"Robust Speech Recognition in Noisy Environment"

A Postgraduate Project Report submitted to Manipal University in partial fulfilment of the requirement for the award of the degree of

MASTER OF TECHNOLOGY

In

Digital Electronics and Advance Communication

Submitted by

Vishal V. Khadake

Under the guidance of

Dr. Samudravijaya K
Scientific officer (f)
Tifr, Mumbai.

&

Dr. Somashekhar Bhat
Professor, E & C Dept.
MIT, Manipal.

DEPARTMENT OF ELECTRONICS AND TELECOMMUNICATION ENGINEERING

MANIPAL INSTITUTE OF TECHNOLOGY

(A Constituent College of Manipal University)

MANIPAL – 576104, KARNATAKA, INDIA

June 2013

DEPARTMENT OF ELECTRONICS AND TELECOMMUNICATION ENGINEERING

MANIPAL INSTITUTE OF TECHNOLOGY

(A Constituent College of Manipal University)

MANIPAL – 576 104 (KARNATAKA), INDIA

Manipal

< Date >

CERTIFICATE

This is to certify that the project titled **ROBUST SPEECH RECOGNITION IN NOISY ENVIRONMENT** is a record of the bonafide work done by **VISHAL V KHADAKE** (*Reg. No. 110915010*) submitted in partial fulfilment of the requirements for the award of the Degree of Master of Technology (MTech) in **DIGITAL ELECTRONICS AND ADVANCE COMMUNICATION** of Manipal Institute of Technology Manipal, Karnataka, (A Constituent College of Manipal University), during the academic year 2012-13.

Dept Guide Name
Prof. Dr. Somashekhar Bhat

Prof. Dr.
HOD, E & C Dept.
M.I.T, MANIPAL

ACKNOWLEDGMENTS

I would like to express my deepest respect and sincere gratitude to my Dept. guide Dr. Somashekhar Bhat, for his constant guidance and encouragements at all the stages of my work. My heartfelt thank to you sir for the solving my silly problems with great patience. Also, I would like to thanks you for the unlimited and valuable support not only during completion of project work but during college time also. Sir, it's just you who encourage me to do my internship in Tifr, Mumbai thank you very much for that sir.

I am extremely grateful of Dr. Samudravijaya sir. I think I am really blessed by destiny and got a chance to work under you sir. Sir, your dedication, hard work and attention have been extreme support to me. Sir, I want to thank you for all your valuable time and your patience towards me. Sir, I am thankful to you for all your valuable technical as well as nontechnical advices. I never thought of completing my thesis work in institute like Tifr, Mumbai, and thank you very much for that sir.

I take this opportunity to thank my lab mates in Tifr, Mumbai and IIT, Bombay, Namrata Karkera, Tejas Godambe, Nikul Prajapati and Jigar Gada for their great support and help during completion of this work. My special thanks to my dear friend Anshu Anand from BARC, Mumbai, for his invaluable advice and help. I would also like to thanks my college friends from MIT, Manipal for constantly helping me both technically and non technically, Rahul Sharma, Raj Goyal, Bhushan Naware, Shraddhye shrivastav and last but not least Ankur Mor.

Finally I would like to dedicate this thesis to my mother, a dedicated teacher and loving mother Mrs. Rajni V. Khadake.

*...**Vishal Vidyanand Khadake***

ABSTRACT

The goal of the project is to explore a novel signal processing technique for improving the performance of Automatic Speech Recognition (ASR) system under noisy environment. We emphasized the samples under glottal closure phase of voiced portions of the speech signals. During glottal closure period, there is no external disturbance to the vocal tract; hence, the acoustic signature of speech signal is better seen in glottal closure period. Glottal closure interval within a pitch period was identified using short time energy profile. Using an appropriate weighting function, we provided special emphasis to the speech samples in glottal closure portion before extracting features from speech signal. We found that such a selective signal processing technique can increase the recognition accuracy of conventional Automatic speech recognition (ASR) system.

Contents

List of Tables

List of Figures

List of Symbols and Abbreviations

List of Symbols	Meaning
S	Speech
E	Energy profile
Em	Smoothed energy profile
EmD	1st derivative of Em
EmS	Smoothed version of EmD
I	Integer reflects sample number
MaxVal	Global maxima of Em
MaxIdxf	Index value of global maxima in EmS
MinVal	Global minima of EmS
MinIdxf	Index value of global minima in EmS
S1	Enhanced speech signal
W	weighting function

List of Abbreviations

ASR	Automatic Speech Recognition
GCI	Glottal Closure Instant
GOI	Glottal Opening Instant
STE	short Time Energy
MFCC	Mel Frequency Cepstral Coefficient
CI	Context Independent
CD	Context Dependent

CHAPTER 1
INTRODUCTION

1.1 Introduction

Speech is the primary means of communication between people. For reasons ranging from technological curiosity about the mechanisms for mechanical realization of human speech capabilities, to the desire to automate simple tasks inherently requiring human-machine interactions. Speech processing, coding, synthesis and recognition are considered to be the most promising application in telecommunication and data transmission. Traditional methods of speech recognition perform well under clean speech conditions, but suffer from large performance degradation under noisy environments. The mismatch in training and testing/deployment conditions essentially causes the performance degradation, and is currently handled in many ways. Some of the approaches extract noise robust features by using techniques such as cepstral mean normalization. In source driven approaches, a speech enhancement algorithm is applied to the noisy speech and then recognition is performed on the enhanced noisy speech using clean speech models. However, performance of the aforementioned approaches is in adequate in real environments.

The aim of our project is the robust speech recognition under noisy environment especially under the additive noise such as white noise, babble noise and vehicle noise. In this project we used short time energy profile estimation method, which concentrates on the voiced speech signal under glottal closure portion. This segmental signal is emphasized as a pre-processing step. Here we found that giving the importance to this segmental portion of signal and passing it through the conventional ASR system will improve the recognition accuracy.
The energy profile further smoothed to suppress high frequency components and differenced smoothed signal is used to get local peaks and valleys which represent glottal closure instances (GCIs) and glottal opening instances (GOIs). An, appropriate weighting function is generated based on GCIs and GOIs. Speech signal is multiplied with an appropriate weighting function for enhancement as a pre-processing step. After enhancement of speech, it passes through conventional ASR system to observe the recognition accuracy before and after enhancement. Also, improvement in recognition accuracy after enhancement is checked when Training and testing speech data were same and also, when they were different.

1.2 Organization of Report

The contents of this project report are organised as follows,

Chapter 2: Background and Literature Review - deals with theoretical background require for the project which includes speech signal and its types viz., voiced, unvoiced and silence. This chapter also throws light towards pitch, glottal closure and glottal open phase.

Chapter 3: Methodology - deals with method titled short time energy profile estimation is used under this project. The main aim of this project is enhancement of samples under glottal closure phase such that recognition accuracy of speech should increase. In this chapter, how Short time energy estimation method reduces additive noise effect is studied. Also, marking instances of glottal closure (GCI) and instant of glottal opening (GOI), generation of proper weighting function and enhancement of speech is studied. The chapter also deals with existing ASR system, mfc feature estimation, training acoustic models (context dependent and context independent), language model estimation, decoding and evaluation of recognition accuracy.

Chapter 4: Result Analysis – discusses percentage recognition results occurs after and before enhancement of speech signal using short time energy estimation method. This chapter also explains about recognition improvement for train =test data i.e., when training and testing data were same and Also, about train! =test data where training and testing data were different. The results are conducted for trainAndTest and also for kfold evaluation.

Chapter 5: Conclusion- summarizes the contributions of project work and discuses future scope for further investigation.

CHAPTER 2
BACKGROUND AND LITERATURE REVIEW

Chapter 02 explores about background theory related to the project viz., speech signal and its types i.e., voiced, unvoiced and silence. The chapter shows some lights towards speech production system and parts helps for the speech production i.e. vocal folds, trachea, larynx, and glottis. Also, in this chapter I will discuss about open and close phase of glottal cycle, importance of glottal closure phase over open phase for speech recognition and generation of weighting function to enhance the speech samples under glottal closure phase.

2.1 Speech Signal

Speech is an acoustic signal produced from a speech production system. From our understanding of signals and systems, the system characteristics depend on the design of the system. For the case of linear time invariant system, this is completely characterized in terms its impulse response. However, the nature of response depends on the type of input excitation to the system. For instance, we have impulse response, step response, and so on for a given system. Each of these output responses lead behavioural understanding of the system under different conditions.

A similar phenomenon happens in the production of speech also. Based on the input excitation phenomenon, the speech production can be broadly categorized into three activities. The first case where the input excitation is nearly periodic in nature, the second case where the input excitation is random noise-like in nature and third case where there is no excitation to the system. Accordingly, the speech signal can be broadly categorized into three regions. They are voiced, unvoiced and silence.

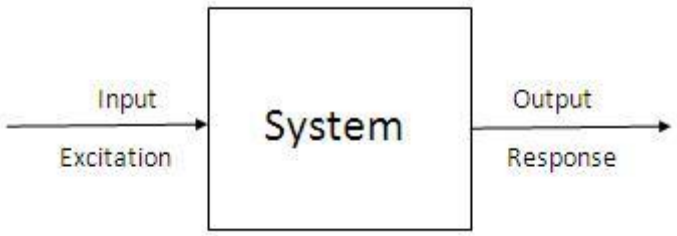


Figure (2.1.1) Block diagram representing relation between signal and system [3]

A classical block diagram shown in Figure 2.1.1 links signal and system concepts. As shown, the system responds to the input signal/excitation and produces output signal/response. For a given design of the system, the output response depends on the type of input excitation. Accordingly, we can have different output responses. The same block diagram can be used for the study of this experiment. For the case of speech, it can be modified as shown in Figure 2.1.1. The speech production system responds to the input excitation by producing speech signal. The schematic of human speech production mechanism is shown in Figure 2.1.2.

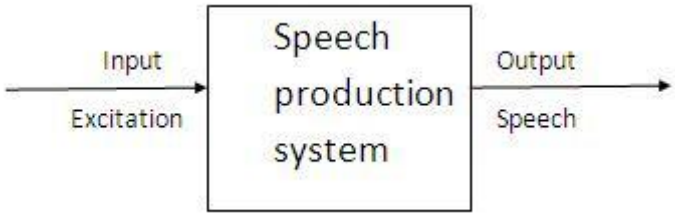


Figure (2.1.2) Speech production process as a signal and system block diagram [3]

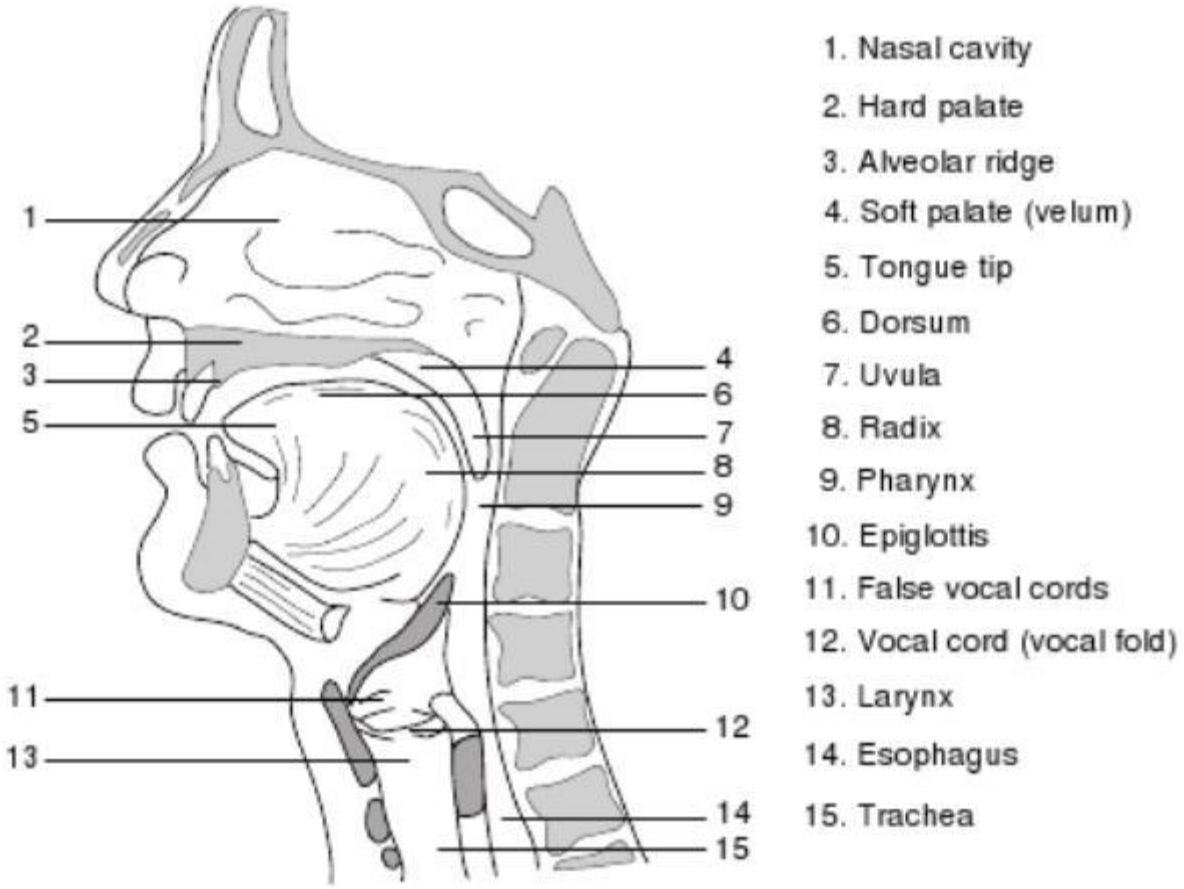


Figure (2.1.3) Schematic diagram of human speech production system [3]

The speech production organs include lungs, larynx, trachea, glottis, pharynx, oral cavity and nasal cavity. The lungs act as air cavity which supplies the required air during exhalation for producing speech. Trachea also termed as wind pipe connects the lungs to the glottis via bronchial muscles. The glottis consists of two thin membranes known as vocal folds or chords and obstructs airflow during specific categories of speech to generate the required excitation signal for speech production.

Larynx

Larynx is also called as voice box. It consists of cartilages, muscles and ligaments. Larynx mainly consists of hypoid bone, epiglottis, and vocal folds. Larynx has cartilages. It connects mouth cavity to the trachea. The function of larynx is three fold, if larynx is completely open it allows breathing. If partially open it creates distortion to the air, results in phonation. The air flows from lungs through trachea and experiences the distortion in glottis. The distortion causes due to successive opening and closing of vocal folds which result in compression and rarefaction of air flow leads to speech production. If larynx is completely closed it protects the respiratory system. We can see the position of Larynx in Figure 2.1.3. The anatomical view of Larynx is shown in Figure 2.1.4.

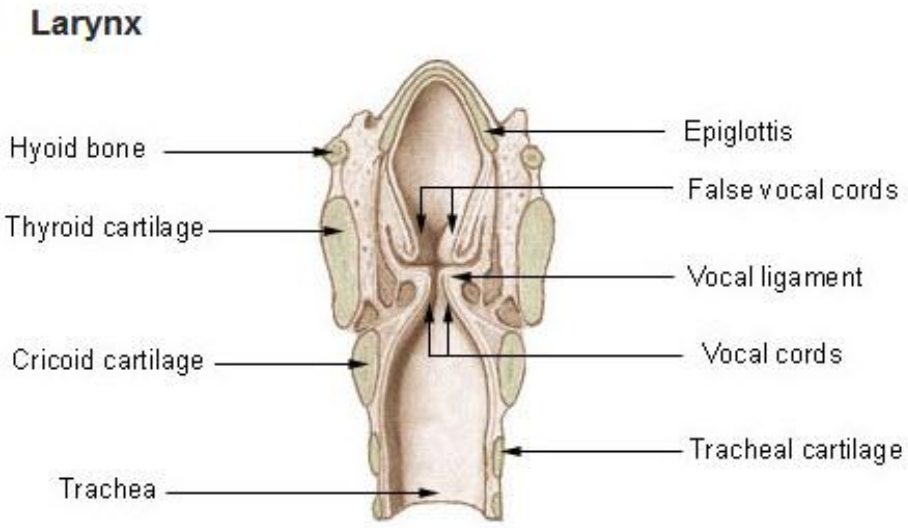


Figure (2.1.4) anatomical view of Larynx [4]

Vocal Cords

The **vocal cords** are also, commonly known as **vocal folds.** Vocal cords are two infolding membranes which are controlled via vagus nerve. Vocal cords performs very important role in functions like breathing and phonation. Normally vocal cords are apart i.e., glottis is open. The two infolding membranes will be stretched to bring vocal cords near. During breathing vocal folds are completely open for inhalation of air. During phonation vocal cords comes closure in such a way that they should creates air pressure beneath a Larynx and due to sub-glottal air pressure vocal cords goes apart which leads to rhythmic oscillations of vocal cords. The rhythmic oscillation of vocal cords are quasi periodic which leads to periodic oscillation. The rate at which vocal folds vibrates known as fundamental frequency. Open and closed positioned of vocal folds and corresponding functions are stated in Figure 2.1.5.

