SEISMIC DATA COMPRESSION IN MULTI-SCALE FRAMEWORK

By

DIVYANSHU PAWAR

ENGG01201801034

Bhabha Atomic Research Centre, Mumbai

A thesis submitted to the

Board of Studies in Engineering Sciences

In partial fulfillment of

Requirements for the Degree of

MASTER OF TECHNOLOGY

of

HOMI BHABHA NATIONAL INSTITUTE

March, 2021

SEISMIC DATA COMPRESSION IN MULTI-SCALE FRAMEWORK

By

Divyanshu Pawar

ENGG01201801034

Bhabha Atomic Research Centre, Mumbai

A thesis submitted to the
Board of studies in engineering Sciences
In Partial fulfillment of
Requirements for the Degree of

MASTER OF TECHNOLOGY

of

HOMI BHABHA NATIONAL INSTITUTE

March, 2021

Homi Bhabha National Institute

Recommendations of the Thesis Examining Committee

As members of the Thesis examining Committee, we recommend that the thesis prepared by
<Divyanshu Pawar > entitled <" Seismic Data Compression in Multi-scale framework" > be accepted as fulfilling the thesis requirement for the Degree of Master of Technology.

	Name	Signature
Member-1	Dr. Smt. Gopika Vinod	
Member-2	Shri S.K. Bahuguna	Bahuguna
Member-3	Dr. Ajith K. John	
Technical advisor	Shri K.K.Bhuyan	Bhuyan
Member-4	Dr. Manoj Kumar	
Guide & Convener	Dr. S. Mukhopadhyay	Mukhopadhyay
Chairman	Dr. Soumitra Kar	SKar

Final approval and acceptance of this thesis is contingent upon the candidate's submission of the final copies of the thesis to HBNI.

I hereby certify that I have read this thesis prepared under my direction and recommend that it may be accepted as fulfilling the thesis requirement.

Date: 16-03-2021 **Place:** MUMBAI

Mukhopadhyay

Signature of the Guide **Name of the Guide : Dr. S. Mukhopadhyay**

ACKNOWLEDGEMENTS

I am deeply obliged to my guide Dr. S. Mukhopadhyay, Head Seismology division, for paving the way for me and giving constant inspiration, courage and energy during the course of my project work.

I am deeply obliged to my technical advisor Shri K.K. Bhuyan, for his intellectual and technical support, enriching my knowledge with new ideas that helped me to enter into the horizon of research work.

I would like to express my special thanks of gratitude to Scientific Officer Shri Abin.C.Joy, Ms. Mahima Shukla and Smt.S.Smruti Rekha for their technical guidance in completing my Project.

I would like to express my sincere thanks to Shri A.Vijay Kumar, Head, Seismic Instrumentation Section, for his kind support and guidance.

Divyanshu Pawar

DECLARATION

I, hereby declare that the investigation presented in the thesis has been carried out by me. The work is original and has not been submitted earlier as a whole or in part for degree/diploma at this or any other Institution / University.

Divyanshu

Divyanshu Pawar

Contents

Synopsis

In this report, a technique using wavelet transform and thresholding is introduced as a potential modeling tool for data compression. Wavelet signal processing is broadly used for analysis of non-stationary data particularly, real-time seismic signals. In the geophysical analysis, numerous wavelet filters are developed to realize the signal characteristics by multi-resolution analysis. However, the selection of optimal wavelet family for seismic wave analysis is a major issue and as of now no rationale has been put forward for choosing the appropriate wavelet filter. As part of this work, a number of seismic signals were analyzed with wavelets namely Haar (db1), Daubechies (db2 to db7) and an attempt was made to find the best wavelet basis based on performance parameters such as mean square error and percentage energy retained.

The original signal corrupted by Gaussian noise is a long established problem in signal or image processing .This noise can be removed by wavelet thresholding. Though numerous threshold models are available, the most widely used universal thresholding which was applied selectively to each level getting high percentage energy retention (PER) along with satisfactory compression ratio (CR).

For validating the code, short term average/ long term average (STA/LTA) and Akaike information criteria (AIC) picker have been used as onset time detection algorithms because in seismic waves if onset time matches then location, origin of time will also match.

This thesis aimed at developing suitable algorithm for seismic data compression using multi-scale resolution framework. The methodology used for selection of wavelet gave us the “db3” as the most suitable wavelet. Using db3, the proposed method achieved high compression percentage (up to 85%) compared to a commercial software with very less variations in magnitude and onset time.

List of figures

Figures *Page number*

List of Tables

List of Abbreviations

AIC	Akaike Information Criteria
AR	Auto-Regressive
DWT	Discrete Wavelet Transform
SPS	Samples Per Second
FT	Fourier Transform
FIR	Finite Impulse response
LPF	Low Pass Filter
LTA	Long Time Average
LQ	Love Wave
LR	Rayleigh Wave
MSE	Mean Square Error
MRA	Multi Resolution Analysis
PER	Percentage Energy Retained
STA	Short Time Average
SNR	Signal To Noise Ratio
WT	Wavelet Transform
SEED	The Standard for the Exchange of Earthquake Data
SEGY	Society for Exploration of Geophysist- “Y”
VSAT	Very Small Aperture Terminal
P Waves	Primary Waves
S Waves	Surface Waves
IRIS	Incorporated Research Institutions for Seismology

1 Introduction

1.1 Seismology

Seismology is the study of earthquakes and seismic waves generated due to various activities (geophysical or manmade) that move through and around the earth. Nearly one percent of total energy released by an earthquake is converted into seismic waves. It is this energy that travels through the earth and is recorded on seismographic stations located on surface of the earth. Seismology deals with detection, location, source identification and magnitude of ground vibration.

1.2 Objective of work

i. The development of compression and decompression algorithm for seismic data.
ii. Validating the magnitude and onset time of reconstruct and original signal. The magnitude of shock should not vary at all and onset time difference should be strictly less than 1 sec (preferably average 0.1 - 0.4 sec).

1.3 Seismic Waves

The two main types of seismic waves are body waves and surface waves. Body waves travel through the earth's inner layers, while surface waves travel along the surface of the planet like ripples in water. At the hypocentre that lies inside the earth (also called as focus of the earthquake), earthquake radiates seismic energy as body waves only and these waves interact with earth's free surface to generate surface waves. Though surface wave arrives after body waves, it is the surface waves that are almost entirely responsible for the damage and destruction associated with earthquakes.

The tremors continue until the plates catch on each other, which causes them to get stuck and stop moving again. The point where the earthquake originates is known as the earthquake focus. The point directly over the focus has the most energy, and it is called epicentre of the earthquake. The energy moves out from the epicentre in waves. While earthquakes are often caused by sudden movements along the fault lines, this is not the only cause of earthquakes. Underground explosions associated with construction work or mining can also cause earthquakes.

Seismic waves can be further divided into two types:

i. Body waves
    - P-Waves
    - S-Waves

ii. Surface waves

1.3.1 P-Waves:

P wave or primary wave is longitudinal like sound wave and it is always the first among seismic waves that reach the recording station. Rock particles affected by propagating ‘P’ wave oscillate backward and forward in the same direction in which the wave propagates.

1.3.2 S-Waves:

The second type of body wave is the S wave or secondary wave. S wave is transverse in nature also called shear waves and moves slower than the P wave. It can travel only through solid rock and not through the liquid medium.

1.3.3 Surface Waves:

Surface waves travel along the surface of the earth and attenuate faster with depth inside the earth. Surface waves are further divided into two types:

i. Rayleigh Waves (LR)

ii. Love Waves (LQ)

Figure 1.1 show a typical seismic signal. The initial part of the signal is called body waves and later part is called surface waves. The body waves are of small amplitude and high frequency. Whereas the surface wave part is of high magnitude and low frequency. Due to different statistical characteristics in both phases, a seismic signal is called as non-stationary signal.

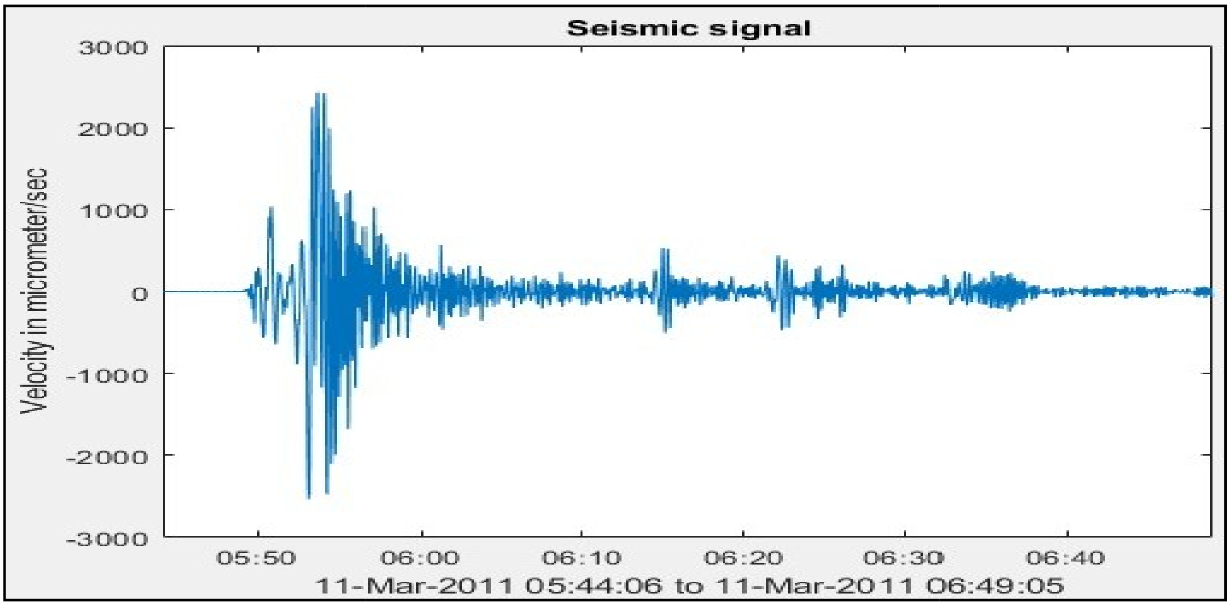

Figure 1.1: A typical Seismic signal

1.4 Instrumentation in Seismology

The connection diagram of seismometer (Sensor), GPS, digitizer is shown in Figure (1.2). The analog signals from sensor and GPS are converted in to digital form by digitizer. Seismometer records the ground vibrations and digitizer converts it to digital form and sends to PC for analysis, processing, event detection and display. Digitizer consists of analog to digital converter (sigma delta of 24 bit resolution) which gives output in counts. These counts are multiplied by resolution (Volt/ Count) of digitizer and divided by sensitivity (Volt/meter/second) of seismometer to get output in velocity (meter/sec). This is the velocity of the ground particles which shook seismometer.

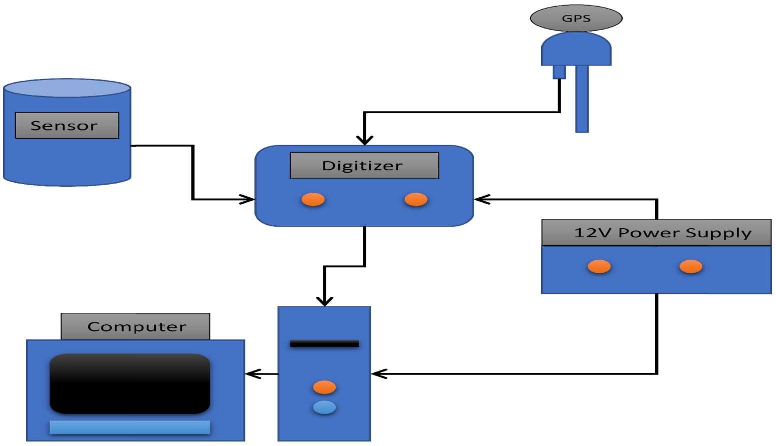

Figure 1.2: Connections of Seismometer, Digitizer and Computer

1.5 Location Estimation in Seismology

Below Figure 1.3 shows the seismic wave propagation diagram from focus to seismic station. First oscillations shows onset of primary-waves and second oscillations shows onset of surface-Waves. Sensor records the ground motion and sends to a seismic data centre via VSAT (Very small aperture terminal) link. GPS is used to append location and timing information.

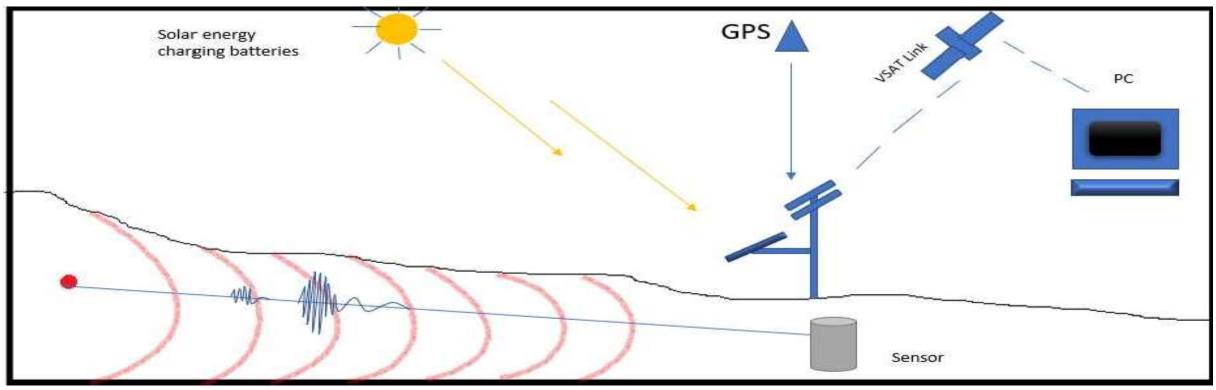

Figure 1.3: Propagation of Seismic data

Δ = Let epicentre to station distance

v_p = P phase velocity

v_s = S phase velocity

T_p = Time at which P wave reaches station

Ts = Time at which S wave reaches station

T_0 = Origin time of event

$$T_p = T_o + \frac{\Delta}{v_p} \tag{1.1}$$

$$T_s = T_o + \frac{\Delta}{v_s} \tag{1.2}$$

$$T_s > T_p \tag{1.3}$$

$$\Delta = (T_s - T_p) \times \left(\frac{v_s v_p}{v_s - v_p}\right) \tag{1.4}$$

Hence if we know time difference of P & S wave and their velocities, we can calculate distance. The direction of incoming earthquake signal is identified using triangulation method. The shear waves (S) travel at about half the speed of compression (P) waves.

$$v_p = 2 * v_s \tag{1.5}$$

$$\Delta = (T_p - T_s) \times (2\, v_s) \tag{1.6}$$

With v_s = 3km/ sec and if we consider uniform delay in P and S waves of original (o) and reconstructed (r) seismic signal, where later is obtained after applying compression and decompression algorithm.

$$\Delta_o = \Delta T_o 2 v_s \tag{1.7}$$

$$\Delta_r = \Delta T_r 2 v_s \tag{1.8}$$

Where Δ_o =epicentre and seismic station distance for original signal

Δr = epicentre and seismic station distance calculated from reconstructed signal

ΔT_o = time difference of P and S wave of original signal.

ΔT_r = time difference of P and S wave of reconstructed signal.

Subtracting (8) from (7) we get,

$$\Delta_o - \Delta_r = (\Delta T_o - \Delta T_r) \times 2vs \qquad (1.9)$$

Let's say time difference between ΔT_o and ΔT_r is 1 second then we get difference in distance of epicentres as 6kms and for time difference of 0.4 seconds we get 2.4 kms as difference which is nominal when we consider teleseismic signals which are thousands of kilometres away.

1.6 Magnitude calculation in Seismology

The magnitude of an event describes the strength of that event. All seismic magnitudes are based on a logarithmic scale (base 10). This means for each whole number one can go up on the magnitude scale, the amplitude of the ground motion recorded by seismograph goes up by 10 times. A magnitude 5 earthquake would result in ten times the level of ground shaking as a magnitude 4 earthquake (and 32 times as much energy would be released).

1.6.1 Richter Scale Magnitude:

There are a number of ways to measure the magnitude of an earthquake. The first widely-used method, the "Richter scale", was developed by Charles F. Richter in 1934. It used a formula based on amplitude of the largest wave recorded on a specific type of seismometer and the distance between the earthquake and the seismometer.

$$\mathbf{M_l = Log_{10}(A) - Log_{10}(Ao)} \qquad (1.10)$$

Where, "A" is the maximum amplitude of "S" wave observed on Wood-Anderson Seismograph (0.8 sec period, 2800 magnification and 0.8 critical damping). "Ao" is amplitude of standard earthquake also called zero magnitude (if A=Ao, M=0) observed on same seismograph. Richter defined Ao as 0.001 mm at 100 km distance. Thus zero magnitude does not mean "No earthquake". The dependency of Richter scale on a type of recording seismograph was the major limitation. This limitation was eradicated by Gutenberg and Richter in defining 'Body wave magnitude (Mb) scale.

1.6.2 Body Wave Magnitude:

$$\mathbf{M_b = \log_{10}(A/T) + Q(\Delta, h)} \tag{1.11}$$

Where, A is P-wave amplitude in microns and T of P-waves is period (< 3 sec) that is frequency range of P-waves is in between 0.5Hz to 2Hz. Q(Δ, h) is the attenuation factor w.r.t distance (Δ) between epicentre and Seismic station location and depth (h).

1.6.3 Surface Wave Magnitude:

$$\mathbf{Ms = \log_{10}(A/T) + 1.66\log_{10}(\Delta) + 3.3} \tag{1.12}$$

Where: A = vertical -component ground amplitude in μm measured from the maximum trace-amplitude of a surface-wave (Rayleigh wave) having a period T between 18 s and 22s. Δ = epicentral distance in degrees, $20° \leq \Delta \leq 160°$.As more seismograph stations being installed around the world, it becomes apparent that the method developed by Richter was strictly valid only for certain frequency and distance ranges. In order to take advantage of the growing number of globally distributed seismograph stations, new magnitude scales that are an extension of Richter's original idea were developed. These included body wave magnitude (Mb) and surface wave magnitude (Ms). Each is valid for a particular frequency range and type of seismic signal. In its range of validity, each is equivalent to the Richter magnitude.

1.6.4 Moment Magnitude Scale:

Because of the limitations of all three magnitude scales (ML, Mb, and Ms), a new, more uniformly applicable extension of the magnitude scale, known as moment magnitude, or Mw, was developed. For very large earthquakes, moment magnitude gives the most reliable estimate of earthquake size. Moment is a physical quantity proportional to the slip on the fault multiplied by the area of the fault surface that slips; it is related to the total energy released in the earthquake. The moment can be estimated from seismograms (and also from geodetic measurements). The moment is then converted into a number similar to other earthquake magnitudes by a standard formula. The result is called the moment magnitude. The moment magnitude provides an estimate of earthquake size that is valid over the complete range of magnitudes, a characteristic that was lacking in other magnitude scales.

$$\mathbf{Mw = Log_{10}\ Mo /1.5 - 10.73} \qquad (1.13)$$

$$\mathbf{Where{:}\ Mo = \mu\ A\ D} \qquad (1.14)$$

Where μ is modulus of rigidity (= 3 x 1019 dyne/cm^2) A is fault area (= L x W) and D is average displacement along fault.

1.7 Stationary Vs Non-stationary signals

A signal is an observation, mathematically it is a recording of a series of events as a result of some process. The stationary signal would have the time period, frequency and spectral content constant, while in not-stationary signals all these fundamental assumptions are not valid. For example a signal whose frequency content do not change in time comes in category of stationary signals. In this case one does not need to know at what times frequency components exists, since all frequency component exist all times. For example:

$$x(t) = \cos(2\pi 5t) + \cos(2\pi 20t) \qquad (1.15)$$

This is a stationary signal because the frequencies of 5, 20, Hz at any given time instant. That means all frequencies are present at all times.

