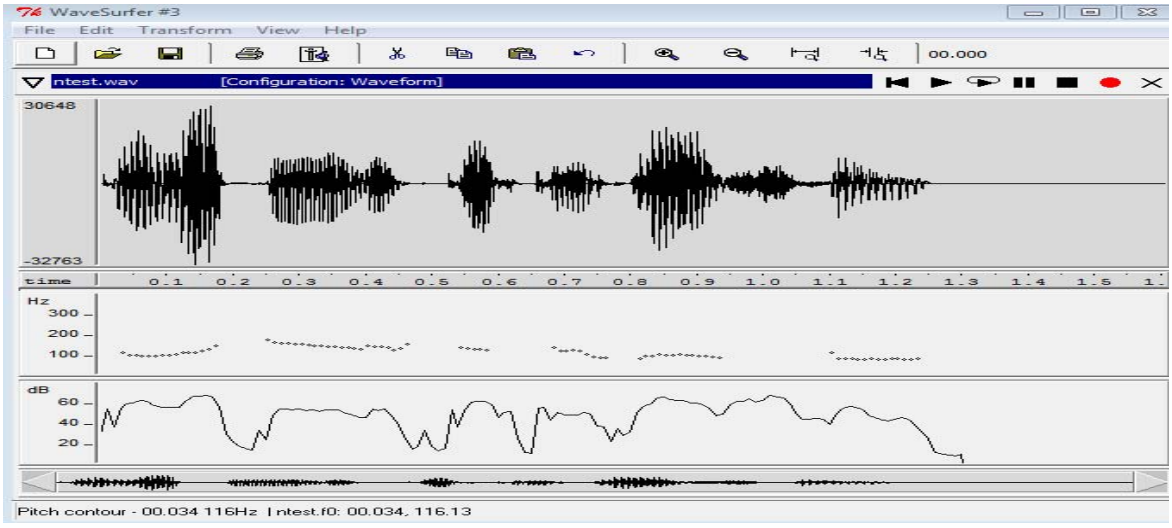

Fig. 8: Silence of Neutral signal

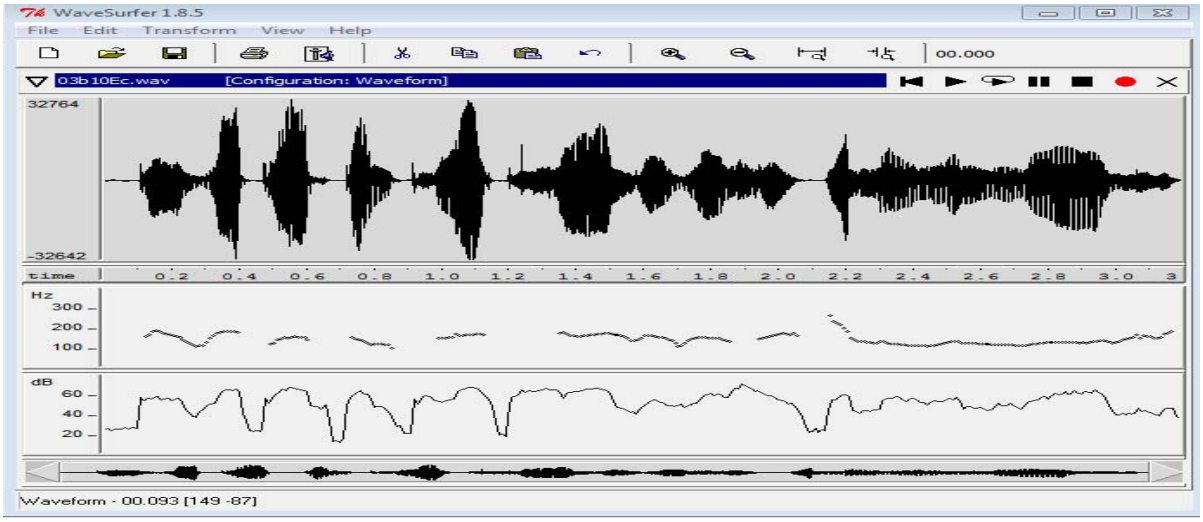

Fig. 9: Disgust signal

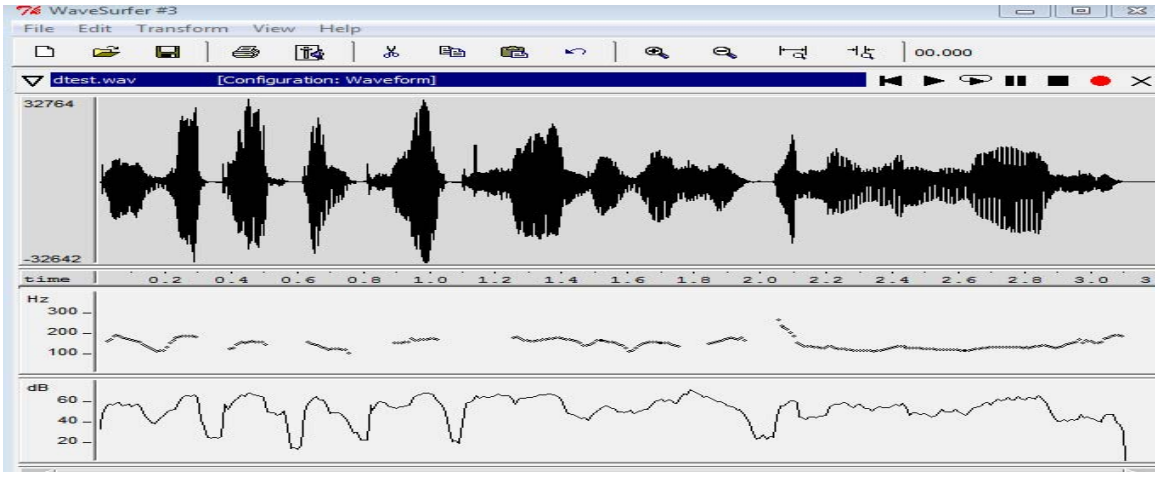

Fig. 10: Silence of Disgust signal

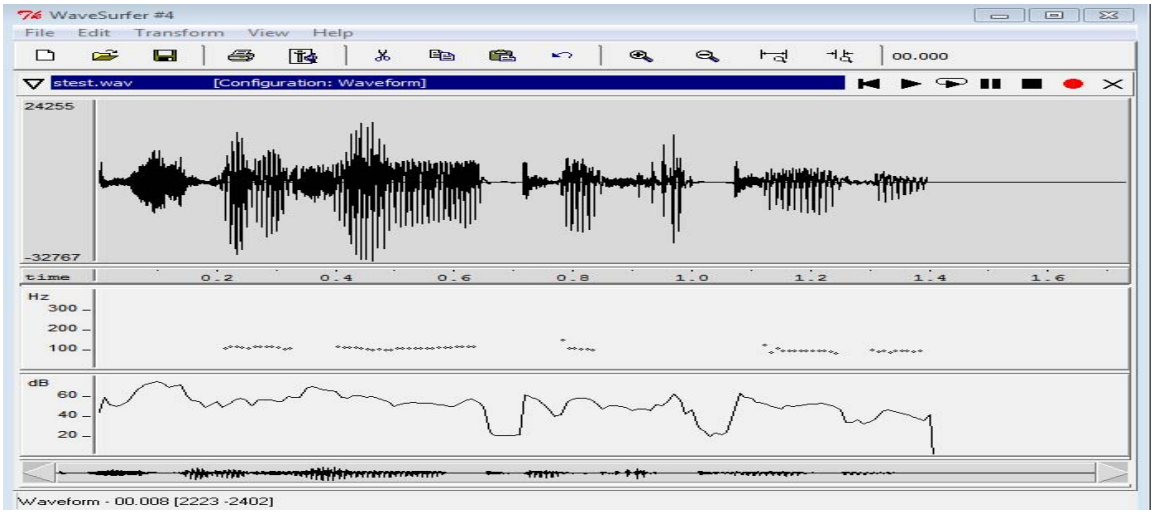

Fig. 11: Sadness signal

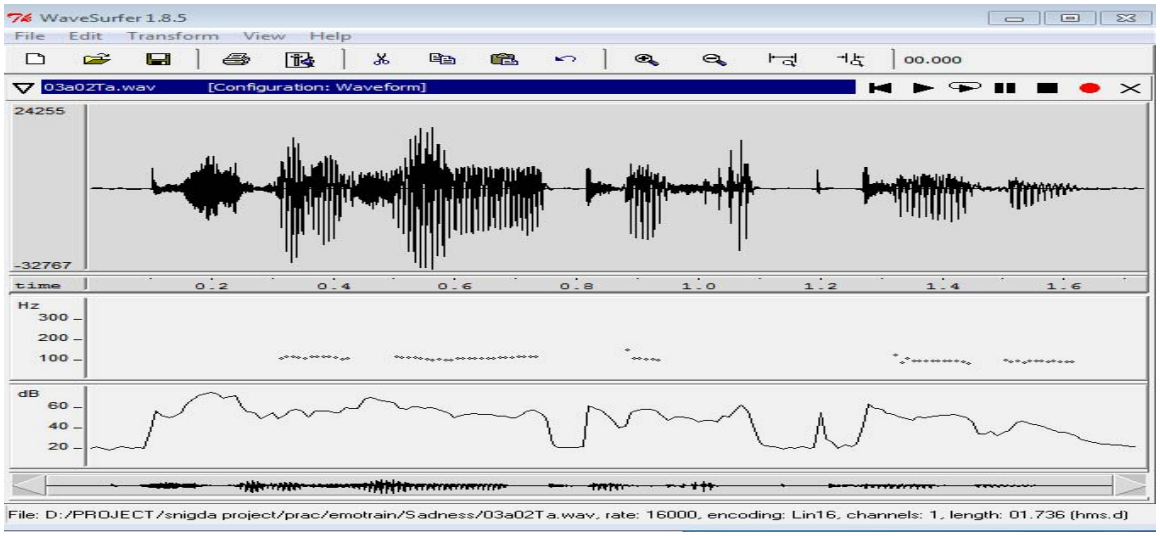

Fig. 12: Silence of Sadness signal

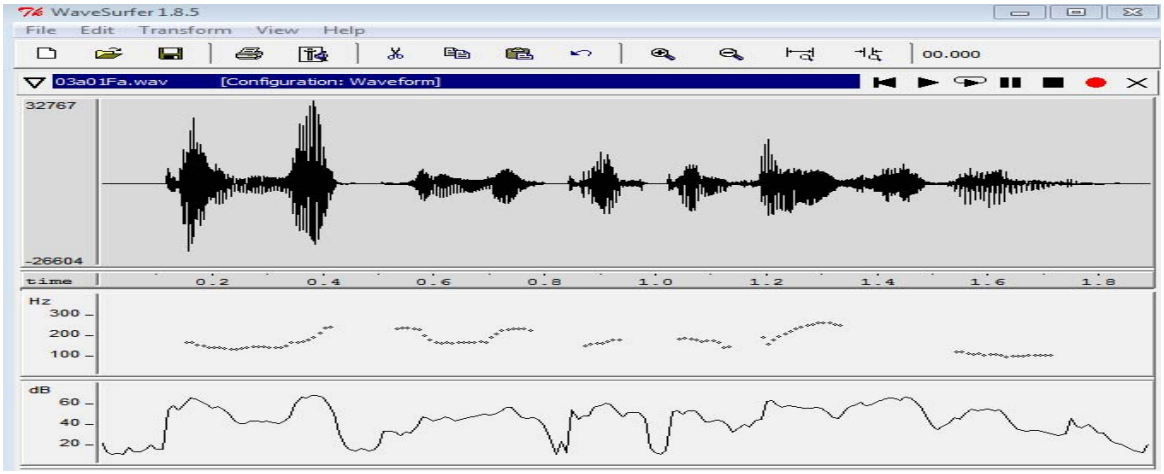

Fig. 13: Happy signal

V. CONCLUSION

The importance of the speech signal in emotion recognition plays an important role to find the accuracy of the system, which is purely dependent on the signal clarity and good features. If the speech signal may contain noise, more silence regions and all other unnecessary information, then using this type of signal we cannot extract a good feature which has proper information to identify the emotion.