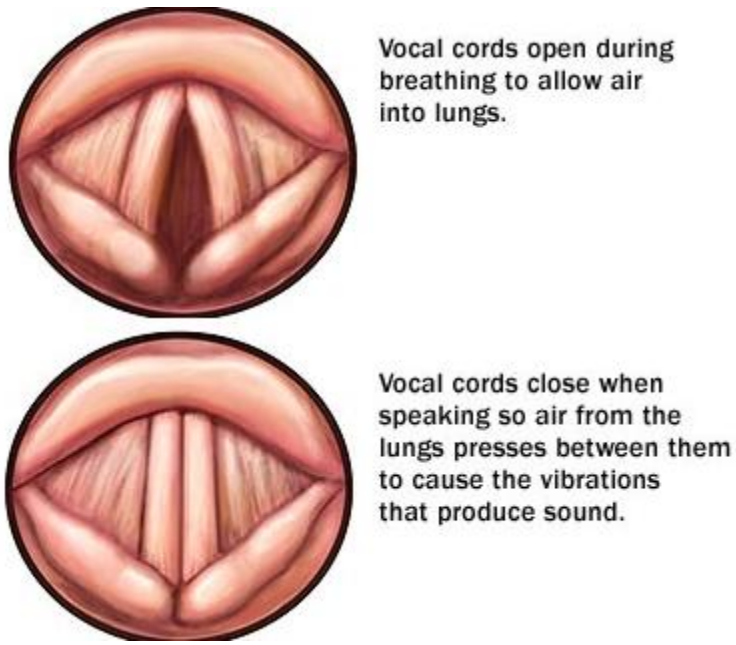


Figure (2.1.5) Different vocal cords positions and corresponding functions [4]

Men and women have different vocal fold sizes. Adult male voices are usually lower pitched and have larger folds. The male vocal folds are between 1.75 cm and 2.5 cm in length which is more than female vocal folds length which is around 1.25 cm to 1.75 cm which helps them to produce high pitch voice.

2.2 Voiced and Unvoiced Speech

This section explores different types of speech signals and explains their characteristics in detail.

➢ Voiced Speech

If the input excitation is nearly periodic impulse sequence, then the corresponding speech looks visually nearly periodic and is termed as **voiced speech**. The speech production process for the voiced speech can be pictorially represented as shown in Figure 2.2.4. During the production of voiced speech, the air exhaling out of lungs through the trachea is interrupted periodically by the vibrating vocal folds. Due to this, the glottal wave is generated that excites the speech production system resulting in the voiced speech. A typical glottal waveform and the corresponding voiced speech are shown in Figure 2.2.4. Thus grossly, when we look at the speech signal waveform, if it looks nearly periodic in nature, then it can be marked as voiced speech.

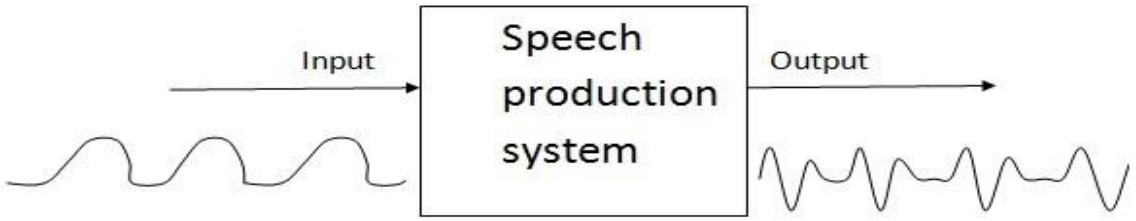


Figure (2.2.1) Block diagram representation of voiced speech production [3]

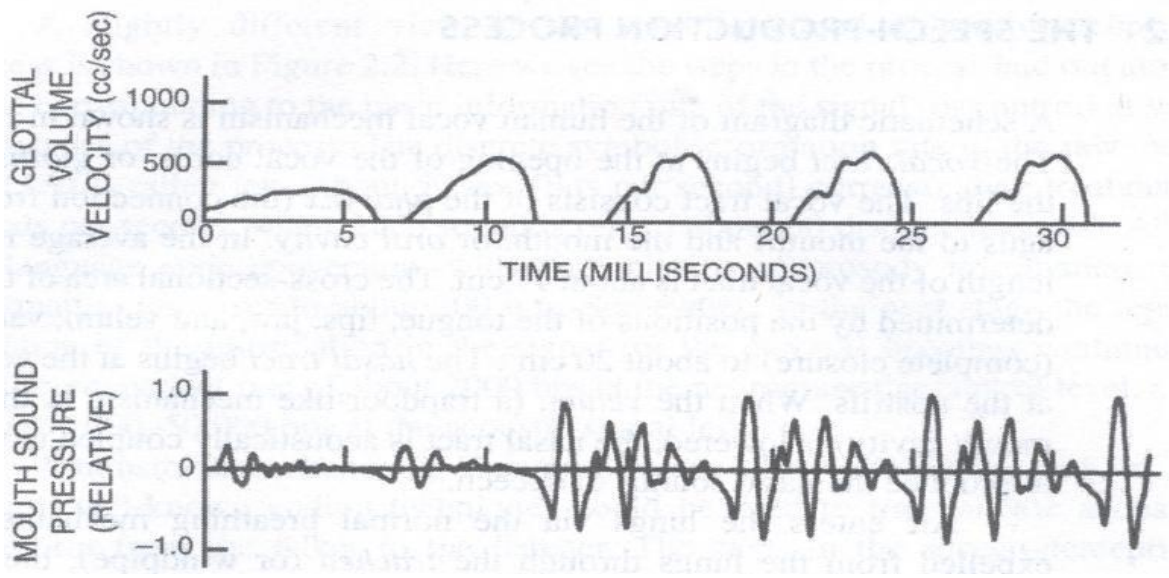


Figure (2.2.2) glottal wave and corresponding speech signal during voiced speech production [3]

The periodicity associated with the voiced speech can be measured by the autocorrelation analysis. This period is more commonly termed as **pitch period [3]**. A segment of voiced speech utterance "a" and its autocorrelation sequence are plotted in Figure 2.2.3a) & 3b).

The distance between the consecutive largest peaks in the autocorrelation sequence from the beginning represents **pitch period**. This is the important and main distinguishing factor for voiced speech. Since voiced speech is periodic in nature, we expect some fundamental frequency and its harmonics in the spectrum of speech. Figure 2.2.3 a) shows a segment of voiced speech and its magnitude spectrum. As it can be observed in the spectrum, there are frequency components repeating at regular intervals indicating the presence of harmonic structure. In the frequency domain, the presence of this harmonic structure is the main distinguishing factor for voiced speech. The fundamental frequency of input excitation is also termed as **pitch frequency** or just **pitch.**

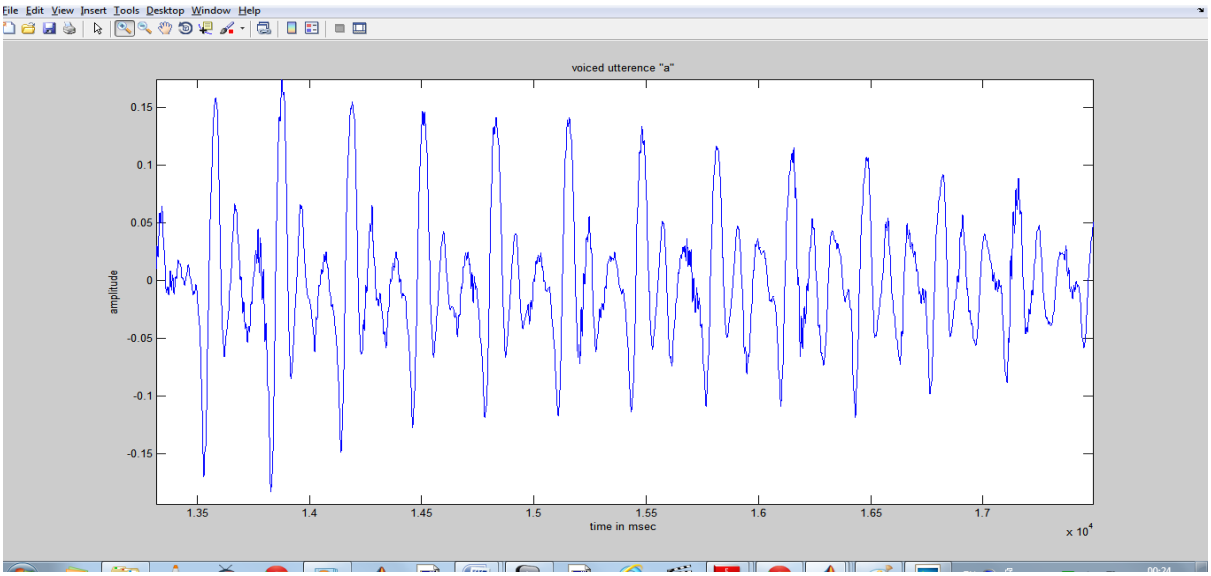


Figure (2.2.3 a) voiced speech utterance "a"

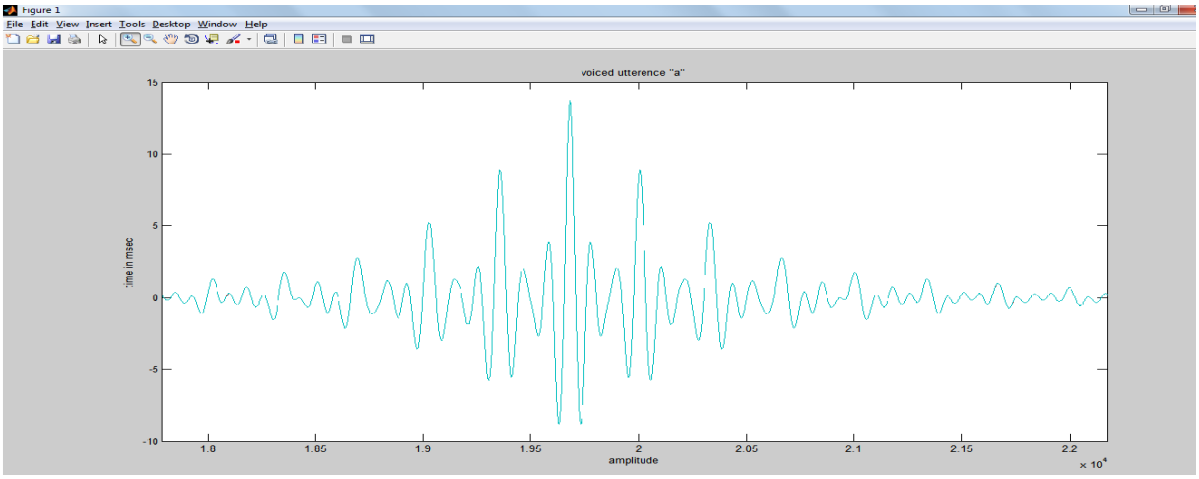


Figure (2.2.3b) autocorrelation sequence of voiced speech utterance "a"

➢ Unvoiced Speech

If the excitation is random noise-like, then the resulting speech will also be random noise-like without any periodic nature and is termed as **Unvoiced Speech**. The speech production process for the voiced speech can be pictorially represented as in Figure 2.2.4.During the production of unvoiced speech, the air exhaling out of lungs through the trachea is not interrupted by the vibrating vocal folds. However, starting from glottis, somewhere along the length of vocal tract, total or partial closure occurs which results in obstructing air flow completely or narrowly. This modification of airflow results in stop or frication excitation and excites the vocal tract system to produce unvoiced speech. The typical nature of excitation and resulting unvoiced speech are shown in Figure 2.2.4 itself. As it can be seen, the unvoiced speech will not have any periodic nature. This will be the main distinction between voiced and unvoiced speech.

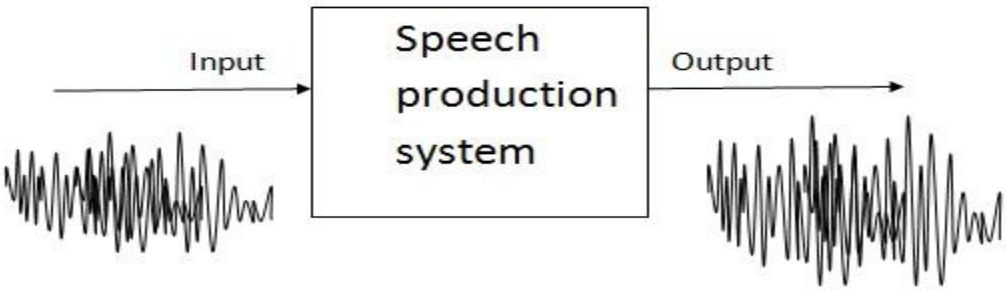


Figure (2.2.4) Block diagram representation of unvoiced speech production [3]

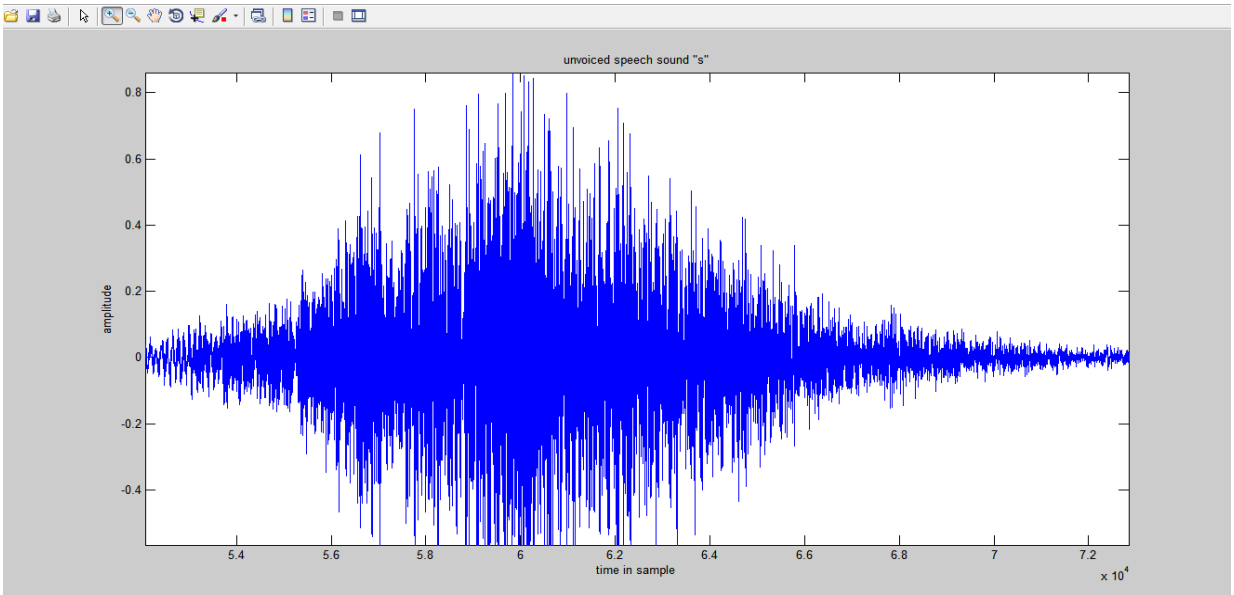


Figure (2.2.5a) Unvoiced speech segment

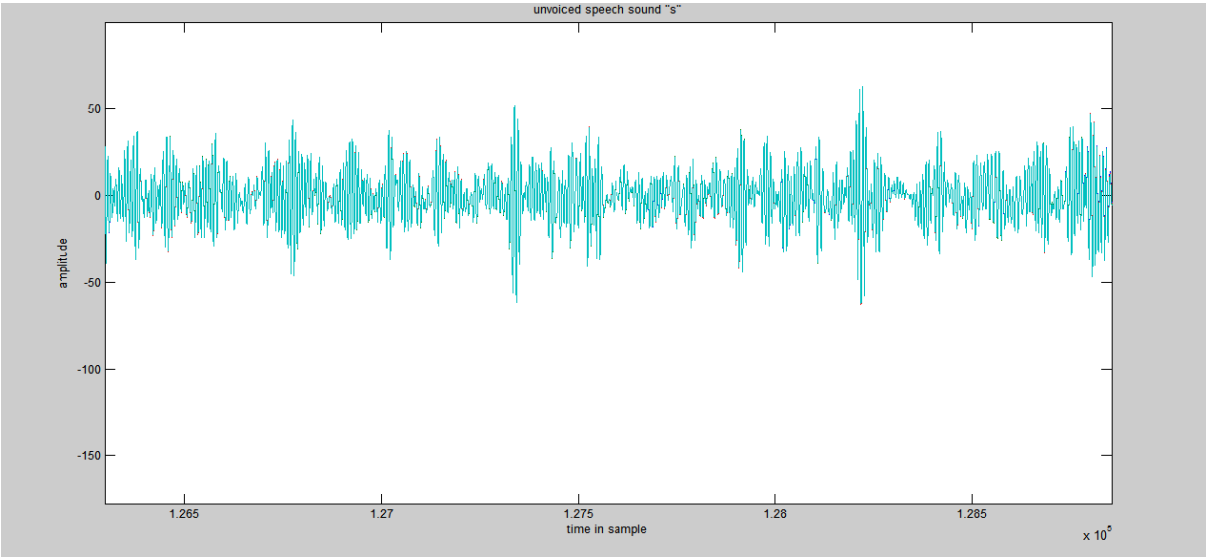


Figure (2.2.5b) Autocorrelation sequence of unvoiced speech utterance "s"

The aperiodicity of unvoiced speech can also be observed by the autocorrelation analysis. A segment of unvoiced speech and its autocorrelation sequence are plotted in Figure 2.2.5a) and 5b). As it can be observed there is no strong peak indicating periodicity. This is the important and main distinguishing factor between voiced and unvoiced speech. As it can be observed in the spectrum, there is no harmonic structure. The absence of this harmonic structure is the main distinguishing factor for unvoiced speech.