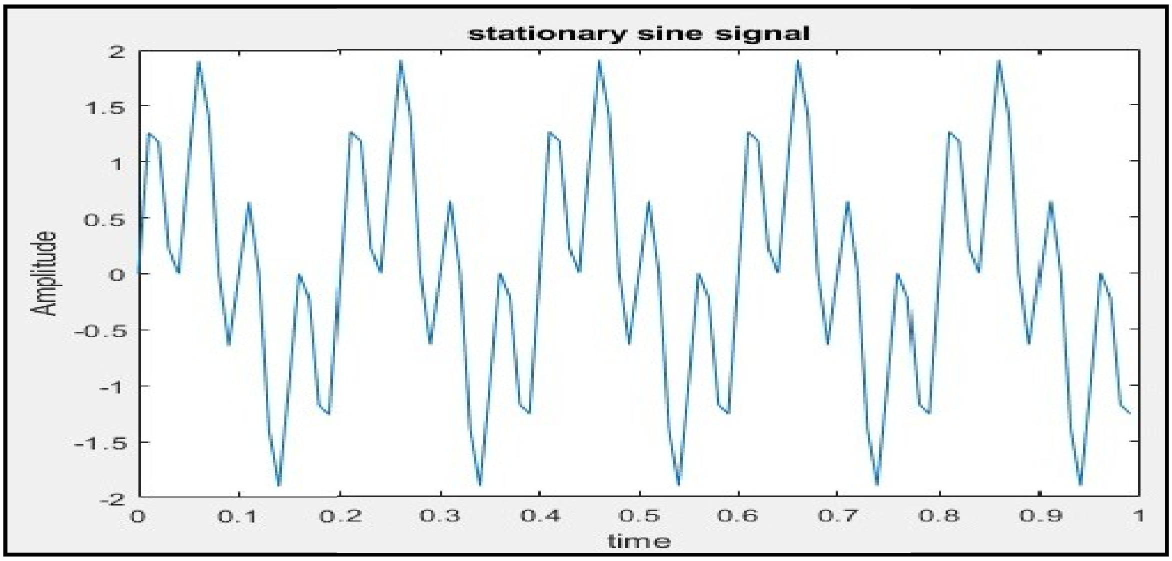

Figure 1.4: Sine signal of stationary frequency 5Hz and 20Hz

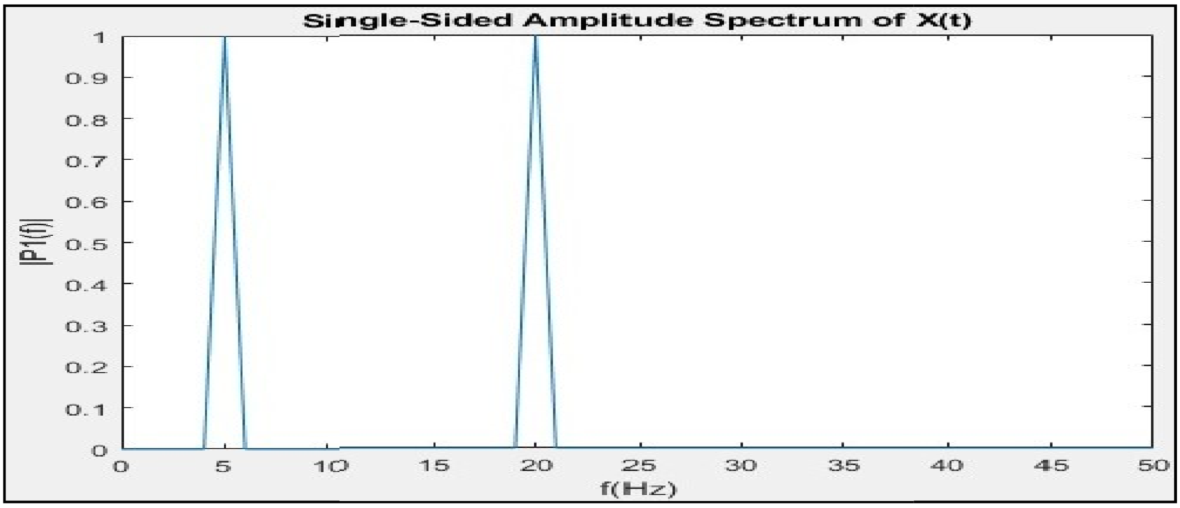

Figure 1.5: Spectrum of sine signal expressed in equation 1.15

Contrary to equation 1.15, the following signal is not stationary. Figure 1.6 shows a signal whose frequency content changes with time i.e. it is non-stationary signal. It consists of two different frequencies 5Hz and 20Hz at two different time interval.

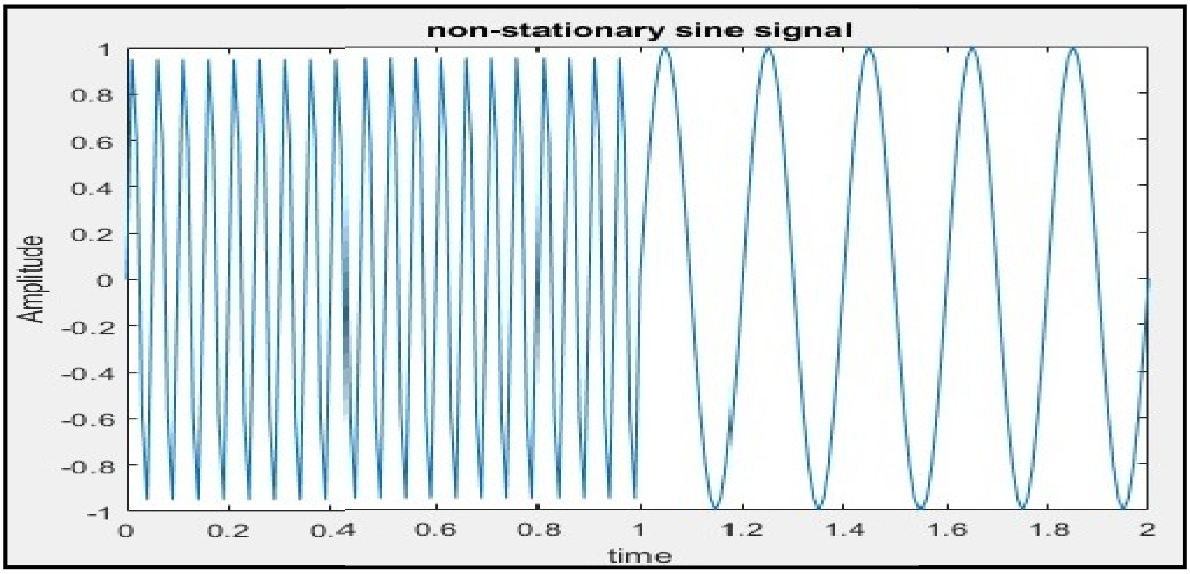

Figure 1.6: Sine signal of frequency 20Hz for first 1 sec and 5Hz for another 1sec

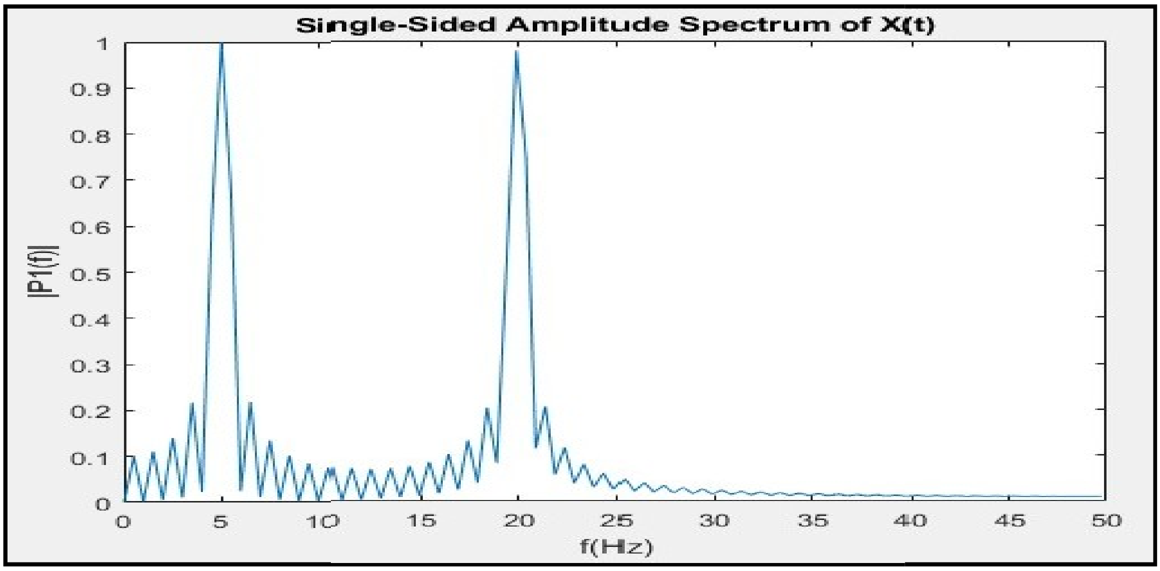

Figure 1.7: Spectrum of above non-stationary sine signal

In case of non-stationary signals also the Fourier transform shows two peaks in figure 1.7 except some ripples due to edge effect. But the difference is by just seeing spectrum content we cannot conclude at what times these frequency component exists. Thus, it can be said that F.T have very good frequency resolution but no time resolution as it does not store timing information of signal. In case non-stationary signals one needs to have timing and frequency information both or it can be said that one need to have time and frequency resolution both. This problem of non-stationary signals is solved by wavelet transform as it reserves information of time and frequency at variable resolution also called as multiresolution.

1.8 Motivation

The amount of data collected in a modern seismic recording/observatory exceed terabytes due to high resolution (24 bit) representation, dense disposition network of sensors and continuous monitoring requirement. Despite recent increase in mass storage capacity problems occurs not only in transmission but also for efficient processing and interpretation. Data compression is a prime solution to the above problems.

Here is a practical problem example:

As a test case scenario, if one has a requirement of 1 hour data file of 3 component seismometers (north-west, east-west, up-down) of 100 stations, then the volume of data would be 1.382GB. It will take at least 11.52 minutes through a dedicated VSAT/broadband with channel capacity 2Mbps, to download the data at user from data centre.

3component*100stations*60minutes*60seconds*40samples/sec*32bits/sample=1.382Gb

The communication channel present via VSAT is = 2Mbps

$$\frac{1382\text{Mb}}{2\text{Mb}} = 691 \text{ seconds} = 11.52 \text{ minutes}$$

Time taken to download the file is 11.52 minutes, which is long time to wait. But if we do compression by 85% then we can reduce this time substantially,

Size of file after 85% compression = 207Mb

$$\frac{207\text{Mb}}{2\text{Mbp}} = 103 \text{ seconds} = 1.7 \text{ minutes}$$

Hence after compression it takes only 1.7 minutes to transmit whole data. This point becomes more significant when we talk about reporting time of an event i.e. whenever a seismic event occurs depending on it's intensity it has to be reported. The reporting time could be reduces substantially.

Moreover, the earthquake early warning detection system is devised for notifying adjoining regions of a substantial earthquake while it is in progress. This is not the same as earthquake predictions which is currently not possible. P waves arrives first, and is detected by sensors. S wave arrives next followed by surface waves which carry most of the energy responsible for damage. For quick estimation of earthquakes/ Tsunamis for disaster management the transmission time of information from seismic station to data centre should be minimum as possible, this is possible with data compression technique.

Let us suppose a teleseismic event occurred and it took "t_1" time to reach nearest Seismic station. Now the data has to be transferred from station to data centre for analysis, let's say it took "t_2" time for this transmission. The analyst at data centre further analysis this event (crosschecking and drawing other conclusions) which takes further "t_3" time.

Total time= $t_1+t_2+t_3$

In practical scenario the typical value of t_1 is 10 minutes for teleseismic events. Mostly the events occur in remote mountainous region or inside the oceans. Considering the event in mid Arabian sea, it will be at distance 4000km (since it extends up to 10,000 km). The time taken by P-Waves to reach nearest Mumbai station will be:

$$T_p = \frac{4000 \text{ km}}{6 \text{ km/sec}} = 666.6 \text{ seconds} = 11.1 \text{ minutes}$$

t_3 is 3 minutes so values of t_2 comes out to be

$t_1 + t_2 + t_3 <= 15$ minutes

$11 + t_2 + 3 <= 15$

$t_2 <= 1$ minutes

In order to save bandwidth and to get fast access of data to meet reporting time and early warning criteria, the file size should be as small as possible which can be easily achieved by compression.

1.9 Compression Techniques

Data compression can be achieved by either predictive or transform coding.

1.9.1 Predictive coding

It involves a predictor circuit which predicts the next sample value. When the original sample arrives the difference of original and predicted value is stored as error. This difference is known as "error". If predictor circuit is well designed then this error value comes out to be a small number hence requires less no. of bits [28].

1.9.2 Transform coding

This technique in which the data is mapped to another domain which has the ability to compress data using a smaller number of coefficients and hence may be proven to be more efficient than predictive coding. Multiscale transformation is one of the most widely used transform coding techniques in case of image, audio, video and geophysical data (latest) compression.

In seismology, a seismogram (seismic data sensed by a seismometer) represents the superposition of seismic waves/signal (in case any) on the background noise. In general, seismic signals are non-inherent characteristics of seismic noise and seismic event they are separable. Moreover, the **coda** (only the seismic phases part of whole signal is called as coda of the event) is composed of many seismic phases distinguishable in time domain and in frequency domain. Owing to all these, seismic data compression is possible through transform coding compromising the quality of seismic event data. Multiscale/wavelet transform of seismic data segregate noise, event signal into different scale (frequency) and time packets. This is equally applicable to different seismic phases (P, S etc.). Moreover, seismic data requires high resolution representation whereas noise are of low resolution. And also, it (multiscale coding) provides a balance that the models become computationally feasible without losing much information. Taking the advantage of all these, modeling of seismic data in multiscale domain is possible to yield a very effective data compression algorithm. Hence, it may be proved/shown that seismic data wavelet transforms and a good coding scheme could result in an improved compression ratio as well as data quality than the existing Compression scheme being used. Due to different inherent characteristics of seismic noise and seismic events, they are separable and can be coded with different number of bits. Hence, seismic data compression is possible through transform coding leading to lossy representation without compromising the quality. The seismic data can be considered as a combination of three types of components:

1.9.3 Geophysical information :

The waves travel through the earth hence their statistical parameters can be used to get information of inner earth surface. For example S (secondary) waves are transverse in nature also called as shear waves. These waves cannot travel through liquid medium hence, if we did not get trace of S wave in seismograph that indicates there is some reservoir or void in between the focus and seismic recorder.

1.9.4 Uncorrelated and broadband noise :

There is always some background seismic noise present in a signal due to wind movement, vehicle movement, people walking etc. Different stations have different sources of noise. Pre study of the site is required to know the noise behaviour and eliminate it by setting proper threshold.

This can be written in a mathematical way as :

Seismic Data = Information + Noise

Hence, different strategies for compressing seismic data aim to eliminate noise.

1.10 Wavelet transform

Wavelet transforms are relatively recent developments that have fascinated the scientific engineering and mathematics community with their versatile applicability. Within geophysics, there have been already numerous applications of wavelet transform, such as in calculations of atmospheric turbulence, remotely sensed hydro meteorological events, data compression, noise reduction, feature extraction and identification of the location of ridge-parallel faulting etc[5]. The reason behind the versatility and attractiveness of wavelets for such diverse applications lies in their unique properties of time-frequency domain. The advantage of analysing a signal with wavelets as the analysing kernels is that it enables one to study features of the signal locally with a detail matched to their scale, i.e., broad features on a large scale and fine features on small scales. This property is especially useful for signals that are non-stationary, have short-lived transient components, have features at different scales, or have singularities. Therefore, wavelets are apt to do time-frequency analysis. Wavelets are seen as elementary building blocks in a decomposition or series expansion akin to the familiar Fourier series. Thus, are presentation of the processing wavelets is provided by an infinite series expansion of dilated (or contracted) and translated versions of a mother wavelet, each multiplied by an appropriate coefficient.

1.11 Multiresolution

Resolution is the ability to distinguish two closely placed quantities/values. Multiresolution implies that this distinguishable ability changes as per our requirement. For example if a signal have some high frequency and some low frequency portions then MRA is designed to give good time resolution and poor frequency resolution at high frequencies and good frequency resolution and poor time resolution at low frequencies. Hence it is possible to analyse a signal at different frequencies at different resolutions.

1.12 Scope of Work

This work includes detailed study of Seismology, seismic signals, stationary v_s non-statioanry signals and wavelet transform. The final objective is to get reduced file size for seismic applications.

The Scope of this project involves:

- Study of Seismic signals
- Selection of suitable Mother wavelet
- Selection of number of decomposition levels
- Selection of Threshold scheme

The Current work of data compression uses differential encoding and wavelet transform. But limitation lies in the fact that there is no rule for selecting Mother wavelet and number of decomposition levels. Different methods proposed by various authors for wavelet selection are discussed in this report but it has been felt that it is largely unexplored area. In this project it has been tried best to work out a method for wavelet and number of decomposition levels selection to achieve compression.

1.13 Organisation of Report

The report is organised as follows:

Chapter 2 -Literature Survey: This chapter introduces different popular compression techniques. It also provides background of wavelet transforms, onset time detection methods and the reasons why they are so popular among researchers.

Chapter 3 -Theoretical background: This chapter explains the detail theory of methods used along with their mathematical formulations and derivations.

Chapter4-Proposed method: It describes the proposed model with flowcharts and detailed explanation of each step.

Chapter 5 - Results: It discusses the results with graphs and tables and screenshots of output window for better understanding and visualisation.

Chapter 7 - Conclusion: This chapter concludes the thesis and discusses future scope.

2 Literature Survey

This chapter introduces techniques that have been used in previous research works for the purpose of data compression. It also discusses the reasons given by various researchers for selecting wavelet transform as the most optimal method for compression.

2.1 Lossy Vs Lossless Data compression schemes

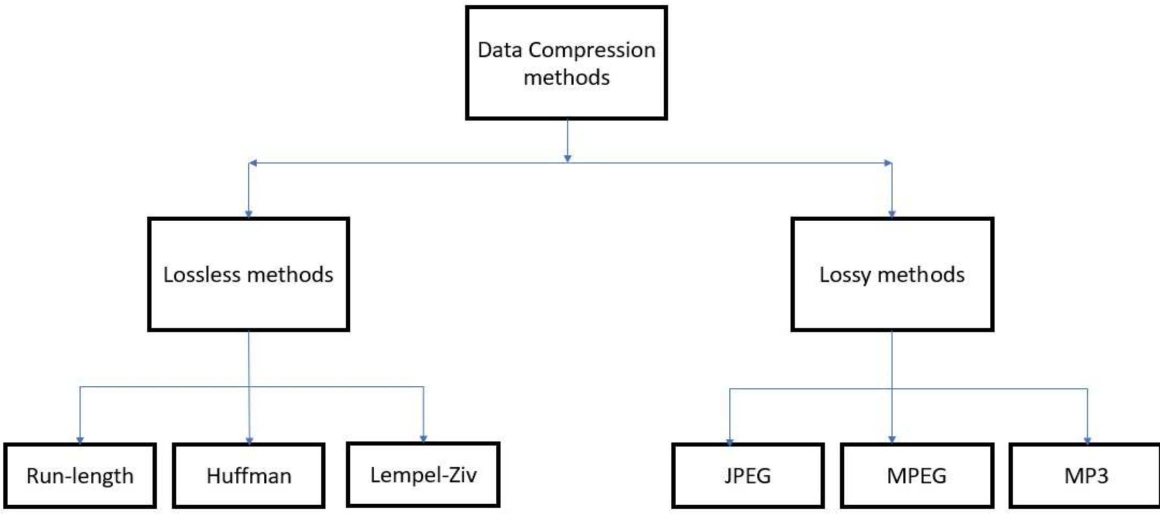

Figure 2.1: Classification of data Compression methods

In lossless data compression, the integrity of the data is preserved i.e. the original data and afterwards data are exactly same. Hence no part of the data is lost in the process. Redundant data is removed in compression and added during decompression. Lossless compression is normally used when we cannot afford to lose data. Thresholding scheme is lossy data compression technique.

2.2 Differential encoding for data compression

GCF (Guralp Compressed Format) is a data compression format mainly used in Guralp devices. It is based on differential encoding meaning difference of successive samples is stored not the actual sample values. Since difference comes out to be a small number so it requires less bits. Each GCF block consists of two parts: a header and a body. For data blocks, this body contains first differences of the sample values, and for status blocks, the body contains text status information. The header is 16 bytes long, split into four 4 byte fields which are:

System ID: A coded 5 or 6 character string representing the originating system Identifier. Coded as a base 36 number.

Stream ID: A coded 6 character string representing the originating stream, that is a unique label identifying device, component and sample rate. The first four characters are the serial number of the originating device, the 5th character the component, and the 6th the output tap or sample rate. Coded as a base 36 number.

Date Code: Coded representation of the date and time when the data in the block begins.

Data format: Contains sub-fields detailing the format of the contents. Each sub-field is 1 byte.

Sample Rate: It stores sample rate of data in the block. This should be constant for a stream. A status block is defined as having a sample rate zero.

The rest of the block contains the data fields:

The first field is the value of the first sample (not a difference). This is not included in the number of record values above. Often referred to as the FIC (Forward Integrating Constant). The difference samples then follow. Note: since the first sample value (FIC) is explicit, the first difference will be zero valued. Finally, as a check, the ending sample value is added. This should match the last decompressed sample value. It is not another sample. Often referred to as the RIC (Reverse Integrating Constant). This is a lossless method of compression. It can maximum reduce a file size by 50% not beyond. That's why researchers explored other methods of compression.

2.3 Wavelet Correlation theory

Purpose of wavelet transform is to de-correlate the data i.e. coefficients corresponding to signal are closer in values and coefficients corresponding to noise are wide spreaded. Wavelet transform provides well recognized advantage over other conventional transforms (discrete cosine and F.T) because of its multiresolution analysis with information both in time & frequency. According to Shannon theorem, a message of n bits (x_1, x_2,……x_n) can be compressed up to n bits, on average where H is called entropy, a quantity that determines the average number of bits needed to represent a dataset. The entropy is calculated as:

$$H = -\sum_{i=1}^{n} p^i * \log_2 p^i$$

Where p^i is the probability of symbol *i,* if all n probabilities are similar, the entropy is larger and entropy is lower when probabilities are at extremes. The concept of introducing entropy into data compression was put forward by *Carlus Fajardo et al.* [2]. Below is a histogram which is obtained from a seismic data set before and after a wavelet transform. It can be seen clearly that the wavelet transform distributes the coefficients such that the signal related coefficients are more correlated and hence forms narrower distribution. In figure 2.2 it's entropy is 7.4 bits/symbol and after wavelet transform it is 6.1 bits/symbol. This shows that wavelet transform have better compression ratio capability. Though the wavelet transform stage reduces the entropy of seismic signals but compression did not occur at this stage because the number of coefficients is same as number of original sample values. To get compression further quantization/ thresholding is required but the step of wavelet transform shows us a way to better compress a signal by removing off noisy part.