Removing the silence regions from the speech is beneficial to identify the correct emotion prediction from the speech signal. Through this analysis, every emotion has own parameters such as anger and happiness have high pitch and energy values compared to others .Also sadness will be identified over the length of the time. From this analysis the actual signal information properly.

REFERENCES

[1] Poonam Sharma1 and Abha Kiran Rajpoot2., Automatic Identification of Silence, Unvoiced and Voiced Chunks in Speech. pp. 87–96, 2013. © CS & IT-CSCP 2013.

[2] Tushar Ranjan Sahoo and Sabyasachi Patra., Silence Removal and Endpoint Detection of Speech Signal for Text Independent Speaker Identification. I.J.Image, graphics and Signal Processing, 2014, 6, 27-35.

[3] J. Meribah Jasmine1, S. Sandhya 2, Dr. K. Ravichandran 3, Dr. D. Balasubramaniam4Silence Removal from Audio Signal Using Framing and Windowing Method and Analyze Various Parameters. Vol. 4, Issue 4, April 2016.

[4] G.Saha1, Sandipan Chakroborty2, Suman Senapati3. A New Silence Removal and Endpoint Detection Algorithm for Speech and Speaker Recognition Applications.

[5] L. R. Rabiner , R. W. Schafer “Digital Processing of Speech Signals”.

[6] http://database.syntheticspeech.de/

[7] Koolagudi. S.G, Reddy. R, Rao. K.S, “Emotion recognition from speech signal using epoch parameters”, IEEE Signal Processing and Communications (SPCOM), 2010 International Conference, ISBN: 978-1-4244-7137-9, Pages: 1-5, July 2010.

[8] Iliou, Anagnostopoulos., "Statistical Evaluation of Speech Features for Emotion Recognition", IEEE Digital Telecommunications, 2009. ICDT '09. Fourth International Conference, ISBN: 978-0-7695-3695-8, pages: 121 – 126, July 2009.
[9] Dmitri Bitouka, Ragini Vermaa, Ani Nenkovab, "Class-level spectral features for emotion recognition", Speech Communication, Volume 52, Issues 7–8, July–August 2010, Pages 613–625.
[10] Iker Luengo, Eva Navas, Inmaculada Hernáez, Jon Sánchez, "Automatic Emotion Recognition using Prosodic Parameters"