➢ Silence Region

The speech production process involves generating voiced and unvoiced speech in succession, separated by what is called **silence region**. During silence region, there is no excitation supplied to the vocal tract and hence no speech output. However, silence is an integral part of speech signal. Without the presence of silence region between voiced and unvoiced speech, the speech will not intelligible. Further, the duration of silence along with other voiced or unvoiced speech is also an indicator of certain category of sounds. Even though from amplitude/energy point of view, silence region is unimportant, but its duration is very essential for intelligible speech. Figure2.2.6 shows waveforms for the word ***pakkaa***. As it can be observed between the two vowels there is silence region representing the sounds *pak-kaa*. Even though the signal energy is lowest or negligible, its duration is important for perceiving it.

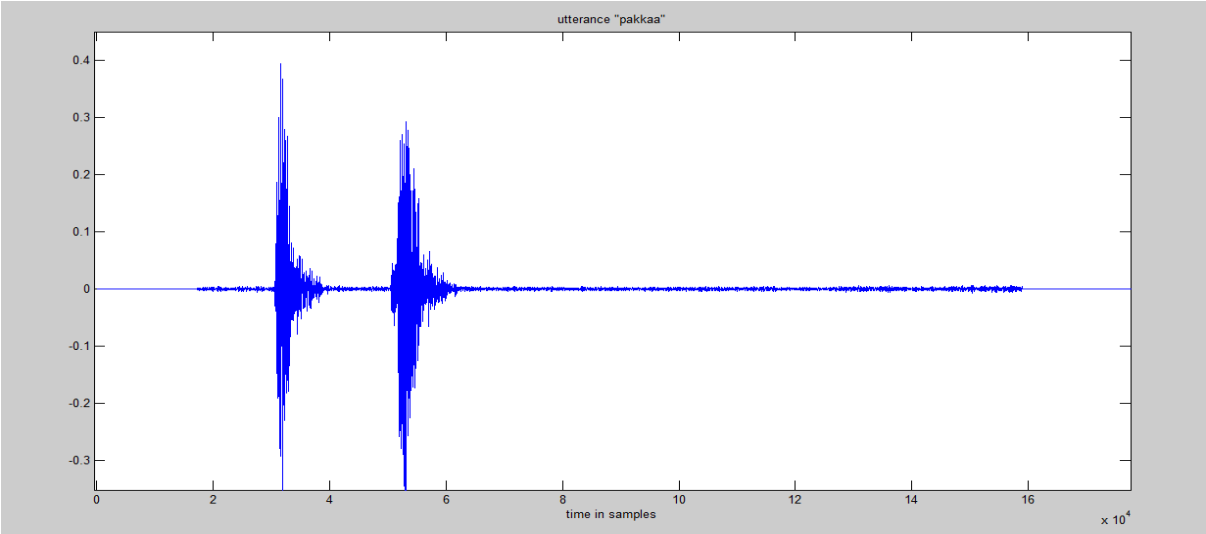


Figure (2.2.6) silence region in between speech utterance “pakkaa”

2.3 Closed and Open Phase of Glottis within Pitch Period

Speech signal can be classified into voiced, unvoiced and silence regions. The near periodic vibration of vocal folds is excitation for the production of voiced speech. The random ...like excitation is present for unvoiced speech. There is no excitation during silence region. Majority of speech regions are voiced in nature that include vowels... semivowels and other voiced components. The voiced region looks like a near periodic signal in the time domain representation. In a short term we may treat the voiced speech segments to be periodic for all practical analysis and processing.

The periodicity associated with such segments is defined as a 'pitch period (‘T_o’) or a ‘glottal cycle’ in the time domain and 'Pitch frequency or Fundamental Frequency ‘F_o' in the frequency domain. Unless specified, the term ‘pitch’ or ‘glottal cycle’ refers to the fundamental frequency ' F_o'. Pitch is an important attribute of voiced speech. It contains speaker-specific information. It is also needed for speech coding task. Thus estimation of pitch is one of the important issues in speech processing.

The glottal cycle is a composite of glottal closure and glottal opening portion as shown in Figure 1.3.1

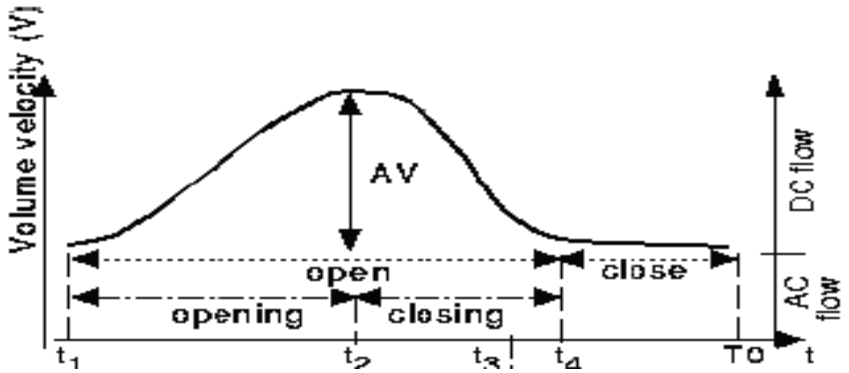


Figure (2.3.1) glottal cycle [3]

It has been seen that under glottal closure portion there will not be an airflow i.e., disturbance to the vocal tract filter. So, under glottal closure portion the acoustic signature of speech signal can be seen well. Thus, in this project we are concentrating on samples under glottal closure portion. We are enhancing the importance of such a sample using a specific weighting function.

2.4 Weighting Function for Enhancement of Speech

The weighting function will enhance the samples under glottal closure portion and a graphical overview is shown in Figure 1.4.1, in this Figure it can be seen clearly that samples which are not under glottal closure portion are not modified. The instant where beginning of glottal closure occurs is termed as **epoch location** which is indicated by red arrow in Figure.1.4.1. The sample under glottal closure portion is enhanced by multiplying samples with **'1+x'**. **x** is a variable quantity which can be fixed by trial and error method i.e., whichever x value provides better recognition result can be selected.

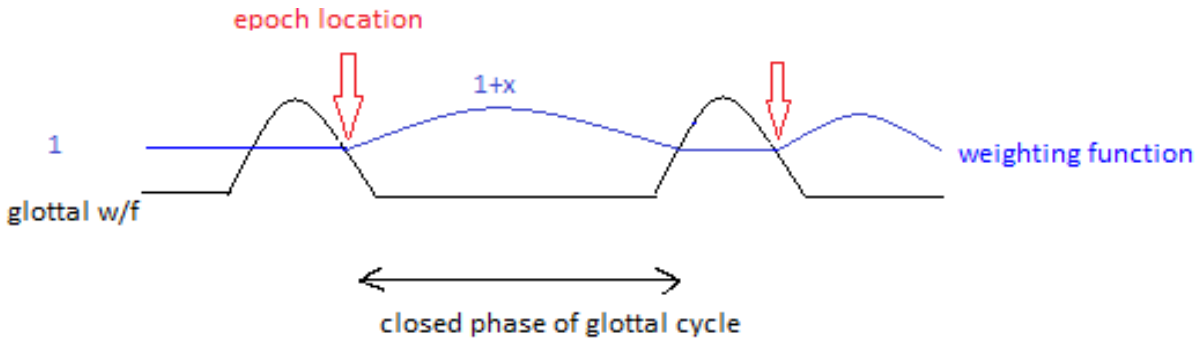


Figure (2.4) weighting function.

CHAPTER 3
METHODOLOGY

This chapter discuss in section 3.1 about the methodology used for speech enhancement in this projects and required important steps for enhancement of speech samples under glottal closure phase. In section 3.2 we will discuss about short time energy estimation method and in section 3.3 we will throw some light towards conventional ASR system.

3.1 Modified Automatic Speech Recognition (ASR) System

The main goal of this project is speech recognition under noisy environment by giving importance to the voiced signal in glottal closure phase. The voiced signals are high energy signals compared to unvoiced or silence signals, hence less affected by the distortion that may be additive or convolution in nature. The voiced signals are near periodic combinations of glottal closure and glottal opening instances. It has been seen that under glottal closure phase, vocal folds comes closer which helps to reduce external disturbances from respiratory system in form of air. So, in this project, we concentrated on the voiced portion of the signal under glottal closure portion, estimation of short time energy profile, capturing impulse like glottal closure effects which leads to estimation of GCI, estimation of GOI, an appropriate weighting function generation for enhancement of speech signal. To do so, we introduce a proper pre-processing block in a regular ASR system as shown below,

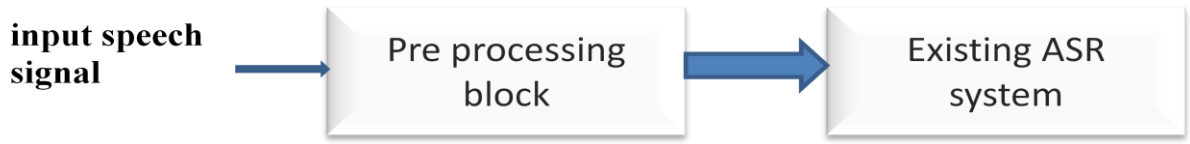


Figure (3.1.1) Modified/proposed ASR system

Here we introduced a proper pre-processing block in a regular ASR system which helps to improve recognition accuracy over existing ASR accuracy. Pre-processing is nothing but enhancing the samples in glottal closure phase by multiplying speech signal with its proper weighting function. The enhancement of original/unenhanced speech leads to improvement in SNR for samples under glottal closure phase. The detailed functioning of pre-processing block is explained in Figure 3.1.2.

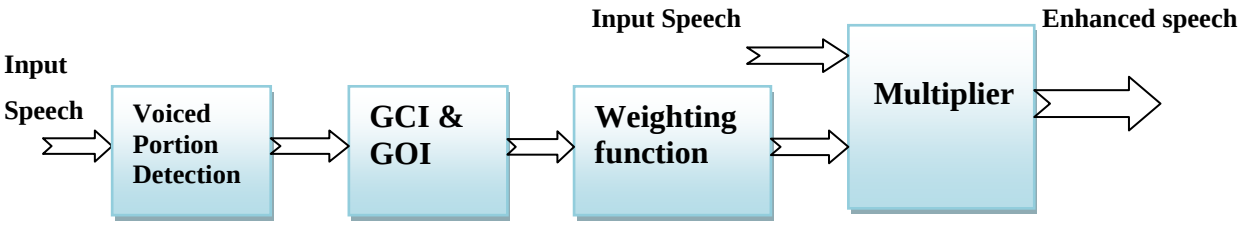


Figure (3.1.2) Functional block diagram of Pre-processing block

To understand the function of pre-processing block one should understand functioning of each step involved in pre-processing block. There are four major steps involved in pre-processing viz.., Voiced Portion Detection, GCI & GOI detection, generating a weighting function and enhancing input speech by multiplying it with proper weighting function. All the steps are explained sequentially as follows,

- **Voiced Portion Detection**

  In this step pre-processing block detects/marks voiced portions present in speech wave file. Voiced samples generate due to rhythmic vibration of vocal cords thus shows quasi periodicity. As, voiced samples are high energy samples, they are less affected by external noise so marking voiced portions is nothing but focusing on high energy low noisy samples.

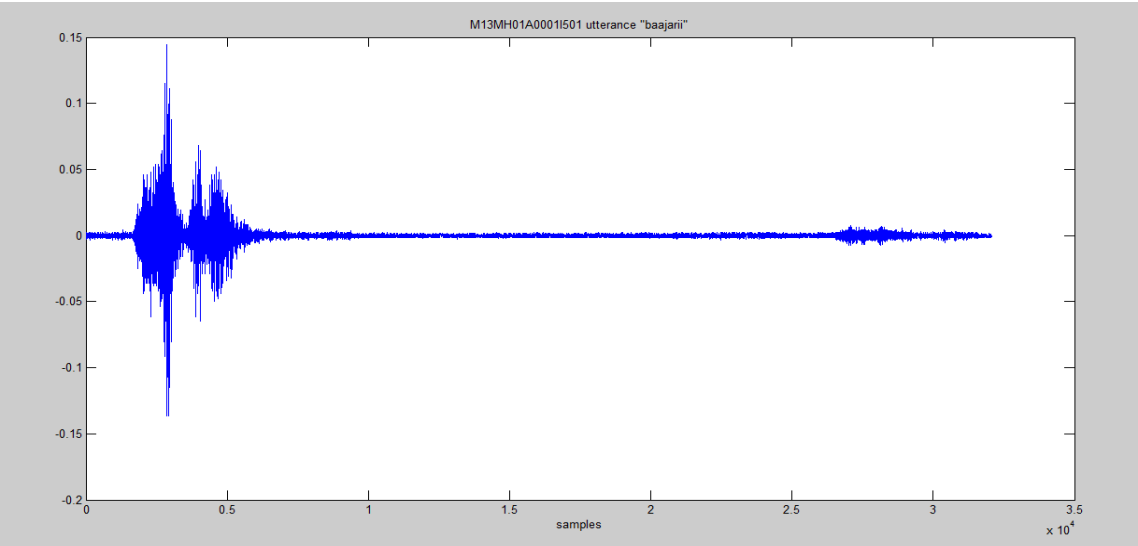


Figure (3.1.3a) speech input wave file

The word “baajarii” contains two high energy vowels ‘aa’ and ‘ii’. As we know that vowels are voiced samples so according to figure (3.1.3b) which is zoomed version of figure (3.1.3a) shows that samples form (1800 to 2100) represents high amplitude regions, Also around 3000 and 3500 we can see high amplitude portions. But, this portion could be a background noise may be a horn and our pre-processing block may consider it as a voiced portion leads to false detection. To avoid false detection we are using short time energy profile estimation method. Short time energy profile voiced detection is explained in section 3.1.