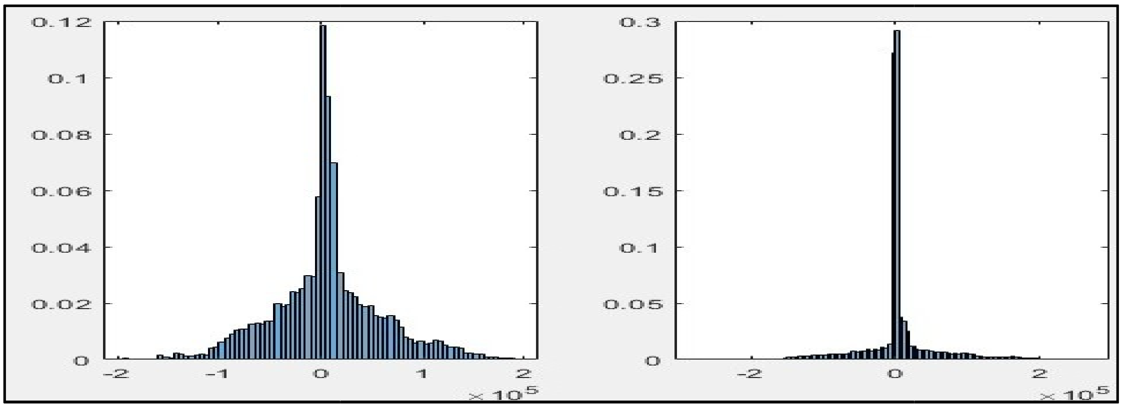

Figure 2.2: Histogram of Original signal and it's Wavelet coefficients

2.4 Selection of Suitable Mother Wavelet

A survey of literature reveals that choosing right wavelet is crucial for a successful wavelet transform application, but choosing the right wavelet for a specific application has been an open question. The main challenge in wavelet transform lies in selecting a mother wavelet, as different wavelets will produce different results hence, one has to find the degree of similarity between mother wavelet and signal. There are various criteria used by researchers to select a wavelet.

i. Degree of correlation between mother wavelet and signal: More a wavelet resembles the signal more easy it is for wavelet to characterize the signal. Normally the shape matching by visual inspection is applied to pick up the most proper mother wavelet. Otherwise cross correlation is calculated between wavelet and signal. The higher cross correlation value means higher similarity.

ii. Maximum energy to entropy ratio:

   In this the ratio of the maximum energy to Shannon entropy value was calculated for each wavelet and the one which give maximum value of this ratio is selected.

   $$R(s) = \frac{E_{energy}(s)}{E_{entropy}(s)}$$

   The high ratio means maximum energy will minimise entropy.

iii. In some applications the properties of wavelet of compact support and vanishing moment were used to select the most optimal mother wavelet.

A wavelet is M-1 times differentiable if it has M vanishing moments mathematically expressed as :

$$\int_{-\infty}^{\infty} t^q \varphi(t)dt \quad 0 \text{ for } q = 0,1,2\ldots M-1$$

When the wavelet $\varphi(t)$, has m vanishing moments then the polynomial of the form $\sum_{0<q<M-1} a_q t^q$can be represented exactly by linear combining scaling wavelet functions only, that is all mother wavelet coefficients will be zero in the representation of a polynomial of order less than or equal to M-1 which in turn tells that a wavelet having good number of vanishing moments is capable of representing a polynomial with less number of coefficients hence ensuring reduction in size of a compressed file.

iv. MSE (Mean Square Error) criteria: MSE of original and reconstructed signal were calculated and the wavelet having minimum MSE is selected as most suitable wavelet. This method is simple and fast.

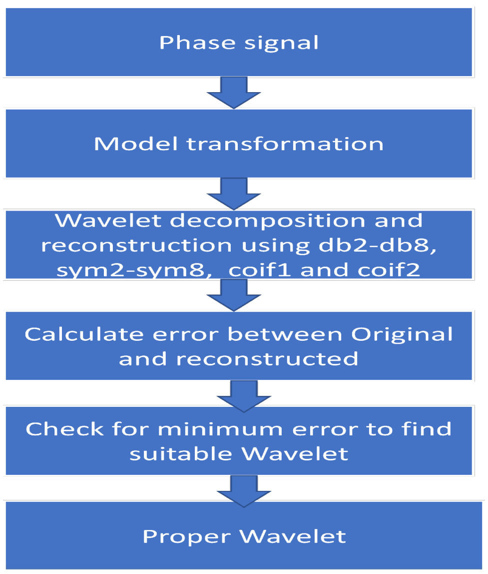

Figure 2.3: Flow chart of selecting suitable Wavelet using MSE criteria

PER (percentage Energy retained) criteria: Amount of energy of original signal that is retained in the reconstructed signal is called as percentage energy retained.

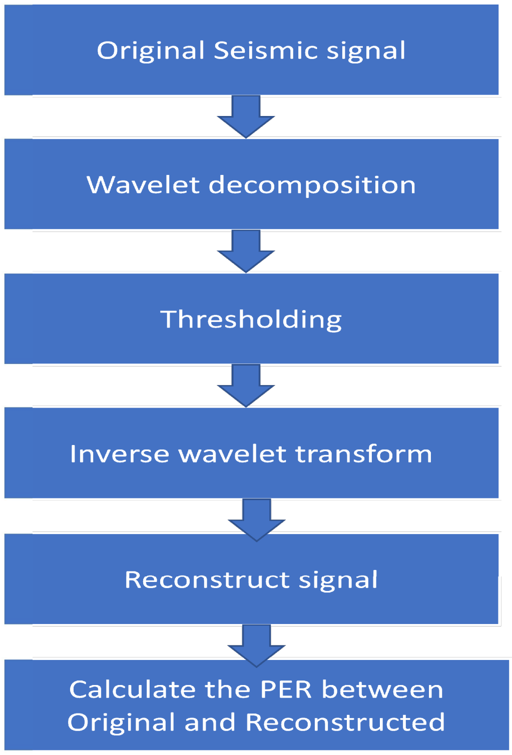

Figure 2.4: Flowchart of selecting suitable Wavelet using PER criteria

2.5 Selection of number of Decomposition levels

An input signal is passed through the filter bank which is an array of filters used to decompose the signal into different frequency bands. As number of decomposition levels increases, the computational cost also increases. Each decomposition level indicates a band of frequency. So, if the decomposition levels are increased, each band will be narrower which means one will have better frequency resolution. We can easily know what exactly the frequency range of a particular level if the sampling frequency (Fs) is known. As per Nyquist criteria the frequency of the signal is half of the sampling frequency.

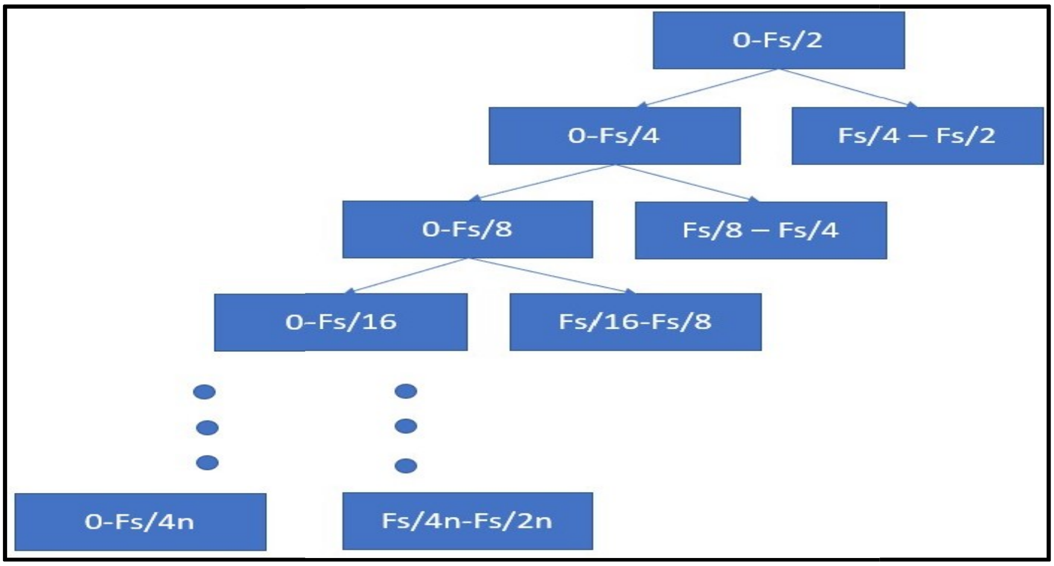

Figure 2.5: frequency band at every decomposition level

The maximum levels up to which a signal can be decomposed is given by following equation:

L_{max}= To_integer($\log_2[\frac{N}{N_f} - 1]$)

Where N= total number of samples of signal.

N_f= total number of filter coefficients.

The To_integer function takes ceil value of the number of levels.

In this study, the value N is 72000 samples as it is 1 hour file of 20 sps (samples per second) and N_f is 6 as wavelet selected is db3. So, maximum number of possible levels comes out to be

= To_integer([log2($\frac{72000}{6}$ – 1)])

= 14 levels.

2.6 Steps for data compression using wavelet transform

Most of the lossy data compression algorithms reported in the literature follows a common approach. The data compression algorithm presented by *Siffuzzaman et. al.*[1] consists of the following steps:

i. Wavelet transform
ii. Quantizing
iii. Encoding
iv. Decompression

2.6.1 Wavelet transform

The data is mathematically transformed to a new representation in which they are more easy to interpret and are better organized for purpose of data compression. The wavelet transform is best chosen transform method as it have various sub bands at different frequency ranges giving information precisely. The largest sub band consists of mostly high frequency data. Because of less correlation coefficient in these sub bands virtually all of the data represent noise of no geophysical significance. The sub bands at lower frequencies contain important information that is not contaminated by significant noise. But these sub bands are mostly at low frequencies consisting far fewer samples.

2.6.2 Quantizing:

This is the second step for leading to the useful data compression algorithm, consisting of quantizing the data samples of wavelets coefficients in such a way that important samples are preserved with high precision to keep their important content intact while unimportant samples i.e. noise ones of high frequency are quantized to zero or very small integers requiring few bits for their representation. The errors introduced in this step are irreversible and loss of information is inevitable and hence, quantization part is responsible for making the compression algorithm lossy.

2.6.3 Encoding

Finally, the data obtained from quantization step is subjected to further encoding using run-length to compare strings of identical values, and entropy encoding to represent frequently occurring values by fewer bits than less frequent values.

2.6.4 Decompression

The decompression steps requires all above steps one by one i.e. decoding the data, inverse wavelet transforms (quantization is irreversible process). The final goal of overall compression /decompression method is to recollect the data points as close as possible to original signal.

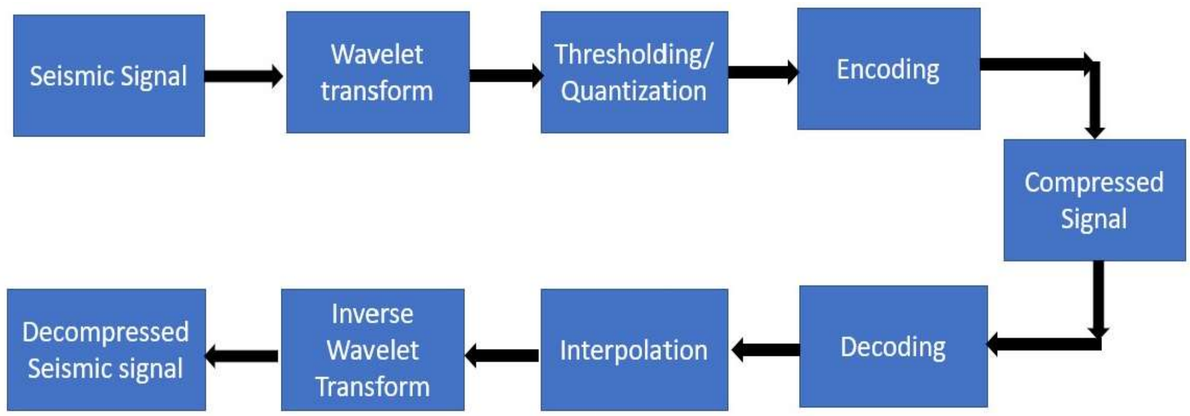

Figure 2.6: General Compression scheme

2.7 Thresholding functions

Thresholding is a common method to remove noise. Threshold is the estimated noise level. The values larger than threshold are regarded as signal, and the smaller ones are regarded as noises.

2.7.1 Hard threshold:

In this method the values which are smaller than threshold are made to zero and larger than threshold are kept as it is. Mathematical representation is :

$$x(t) = \begin{cases} x(t), & \text{if } |x| \geq T \\ 0\,, & \text{if } |x| < T \end{cases}$$

2.7.2 Soft threshold:

In this method the values smaller than threshold are made zero and larger than threshold are reduced in amplitude by the threshold value. It's mathematical representation is:

$$x(t) = \begin{cases} x - T, \text{if } x > T \\ x + T, \text{if } x < -T \\ 0, \quad \text{if } |x| < T \end{cases}$$

For cases where exact recovery of signal amplitude is not required for example image denoising the soft thresholding can be used as it ensures the regularity of signal but, the SNR of the signal is better in hard thresholding since actual magnitude of coefficients is retained.

Here is a graphical presentation of hard and soft thresholding:

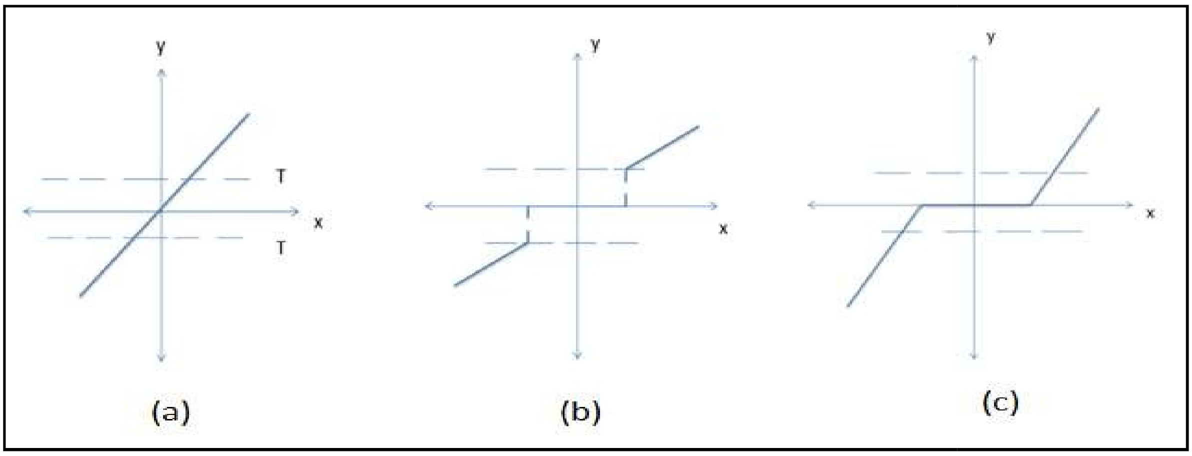

Figure 2.7: Thresholding function (a)Original Signal (b)Hard threshold (c)Soft threshold

The risk of introducing errors after thresholding is closely related to the thresholding function. The proper choice of threshold T is important to minimize the risk of estimation.

Donoho and Johnstone (1994)[24] proposed a universal threshold T .They proved that the risk of thresholding, no matter hard or soft, is small enough to satisfy the requirements of most applications. A approach to relate threshold to the dispersion of coefficients of DWT vector. One such threshold, derived under the assumption that the noise is white with variance. The mathematical formulation is given as :

$T=\sigma_i\sqrt{2\log_2 N_i}$

σ_i = Standard deviation of detail coefficients at each level

N_i = Number of detail coefficients at each level

2.8 Time domain technique for onset detection

In 1965, *Vanderkulketal*[25] used the ratio of short term average to long term average (STA/LTA) algorithm on absolute values of data offering significant computational savings. The algorithm based on STA/ LTA principal is found to be more suitable for earthquakes and strong motion detection. It depends on the amplitude variations of seismic signals rather than the other parameters like signal polarization and frequency. For a uniform amplitude signal, the moving average is constant. But if input signal changes rapidly, the average will also change accordingly. This concept has been used for seismic signal detection. If the current average is greater than the previous average value, the incremental change can be assessed. For seismic signals, the current average over a short period and previous average over a long period will be different whenever an event occurs. To use an event detections algorithm based on this technique, three parameters will be required:

i. Short term average (STA)
ii. Long term average (LTA)
iii. Threshold value (α)

For true event the following condition must be satisfied.

$$\text{Ratio} = \frac{\text{STA}}{\text{LTA}} > \alpha$$

The short period average represents the average of the shortest period over which an event of interest should occur. The long period average represents the average of the longest period to assess the background noise behaviour α is the threshold value based on characteristics of the seismic station area. Threshold value is compared with the ratio of the short period average to the long period average that will cause event discrimination. Selection of α is based on location and preference for detecting local, micro and distant earthquakes as per parameter selection. The averaging process is simple to compute and takes less time, making it suitable for on-line seismic applications. In 1978, *Allen* used envelop of data which is the sum of squares of the data and the weighted square of the first derivative with STA/LTA algorithm. This envelope includes components of both the unfiltered and high-pass filtered data. The processed data stream is then subjected to a set of logical and mathematical tests for phase identification and timing. This scheme works well for short period data (frequencies > 1Hz) but incorporated high pass filter making it inappropriate for teleseismic and volcanic event detection. The amplitude threshold trigger simply searches for any amplitude value exceeding a preset threshold. Recording starts whenever the threshold is reached. The algorithm is normally used in strong motion seismic instrument, where high sensitivity is not an issue, and where consequently cultural and natural seismic noises are not critical. With improvement in this domain, in 1984, *Houlistonet et. al.*[26] described STA/ LTA algorithm for a multi-channel seismic network system. Limitation with STA/ LTA was that it does not function well with sites with high, irregular man-made seismic noise.

2.9 Statistical Methods

In 1992 *Sleeman and Eck*[27] introduced an autoregressive techniques which are based on the assumption that the seismogram can be divided into locally stationary segments as an autoregressive process (AR) i.e. the intervals before and after onset are two different stationary processes. On the basis of this assumption, an autoregressive Akaike Information criteria (AR-AIC) method has been used to detect P or S phases. For the AR-AIC picker, the order of the AR process must be specified by trial and error and the AR coefficients have to be calculated for both intervals. In contrast to the AR-AIC picker, *Maeda* [11] (1985) suggested a different AIC picker, which can be calculated directly from the records without fitting them with the AR processes. However, when the signal to noise ratio (SNR) is low and the arrival is not evident, the AIC picker does not perform well. Further, for the AIC picker to identify the proper onset, a limited time window of the data must be chosen.

To summarize, all work carried out in the field of Seismology includes predictive or transform encoding for data compression. Predictive encoding though lossless can provide maximum 4:1 compression only, whereas transform coding though lossy can provide very high compression ratio up to 100:1. The few unanswered questions left in transform encoding are regarding selection of wavelet, selection of number of levels, selection of thresholding or quantizing scheme. Different methods proposed by different scholars were studied and discussed here extensively.

3 Theory

This section explains the theoretical background that is required for understanding each and every step in data compression. It consists of diagrams, mathematical formulations, derivations of onset triggering algorithms and transform method.

3.1 Discrete wavelet transform

3.1.1 Wavelet Definition

The term wavelet means a small wave. The smallness refers to the condition that this (window) function is of finite length (compactly supported). A wavelet is a wave-like oscillation with an amplitude that begins at zero, increases, and then decreases back to zero. It can typically be visualized as a "brief oscillation" like one recorded by a seismograph or heart monitor.

3.1.2 Importance of wavelet transform

The key point of wavelet transform is the ability to extract the transient components from the sampled waveforms and then process these to detect quickly the occurrence of a transient. We know, Fourier transform (FT) gives the frequency information of the signal, which means that it tells us how much of each frequency exists in the signal, but it does not tell us when in time these frequency components exist. FT is good for analyzing stationary signals (having fixed frequency components throughout). FT gives only spectral content of the signal with no timing information. Therefore, FT is suitable technique for stationary signal. More commonly we find non-stationary signals in nature and sometimes a particular spectral component occurring at any instant can be of particular interest. In such case wavelet transform is capable of providing the time and frequency information simultaneously, hence giving a time-frequency representation of the signal. We pass the time-domain signal from an array of high pass and low pass filters, which filters out either low frequency or high frequency portions of the signal. This procedure is repeated and every time the signal gets segregated into sub bands. As explained by *Robipolikar*[18], suppose we have a signal for demonstration having frequencies up to 2000 Hz. In the first level we split up the signal into two parts by passing the signal from a high pass and a low pass filter (filters should satisfy some certain conditions, so-called admissibility condition) which results in two different versions of the same signal: portion of the signal corresponding to 0-1000 Hz (low pass portion), and 1000-2000 Hz (high pass portion). Then, we take either portion (usually low pass portion) or both, and do the same thing again. This operation is called decomposition as shown:

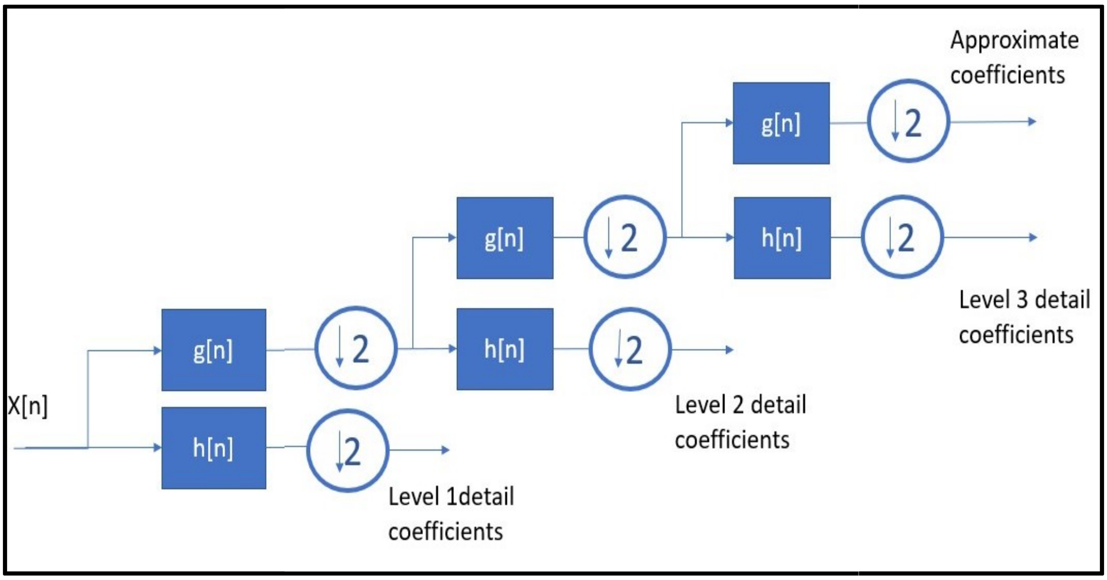

Figure 3.1: Filter bank decomposition structure of Discrete Wavelet transform

The output at each stage is convolution of input signal and filter after down sampling it by 2. The mathematical equations are:

$$y_{high}[k]=\sum_{n=-\infty}^{\infty} x[n].\, g[2k-n] \tag{3.1}$$

$$y_{low}[k]=\sum_{n=-\infty}^{\infty} x[n].\, h[2k-n] \tag{3.2}$$

Where g[n]= low pass filter

h[n]=high pass filter

x[n]=input signal

y[n]=Output Signal

Then we have a bunch of signals, which actually represent the same signal, but all corresponding to different frequency bands. The frequencies that are most prominent in the original signal will appear as high amplitudes in that region of the DWT signal that includes those particular frequencies. The difference of this transform from the Fourier transform is that the time localization of these frequencies will not be lost. However, the time localization will have a resolution that depends on which level they appear. If the main information of the signal lies in the high frequencies, as happens most often, the time localization of these frequencies will be more accurate, since they are characterized by more number of samples. If the main information lies only at very low frequencies, the time localization will not be very precise, since few samples are used to express signal at these frequencies. This procedure in effect offers a good time resolution at high frequencies.