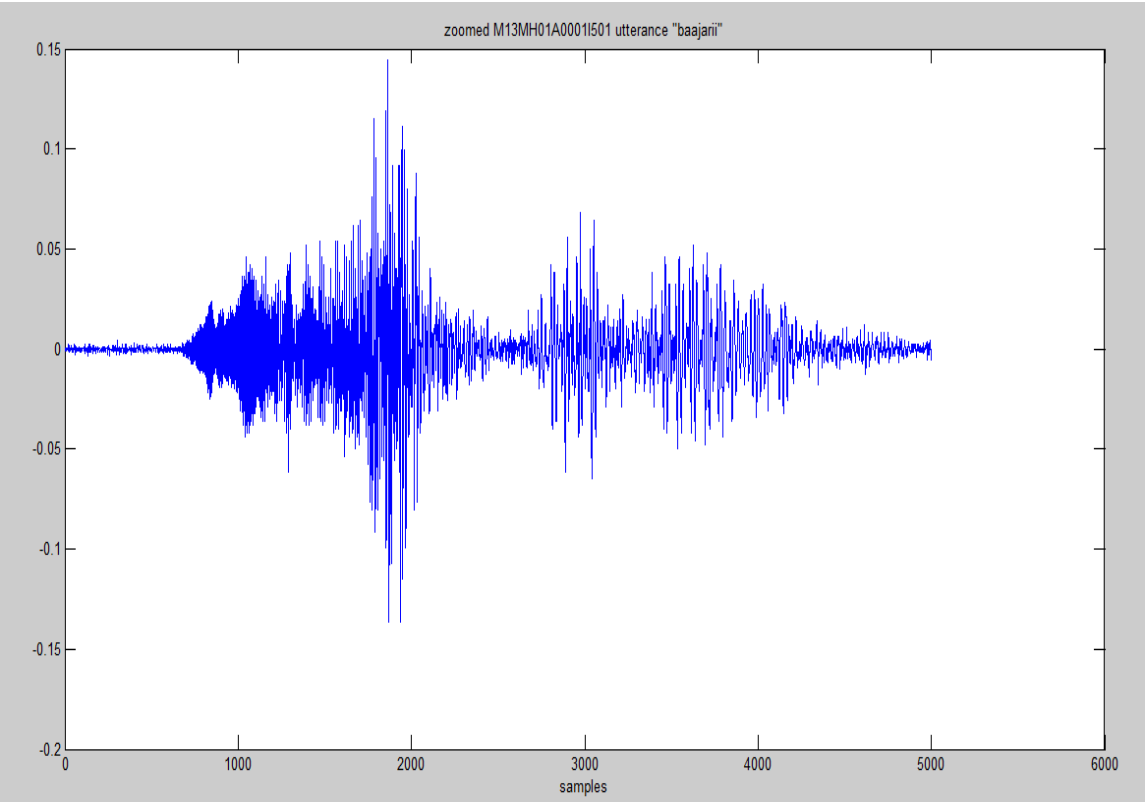


Figure (3.1.3b) zoomed speech input wave file

- **Instances of Glottal Closure and Glottal Opening**

  Once we get the voiced portion present in speech wave form, next step is to determine GCI and GOI in voiced marked portion so that we should get instances of samples present on glottal closure phase. Again, short time energy profile estimation method is used to mark the GCIs and GOIs explained in section 3.2. For examples ‘*’ and ‘+’ represents GOIs and GCIs respectively as shown in figure 3.1.4. This ‘*’ and ‘+’ actually represents

instances of samples under glottal closure phase which helps to design an appropriate function to enhance the input speech.

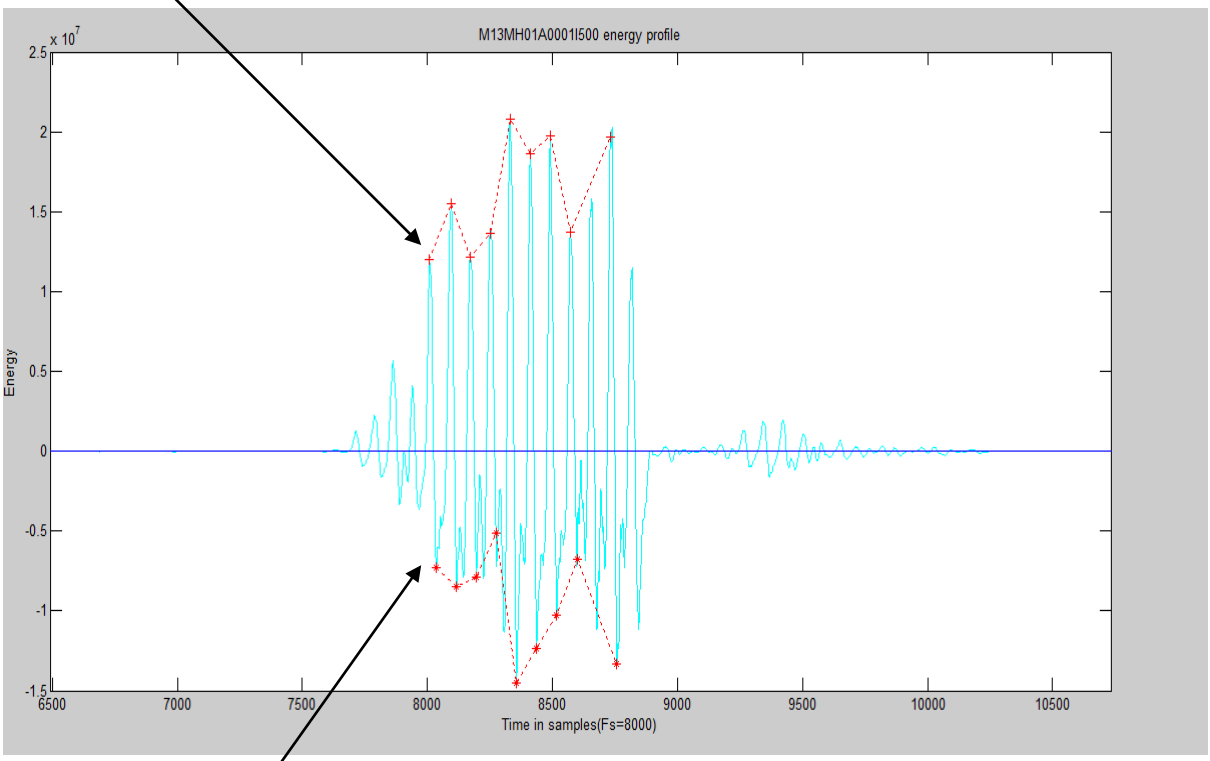


Figure (3.1.4) zoomed energy profile showing GCIs and GOIs.

- **Weighting Function**

After getting GCIs and GOIs we will generate a proper weighting function based on alternative GCIs and GOIs. The weighting function starts from GCI and ends at successive GOI which ultimately represents portion under glottal closure. For example weighting function generated for a speech input is as shown 3.1.5. The weighting function used here is starts from successive GCIs and ends at corresponding GOIs. The shape of weighting function is selected randomly, as it shows increasing amplitude nature for

Samples nearer to GCI i.e., starting samples under glottal closure phase. The maximum value of weighting function could reach up to value of 2 and minimum value will be 1. The minimum value is selected 1 so that after multiplication of input speech with weighting function should not affect voiced samples which are not under glottal closure phase and also unvoiced and silence samples.

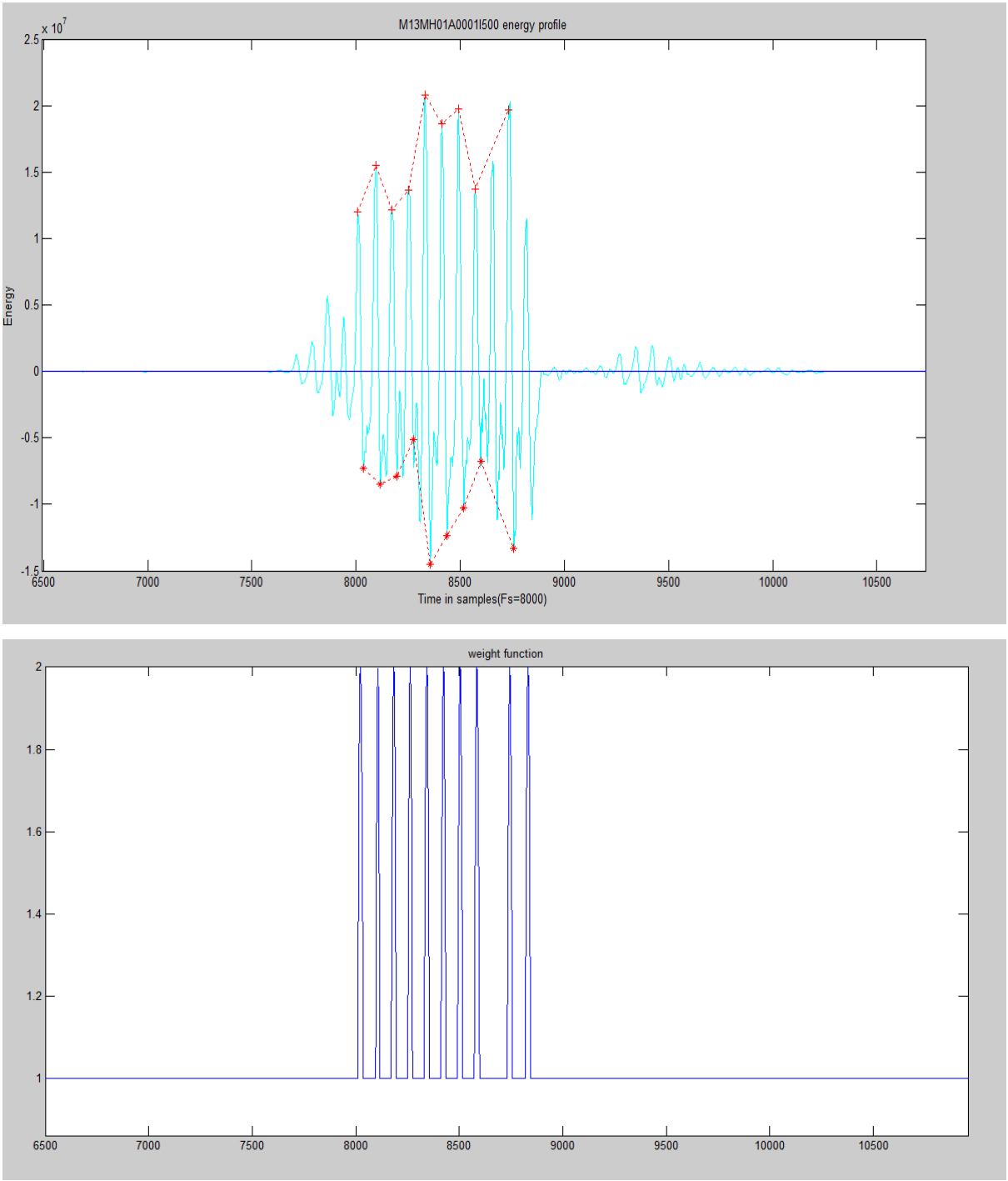


Figure (3.1.5) weighting function matching with marked local peaks and local valleys

- **Multiplier**

  After all three preliminary steps speech signal multiplied by its proper weighting function gets enhanced. This enhanced speech signal passed through the conventional ASR system for finding the improvement in recognition accuracy after enhancement.

3.2 Short Time Energy Profile Estimation Method

To design pre-processing block we propose short time energy profile estimation method which helps in first three functional steps involved in pre-processing block. The Energy profile estimation method is used to find the epoch locations and instances of glottal closure phase in a glottal cycle which helps to distinguish between voiced and other parts of input speech signals.

The strength of excitation of vocal tract can be considered to be significant in glottal activity region. In the absence of vocal folds vibration (non glottal region), the vocal tract system can be considered to be excited by random noise, as in case of fricatives. In the glottal activity region, the energy of the impulse is distributed uniformly in frequency domain but highly concentrated in the time domain; in absence of vocal folds vibration, it is distributed uniformly both in time and frequency domain. Hence short time energy profile signal shows lower amplitude to the random noise excitation compared to the impulse like excitation. Thus we can get the glottal activity region where high amplitude occurs. The steps involved in short time energy profile estimation method are shown in figure (3.2.1).

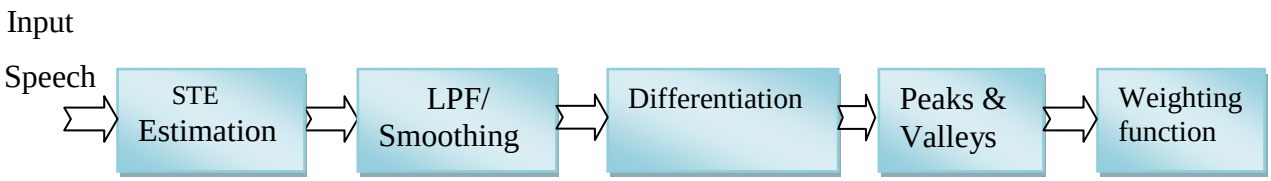


Figure (3.2.1) STE block diagram

In energy profile estimation method, the steps involved are short time energy profile estimation, Low pass filtering (LPF/Smoothing), Differentiating smoothed version of energy

profile, Estimation of local peaks and local valleys which represents GCIs and GOIs, generation of weighting function for enhancement of input speech.

- **STE Estimation**

During the estimation of short time energy profile window size plays a very vital role which decides how many samples will be considered to estimate the energy profile. As we know, pitch period among normal human beings vary in between 3.3 ms to 10ms i.e., pitch frequency of 100 Hz -300 Hz. As per our goal to enhance the samples under glottal closure portion of a glottal cycle, we must know opening and closing instants of glottal closure portion or we should have information regarding pitch frequency. Unfortunately, we do not have both of them. From the energy profile we should get these instances. According to [1] epoch locations occur at maximum positive slopes of energy profile signal. Also, wherever maximum positive slope occurs it reflects a beginning of glottal closure instant and maximum negative slope provides the closing of glottal closure instant.

So, frame size of 16 samples (i.e., nearly a frequency of 60 Hz/2msec) is selected since pitch frequency cannot be less than 60 Hz. So, it is expected that within a period of 2 ms we should get at least 1 maximum positive slope and 1 maximum negative slope instants. Here we start with a ground truth that speaker cannot have frequency less than 60 Hz. Whenever glottal closure instance occurs an impulse like excitation produces i.e., sudden change of energy occurs. Also, it has been seen that voiced portions of speech signals have higher energy.

Thus, we estimate short time energy for a specific (2msec) window sizes such that higher amplitude signals (voiced speech samples) gets highlighted. The speech signal and its corresponding short time energy estimated signal is shown in 3.2.2. The utterance in wave file is word ‘jvaarii’ which has two high energy vowels ‘aa’ and ‘ii’. But actual utterance in speech wave file is horn noise + utterance ‘jvaarii’. If we see carefully speech wave file shown in 3.1.2 samples up to 500 do not belong to speech but it is horn noise samples. Instead of horn noise it could be any additive noise present in wave files. So, first step in pre-processing block is detecting voiced portions from given speech signal. To do so, while

computing energy profile precaution has been taken that after estimating short time energy, samples are placed at mean position i.e., let us say we are estimating energy profile or first frame of 2 msec (Fs=8000, 16 samples) estimated 1st energy sample is placed at 8th position and continues for remaining windows. Thus, if we see estimated energy profile carefully, it shows higher amplitudes for speech portion and initial portion of noisy samples has less amplitude.