This decomposition halves the time resolution since only half the number of samples now characterizes the entire signal. However, after this operation, the frequency band of the signal now spans only half the previous frequency band, effectively reducing the uncertainty in the frequency by half. The above procedure, which is also known as the sub band coding, decomposition. At every level, the filtering and sub samples (and hence half the time resolution) and half the frequency band spanned (and hence double the frequency resolution).

3.2 Importance of discrete wavelet transform for reduction

For a given image, the DWT coefficients are computed of say each row, and discard all values in the DWT that are less then certain threshold. We then save only those DWT coefficients that are above the threshold for each row, and when row with as many zeros as the number of discarded coefficients, are used in the inverse DWT to reconstruct each row of the original image. We can also analyze the image at different frequency bands, and reconstruct the original image by using only the coefficients that are of a particular band. In the figure 3.2, a 2 sec sine signal of sampling rate 100 sps of frequency 5 hz is generated. The figure 3.3 shows the plot of DWT coefficients, in different colours corresponding to a particular frequency band, decomposed up to 6 levels. Each detail coefficients are shown in different colour in figure (3.3). Due to multiresolution property the number of coefficients at each level are different. For first level after filtering and down sampling, 100 samples are generated. For second level after filtering and down sampling 50 samples are generated. Total number of DWT coefficients will sum up to 200 samples (same as original number of samples). It should be noted that only the first 60 samples, which correspond to lower frequencies of the analysis, carry relevant information. Therefore, all but the first 60 samples can be discarded without any loss of information. This is how DWT provides a very effective data reduction scheme.

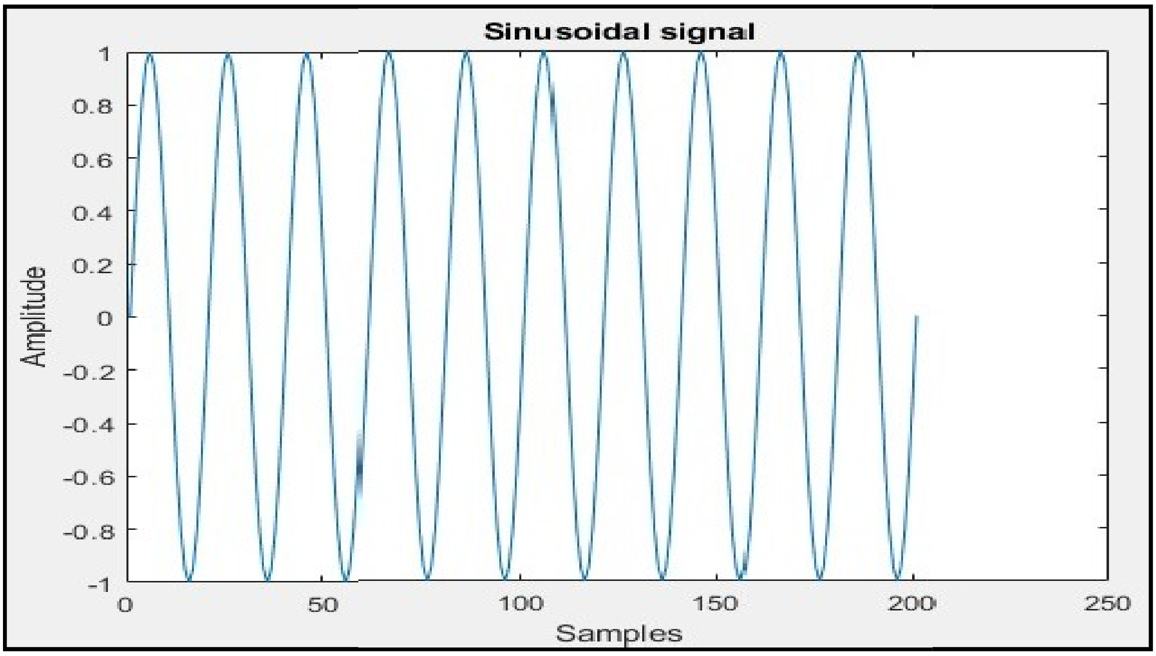

Figure 3.2: Sine signal of 5Hz

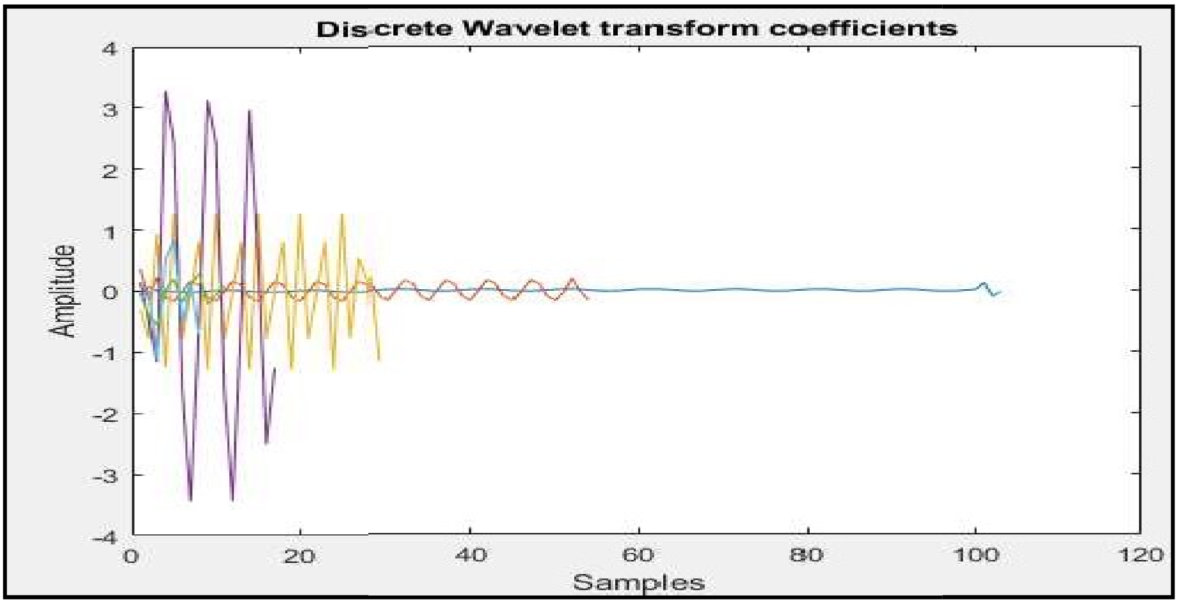

Figure 3.3: Plot of DWT coefficients

A seismic network operating continuously at high sampling frequency produces an enormous amount of data, which is often difficult to store (and analyze) locally or even at the recording centre of a network. Processing software a trigger algorithm serves for the detection of typical seismic signals (earthquakes, underground nuclear explosion signals, etc.) in the constantly present seismic noise signal. Once an assumed seismic event is detected, recording and storing of all incoming signals starts. It stops after trigger algorithm 'declares' the end of the seismic signal.

3.3 Stationary Wavelet Transform

The basic idea of the stationary wavelet transform is to 'fill in the gaps' caused by the decimation step in the standard wavelet transform. This leads to an over-determined, or redundant, representation of the original data which helps in understanding behaviour of the signal. The SWT algorithm is very simple and is close to the DWT one.

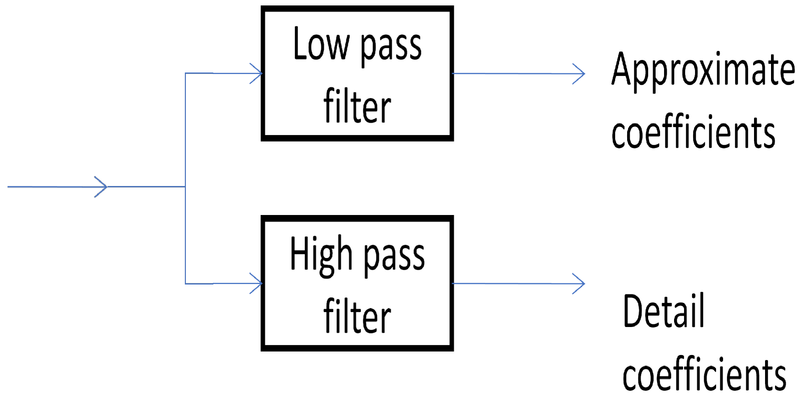

Figure 3.4: Iterative structure of filter bank

More precisely, for level 1, all coefficients of SWT for a given signal can be obtained by convolving the signal with the appropriate filters as in the DWT but without downsampling. Then the approximation and detail coefficients at level 1 are both of size N, which is the signal length. This can be visualized in the following figure (3.5).

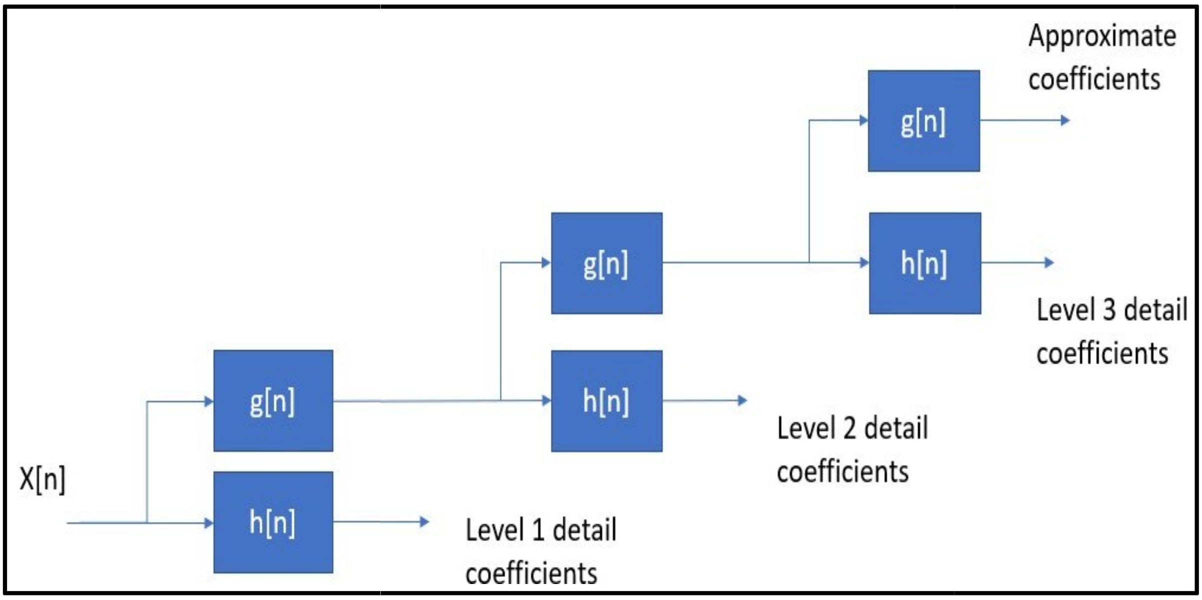

Figure 3.5: Filter bank for sub band decomposition

The output equation at each level is written as :

$$y_{low}[k]=\sum_{n=-\infty}^{\infty} x[n].\, g[k-n] \qquad (3.3)$$

$$y_{high}[k]=\sum_{n=-\infty}^{\infty} x[n].\, h[k-n] \qquad (3.4)$$

Where g[n]=low pass filter

h[n]=high pass filter

x[n]=input signal

y[n]=output signal

3.4 Onset Trigger Algorithms

The simplest trigger algorithm is the amplitude threshold trigger. It simply detects an event and start recording whenever it's amplitude exceeds a pre-decided threshold. This algorithm is rarely used in weak motion seismology. It is mostly used in strong motion seismology or in strong motion seismic instruments, that is in systems where high sensitivity is mostly not an issue, and where consequently man-made and natural seismic noise amplitudes are much smaller than the signals which are supposed to trigger the instrument. Today, For weak motion seismology the 'short-time-average through long-time-average trigger' (STA/LTA) is the most broadly used algorithm. It continuously calculates the average values of the absolute amplitude of a seismic signal in two consecutive sliding time windows. The short time average window (STA) is sensitive to seismic events while the long time average window (LTA) is captures the temporal amplitude of seismic noise at the site. When the ratio of STA/ LTA exceeds a pre-set value, an event is said to be 'declared' and data starts being recorded in a file. Successful capturing of seismic events depends on proper settings of the trigger parameters.

3.4.1 Steps of STA/LTA trigger algorithm are as follows:

First, the absolute amplitude of each data sample of an incoming signal is calculated. Next, the average of absolute amplitudes in both windows is calculated as shown in equations below (3.5) and (3.6). In a further step, a ratio of both values — STA/ LTA is calculated. This ratio is continuously compared to a user selected threshold value, STA/ LTA trigger threshold level. If the ratio exceeds this threshold, a channel trigger is declared. Figure 3.6 depicts graphically the calculation of STA and LTA values for a signal. "i" is the point where window ends and $i-l_1$ is the point of STA window starting (blue colour line) and $i-l_2$ is the point of LTA window starting. l_1 is the length of STA window and l_2 is the length of LTA window. Ratio of the value obtained from

STA/ LTA division is compared with the threshold value and if this ratio exceeds threshold then event is said to be triggered else not.

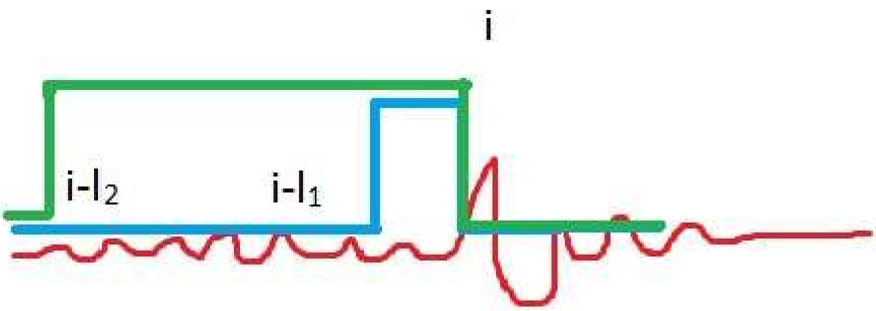

Figure 3.6: Definition of STA/ LTA ratio at test point i of an example signal

Let 'x' be the time domain signal then:

$$STA = \left(\sum_{j=i}^{i-l1} |x_j|\right) / l_1 \quad (3.5)$$

$$LTA = \left(\sum_{j=i}^{i-l2} |x_j|\right) / l_2 \quad (3.6)$$

$$Ratio = \frac{STA}{LTA} \quad (3.7)$$

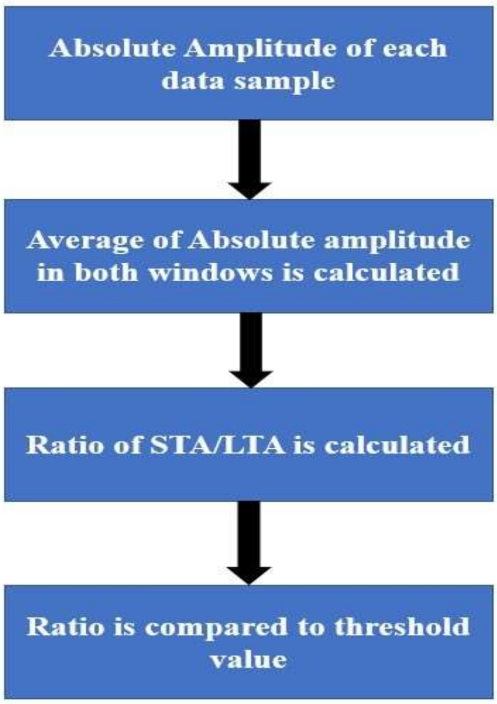

Figure 3.7: Flow chart of STA/LTA algorithm

A channel trigger does not necessarily mean that a data logger or a network actually starts recording of the seismic signals. All seismic networks and most seismic recorders have a 'trigger voting' mechanism built in that defines how many and which channels have to be in a triggered state before the instrument or the network actually starts to record data.

To simplify the explanation, we shall observe only one signal channel. We will assume that a channel trigger is equivalent to a network or a recorder trigger. After the seismic signal gradually terminates, the channel de-triggers. This happens when the current STA/ LTA ratio falls below another user-selected parameter - STA/ LTA detrigger threshold level. Obviously, the STA/ LTA de-trigger threshold level should be lower (or rarely equal) than the STA/ LTA trigger threshold level.

To set the basic STA/LTA trigger algorithm parameters one has to select the following:

•STA window duration

•LTA window duration

•STA/ LTA trigger threshold level

•STA/ LTA detrigger threshold level

3.5 Selection of Short time average (STA) window

Short-time average window measures the 'instant' value of a seismic signal or its envelope. To some extent the STA functions as a signal filter. The shorter the duration selected, the higher the trigger's sensitivity to short lasting local earthquakes compared to long lasting, low frequency distant earthquakes. For teleseismic signals as they travel longer distance, so most of the high frequency content are attenuated and low frequency are left. The longer the STA duration selected, the less sensitive it is for short local earthquakes. Therefore, by changing the STA duration one can, to some extent, prioritize capturing of distant or local events.

The STA duration is also important with respect to false triggers. By decreasing the time duration of the STA window, triggering gets more sensitive to spike type man-made seismic noise and vice versa. Although such noise is usually because of instrumental nature, it can also be seismic. At the sites highly polluted with spike type noise, one will be frequently forced to make the STA duration significantly longer than these spikes, if false triggers are too numerous. Unfortunately, this will also decrease the sensitivity of the recording to very local events of short duration. Figure 3.8 explains the effect of STA duration on local events and spike type noise. On graph a signal with an instrumental spike on the left and with a short local earthquake on the right side is shown. It is shown that STA, LTA, STA/ LTA ratio, and trigger active states. The STA/ LTA trigger threshold was set to 10 and de-trigger threshold to 2. One can see that when using a relatively long STA of 3sec, the earthquake did trigger the system, but only barely.

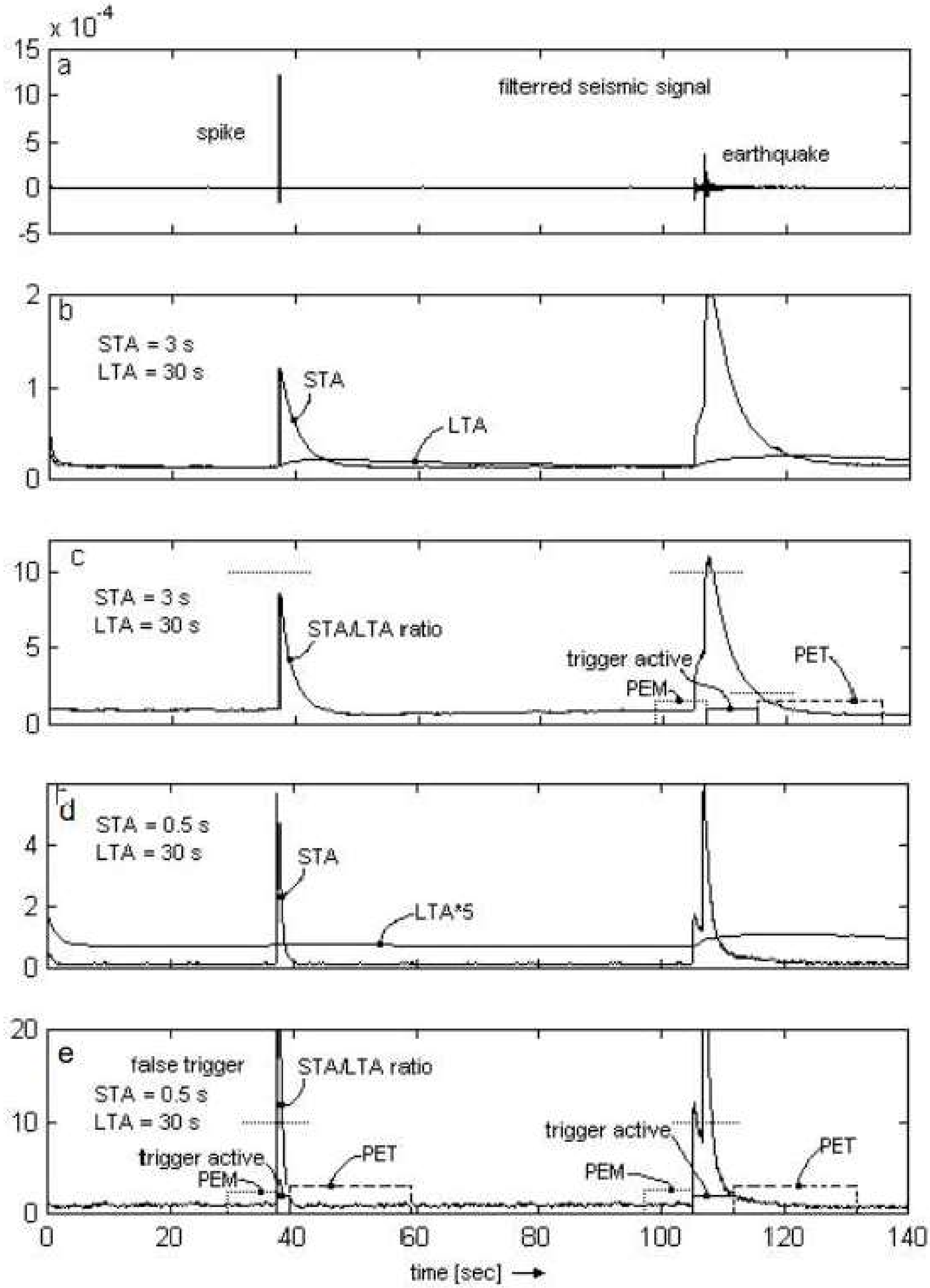

Figure3.8: STA window duration effect of triggering event

However, a much bigger amplitude (but shorter) instrumental spike did not trigger it. The STA/LTA ratio did not exceed the STA/LTA threshold and there was no falsely triggered record due to the spike. The lower two graphs show the same variables but for a shorter STA of 0.5 sec. The spike clearly triggered the system and caused a false record. Of course, the earthquake triggered the system as well.

For regional events, a typical value of STA duration is between 1 and 2 sec. For local earthquakes shorter values around 0.5 to 0.3 s are commonly used in practice.

3.6 Selection of Long time average (LTA) window

The LTA window measures average amplitude of seismic noise. It should last longer than a few 'periods' of typically irregular seismic noise fluctuations. By changing the LTA window duration, one can make the recording more or less sensitive to regional events. These events typically have the low amplitude emergent waves as the first onset. The shorter LTA duration allows the LTA value more or less to adjust to the slowly increasing amplitude of emergent seismic waves. A shorter LTA duration is needed to exclude emergent regional events from triggering. A short LTA will successfully accommodate recorder sensitivity to gradual changes of 'continuous' man-made seismic noise. Such 'transition' of man-made seismic noise from low to high is typical for night-to-day transition of human activity in urban areas. Sometimes, using a short LTA can mitigate false triggers due to traffic. Examples of such cases could be a single heavy vehicle approaching and passing close to the seismic station on a local road, or trains on a nearby railway. Thus the STA/LTA ratio remains low in spite of increasing STA (nominator and denominator of the ratio increase). This effectively diminishes trigger sensitivity to such events. In the opposite case, using a long LTA window duration, trigger sensitivity to the emergent earthquakes is increased because the LTA value is not so rapidly influenced by the emergent seismic signal, allowing surface waves to trigger the recording.

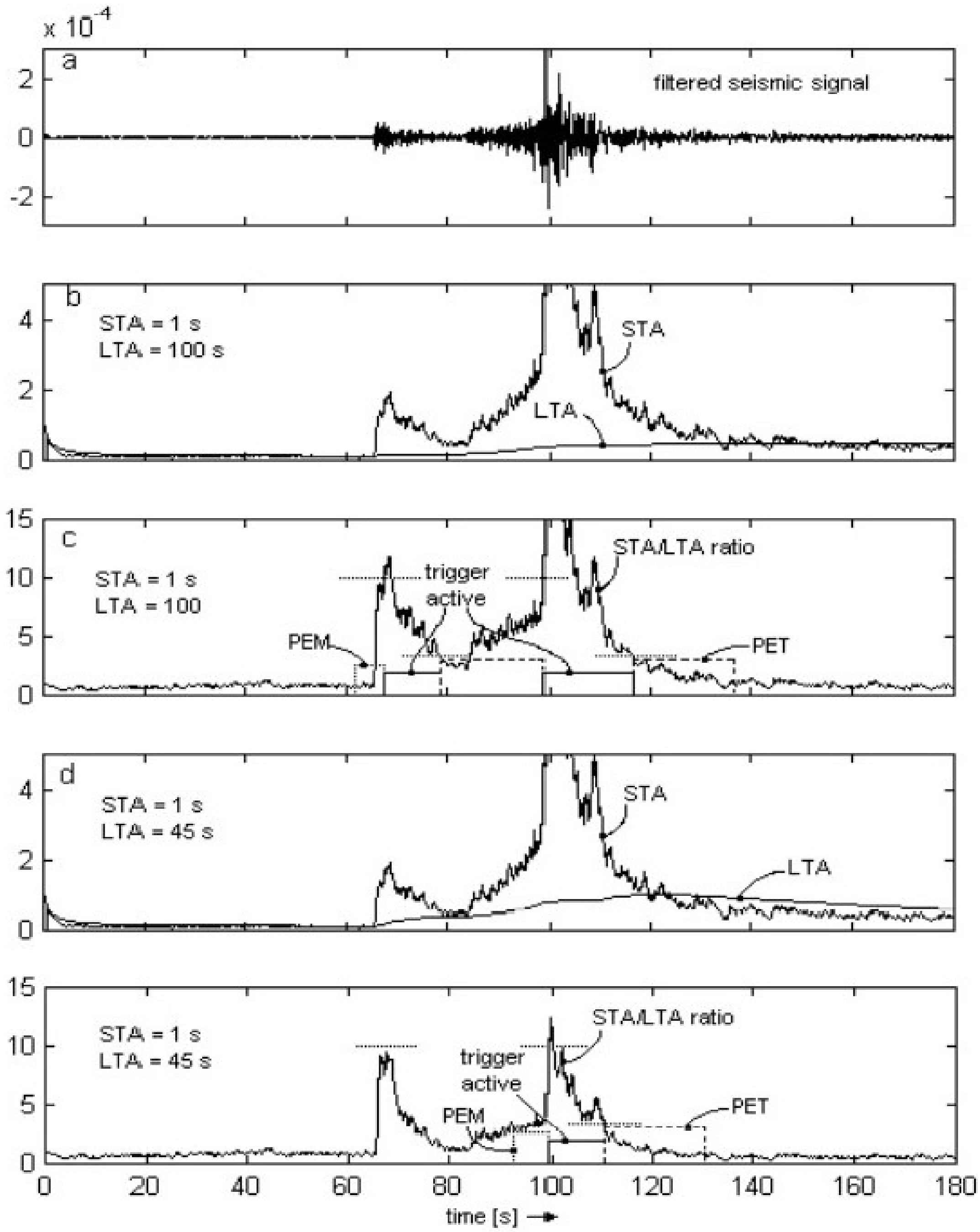

Figure 3.9: LTA window duration effect on event triggering

Figure 3.9 represents such a case. It shows a typical event with significantly bigger later phase waves than P waves. Graphs b) and c) show trigger parameters for a long LTA of 100 s. P wave packet as well as S wave packet trigger the recorder. Graphs d) and e) show the same situation but for a shorter LTA duration of 45 sec. One can see that the P waves did not trigger at all, while the S waves barely gets triggered. The STA/LTA ratio hardly exceeds the STA/LTA trigger threshold. As the result, the recorded data file is too much short. P waves and information about seismic noise before them are missing in this record. A slightly smaller event would not trigger at all. The LTA duration of 60 seconds is a common initial value. A longer LTA can be used for distant regional events with long S, P times.

3.7 Selection of STA/LTA trigger threshold level

The STA/LTA trigger threshold level to the greatest extent determines which events will be recorded and which will not. With high threshold value there will be very less occurrence of false trigger; however actual earthquake may also be passed over . The lower the STA/ LTA trigger threshold level is selected, the more sensitive the seismic station will be and the more events will be recorded. However, more frequent false triggers also will occupy data memory and burden the analyst. An optimal STA/ LTA trigger threshold level depends on seismic noise conditions of the site and on one's tolerance to falsely triggered records. Not only the amplitude but also the type of seismic noise influence the setting of the optimal STA/ LTA trigger threshold level. A statistically stationary seismic noise (with less irregular fluctuations) allows a lower STA/ LTA trigger threshold level, completely irregular behaviour of seismic noise demands higher values.

An initial setting for the STA/ LTA trigger threshold level of 4 is common for an average quiet seismic site. Much lower values can be used only at the very best station sites with no man-made seismic noise. Higher values about 8 and above are required at less favourable sites with significant man-made seismic noise. In strong-motion applications, higher values are more common due to the usually noisier seismic environment and generally smaller interest in weak events.

3.8 Akaike Information Criteria

Akaike information criteria (AIC) is one of the most widely used onset time determining algorithms. Modeling the signal as an autoregressive (AR) process is another approach for onset time determination. It is based on the so-called Akaike Information Criterion (AIC) picker. In this case, the intervals of the signal before and after the onset time are assumed to be two different stationary datasets. For a fixed order AR process, the point, at which the AIC is minimized, determines the separation of the two time series.

AIC which was derived by [29] is defined by the following equation:

$$AIC = -2 * \ln(L) + 2k \quad (3.1)$$

Where k is the number of parameters in statistical model of the signal, L is the maximized value of the likelihood function for the estimated model. Generally a model, with minimum AIC value is thought to be the most suitable one. The technique assumes that we have a time series $X_n(X_1, X_2, \ldots X_n)$ which includes the onset of an acoustic signal and a first estimate of the onset time. The intervals before and after the onset time are assumed to be two stationary time series.

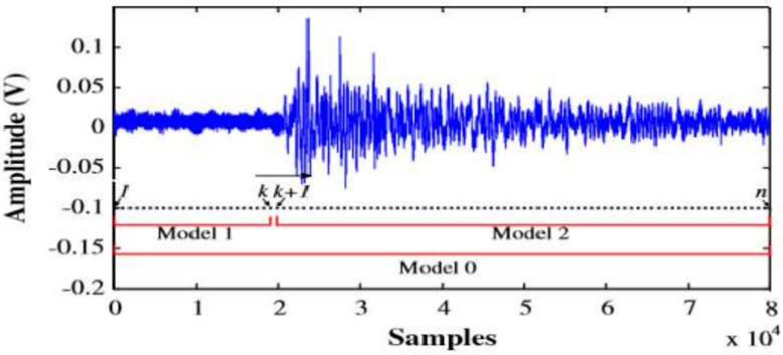

Figure 3.10: An illustration to showing signal division and MODEL assigning

As illustrated the section is divided into two parts, MODEL1 (fitted to first k points) and MODEL 2 (fitted to $(k+1)_{th}$ to nth point) with the maximum likelihood estimation. In each interval i=1,2 the one preceding and the one including phase onset, we model a window in which we fit the data to an AR model of order M with coefficients a^i_m where m=1,2,3….M

$$x_t = \sum_{m=1}^{M} a^i_m x_{t-m} + e^i_t \quad (3.2)$$

With t=1…..k for interval 1 and t=k+1……n for interval 2. The model window divides the time series into a deterministic and non- deterministic part. The non- deterministic part i.e. noise is assumed to be Gaussian with mean=0 and variance = σ_i^2 and uncorrelated with deterministic part of time series. The maximum likelihood function (MLE) is used to extract the non-deterministic part in interval [1,k] and [k+1,n] using Equation (2) where k is the division point. As we assume the non-deterministic part to be Gaussian, we can express the approximate likelihood function L for the two non-deterministic time series in the interval [1,k] and [k+1,n] as :

$$L(x;k,M,\Theta_i)=\prod_{i=1}^{2}\left(\frac{1}{\sigma_i^2 2\pi}\right)^{\frac{n_i}{2}} * \exp\left(-\frac{1}{2\sigma_i^2}\sum_{j=p_i}^{n_i}\left(x_j - \sum_{m=1}^{M} a_m^i x_{j-m}\right)\right) \tag{3.3}$$

here $\Theta_i=\Theta(a_1^i\ldots\ldots a^i, \sigma_i^2)$ represents the model parameters and p1=1, p2=k+1, n1=k, n2=n-k.

Taking the logarithm of Equation (3) and searching for maximum likelihood estimation of the model parameters we get:

$$\frac{\delta \ln L(x\,;\,k,M,\Theta_i)}{\delta\Theta_i} = 0 \tag{3.4}$$

Which has the solution:

$$\sigma_{i,max}^2 = \frac{1}{2}\sum_{j=p_i}^{n_i}\left(x_j - \sum_{m=1}^{M} a_m^i x_{j-m}\right)^2 \tag{3.5}$$

The calculation of Autoregressive coefficients depends on statistical parameters of the signal.

The maximum of logarithmic likelihood function for the two models as function of k becomes:

$$\log L(x;k,M,\Theta_1,\Theta_2)=-\frac{n1}{2}\log(\sigma^2_{1,max}) - \frac{n2}{2}\log(\sigma^2_{2,max}) - \frac{n}{2}(1 + 2\pi) \tag{3.6}$$

$$\log L(x;k,M,\Theta_1,\Theta_2) = \frac{k}{2}\log(\sigma^2_{1,max}) - \frac{(n-k)}{2}\log(\sigma^2_{2,max}) + C \tag{3.7}$$

Where C is a constant and thus it's influence can be neglected when n is large enough. The AIC of two signal comes out to be:

$$AIC(k) = k*\log(\sigma^2_{1,max})+(n-k)*\log(\sigma^2_{2,max})+2C \quad (3.8)$$

Point 'k' where the joint likelihood Eq. (7) is maximized, or AIC(k) in Eq. (8) is minimized, determines the optimal separation of the two stationary time series. AIC picker gives high quality if the AIC is applied only to a pre-selected window of the time series containing the onset-time so a proper time window should be chosen. Only issue is the calculation of autoregressive coefficients which becomes a tedious task So, Maeda [11] suggested a different AIC picker, which can be calculated directly from the records without fitting them with the AR processes. It is given as:

$$AIC(k) = k*\log(variance(x[1,k])) + (N-k-1)*\log(variance(x[k+1,N])) \quad (3.9)$$

Where k ranges through the entire seismogram samples.

However, the limitation of AIC method lies in the quality of the signal i.e. when the signal to noise ratio (SNR) is low and the arrival is not evident, the AIC picker does not perform well. Further, for the AIC picker to identify the proper onset, a limited time window of the data must be chosen.

4 Proposed Method

This section explains the proposed model for the compression and decompression algorithm using flowchart, diagrams and definitions. It further elaborates the importance of wavelet transform and challenges faced in it's implementation.

4.1 Flow Chart of Compression Scheme

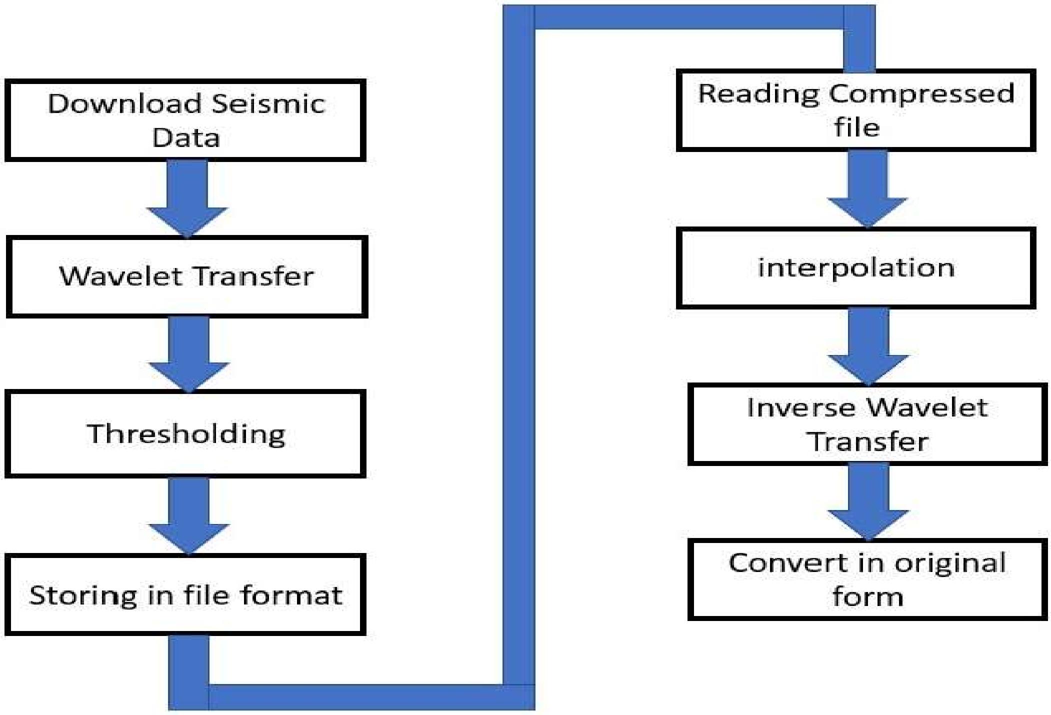

Figure 4.1: Flow chart of Compression scheme followed in this project

4.1.1 Download Seismic data:

Seismic data was downloaded in SAC (Seismic Analysis Code) format from IRIS (Incorporated research Institute of seismology) website. The downloaded files were of 65 minutes duration consisting of P, S and surface waves. The IRIS web interface provides following options for selecting events:

i. Azimuth and depth of event

ii. Magnitude of event

iii. Number of channels of seismometer

iv. Seismic network

v. Arrival time of P waves

vi. Seismic stations.

4.1.2 Depth of Event:

In seismology, the **depth of focus** or **focal depth** refers to the depth at which an earthquake occurs. Earthquakes occurring at a depth of less than 70 km (43 mi) are classified as shallow-focus earthquakes, while those with a focal depth between 70 km (43 mi) and 300 km (190 mi) are commonly termed mid-focus or intermediate-depth earthquakes. Earthquakes occur in the crust or upper mantle, which ranges from the earth's surface to about 800 kilometres deep (about 500 miles). The strength of shaking from an earthquake diminishes with increasing distance from the earthquake's source, so the strength of shaking at the surface from an earthquake that occurs at 500km deep is considerably less than if the same earthquake had occurred at 20 km depth.

4.1.3 Magnitude of event:

The magnitude is a number that characterizes the relative size of an earthquake. Magnitude is based on measurement of the maximum motion recorded by a seismograph. Several scales are used for magnitude characterization as explained in section [1.7].

4.1.4 Number of channels of seismometer:

Modern seismometers include 3 elements to determine simultaneous movement in three directions: up-down, north-south, east-west. Each component gives information about earthquake in that direction.

4.1.5 Seismic Network:

A seismic network is defined as a group of stations working together jointly for data collection and analysis. Seismic stations operating independently can be considered a network if the data from these stations is joined and processed together.

4.1.6 Arrival time of P-Wave:

This is also called as onset time. In absence of earthquake there is some background noise which is always present in environment, whenever an event occurs the seismometer records it in seismogram which can be seen as sudden increase in amplitude of signal.

4.1.7 Seismic stations:

A seismic station is considered to be a permanent installation of a seismic sensor, digitizer and some transmission lines with the recorder elsewhere or the station can be complete with recorder and a communication facility. A seismic station is considered to be a permanent installation of a seismic sensor.

A wide variety of signals were downloaded from IRIS consisting of magnitude from 3 to 9 for example Tsunami seismic signal of Japan (2011), Alaska, California and Siberia earthquakes. The National Earthquake Information Centre (NEIC) records an average of **20,000 earthquakes** every year (about 50 a day) around the world. There are, however, millions of earthquakes estimated to occur every year that are too weak to be recorded. Magnitude greater than 4 travels throughout the earth and magnitude less than 3 is hardly felt.

4.1.8 Wavelet transform:

Discrete wavelet transform is very good technique for data compression primarily due to 4 reasons:

i. Time-frequency information
ii. Multi-resolution
iii. Histogram shows data compression is possible

iv. Ability to distinguish between significant and non-significant coefficients.

4.1.9 Time-frequency information:

Using wavelet transform we not only get frequency components present in that signal but also where they are located in time.

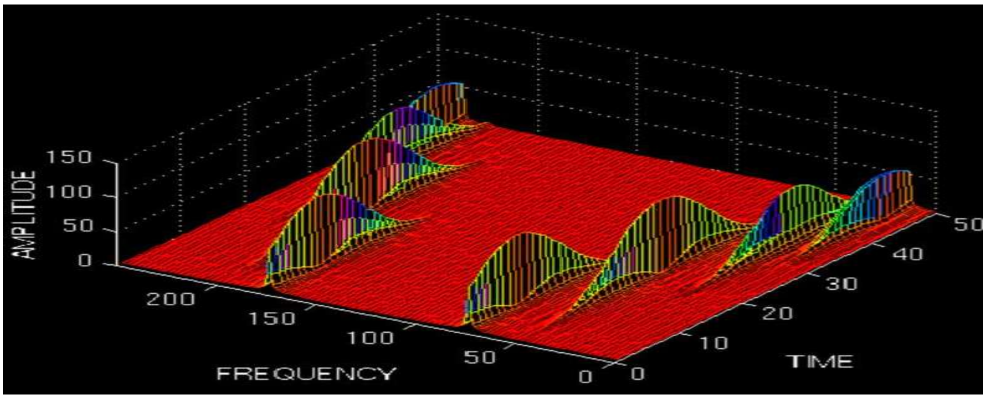

Figure 4.2: Time frequency representation of Wavelet transform

The figure is shown from a top bird-eye view with an angle for better interpretation. All peaks are separated in time and frequency axis. These peaks correspond to different time interval in time-domain signal. This is also known as space-frequency localization meaning that at any specified location in space, one can obtain its details in terms of frequency.