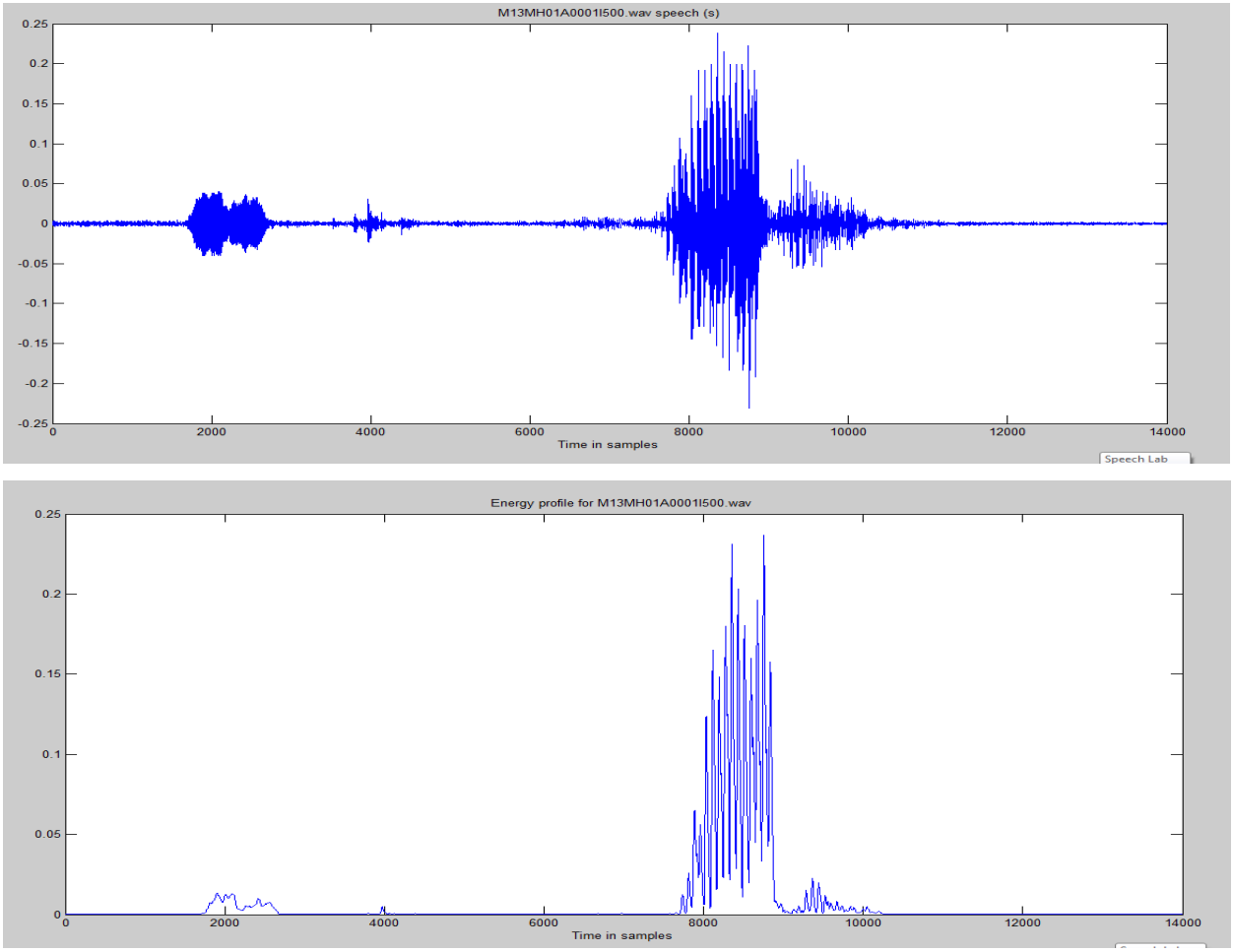
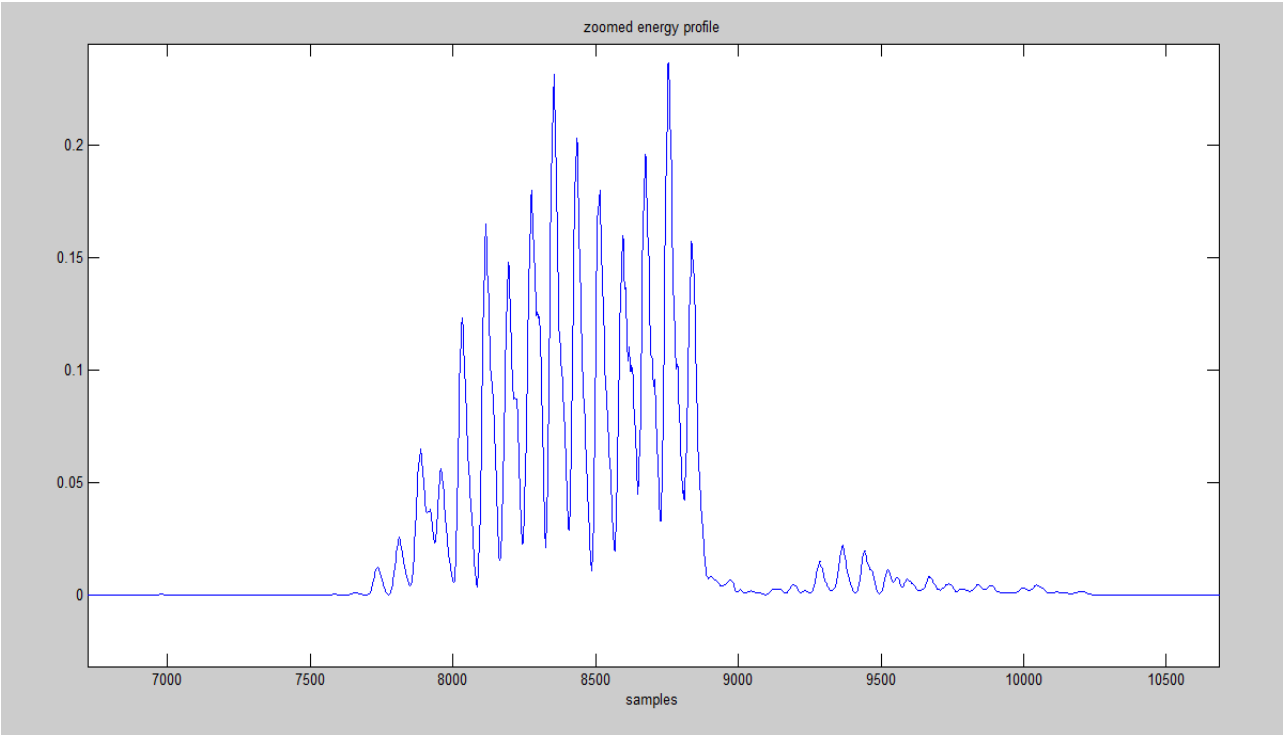


Figure 3.2.2 M13MH01A0001I500.wav speech (s) and short time energy profile (E)

- **Low Pass Filtering (LPF) / Smoothing**

Although placing energy samples at their means reduces effect of noise up certain level their still be a possibility that high frequency samples will present. Now one of our main goals is estimation of maximum positive and maximum negative slopes which reflects GCIs and GOIs using which we will estimate the weighting functions. If we look carefully to the zoomed version of energy profile shown in figure 3.2.3, we will find presence of high frequency component. In figure 3.2.3 samples from 7500 to 9000 shows high frequency

component i.e., rapid amplitude variation. This high frequency component has to remove to get proper maximum positive and maximum negative slopes.



3.2.3 zoomed energy profile (Em)

So, second step in energy profile estimation method is low pass filtering (LPF) which is nothing but smoothing energy profile. The higher vocal frequency could be 300 Hz. Thus considering higher frequency range, signal will be smoothed by eliminating higher frequency i.e., greater than 300 Hz. Figure 3.2.4 and 3.2.5 shows smoothed energy profile and zoomed smoothed short time energy profile respectively.

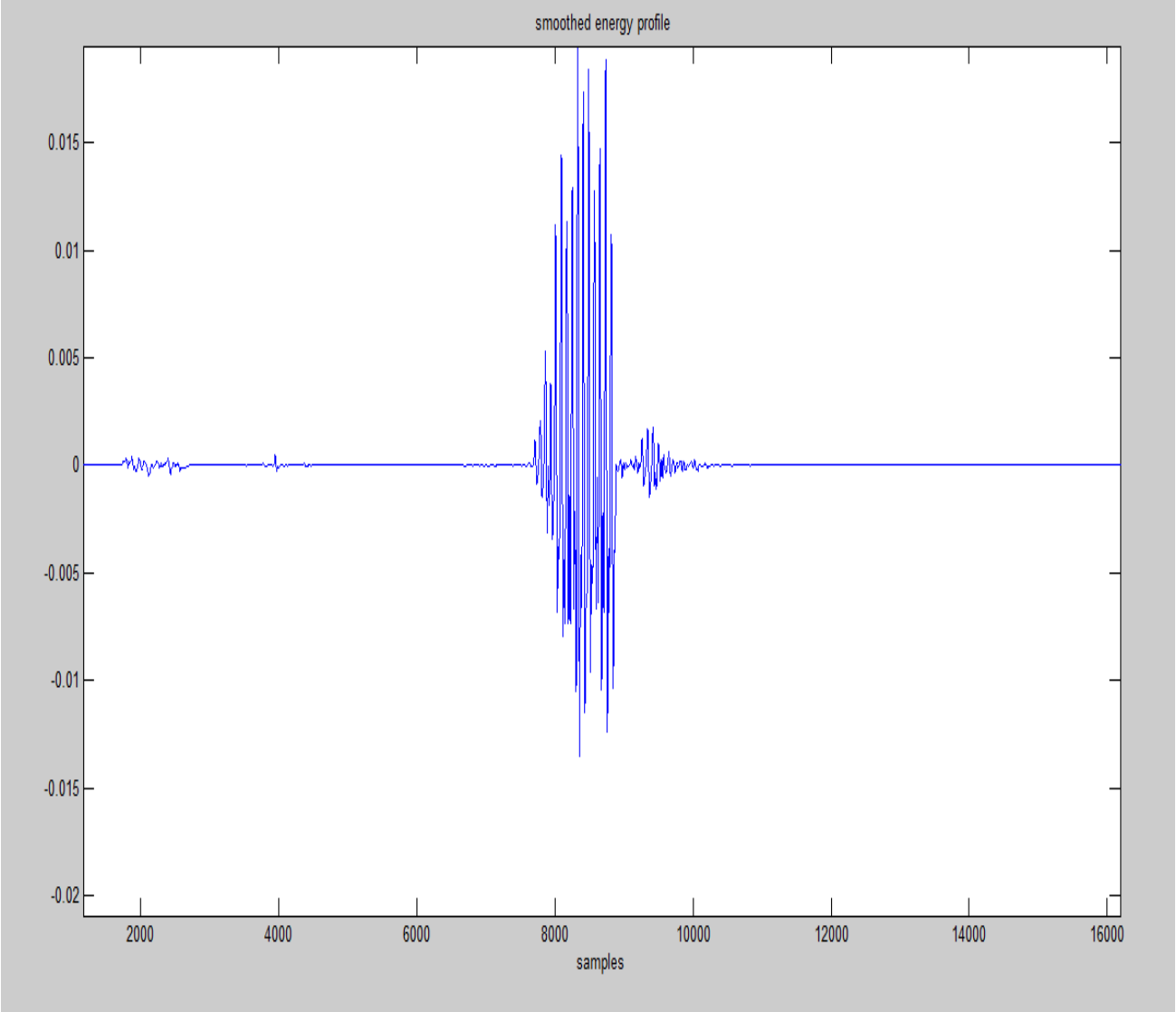


Figure 3.2.4 smoothed energy profile

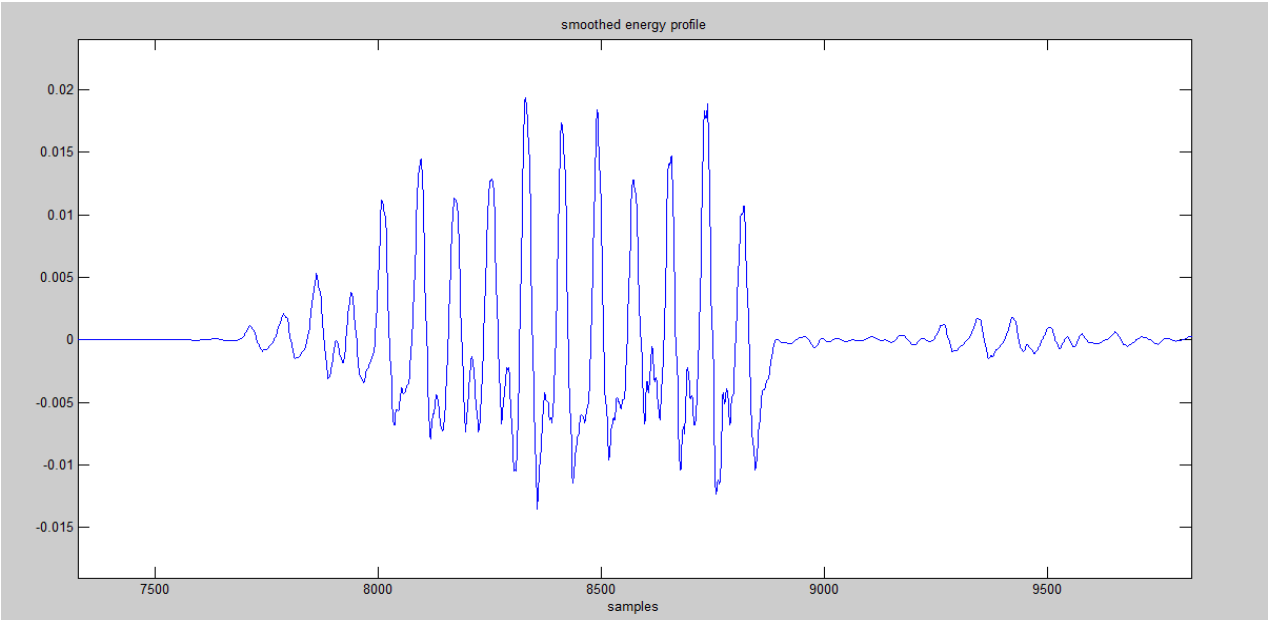


Figure 3.2.5 zoomed smoothed energy profile

It can be seen by comparing energy profile before and after smoothing that peaks and valleys are clearly seen in figure 3.2.5. Now, our next step is to mark these local peaks and valleys which ultimately provide GCIs and GOIs.

- **Location Glottal closure instances and Glottal Opening Instances**

As we know, speech signal is highly complicated non-stationary signal and can be seen in Figure. 3.2.2. So we smoothed energy profile. Now, we need to find out local peaks and valleys which occur at maximum positive and maximum slope of energy profile. To estimate the positions/indexes of maximum positive and maximum negative slope we are differentiating the smoothed energy profile. Here we used a small mathematical trick of differentiation to get maximum positive and negative slopes. Following figure 3.1.6 explain how differentiation helps to estimate maximum positive and maximum negative slopes. Let us consider that we have sinusoidal signal and we want to find out its maximum positive and maximum negative slopes. Since, maximum positive slope occurs at positive zero crossing and maximum negative slope occurs at negative zero crossing.

Either we can trace this zero crossing or we can use differentiation method. Direct tracing for sinusoidal case is very easy. But, for our actual speech input due to its highly non-stationary nature it is not that easy. If we differentiate our sinusoidal signal, output will be a co-sinusoidal signal. One can see in figure 3.2.6 that positive zero crossing in sinusoidal signal is correspond to maximum peak in 1^{st} derivative of sinusoidal signal and negative zero crossing correspond to valley in co-sinusoidal signal which represents GCI and GOI respectively. Thus, by marking successive peaks and valleys we will get successive GCIs and GOIs. The corresponding GCI and GOI for input speech waveform having utterance 'jvaarii' is shown in Figure 3.2.7.