4.1.10 Multi-Resolution:

It is our common observation that the level of details within an image/signal varies from location to location. Some locations contain significant details, where we require finer resolution for analysis and there are other locations, where a coarser resolution representation suffices. A multi-resolution representation of an image/signal gives us a complete idea about the extent of the details existing at different locations from which we can choose our requirements of desired details. Multiresolution representation facilitates efficient compression by exploiting the redundancies across the resolutions. Wavelet transforms is one of the popular, but not the only approach for multi-resolution image analysis. One can use any of the signal processing approaches to sub-band coding, such as quadrature mirror filters (QMF) in multi-resolution analysis.

4.1.11 Histogram shows data compression:

Histograms built on the cumulative data values gives very good idea related to spread of signal with limited space usage. According to wavelet correlation theory a useful signal have a strong correlation in various decomposition scales, whereas the wavelet coefficients of noise are weakly correlated or uncorrelated. Therefore, the histogram distribution of wavelet coefficients is narrower compared to original signal's distribution hence confirming that wavelet transforms can provide us with improved compression ratio. Here are some plots showing the histogram of original and DWT coefficients.

4.1.12 Ability to distinguish between significant and non-significant coefficients:

The plot of DWT shows the significant and non-significant coefficients clearly. The significant coefficients are reserved and non-significant coefficients are removed.

4.1.13 Thresholding:

Wavelet threshold denoising method was proposed by American scholar Donohue. The thresholding methods can be used to cut-off the undesired coefficients that represents the noise in the signal or images. WT is capable to study the signal or image at various scale and across time (time-scale approach). Denoising via thresholding is the key to signal processing technology, that is the process of removing noise to the maximum extent to restore the original signals. It has became an indispensable link in signal processing. In seismic signals the background noise is of small amplitude in wide frequency spectrum is always present. The threshold value is chosen as per the application and the values less than threshold are made zero.

4.1.14 SAC file format:

HEADER Start Word=0 Word Length=158 (32 bit each)	DATA SECTION Start word=158 Word Length=NPTS(no. of points, 32 bit each)

Figure 4.3: SAC file format

SAC file consists of header and data section arranged as shown in figure above. The header is of 158 word length each of 32 bit in size. Header consists of various information regarding signal, origin, location, seismic station, network Id, timing information etc. Data section starts with 158th word. The size of data section depends on the file duration that is more duration, more number of sample points hence large size data section.

4.1.15 Compressed file format:

Header	Length of Wavelet coefficients by Level	Total no. of Elements	Coeff	Pos	Coeff	Pos	— — —	Coeff	Pos

Figure 4.4: Compressed File format

The header information of the original file is preserved in the compressed file. Length of all wavelet detail and approximation coefficients is stored in an array.

L(0)= length of approximation coefficient

L(1)=length of detail coefficient at Nth level

L(2)=length of detail coefficient at (N-1)thlevel

L(n)=length of detail coefficient at 1st level.

Total no. of coefficients required to have information regarding how many sample points are present in the original signal so the same number of sample points is required in reconstructed signal also. Next are the non-zero coefficients left after thresholding along with their positions. Position storage is required so as to get original array back (after thresholding) as in transmitter side.

4.1.16 Decompression

At receiver side the compressed file is read and coefficients are restored to their original positions. Inverse wavelet transform is applied to this array to get sample points. These are converted to original file format in which the data was downloaded. Thus, called reconstructed file.

It includes the following steps:

i. Read compressed file
ii. Interpolate
iii. Inverse wavelet transform
iv. Convert to original file format

At receiver's side the compressed file is read and all the consecutive places are restored with zero value. Inverse wavelet transform is taken of this array using the same wavelet function and same number of levels. This is then converted to original file format i.e. SAC now called as reconstructed file.

4.2 Similarity Check

The following parameters are used to check the similarity

4.2.1 Percentage energy retained:

This is defined as the amount of energy of original signal that is present in reconstructed signal.

PER = $\frac{\sigma_r^2}{\sigma_x^2}$ x 100, where σ_r^2 is the variance of the reconstructed signal and σ_x^2 is the variance of the original signal.

4.2.2 Cross correlation coefficient:

Cross-correlation is a measure of similarity of two different series as a function of the displacement of one relative to the other. The possible range for the correlation coefficient of the time series data is from -1.0 to +1.0. The closer the cross-correlation value is to 1, the more closely the sets are identical.

$\rho_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - x)^2 \sum(y_i - y)^2}}$, where 'x' is original signal and 'y' is reconstructed signal.

4.2.3 Percentage Compression:

It tells how much compression is achieved when compared to original file.

Percentage Compression: $\frac{(\text{size original} - \text{size compressed})}{(\text{size original})}$ x 100

4.3 Challenges in the proposed algorithm

There following are the challenges in using Wavelet transform for compression.

i. Selecting mother wavelet
ii. Selecting number of decomposition levels.
iii. Thresholding function

4.3.1 Selecting Mother Wavelet:

Choosing or designing the right wavelet is crucial for a successful wavelet transform application but choosing the right wavelet for a specific application has been an unanswered question. The chosen wavelet must be close to the analysed signal. Hence, it gives a prefect reconstruction with few decomposition levels. The error between original signal (x_s) and reconstructed signal (x_r) should be the smallest for prefect reconstruction. The reconstruction criterion is evaluated using the most common method, which is the mean square difference method given in equation:

$$E_{error} = ||x_s - x_r|| = \sum_{n=0}^{N-1} |x_s - x_r|^2$$

Another parameter chosen for selecting mother wavelet was PER (percentage energy retained) between reconstructed and original signals. The formula used was:

$$PER = \frac{\sigma_r^2}{\sigma_x^2} * 100$$

The signal flow diagram for calculating PER and MSE was:

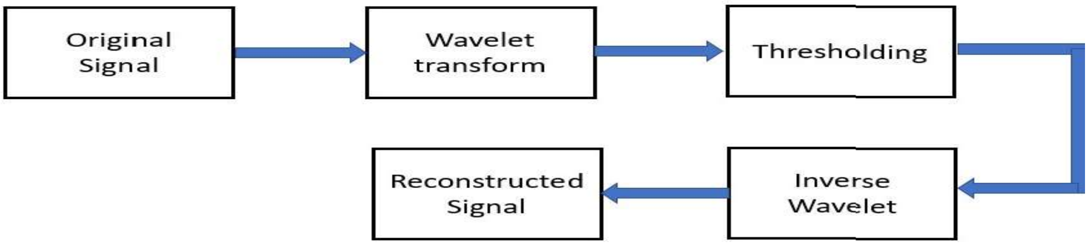

Figure 4.5: Block diagram for MSE or PER calculation

The Result of this experiment was:

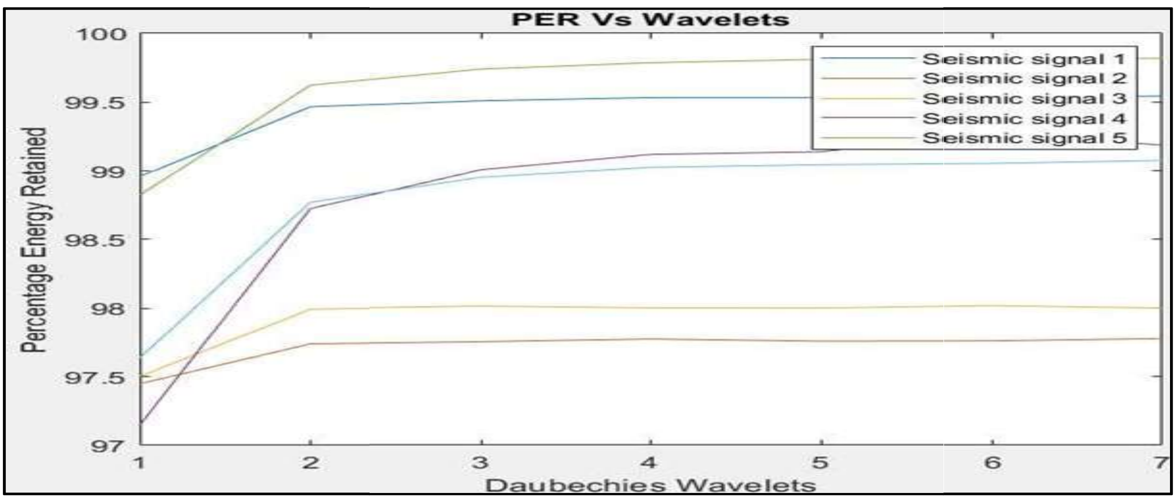

Figure 4.6: PER Vs Daubechies Wavelet

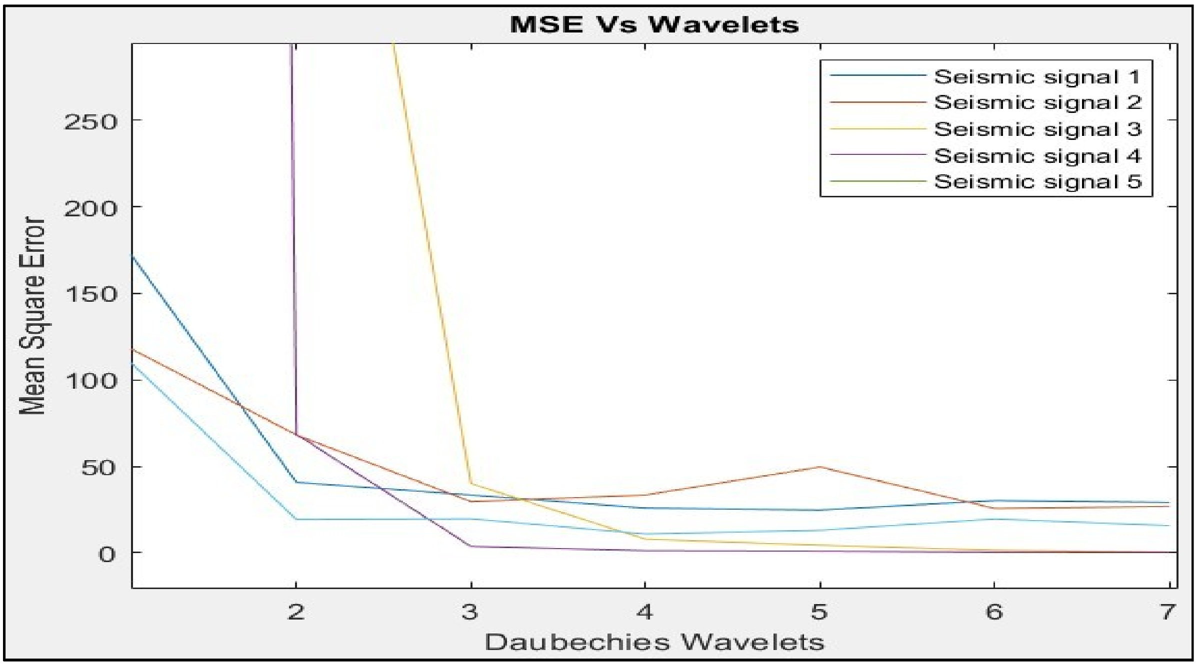

Figure 4.7: MSE Vs Daubechies Wavelet

4.4 Selection of Decomposition Levels:

Again there is no hard and fast rule to select levels of decomposition. It all depends on application. As the levels increase the computational cost increases and also that each decomposition level is a filter for which the initial signal must pass through. Considering proper decomposition level and wavelet function to reconstruct original signal and extract desired features has an undeniable importance in signal processing. The summation of all levels decomposition coefficients has to be equal with original signal. Lower decomposition levels keeps high frequency information at high resolution and higher decomposition levels keeps low frequency information with low resolution. The calculated percentage energy retained for all levels (up to 12^{th} stage of decomposition.) is shown in the graph.

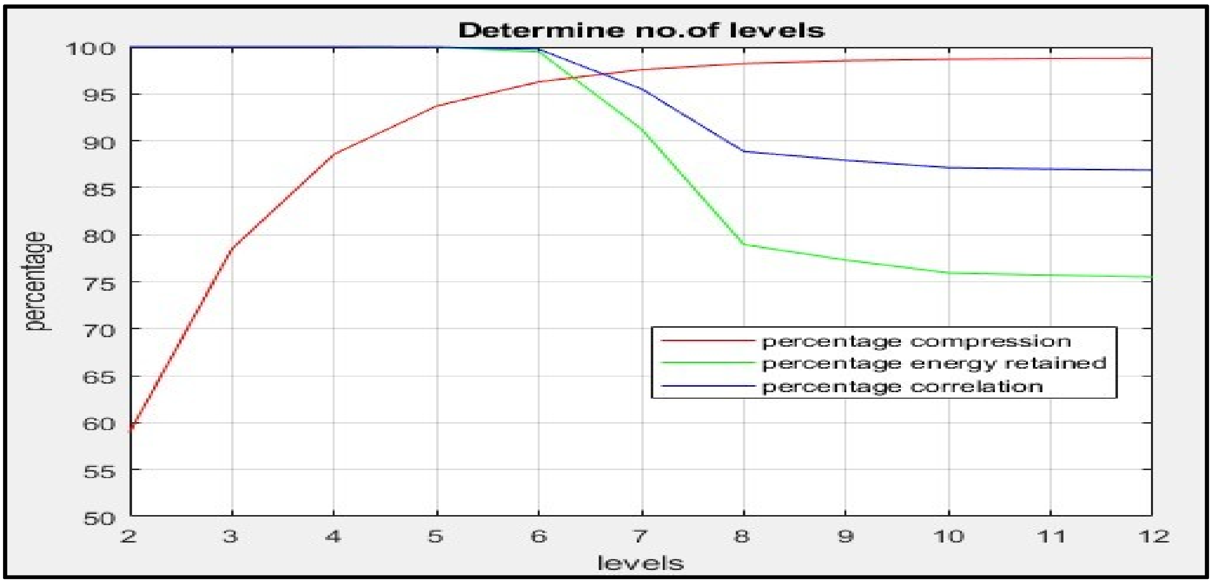

Figure 4.8: Percentage of parameters Vs Levels of Seismic signal of Tsunami data of Japan station(9).

Figure 4.8 shows is of tsunami in Japan i.e. a high magnitude signal (6 above) PER, percentage compression and percentage correlation between original and reconstructed wave is calculated as shown in different colours. It is observable that as number of levels increase the more and more coefficients gets thresholded/filtered hence giving a good compression but simultaneously decreasing correlation and PER values. So, the idea is to select a optimum level number to balance all parameters.

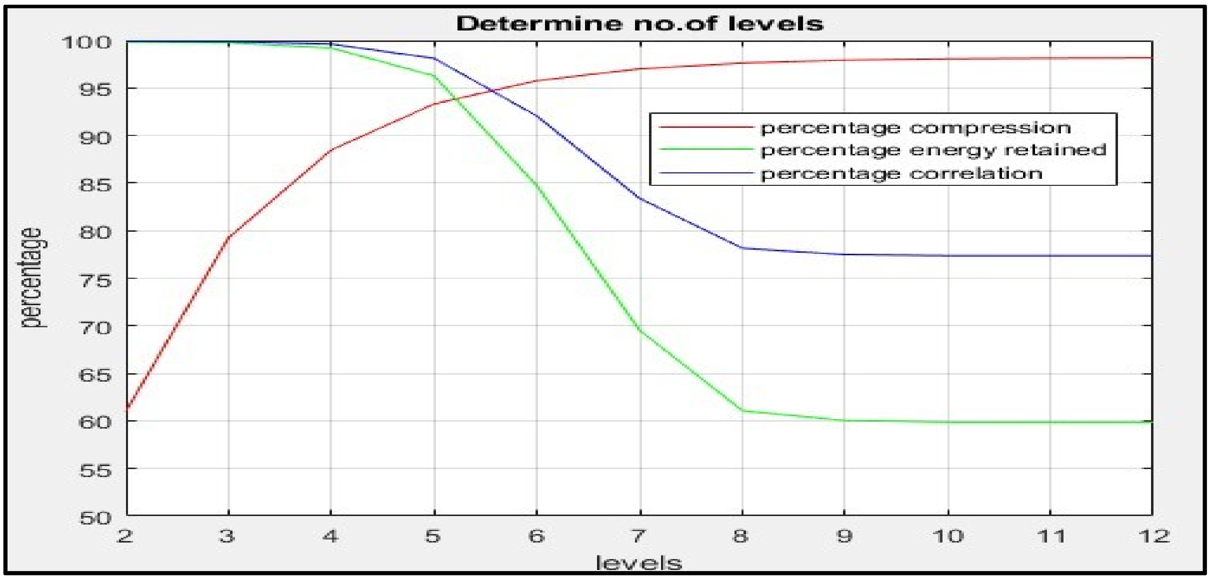

Figure 4.9: Seismic event data of mona passage (5.5)

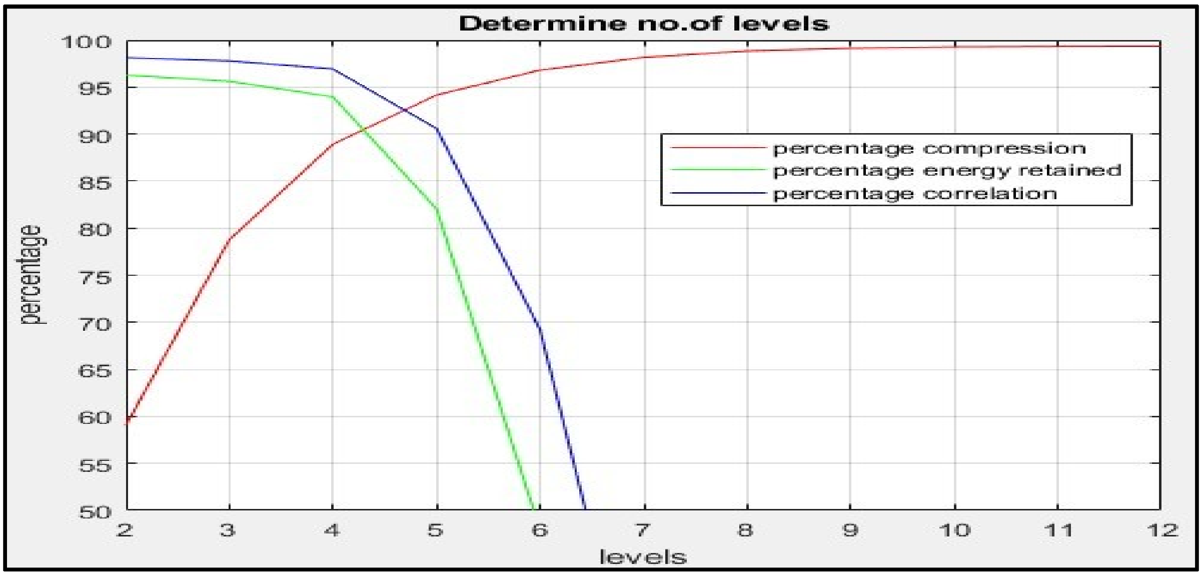

Figure 4.10: Seismic event data of northern California region

Figure 4.10 the optimized levels are between 4 to 7 depending on the magnitude of the signal. In case of high magnitude signal it is 7 and in case of noise it is 4 so we have taken 6 as the average value.

4.5 Thresholding Function

Owing to simple calculation and good denoising effect, wavelet threshold denoising method has been widely used in signal denoising. In this method threshold value is an important parameter that affects denoising. In a noisy signal noise energy is generally concentrated in high frequency region, and the spectrum of useful signal is distributed in the low frequency region. However, in some signals the high frequency region not only contains noise but also possesses a lot of useful information. Therefore, directly filtering out high frequency information is unreasonable. Recently because of multiresolution and low entropy, wavelet transform has become a popular research topic in signal denoising. The essence of these methods is non-linear processing on the wavelet coefficients and then using the processed coefficients to reconstruct signals. The denoising effect of wavelet threshold method depends on threshold determination. If the selected threshold is too large, then some useful information is filtered out, and if the threshold is too small, then a certain amount of noise is retained. In order to solve this problem many researchers studied the threshold determination methods. Donoho and Johnstone[24] presented a universal threshold by analysing white noise as expression given in equation given in section [2.7]. The following graph shows the effect of thresholding.

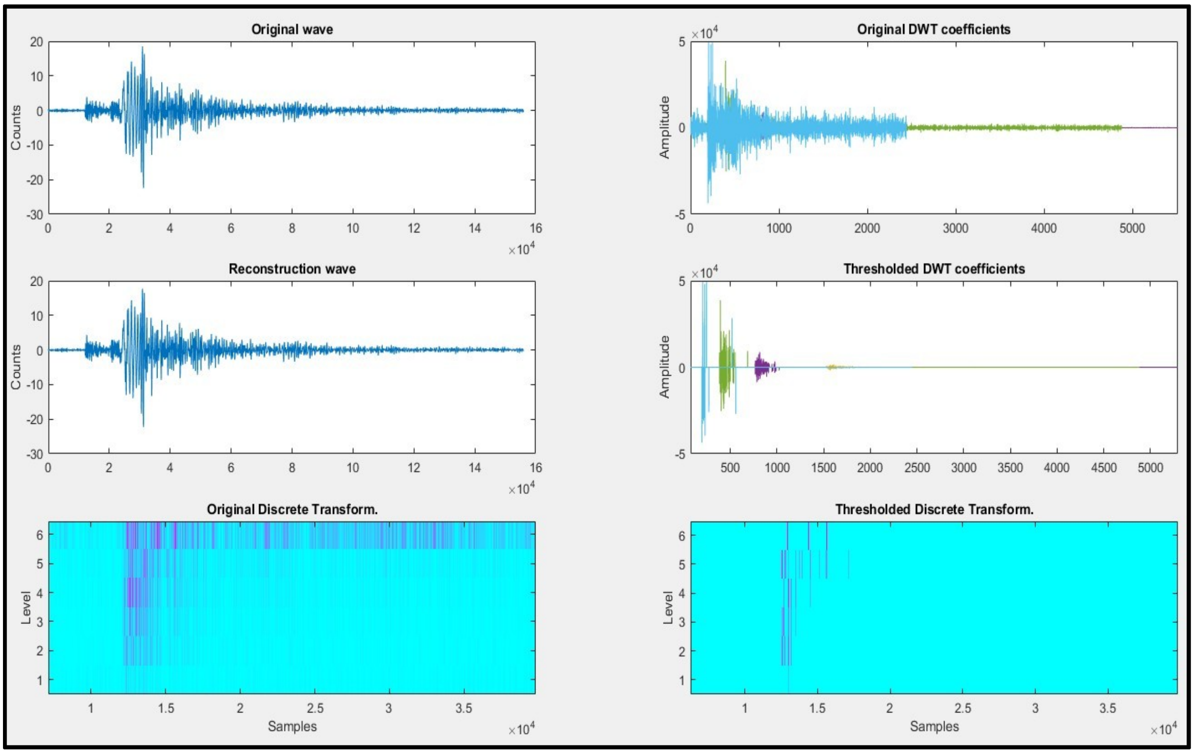

Figure 4.11: Graphs showing the effect of thresholding

Figure 4.11 shows the original detail coefficients are actual ones of the original signal and the thresholded DWT coefficients shows only significant higher ones are left rest are made to zero.

4.6 Decomposition graphs at different levels.

In figure 4.12 a seismic signal of COCO islands (near Myanmar) is taken from IRIS website as shown in red colour. Then DWT is applied which breaks down signal to many sub- bands called detailed (d1 to d8 shown in green colour) and approximate coefficients (shown in violet colour) Here we have used db3 up to 10 levels of decomposition. As we can see here approximation plot a8 contains low frequency and detail plot d1 to d8 contains high frequencies. “d1” level contains the highest frequency band which can be eliminated completely without any loss to the signal. Rest detail coefficients are threshold to certain level to achieve compression.

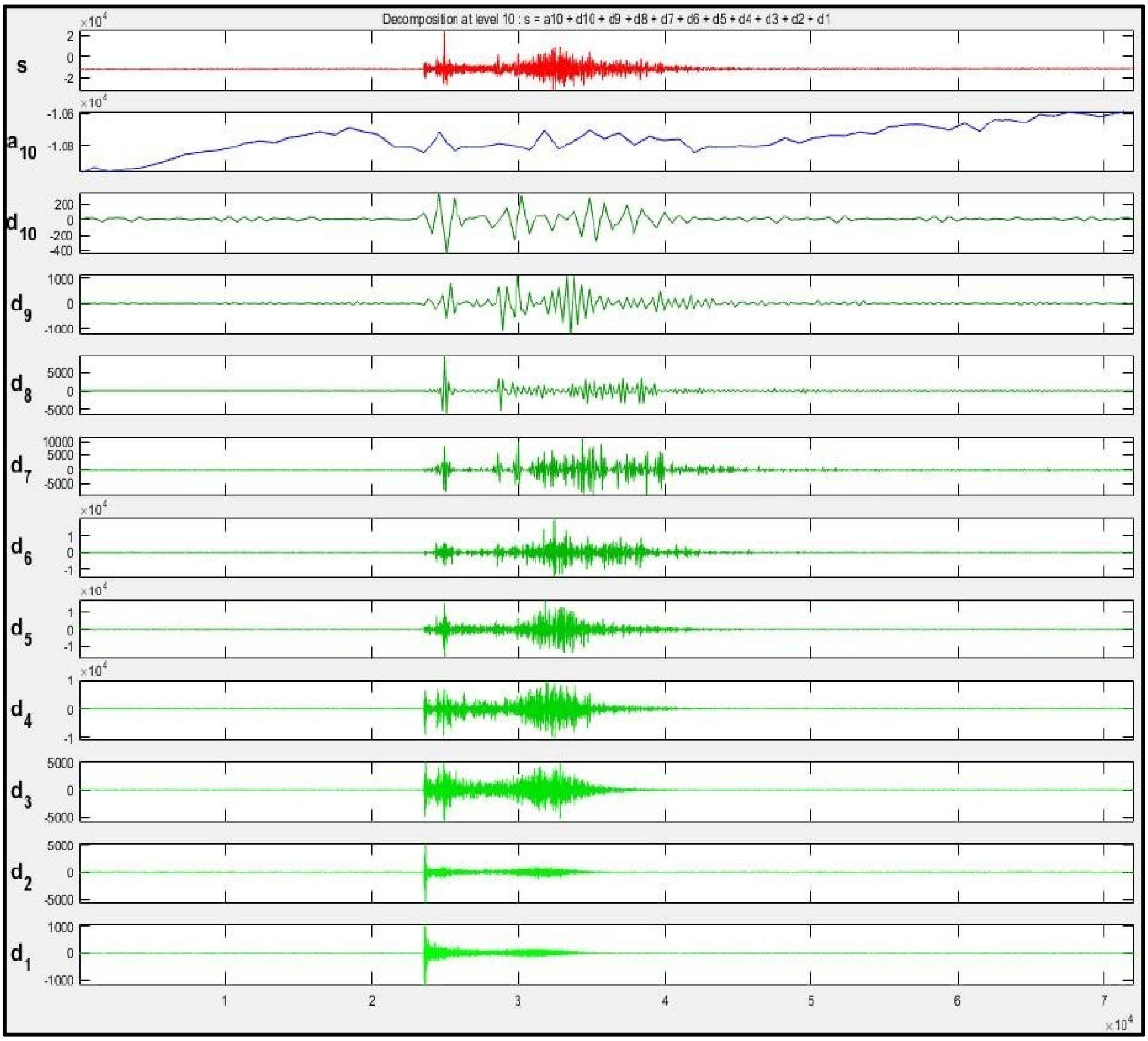

Figure 4.12: Decomposition Graphs of a Nepal Earthquake Signal

4.7 Autocorrelation of Detail and approximation coefficients

DWT sub-samples a signal by 2 hence eliminating redundancy. In SWT this down sampling is not done, so redundancy is there hence we can visualize behaviour of the coefficients properly. Detail coefficients mostly contain high frequency, the same is shown by taking their autocorrelation which comes like delta function (autocorrelation of white noise is delta). As per Wiener khinchin's theorem, the Fourier transform of autocorrelation is power spectral density of the signal and vice-versa as shown in equations below:

$$R_{xx}(r) = \int_{-\infty}^{\infty} S_{xx}(f) e^{i2\pi fc} df \tag{4.1}$$

$$S_{xx}(f) = \int_{-\infty}^{\infty} R_{xx}(r) e^{-i2\pi f} dr \tag{4.2}$$

$R_{xx}(\tau)$ is auto-correlation of signal and $S_{xx}(\tau)$ is power spectral density of signal.

The autocorrelation equation formula of a random signal 'y' is given as:

$$\rho_h = \frac{\sum_{t=h+1}^{T} (y_t - \bar{y})(y_{t-h} - \bar{y})}{\sum_{t=1}^{T} (y_t - \bar{y})^2} \tag{4.3}$$

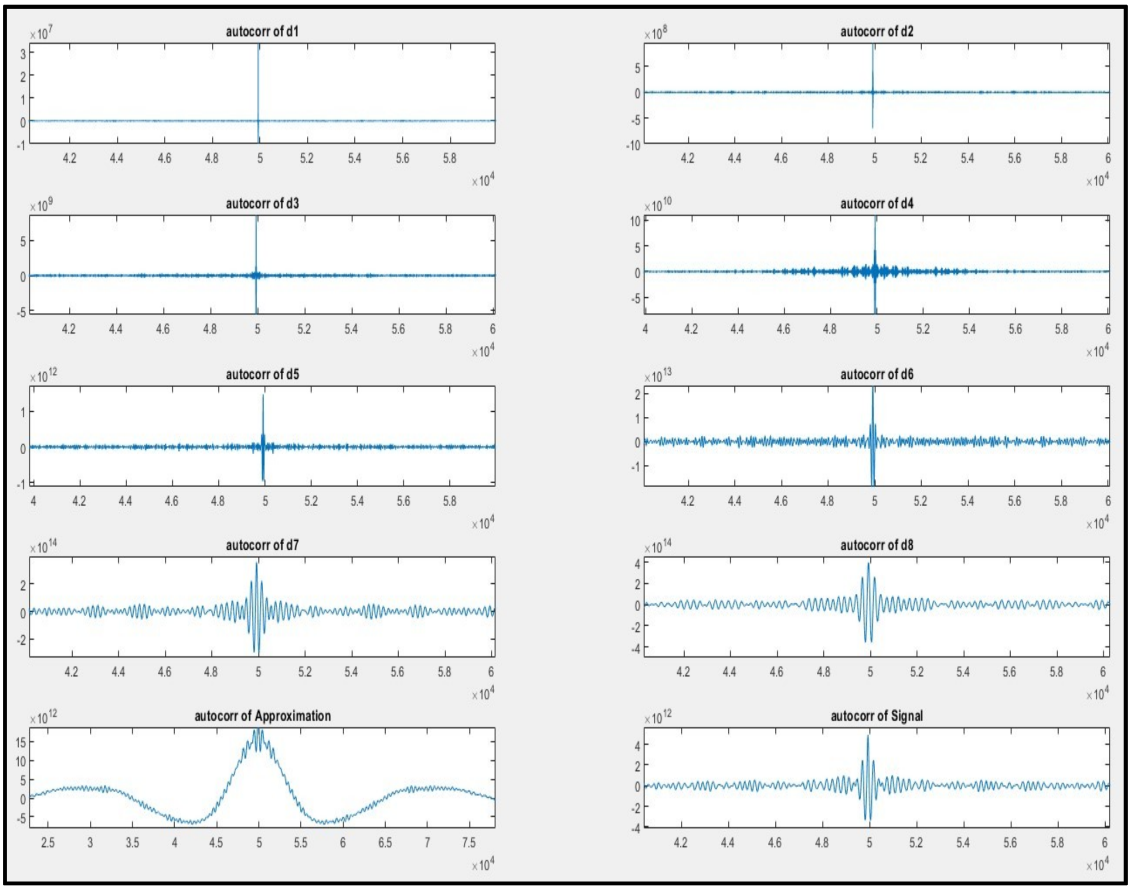

Figure 4.13: Autocorrelation of Detail and approximate coefficient

Figure 4.13 shows that the autocorrelation of detail coefficients at lower levels comes out to be a delta function and this indicates their power spectrum density will be white noise. Whereas the auto correlation at higher detail levels and approximate level is more like signal.

The thresholding function used will not be applied on approximation coefficients as it contains most signal part instead it will be applied on detail coefficients as a function of detail levels because each detail level holds different frequency range.

This chapter explained each step of proposed algorithm and also elaborated the methods used for selection of wavelet and number of decomposition levels.

5 Results

In this section, the results and conclusions of the proposed algorithm are presented. For validating the reconstructed and original signal their onset time and magnitude are tabulated. Test signals are used for understanding the working of the STA/LTA and AIC picker techniques.

5.1 Comparison Table showing PER, cross correlation and Compression percentage.

The seismic files used for this project are of earthquakes recorded at various seismic stations within 20 degrees(1 degree = 111km) of the epicentre downloaded from IRIS website (https://www.iris.edu/hq/). The earthquake events taken for the analysis are of Japan, Siberia Russia, Alaska, Hawaii, Australia, Mona passage, Coco islands etc. All files are of duration 65 minutes with file size 312KB (20 sps) and 624KB (40sps) in SAC format. These files are given as input to the proposed compression algorithm which gives output as a compressed file (Column 5 of table 5.1 showing compression percentage). Table 5.1 shows epicentre, magnitude, original file size (all files are of duration 65 minutes in SAC format), compressed file size, percentage compression, cross correlation, PER and compression by a commercial software (ZIP). Parameters like PER, is used to check the amount of energy retained in reconstructed signal and cross correlation checks similarity between original and reconstructed signal (as explained in section 4.2).To show a comparison with lossless scheme, ZIP software was used to compress the seismic file (Column 8 of table 5.5)

Table 5.1							
Event	**Magnitude**	**Original size (KB)**	**Compressed size (KB)**	**Percent age Compression**	**Percentage Energy retained**	**Correlation Coefficient *100**	**Compression % by commercial software(ZIP)**
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
East Coast Japan	6.1	312	12.5	95.9	97.7	98.8	42.9
		312	11.6	96.2	99.8	98.3	39.7
		624	26.4	95.7	97.8	98.9	27.7
		312	13.9	95.5	99.0	99.5	29.1
		624	25.7	95.8	98.9	99.4	29.9
Eastern Siberia Russia	6.4	624	22.7	96.3	99.4	99.7	41.8
		624	19.5	96.8	98.7	99.3	42.6
		624	19.8	96.8	99.3	99.6	27.2
Alaska Peninsula	7.8	624	26.9	95.6	97.7	98.8	28
North Korea Nuclear explosion	6.3	624	22.2	96.4	98.1	94.9	48.8
Japan Tsunami	9.1	312	10.7	96.5	99.2	99.6	20.8
Average				**96**	**98.3**	**99.1**	**34.4**
Mona passage	5.5	624	19.1	96.9	96.6	96.2	46.9
		624	19	96.9	96	97.9	55.7
		312	12.8	95.9	96.5	93	38.1
		624	25.3	95.9	98.4	97.7	50.6
		624	19.8	96.8	99.3	99.6	47.7
		624	25	95.9	96.1	98.0	50.9
		624	24.5	96	97.4	95.6	46.9
		312	11.3	96.3	96	94.3	48.7
Hawaii	5.1	624	24.3	96	96.3	90.7	40.2
		624	21.5	96.5	98	95.9	57.6
		624	20.9	96.6	98.2	95.5	79.4
Average				**96.33**	**97.16**	**95.8**	**51.1**
Noise		362	11.9	96.6	79	88.9	45.0
		181	6.8	96.2	17	41.2	41.4
		390	12	96.9	34.3	58.6	77.1
		389	11.8	96.9	47	69	48.0
		182	6.9	96	13.1	36.1	40.6
		366	12.7	96.5	74	86	46.7

Table 5.1:Comparison table showing PER, Cross correlation and Compression percentage for different magnitude earthquakes at different regions.

Table 5.1 shows that the percentage compression is 95-96% and the corresponding cross correlation values are 98% or more and PER is also high up to 96% and above for high and medium magnitude signals, On the other hand the commercial loss less ZIP software gives around 50% compression only Hence, the developed lossy compression algorithm gives very high percentage compression with less variation in PER and cross correlation. In case of noise signal the PER and cross correlation values are very poor because the noise is a random signal existing in all frequency bands which can't be compressed. But PER and cross correlation are not the sufficient parameters for validation of seismic signals. There is a need to introduce other validation parameters to verify this reconstruction process.

5.2 Validation of Results

Compression and decompression must be followed with validation process because it helps in qualifying the algorithm. Decompression algorithm generates reconstructed signal which will be compared with the original signal. In seismic signals quality of the signal is quantified based on onset of the arrival of different seismic phases in the seismogram and their magnitude. The important parameters to qualify our proposed algorithms are:

i. Onset time
ii. Amplitude

Onset time is defined as the first arrival of a seismic signal. It is usually marked by sudden change in signal's amplitude and frequency content. If onset time differs there will be a delay/ error in phase added to the signal which will introduce errors in the further calculations of estimating origin-time, location and magnitude of earthquake.

Similarly, amplitude variation is also important as it tells about the magnitude of earthquake.

The relation between amplitude, time period, distance and depth of focus is given as:

$$\mathbf{M_b = \log_{10}(A/T) + Q(\Delta, h)} \tag{5.1}$$

Where

A = It is the maximum amplitude observed on seismograph.

T = It is the time period of P-waves,

$Q(\Delta, h)$ = It is the attenuation factor w.r.t distance(Δ) between epicentre and seismic station and depth(h) of focus from earth's surface.

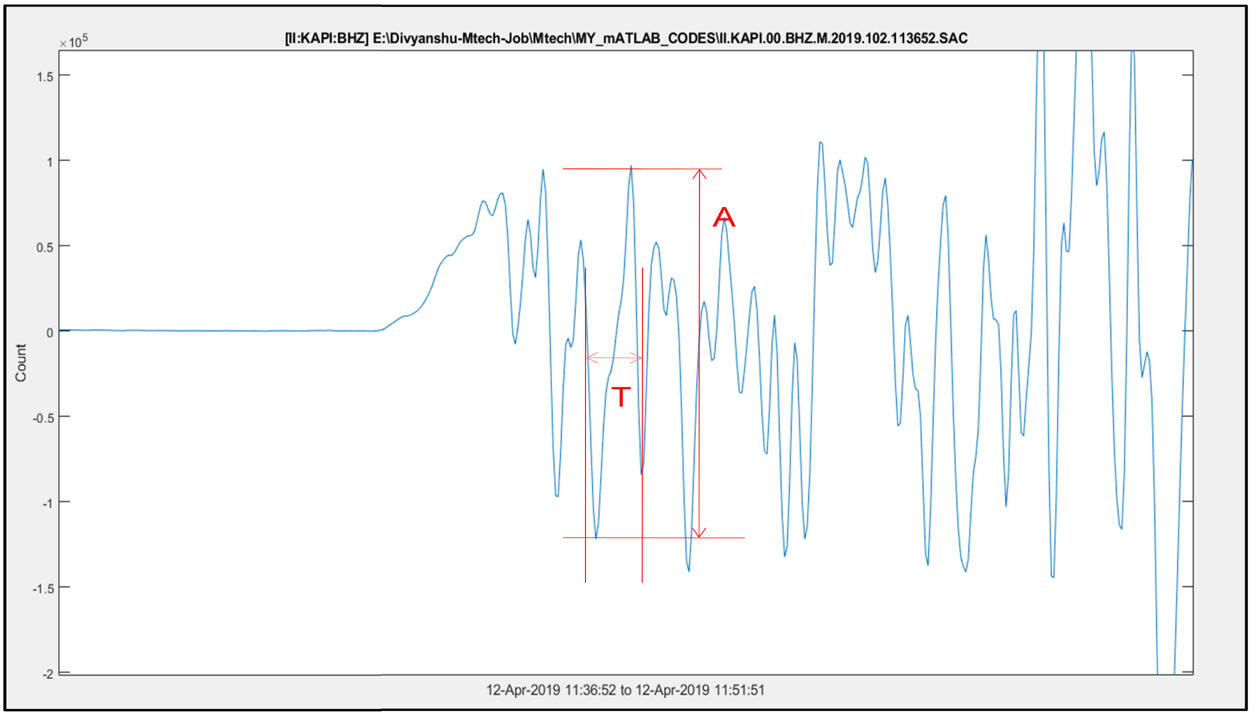

Figure 5.1: P-wave in Seismic signal

The figure 5.1 shows the amplitude 'A' and time period 'T' in P-wave which are used in the magnitude calculation. Up scaling/ downscaling of magnitude (M_b) by 1 will happen if $\frac{A}{T}$ changes by 10 times. Since, the magnitude mainly depends on amplitude of the signal hence the same will be verified from the figure 5.15 by overlapping both original and reconstructed signal.

5.3 Onset time validation

5.3.1 Filtering

Pre-filtering is required before applying STA/ LTA to avoid the false trigger due to spikes (unwanted high frequency noise) present in a signal. The seismic wave bandwidth for teleseismic signals lies between 1/40 seconds to 2Hz posing a requirement of lowpass filter of bandwidth 2Hz as it will select only the signal portion. Figure 5.2 shows the magnitude response in frequency domain of lowpass FIR 100th order with 3db bandwidth of 2Hz slope of 38db/octave.

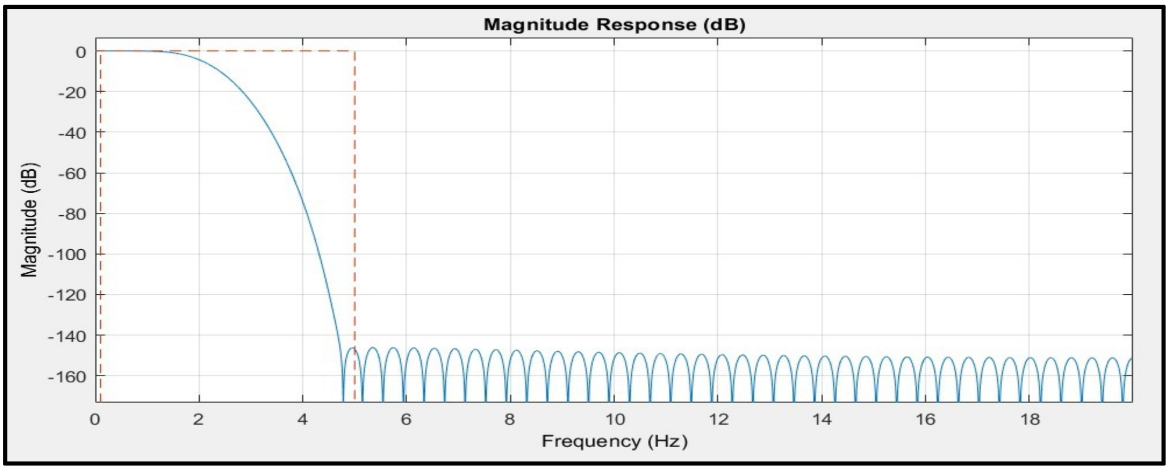

Figure 5.2: Lowpass FIR filter used

This magnitude response differs (blue line in figure 5.2) from the magnitude response of the ideal low pass filter (red line in figure 5.2). It is a close approximation of the ideal low pass filter. There is still a pass band – an interval of frequencies passed by the filter, and a stop band – an interval of frequencies stopped by the filter. There is, however, usually a transition band (in figure 5.2 it extends from 2Hz to 5Hz), as the filter does not immediately move from the pass band to the stop band, but rather does that smoothly over an interval of frequencies around the transition frequency. Digital filters characteristically have ripples that increase the closer we get to the transition band; for example the stop band frequencies are not completely "stopped" and so there is "stop band attenuation": the amount of decrease in the amplitude of the stop band frequencies. In Figure 5.2 the stop band ripples are even smaller than -140db.Frequency spectrum of the signal decides design parameters of the filter. A signal can be represented as linear composition of different sinusoids with varied amplitude, frequency and phase. Distinguishing this composition requires analysing the corresponding frequency, amplitude and phase spectrum of the signal. In theory, a signal's frequency spectrum is its presentation in the frequency domain based on the Fourier transform of its time domain function. Figure 5.3 shows single sided frequency spectrum of four seismic signals recorded at four different seismic stations Macquarie island in Australia, Barbuda island in Caribbean Sea, Florida USA, Coco islands in Bay of Bengal corresponding to different earthquakes.