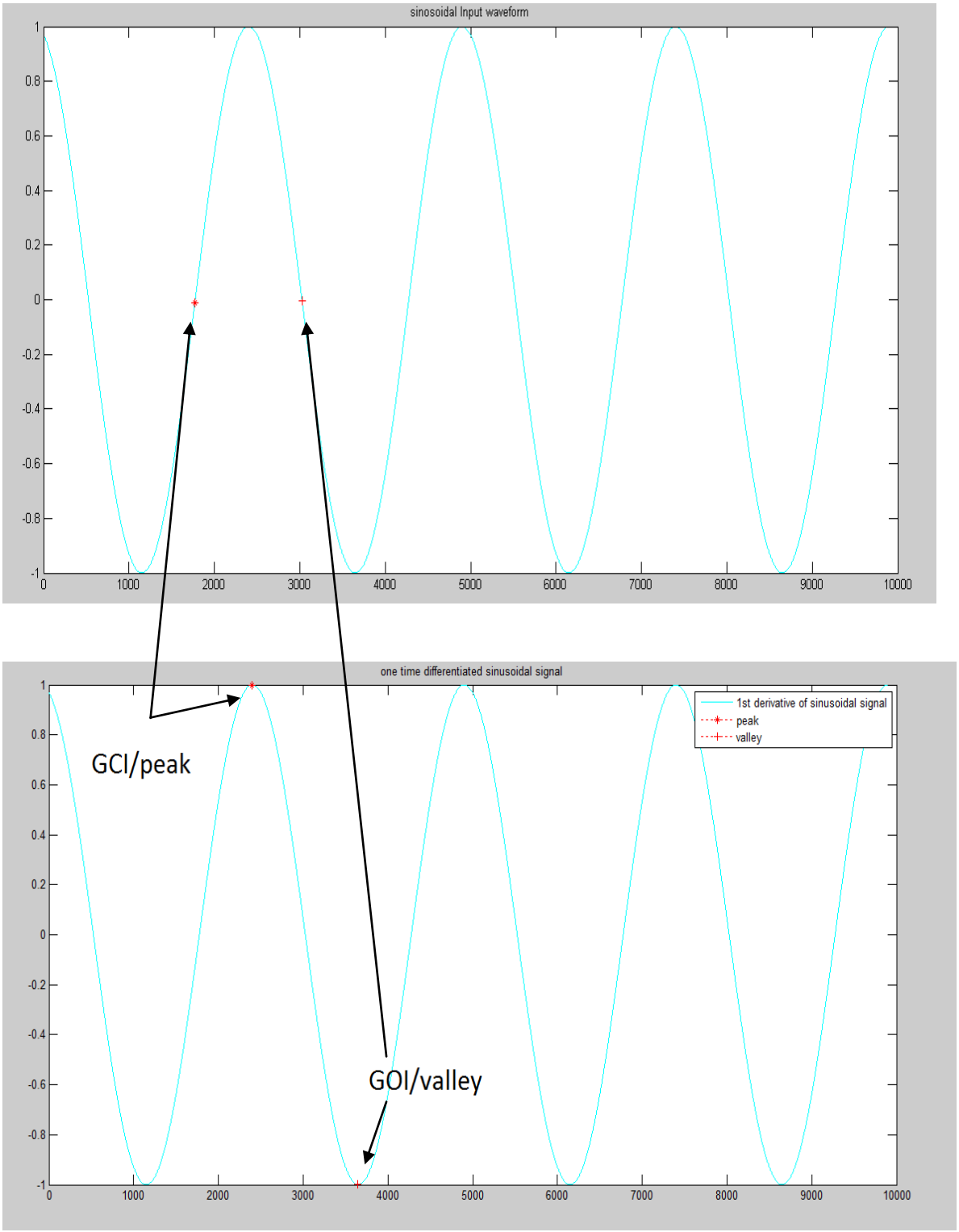


Figure 3.2.6 sinusoidal signal and its 1st derivative signal

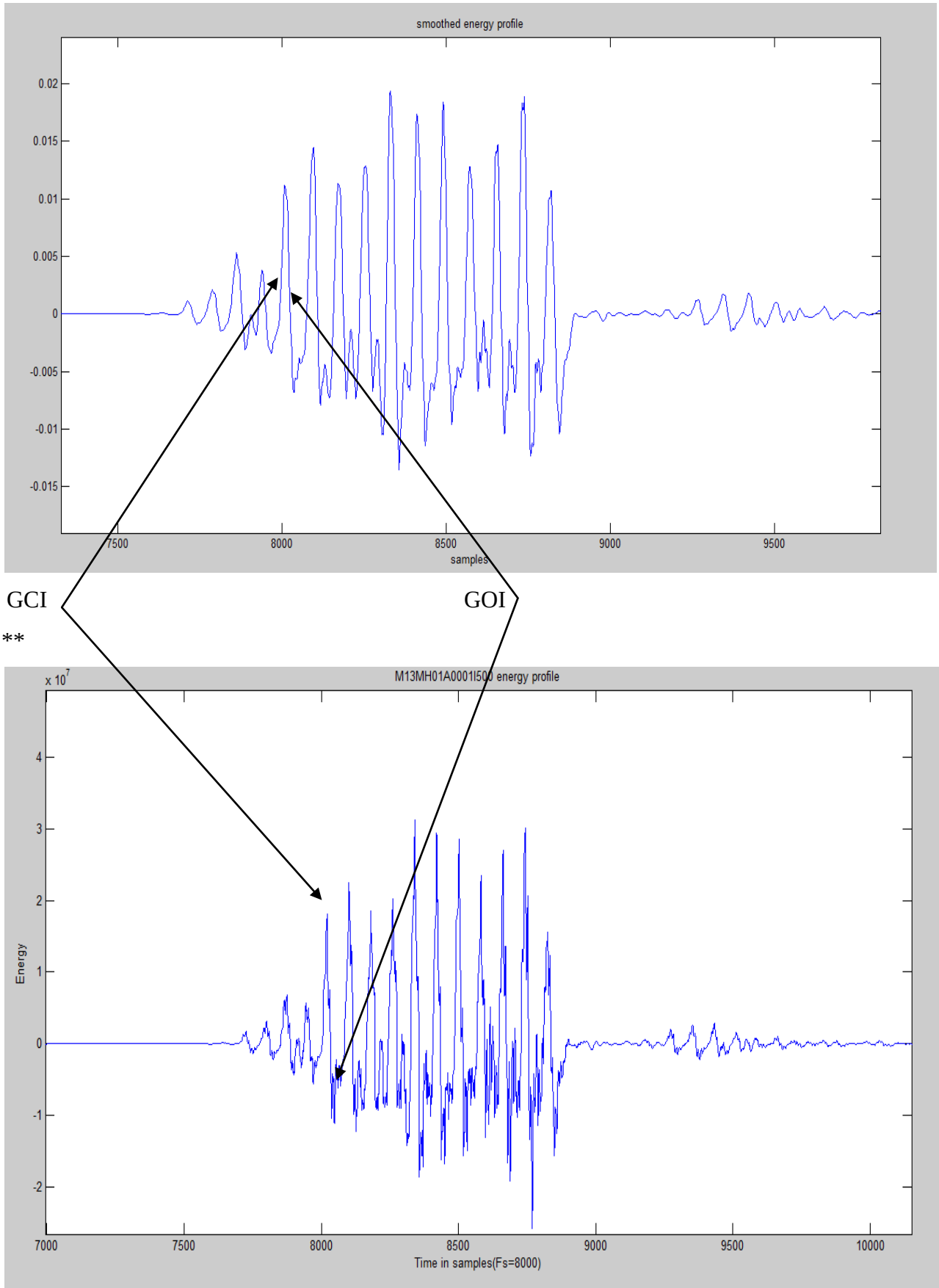


Figure 3.2.7 smoothed energy profile and its 1st derivative, for input speech waveform having utterance "jvaarii"

- **Local Peaks and Valleys**

As, discussed earlier that local peaks and valleys in derived version (first derivative) of smoothed energy profile matches with the maximum positive slope of energy profile and local valleys with the local maximum negative slope occurs in smoothed energy profile waveform. The local peaks represent GCIs and valleys GOIs. Figure 3.2.8 represents local peaks and local valley marked.

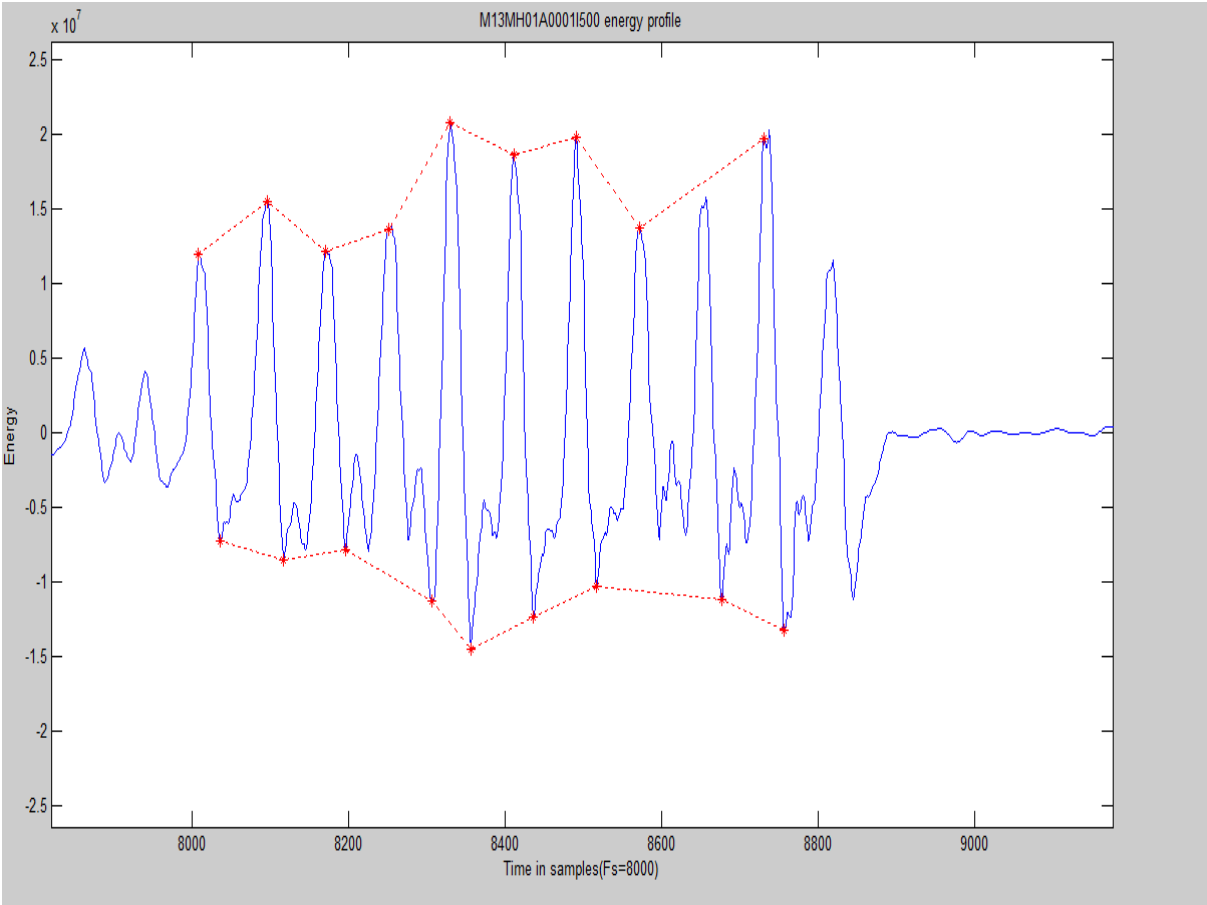


Figure 3.2.8 Local peaks and valleys

According to findpeaks.m, the local peaks will be a peak if and only if it is greater than its neighborhood sample on either side, Also its value has to be greater than half a global peak value. 'max' Matlab command provides global peak and its location. Similarly for global peak I used 'min' command. Following steps explain local peaks and valleys selection more

clearly. As explained already, a sample is declared as a peak if its sample value is higher than its previous and next sample i.e., for ith sample to be declared as a peak,

EmS (i-1) < EmS (i) > EmS (i+1) **(1)**

&

EmS (i) >= (0.5*Maxval) **(2)**

Similarly the condition for a sample to be a valley is,

EmS (i-1) > EmS (i) < EmS (i+1) **(3)**

&

EmS (i) < = (-0.5* Maxval) **(4)**

If we see carefully all valleys occurs after successive peaks positions so that generation of weighting function should not face any problem. But it is not the case all the time i.e., after occurrence of peak their might be another peaks before successive valley which we should discard. After discarding extra valleys and peaks we can generate proper weighting function.

- **Weighting Function**

Using local peaks and local valleys we will generate proper weighting function for speech files. Figure 3.2.9 shows weighting function estimated for speech input having utterance "jvaarii". Weighting function starts from successive GCI and ends at GOI.

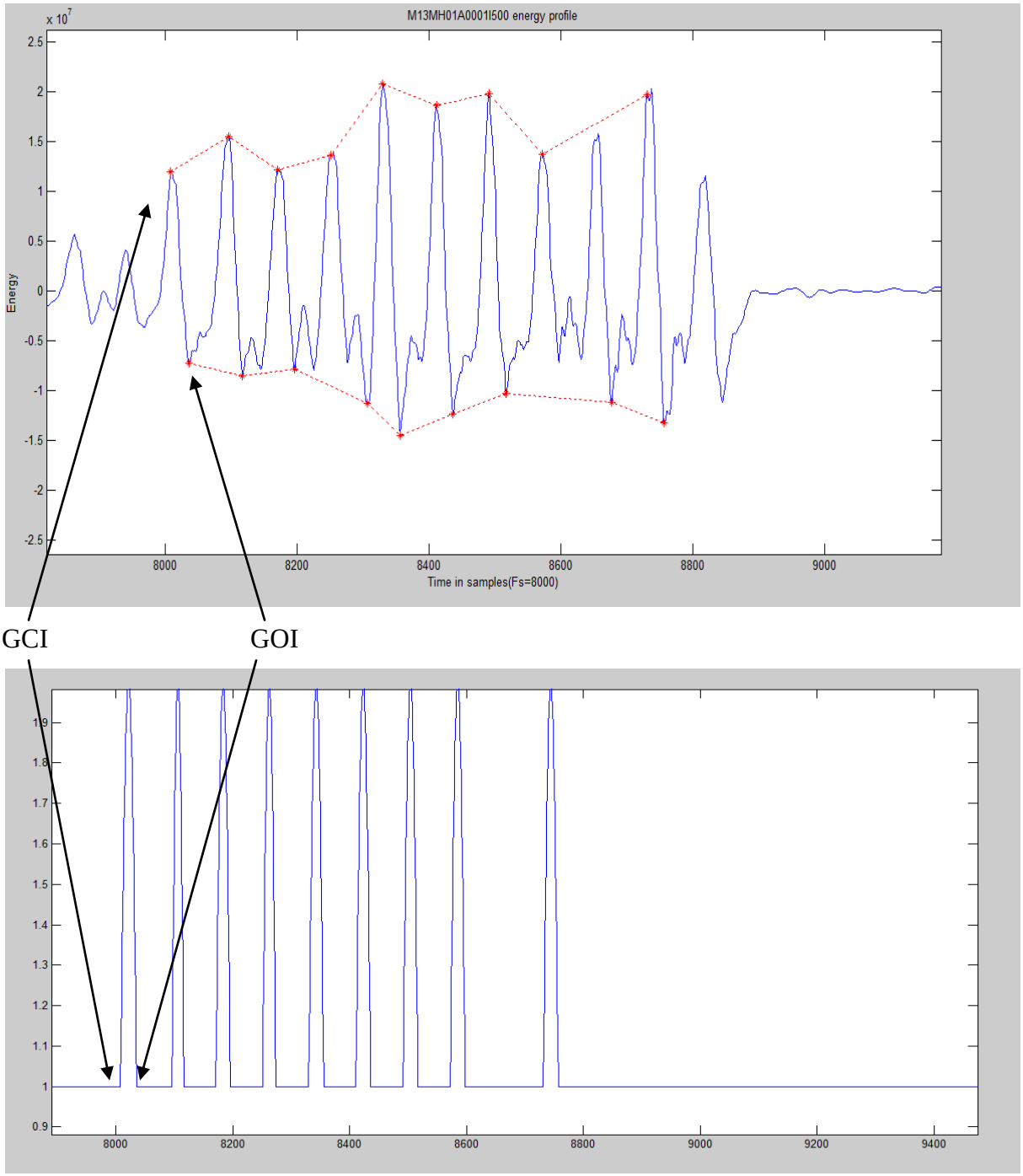
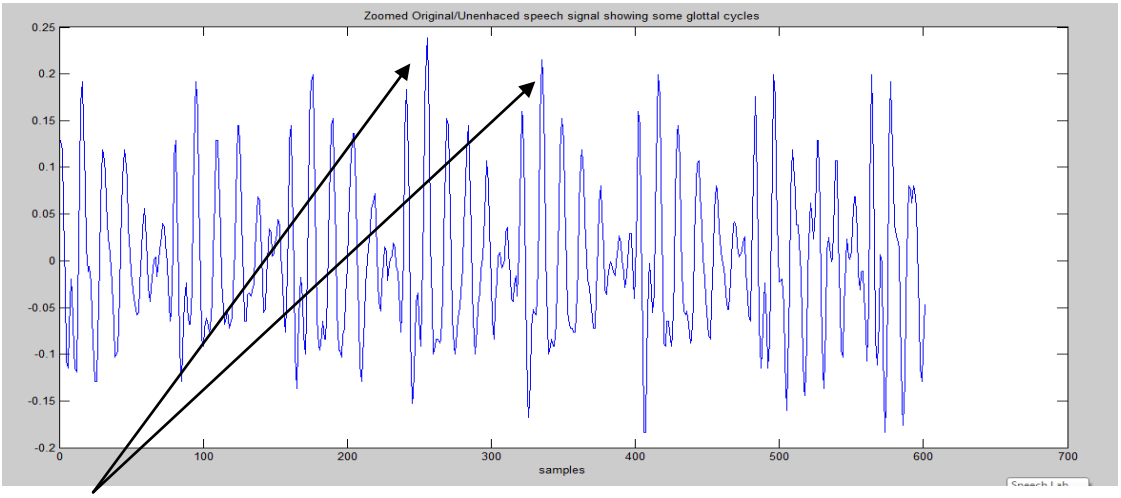


Figure 3.2.9 weighting function estimated from local peaks and valleys location

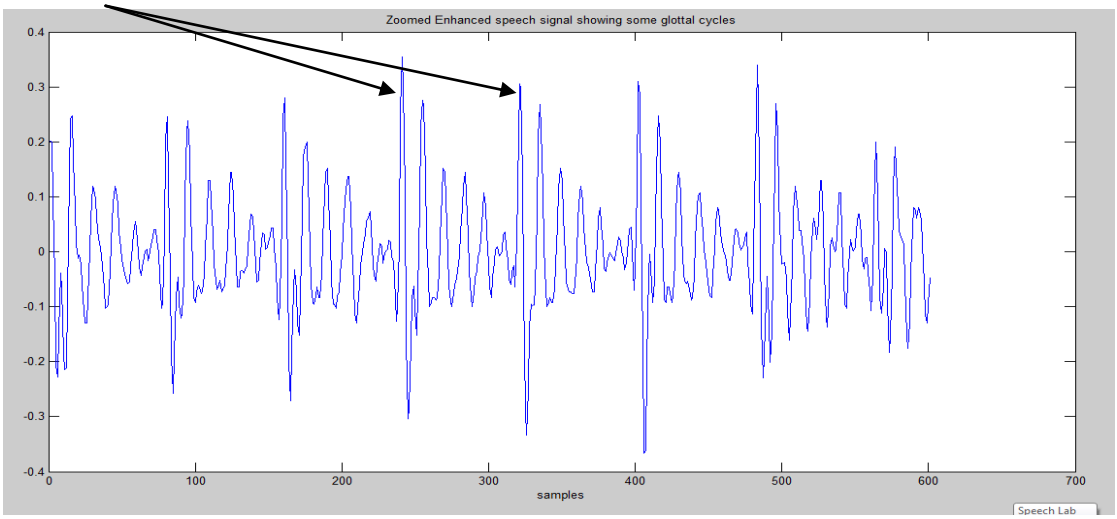
Enhanced Wave File

Multiplying input speech wave file with its appropriate weighting function for enhancement. As, weighting function is greater than 1 for samples under glottal closure phase after multiplication we get enhanced speech. Thus, we significantly enhanced input speech signal

and is shown in Figure 3.2.10. We can see in figure 3.2.10 which shows original as well as enhanced file that voiced samples under glottal closure phase are enhanced.



Amplitude comparison Original speech



Enhanced speech

Figure 3.2.10 zoomed original and enhanced speech wave file.

If we observe Y axes of both graphs carefully, we observe that local maximum values for enhanced one is double that of the local maximum values in original speech signals. Amplitude comparison directly shows that enhanced speech file has higher SNR under glottal

closure phase. Also voiced portion in enhanced file can be seen clearly in enhance file. Now, this enhanced file will feed to conventional ASR system to estimate recognition accuracy.

3.3 Automatic Speech Recognition

ASR is one of the most promising research interests in speech processing domain. ASR functions in most of daily usages viz., voiced control security systems, voice controlled embedded systems like car, music systems etc. The ASR system used for this project is actually a telephone help centre project for farmers sponsored by DIT (department of information technology), India. The project is about building a telephonic help line to Indian farmers where they will come to know commodity prices in all mandis/markets in Maharashtra. So that he can decide where to sell his commodity to get more profit. The data base used for ASR is Marathi speaker isolated words data base collected over 34 districts of Maharashtra, from 1500 speakers i.e., around 45 speakers per District. Total file are around 45000. [6] For building ASR system, sphinx-3 ASR tool kit is used.

Automatic speech recognition system involves two broad steps viz., training and decoding as shown in Figure 3.3.1. Also, signal processing is an important step in ASR system.

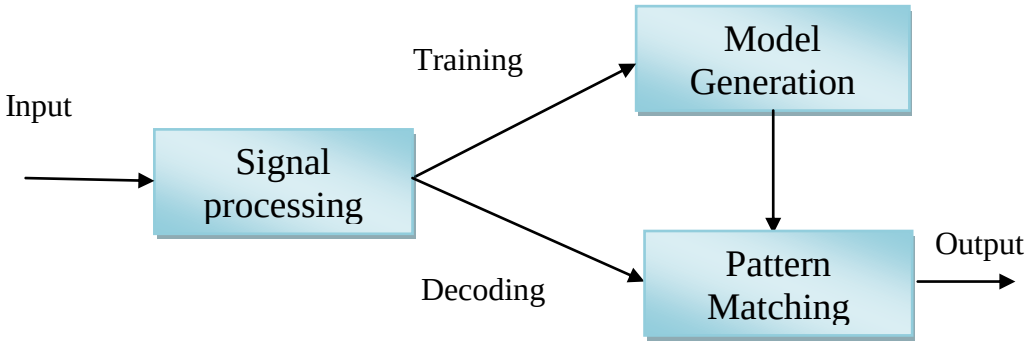


Figure 3.3.1 ASR system block diagram

Signal Processing

In this step signal is first processed then signal is feeded to Model generation block for estimation of coustic models and features are feeded to pttern matching block for decoding. Signal processing step also involves two important steps viz., silence truncation and mfcc feature extraction as shown in Figure 3.3.2. In silence truncation, silence more than 30 frames is truncated to 30 frames. Each frame is of size 10 msec, 30 frames are of 0.3 msec. Thus, silence truncation program truncated silence more than 0.3 msec to 0.3 msec. After silence

truncation speech signal is feeded to MFCC estimation block. Here speech signal is processed to estimate 39 dimentional MFCC features vectors (mfcc, mean and variance vectors).

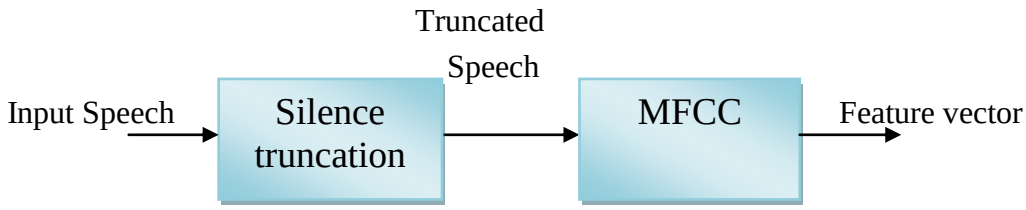


Figure 3.3.2 Signal processing block diagram of ASR system.

Training

In this step silence truncated speech signal is use as input. Training of acoustical/statistical and language model is a primary goal of this step. The sequential steps occurs during acoustical model estimation are shown below in Figure 3.3.3. and for language models is shown in Figure 3.3.4.

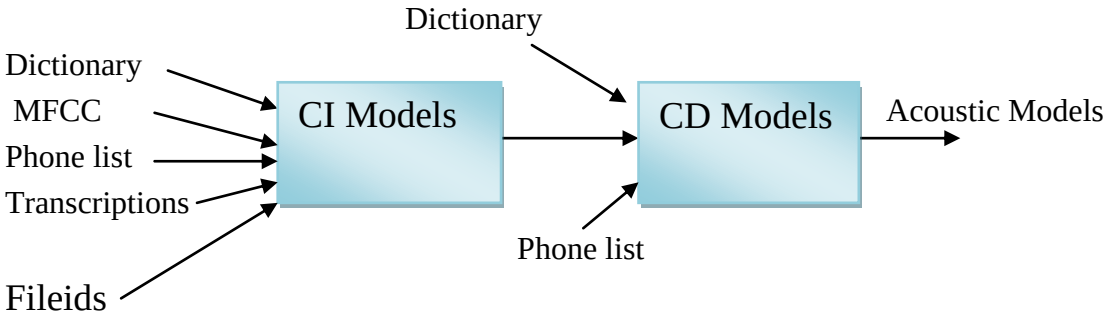


Figure 3.3.3 Training Context independent models and Context dependent models

The MFCC, phone list, dictionary, transcriptions and file ids are essential inputs to for generation of context independent (CI) models. The CI models are generated using forward and backward algorithm called Baum-Welch algorithm. As per this algorithm, after several iterations, it tries to match models having maximum likelihood with the utterance only which will not be a universal model. So we need to put cut-off limit for maximum number of iterations which is kept 10 here. Also, upper limit is there on a ‘convergence ratio’. Convergence ratio is nothing but ratio of maximum likelihood entry to its maximum likelihood. Once context independent models are generated the next step is to generate context dependent models.

In context dependent models the first step is to list all tri-phones possible in vocabulary. Vocabulary is generated from provided dictionary. Then, in next step we will find number of times each of tri-phones occurred in dictionary. On the basis of number of occurrences more

than a specific threshold, tri-phones are short listed. The threshold is being kept according to convenience of memory available and need of number of tri-phones. The CI phones and short listed phones are used to estimate untied CD model definition file. After flat initialization actual CD untied models will be generated using same Baum-Welch algorithm. Once CD untied models are computed next step is decision tree building. Decision trees are used to decide which of the HMM states estimated from all tri-phones are similar, so that data from these states can be collected together and tied to get one global state called 'senone'. Many of such states can be tied. The number of senones are user defined here we used 1000 senones. The cd models are generated using senones and decision tree. [5]

Language Model

The n-gram Language model (LM) is used in this project. The language model is generated using occurrences of phones in vocabulary generated from dictionary and transcriptions.
In n gram language model is generated by considering occurrence of n possible phone entries based on linguistic probabilities.

Decoding

Decoder provides best possible hypothetical phoneme sequence based on output of viterbi algorithm. Decoder uses lexical tree structure for decoding. Decoder uses language models, acoustic models, lexical models and input transcription (train) file as input to generate hypothetical transcriptions file.

Evaluation

Recognition accuracy is estimated using dynamic programming. The generated hypothetical transcriptions are compared with input (train) transcriptions which results out recognition accuracy in percentage form.

CHAPTER 4
RESULT ANALYSIS

This chapter discuss about performance after and before introduction of preprocessing block in regular ASR system. Recognition accuracy estimated via TrainAndTest and kfold evaluation. In TrainAndTest evaluation training and testing data is same, while in case of kfold evaluation train and test data are different. Above mention evaluation approaches are used for both enhanced and unenhance/original wave files. During both the evaluations, three kind of errors occurred viz., substitution, deletion and insertion.

To explain errors let us consider an example let us say we have training speech file having utterance ‘krushi utpanna baazaar samiti’. In case of subtituion errors, due to some problem hypothetical words which is a text output of decoder is substituted instead of correct word i.e., in this case instead of ‘krushi **utapanna** baazaara samiti’ if we get out output as ‘krushi **nagar** baazaara samiti’ i.e., decoder has substituted word **nagar** instead of **utapanna** leads to substitution error. In case of deletion error some words will be deleted from original train transcription i.e., in ‘krushi utapanna baazaara samiti’ if we may get output ‘krushi utapanna samiti’ i.e., **baazaara** word is deleted by decoder and in evaluation it shows deletion error. Similarly in case of insertion extra word is inserted in original transcription i.e., for ‘krushi utapanna baazaara samiti’ if we get out ‘krushi utapanna baazaara samiti **pusad**’ i.e., here extra word **pusad** is inserted which leads to insertion error.

4.1TrainAndTest Evaluation

In this evaluation approach training and testing data are same i.e., train=test. All 45521 files are used for training (discussed in section 3.3) and same (45521) files are used for decoding. The results are estimated for both Marathi Enhanced (M.E.) and Marathi Original (M.O.) data seperartely. Table 4.1.1 shows sentence recognition accuracy and table 4.1.2 shows word recognition accuracy for both Marathi Enhnaced and Marathi Original data. We can see improvement in sentence recognition that sentences with total errors (i.e., deletion, substitution, insertion) are 22.5% (10022) for Enhanced data and 23.5% (10465) i.e., 443 extra transcriptions are correctly recognized after enhancing the speech date using proposed algorithm. Also for enhanced files, individually % substitutions, insertion and deletion are reduced. The % substitution reduced from 16.9 % (7537) to 16.0 % (7110) i.e., 425 lesser

substitution occurs after enhancement. The % deletion reduced 4.3 % (1898) to 4.0 % (1802) reflects 96 lesser deletions and In case of insertion also there is reduction of 0.1% (62).

Table 4.1.1 TrainAndTest sentence recognition performance for complete data base.

(train=test)	Correctly Recognised Words	Total Error	Substitutions	Deletion	Insertion
Marathi Original	77.9% (35460)	22.1% (15679)	11.8% (8417)	3.8% (2714)	6.4% (4548)
Marathi Enhanced	79.0% (35961)	21.0% (14926)	11.1 % (7110)	3.7% (1802)	6.2% (4406)

In word regonition also, recognition accuracy has been increased by 1.1 % i.e., (500) more word transciptions are recognised in enhanced case. Total % of errors (deletions, substitutions and insertion) are also reduced from 22.1 % to 21.0 %. Indivisually %substitutions reduced by 0.7 %, deletions by 0.1% and insertio by 0.2 %. Thus, we can say proposed algorithm works well for TrainAndTest (train=test) evaluation.

Table 4.1.2 TrainAndTest word recognition performance for Marathi Original (M.O.) and Marathi Enhanced (M.E.)

(train=test)	Sentences with errors	Substitution	Deletion	Insertion	Sentence with Accuracy
Marathi Original (M.O.)	23.5% (10465)	16.9% (7537)	4.3% (1898)	7.5% (3336)	76.5%
Marathi Enhanced (M.E.)	22.5% (10020)	16.0 % (7110)	4.0% (1802)	7.4% (3274)	77.5%

4.2 Kfold Evaluation

To check the robustness of our algorithm we did kfold evaluation also. In kfold evaluation training and testing data bases are different i.e., ASR system is trained with oneset of data and test with another set of data to observe system perfornamnce with unseen data. Here for this project value of k is 3 i.e., it is 3-fold evaluation. For this approach of evaluation Marathi Original and Marathi Enhanced data is divided in to nearly three equal parts. During partitions caution is taken care that number of male and female speakers from each district are partitioned in nearly three equal parts i.e., not only total number of speakers present in 34 districts are partitioned but male female ratio (districtwise) is also mentain. For e.g. in district 01 we have 1308 male speakers files and 28 female speakers files so each of the three parts will have around 436 male speakers files and around 9 female speakers files from district 01. Thus, we divide original data in to three nearly equal parts titled as marathiAgmark1, marathiAgmark2 and marathiAgmark3. Two of this three parts are combined to create a data base for kfold evaluation resulting in three sets titled marathiAgamark12 which is a combination of marathiAgmark1 and marathiAgmark2 like wise marathiAgmark13 and marathiAgmark23 will be generated. Now,in kfold evaluation one of above mention three parts (i.e., marathiAgamark12/13/23) will be used for training and another part (except train part) is used for testing. Also, we performed train=test evaluation for kfold partitioned data i.e., for kfold data we did two checks viz., train=test and train!=test and results are shown in table 4.2.1 and 4.2.2 respectively.

Table 4.2.1 3-fold (train=test) decoding and evaluation results for Marathi Enhanced (M.E.) and Marathi Original (M.O.) data.

(train=test)	Accuracy (%)		Substitution (%)		Insertion (%)		Deletion (%)	
	M.O.	M.E.	M.O.	M.E.	M.O.	M.E.	M.O.	M.E.
Marathiagmark12	82.7	82.1	8.6	8.8	5.3	5.6	3.4	3.5
MarathiAgmark13	82.7	82.5	8.6	8.7	5.3	5.3	3.4	3.4
Marathiagmark23	82.4	82.4	8.6	8.6	5.6	5.6	3.4	3.4
Average	82.6	82.33	8.6	8.7	5.4	5.5	3.4	3.43

From tabular results it can be seen cleary that for train=test recognition accuracy of enhanced data is nearly equal to Marathi Original data. Also %substituion, %deletion and %insertion also almost equal. Thus, we can say that this algorithm stands well for 3-fold (train=test evaluation).

Table 4.2.2 3-fold (train!=test) decoding and evaluation results for marathi enhanced and original data

(train!=test)	Accuracy (%)		Substitution (%)		Insertion (%)		Deletion (%)	
	M.O.	M.E.	M.O.	M.E.	M.O.	M.E.	M.O.	M.E.
Marathiagmark12	64.8	65.3	21.9	21.8	7.8	7.4	5.5	5.9
MarathiAgmark13	65.7	65.8	21.7	21.9	7.2	7.1	5.4	5.2
Marathiagmark23	64.7	64.2	22.5	22.7	7.7	7.9	5.1	5.2
Average	65.06	65.10	22.0	21.13	7.56	7.46	5.33	5.43

Table 4.2.2 also, shows that percentage Average Accuracy is higher in case of enhanced case. Also indivisual errors viz., substituion, insertion are comparatively less. Thus, algorithm passes both the kfold evaluation tests and shows better results.

CHAPTER 5
CONCLUSION AND FUTURE SCOPE

In this project, we examined whether enhancement of speech samples during glottal closure interval improves speech recognition accuracy over original/unenhanced data. In this chapter, a summary of findings is presented and throw some light toward future scope of work.

5.1 Summary

In this project, we proposed and examined a method based on short time energy profile to enhance speech samples under glottal closure phase. The algorithm describes a method of detecting voiced portions present in speech signal. Also, the algorithm explains detection of instance of glottal closure (GCI) and instant of glottal opening (GOI). In order to identify the glottal closure interval in each pitch cycle of voiced speech segment. By emphasizing speech samples in glottal closure interval, we observed significant improvement while decoding training data, and marginal improvement while decoding unseen (test) speech data.

5.2 Contributions of this Work

The important contributions of discussed in this research project are mention below :

1. Estimation of voiced portions present in speech signal using short time energy profile.

1. Estimation of instance of glottal closure (GCI) i.e., epoch location and instance of glottal opening (GOI).

2. Generation of an appropriate weighting function to emphasize speech samples diring glottal closure interval .

3. Examination of potential inprovement in speech recognition accuracy due to the proposed speech enhancement method.

5.3 Future Scope of Work

- In this project work we used energy profile for several purposes viz., voiced portion detection, estimation of instance of glottal closure (GCI) and instance of glottal opening (GOI) within pitch period. Recently, a method based on zero frequency filtering (ZFF) proposed [1,2] and was shown to detect GCI and GOI with higher accuracy. Usage of ZFF method instead of short time energy profle to estimate glottal closure interval may result in better improvement.

- Speech segments whose time energy values were greater than or equal to a threshold were detected as voiced speech segments. In this work, the threshold was set as half of the maximum energy value in the speech file. Determination of optimal threshold based on experimentation may lead to better voiced portion detection for multi speaker data base.

- Speech samples during glottal closure interval were enhanced using a weight function.Weighting function used in this project is positive half cycle of sinusoidal function generated from GCI and ends at respective GOI. Several other shapes of weighting functions can be used to generate different weighting functions and enhance speech using such different weighting functions. This approach may lead toward better speech recognition accuracy.

REFERENCES

1. B. Yegnanarayana, K. Sri Rama Murty and M. Anand Joseph, "***Characterisation of glottal activity from speech signal***," *IEEE SIGNALS PROCESSING LETTERS*, VOL. 16, NO. 6, J UNE 2009\
2. R. Smits and B. Yegnanarayana***, "Determination of instants of significant excitation in speech using group delay function,"*** *IEEE Trans.Speech Audio Process.*, vol. 3, no. 5, pp. 325–333, Sep. 1995
3. Rabiner L.R., Juang Biing-Hwang, B. Yegnanarayana, **"*Fundamentals of speech recognition*",** India 1st Edition, published by Pearson Education.
4. Anatomical views of Larynx and vocal folds <www.mayoclinic.com>
5. sphinx manual <www.speech.cs.cmu.edu/sphinxman/scriptman1.html>
6. T. Godambe and Samudravijaya K, "***Speech Data Acquisition for Voiced Based Agricultural Information Retrieval***," Tifr, Mumbai.

BIBLIOGRAPHY

1. Thomas Drugman and Thierry Dutoit, "***Glottal Closure and Opening Instant Detection from Speech Signals,"*** *TCTS Lab, Faculty Polytechnique de Mons-31, Boulevard Dolez, 7000, Mons, Belgium.*

2. K. Sri Rama Murthy and B. Yegnanarayana, ***"Epoch Extraction From Speech Signals"*** *IEEE transactions on audio, speech, and language processing, vol. 16, no. 8, November 2008.*

3. R Smiths and B. Yegnanarayana, **"*Determination of instants of Significant* excitation in speech using group delay function,"** *IEEE Trans.1.Speech Audio Process, vol.3, no. 5, pp 325-333, sept 1995.*

4. B. Yegnanarayana and R. N. J. Veldhuis, **"*Extraction of vocal-tract system characteristics from speech signal,"*** *IEEE Trans. Speech Audio Process, vol. 6, no. pp. 313-327, Jul. 1998.*

5. B. Yegnanarayana and P. Murthy, ***"Enhancement of reverberant speech using LP residual signal,"*** *IEEE Trans. Speech Audio Process, vol.8, no.3,pp 267-281, May 2000.*

6. H.W. Strube, ***"Determination of the instant of glottal closure from the speech wave,"*** *J. Acoust. Soc. Amer., vol.56, pp.1625-1629, 1974.*

7. Y. M. Cheng and O'Shaughnessy, "***Automatic and reliable estimation of glottal closure instant and period,"*** *IEEE Trans. Acoust., Speech, Signal Process., vol. 27, no. 12, pp. 1805–1815, Dec. 1989.*

8. B. Yegnanarayana and R. L. H. M. Smits, ***"A robust method for determining instants of major excitations in voiced speech,"*** *in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., Detroit, MI,May 1995, pp. 776–779.*

9. P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, ***" Estimation of glottal closure instants in voiced speech using the DYPSA algorithm,"*** *IEEE Trans. Audio, Speech Lang. Process., vol. 15, no. 1, pp. 34–43, Jan. 2007.*

1. ANNEXURE

Matlab programs used for enhancement of speech samples under glottal closure portion 1.Main program, 2.short time energy (STE) estimation, 3.Peaks and 4.valleys detection, weighting function are given below,

1 Main program

Here this program calls all subroutine programs used.

```
% Enhancement of speech sample under glottal closure phase %

tic;
WaveFileList = textread('wavList.txt','%s'); % WaveFileList is a cell array
which contains list of all the wave files with locations
lwls=length(WaveFileList); %length of wave list
for i=1:lwls %counter set to read all the wave files
      clc;
      celldisp(WaveFileList(i)) %displays current file name
      i
      display('file is under process.....');
      [s]=wavread(WaveFileList{i}); % Reading through the a cell array
      [EmS,Em,E]=STE(s);%estimating the smoothed short time energy profile
      %estimating peaks and valleys
      [MaxIdxf,MinIdxf,Lp,Lv,a,b]=PV(EmS);
      %weighting function estimation and
      if min(Lp,Lv)>2
            [W]=BP(EmS,MinIdxf,MaxIdxf);%finding appropriate weighting
            function
            W=W+1;
            for j=1:length(W)

                  s1(j)=(s(j)*W(j)); %multiplying a waveform by its
                  appropriate weighting function

            end
            %WaveFileList_Final= strcat(WaveFileList,'e1');
            Y=s1;
            M=strcat(WaveFileList(i,:),'e2');
            AM=cell2mat(M);
            WAVWRITE(Y,AM);
      end
 end
 toc
```

==========

2. Short time energy (STE) estimation

```
function [EmS,Em,E]=STE(s)
M=16; %window lenght
x=double(s);
[nr,~]=size(x); % nr and nc contains total number of samples i.e. total
number of rows and columns in d respectively
E = zeros(1,nr); % initialising energy vector
```

```
K=0;
for n = M+1:nr
    for i = 1:M
      K=(x(n-i)*x(n-i));
      E(n)=E(n)+K;
    end
 end
Em=zeros(1,nr);
%%%for smoothing the energy waveform%%%%%
for j=1:length(E)-(M+1)
      A=0;
      for k=j:(M+j-1)
          A=A+E(k);
          Em(j)=A/M;
      end
end
%%%%%1st derivative of smoothed energy profile%%%%%
for k=1:length(Em)
      if k==1 %%%%%% for 1st sample derivative is next sample-present
      sample
                  EmD(k)=Em(k+1)-Em(k);

      elseif k==length(Em) %for last sample Em(i+1) not allowed

                  EmD(k)=Em(k);
      else
                  EmD(k)=Em(k+1)-Em(k-1);
          end
end
%%%smoothing the differentiated w/f to get the local peaks%%%

for j=1:length(EmD)-(M+1)
      A=0;
      for k=j:(M+j-1)
          A=A+EmD(k);
          EmS(j)=A/M;
      end
end
end % function
```

==========

3. Peaks and valleys detection

```
function [MaxIdxf,MinIdxf,Lp,Lv,a,b]=PV(EmS)
%the following program is to get desire peaks and desire valleys

[Maxima,MaxIdx] = findpeaks(EmS);
DataInv = 1.01*max(EmS) - EmS;
[~,MinIdx] = findpeaks(DataInv);
Minima = EmS(MinIdx);
%%following two 'for' loops used to get desired peaks and valley samples
instances%%
a=min(EmS)/2;
for i=1:length(Minima)
      if Minima(i)<a
            Minima(i)=Minima(i);
            MinIdx(i)=MinIdx(i);
      else
            MinIdx(i)=0;
```

```
            Minima(i)=0;
      end
end
b=max(EmS)/2;
for i=1:length(Maxima)
      if Maxima(i)>b
            Maxima(i)=Maxima(i);
            MaxIdx(i)=MaxIdx(i);

      else
            MaxIdx(i)=0;
            Maxima(i)=0;
      end
end
%Keep non-zero values from MaxIdx,Maxima, Minima and MinIdx%
MaxIdxf= MaxIdx(MaxIdx~=0);
MinIdxf= MinIdx(MinIdx~=0);
MaxIdxf=sort(MaxIdxf,'ascend');
MinIdxf=sort(MinIdxf,'ascend');
Lp=length(MaxIdxf);
Lv=length(MinIdxf);

%%% extraa unwanted peaks and valleys which are near to each other less
%%% than 16 samples will be erased
if min(Lp,Lv)>1
MaxIdxf1(1,1)=MaxIdxf(1,1);

for i=2:Lp
    if MaxIdxf(i)-MaxIdxf(i-1)>16
        MaxIdxf1(i)=MaxIdxf(i);
    end
end
MaxIdxf1=MaxIdxf1(MaxIdxf1~=0);
MinIdxf1(1,1)=MinIdxf(1,1);
for i=2:Lv
    if MinIdxf(i)-MinIdxf(i-1)>16
        MinIdxf1(i)=MinIdxf(i);
    end
end
MinIdxf1=MinIdxf1(MinIdxf1~=0);
V=sort(MinIdxf1,'ascend');
K=sort(MaxIdxf1,'ascend');
Lp=length(K);
%peaks   => MaxIdxf(i);
%valleys => MinIdxf(i);
% following program is to filter out unwanted peaks and valleys coming in
between. The program places peaks and valleys in odd and even positions in
A respectively
%The aim here is, peak should be greater than its previous valley and
%current valley should be greater than current peak, which helps us to get
%alternate peaks and valleys that too desired one.

V=V(V>K(1));%valley vector which should have valley(1) > peak(1)
Lv=length(V);
A(1,1)=K(1);
A(1,2)=V(1);


%1st valley position has to be greater than 1st peak position%
for i=2:min(Lp,Lv)
   l=length(A);
     %if l<min(Lp,Lv)
     if    K(i)>A(1,l)
```

```
            A(1,(l+1))=K(i);
      else
            p=K(i);
            k=i;
            l=length(A);
            while  p < A(1,l)
                   k=k+1;
                   if k <= Lp
                      p=K(k);
                      A(1,(l+1))=p;
                    else
                      break;
                   end
            end
      end

      l=length(A);
       if    V(i)> A(1,l)
        A(1,(l+1))=V(i);
      else
             v=V(i);
             j=i;
             l=length(A);
             while v < A(1,l)
                   j=j+1;
                   if j<=Lv
                   v=V(j);
                   A(1,(l+1))=v;

                   else
                       break;
                   end

             end
      end

end


% above program may generate the unwanted values of peaks and valleys to
filter it we are keeping precautions by retaining proper values

A1(1)=A(1);
for i=2:length(A)
    if A(i)<=A(i-1)
        break;
    else
        A1(i)=A(i);
    end
end
K=A1(1:2:end);
V=A1(2:2:end);

%%% adjusting relative positions of peaks and valleys%%%%
%Explanation- The relative distance between each peaks and valleys should
%not to be more than 2 times diff between previous valley peak distances.

Lp=length(K);
Lv=length(V);
for n=1:min(Lp,Lv)
   if Lp==1
       break;
```

```
   end
   if Lp==2
      if n==2
        diff56=V(n-1)-K(n-1);
        V(n)=K(n)+diff56;
      end
   end
    if Lp>2 && Lv>2  % atleast 1 peakvalley combo should be there
%%for n==1%%
    if n==1
    diff01=V(n+1)-K(n+1);
    diff02=V(n+2)-K(n+2);
    d11=((diff01+diff02)/2);
    end
%%for n==2%%
    if n==2
    diff12=V(n-1)-K(n-1);
    diff22=V(n+1)-K(n+1);
    d21=((diff12+diff22)/2);
    end
%%for n>3%%
   if n>3
    diffa=V(n-1)-K(n-1);
    diffb=V(n-2)-K(n-2);
    d31=((diffa+diffb)/2);
   end
%%%%%%%%%%%%%%%%
   if n==1 && ((V(n)-K(n))>(2*d11))
   V(n)=K(n)+d11;
   end
   if n==2 && ((V(n)-K(n))>(2*d21))
   V(n)=K(n)+d21;
   end
   if n >3 && ((V(n)-K(n))>(2*d31))
   V(n)=K(n)+d31;
   end
V(n)=round(V(n));
    end
  MaxIdxf=K;
MinIdxf=V;
end
end
end %function
```

==========

4. Weighting function generation

```
function [BP1]=BP(EmS,MinIdxf,MaxIdxf)
%%%%plotting weighting function using special V trick %%%%%%%%
%technique to plot sine function of appropriate time period i.e.
 %Require zero crossings is shown below
vEmSf=MinIdxf;
pEmSf=MaxIdxf;
Lp=length(pEmSf);
Lv=length(vEmSf);
%Lv=length(vEmSf);
BP=zeros(1,length(EmS)); %base pointer matrix (which for j=1:min(Lp,Lv)
T=2*(vEmSf(j)-pEmSf(j));
K1=sind(2*180*1/T*(1:T));
    for m=1:length(K1)
```

```
        K(j,m)=K1(m);
    end
  %appropriate value of R2 has to be chosen so that shifting of sample can
be done
    R2=(zeros(1,length(K1)+pEmSf(j)));
  %shifting the values as per require index i.e.for i=1; shifting K1 (1) to
required values R2(i+pEmSf(j))
  for i=1:length(K1)
      R2(i+pEmSf(j))=K1(i);
  end
   %%%plotting only positive values i.e. (pEmSf (i):v2ab (i)) %%%
    for i=1:length(R2)
      if R2(i)>0
           BP(1,i)=R2(i);       % important step to store all the values
safely with diff vectors in a matrix
      else                      % plotting a sine wave for (7:16)
i.e.(pEmSf(2):v2ab(2))
           R2(i)=0;           % access just second row of the matrix
      end
    end
 end

W=BP(1,:); weighting function matrix

end %function
```

PROJECT DETAILS

Student Details			
Student Name	**Vishal V. khadake**		
Register Number	110915010	Section / Roll No	110915010
Email Address	Vishalhadake.30@gmail.com	Phone No (M)	
Project Details			
Project Title	**'Robust Speech Recognition in Noisy environment'**		
Project Duration	11 months	Date of reporting	
Organization Details			
Organization Name	**Tata Institute of Fundamental Research (Tifr), Mumbai**		
Full postal address with pin code	Dr. Homi Bhabha Road, Navy Nagar, Colaba Mumbai, Maharashtra 400005		
Website address			
Supervisor Details			
Supervisor Name	**Dr. Samudravijaya K.**		
Designation	Scientific Officer (F)		
Full contact address with pin code	School of technology and Computer science, Tata Institute of Fundamental Research, 1 Homi Bhabha Road, Colaba, Mumbai-400005, India.		
Email address	samudravijaya@gmail.com	Phone No (M)	91-22-22782322
Internal Guide Details			
Faculty Name	**Dr. Somashekhar Bhat**		
Full contact address with pin code	Dept of Electronics and Telecommunication Engg., Manipal Institute of Technology, Manipal – 576 104 (Karnataka State), INDIA		
Email address	somabhatmit@gmail.com		