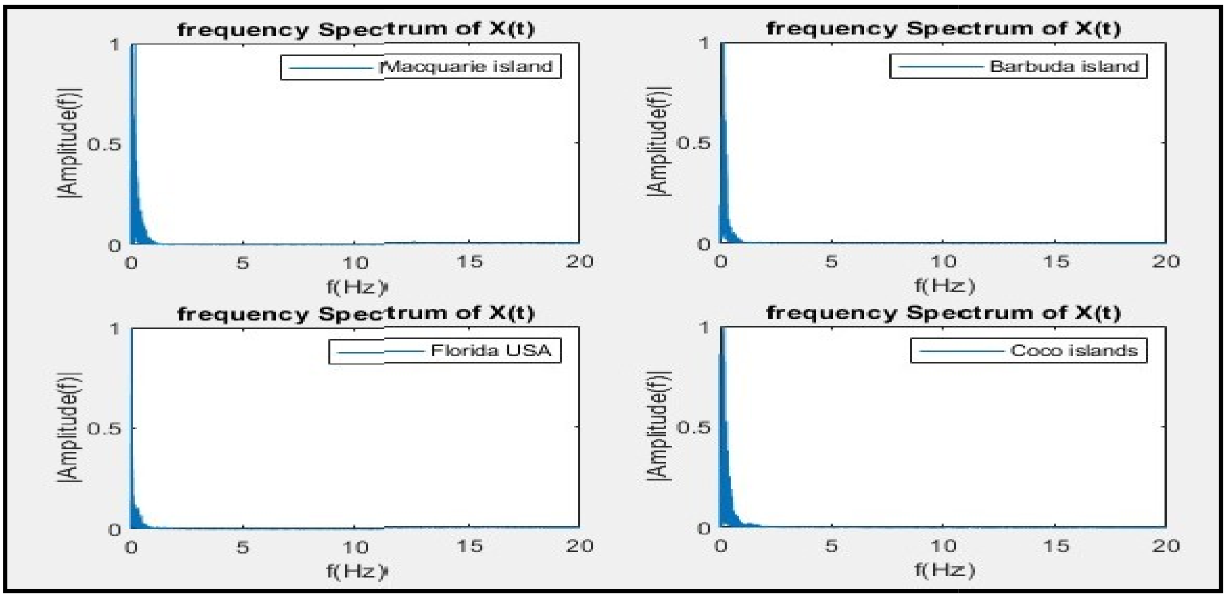

Figure 5.3: Frequency spectrum of Seismic signal.(a) top left image is of Macquarie island, Australia,(b) top right is the frequency spectrum of Barbuda island.(c)bottom left is frequency spectrum of Florida,USA (d)botton right is frequency spectrum of Coco islands of Myanmar in bay of bengal

The figure 5.3 depicts that the spectrum of all signals is concentrated in 1/40 seconds to 2Hz range. The horizontal axis represents the discrete frequencies and the vertical axis shows absolute amplitude values normalized with the maximum amplitude value. The amplitude values are calculated using "FFT" technique. Only positive frequency values are plotted here.

The effect of filtering can be seen clearly on the following signal in figure 5.4 shows the seismic signal of Macquarie Island, Australia date July 22 2020, time= 5:57:32. A spike has been intentionally added to the signal at 06:18:30.

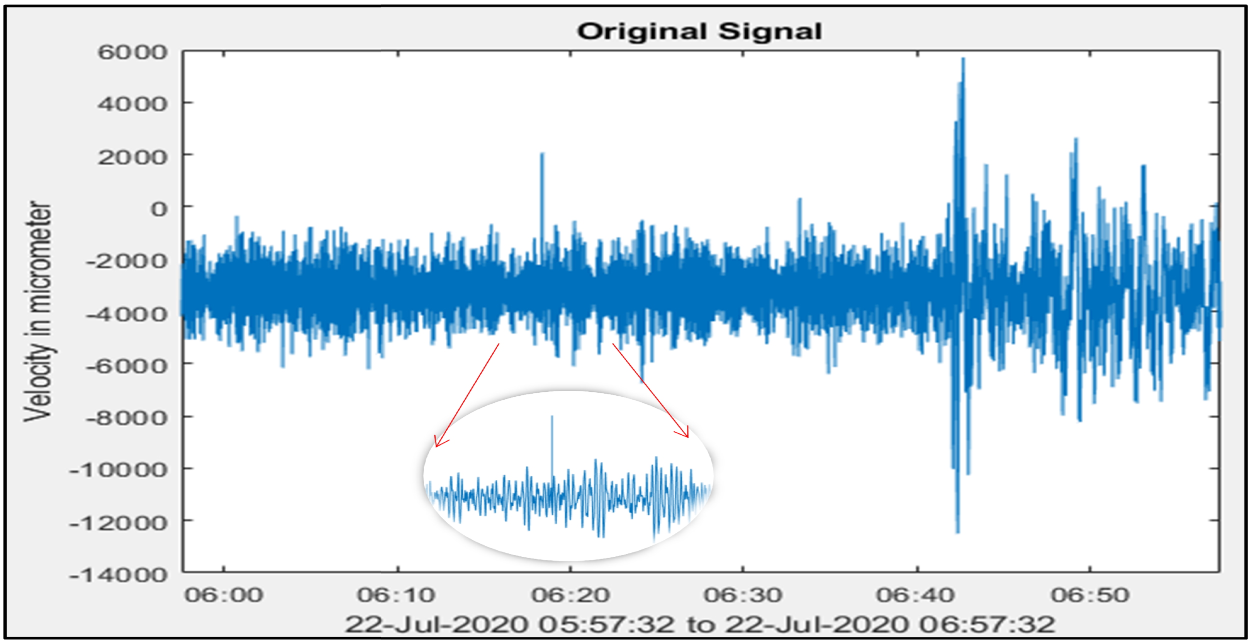

Figure 5.4: Seismic signal with added spike

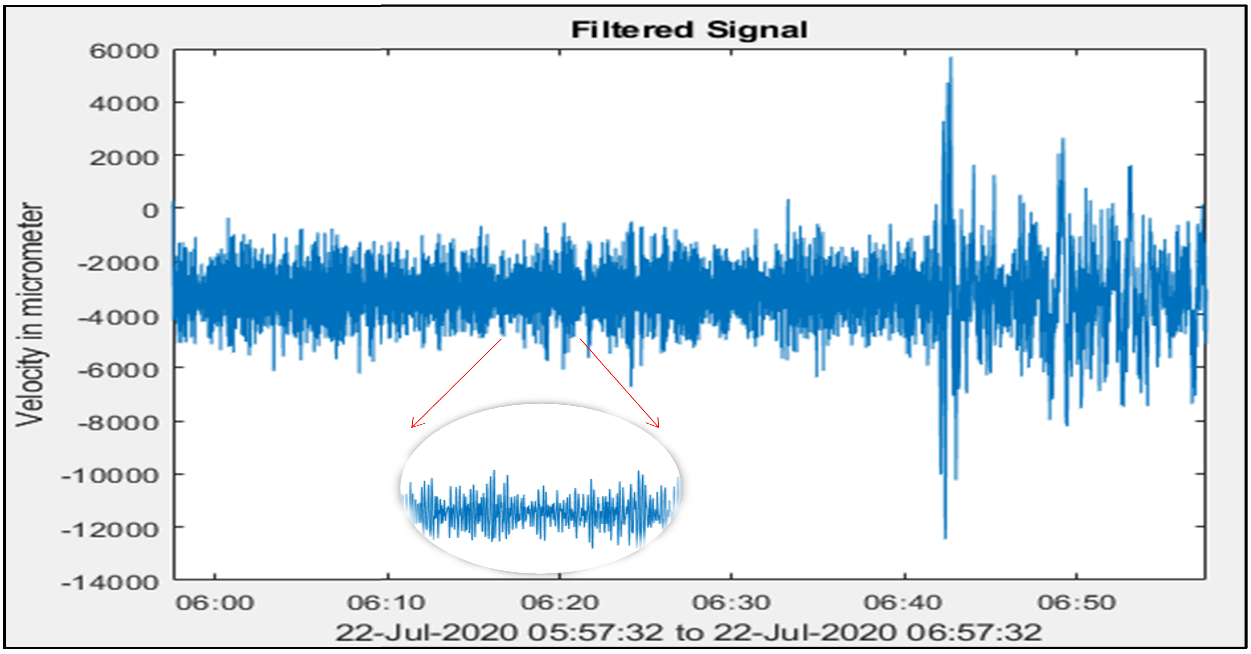

Figure 5.5: Seismic signal After filtering

After filtering as shown in figure 5.5 the spike at 6:18:30 got removed and rest of the signal is same. Spike added is shown as below:

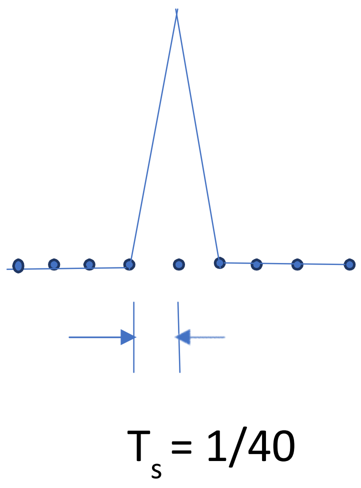

Figure 5.6: Spike signal of frequency 20Hz

The sampling frequency of the seismic signal is fs = 40 samples per second and hence T_s is 1/40 which is the time difference between two samples. For the spike signal, which goes up to a point and comes back, the time duration is 2Ts = 2/40 = 1/20 so frequency of spike will be 20 Hz. We can say that we added a high frequency signal of 20Hz called as spike henceforth.

5.3.2 STA/LTA method

The STA/LTA ratio method compares the average energy in a short time average (STA) leading window to that in a long time average (LTA) trailing window. If the average value captured in the STA is larger than the background levels in the LTA, this will produce an STA/LTA ratio greater than 1 and vice versa. When the ratio between the two averages are found to be higher than a specified limit (called threshold), a single channel will trigger. On the other hand if the average value captured in STA is smaller than the background levels in the LTA, this will produce an STA/ LTA ratio smaller than 1. When the ratio between the two averages are found to be lesser than a specified limit, a single channel will not trigger.

5.3.3 Effect of window length

Initially, the long time window length (L_{ltw}) was kept12 seconds and short time window (L_{stw}) length = 0.25 seconds. Small window of STA implies averaging over small interval which

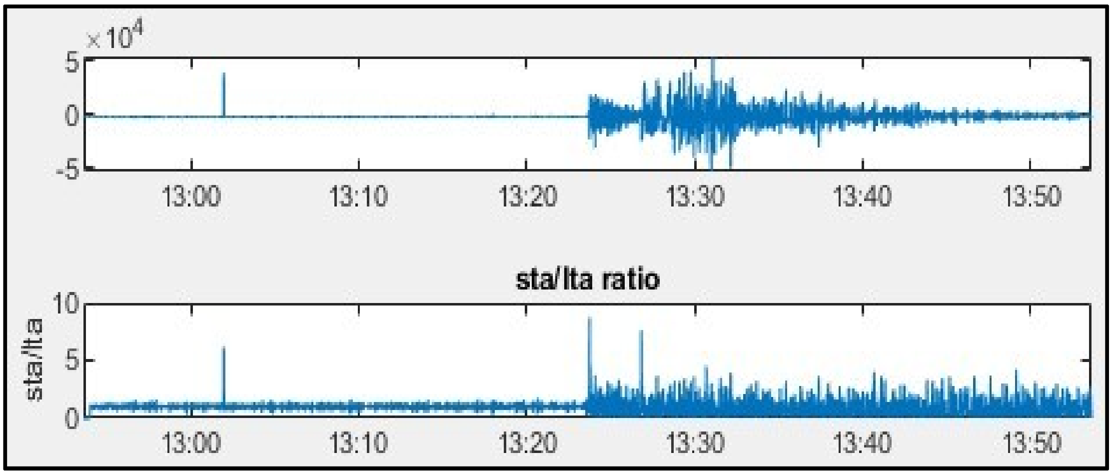

Figure 5.7: Original signal having spike and detected by STA/LTA algorithm

makes triggering more sensitive to high frequency contents (spike-type man-made seismic noise) and the longer the STA duration selected the less sensitive for high frequency content. Such a short window length will result in generating trigger to high frequency content as onset.

In figure 5.7 the above signal is a medium magnitude earthquake event of Japan Okinawa island .A spike (high frequency signal) has been added intentionally to the signal at 13:02:35. The bottom signal in figure 5.6 is the plot of STA/ LTA ratio vs time. We can observe that if threshold is kept 5 then the spike at 13:02:35 will be detected as onset time which would result in a false trigger.

In second instance we kept short term window length is increased to 3 seconds keeping long term window length same, then it is visible in figure 5.7 that noise is not picked up thus detecting correct onset time that is at 13:24:25 .

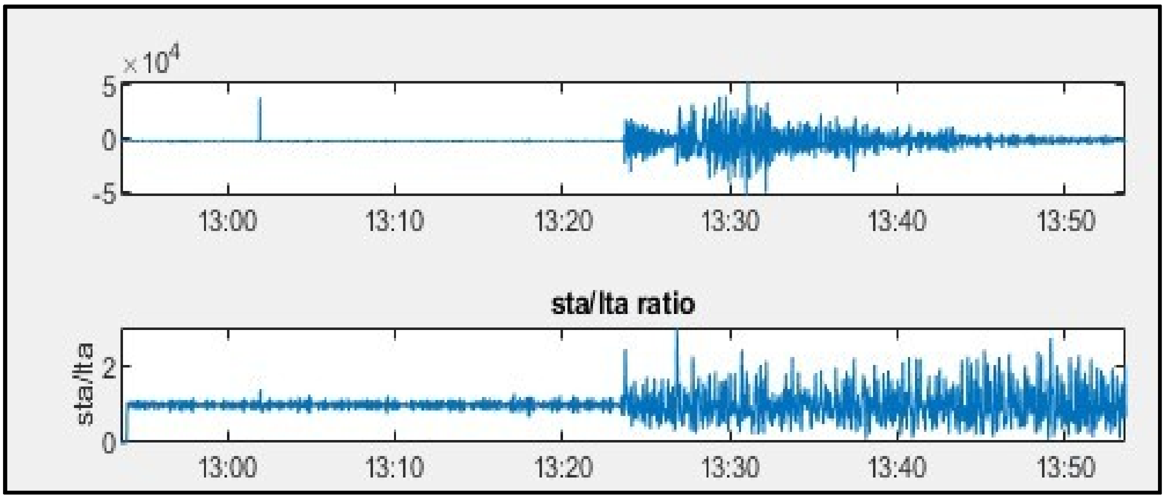

Figure 5.8: Original signal having spike and not detected by STA/LTA algorithm

To some extent the STA functions as a signal filter. The shorter the window length selected, the higher the trigger's sensitivity to short lasting local earthquakes sometimes spikes also compared to long lasting and lower frequency distant earthquakes. The longer the STA the window length selected, the less sensitive it is for short local earthquakes. Therefore, by changing the STA duration one can, to some extent, prioritize capturing of distant or local events.

5.3.4 AIC method

Prompt detection and accurately picking of the first-arrival of a P-Wave is of great importance in locating earthquakes and characterizing velocity structure, especially in the eras of large volumes of digital and real-time seismic data. The detector should be capable of finding the onset of the P-Wave arrival against the background of micro seismic and cultural noise. Normally P-wave onset is characterized by a rapid change in the amplitude and/or the arrival of high-frequency content. The AIC picker has been used to detect and pick up the P-wave arrival. But AIC requires an appropriate time window, or else it will detect the wrong P-Wave arrival. The STA/LTA method is used to trigger P-Wave arrival from which a time window can be chosen for the AIC picker.

5.3.5 Test Signals:

A pseudo random square was generated and it's onset was calculated using STA/LTA method. STA and LTA window duration was of 1 second and of 5 seconds respectively. This signal consists of 50 samples of randomly generated pulse with variable width. Overlapped type windows are used with threshold value = 2.6

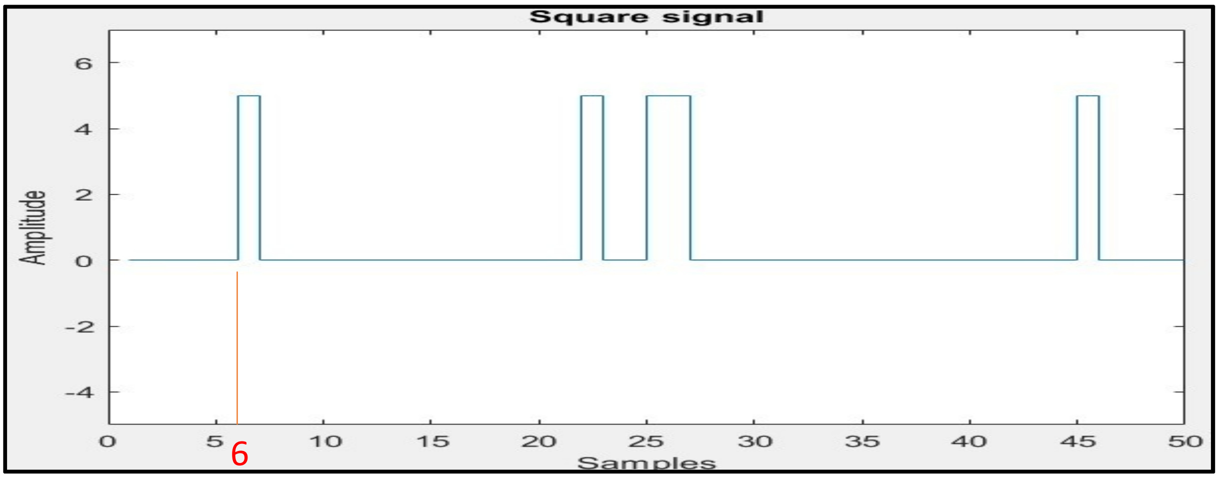

Figure 5.9: Pseudo Random Square Wave

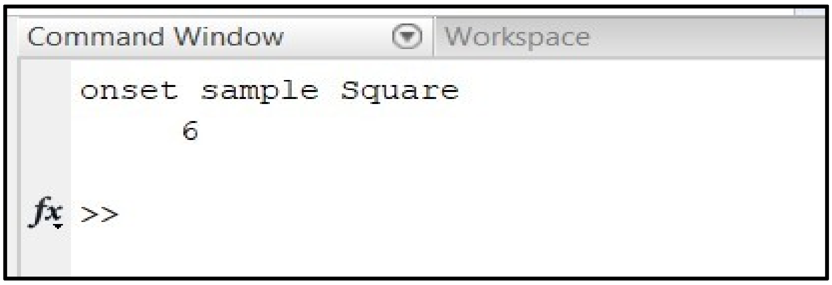

Figure 5.10: Onset time value calculated by STA/LTA algorithm

Figure 5.10 is the screenshot of command window showing the value of onset time of square wave. The first onset occurred at 6th sample in figure 5.9 square pulse and the same is detected by the STA/ LTA algorithm.

5.3.6 Test Signal 2: Noisy Sine Wave

A sine signal added with uniformly distributed noise was generated as shown in figure 5.11. The starting part of the signal is a uniformly distributed noise and later part from 400 to 600 samples is 2Hz sine signal sampled at 100sps (samples per second). This signal was made by first generating a sine wave of 2 Hz at 100 samples per second and then right shifting this sine wave by 400 samples. Lastly, uniformly distributed noise was added to the whole signal. Figure 5.11 shows original noisy sine wave and reconstructed signal. The reconstructed signal was generated by pushing the original signal into the same compression and decompression algorithm.

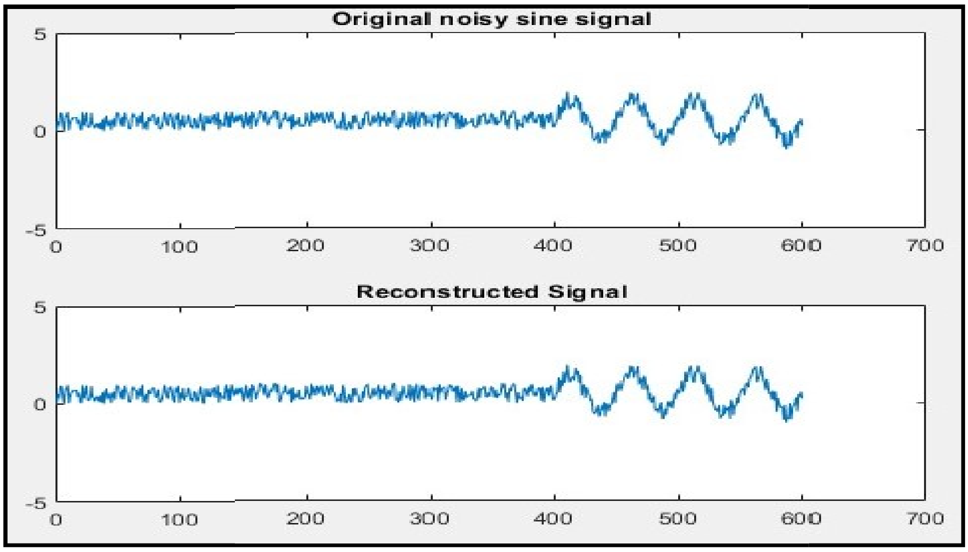

Figure 5.11: Noisy Sine Wave

Figure 5.11 is the screenshot of output window showing onset time of original and reconstructed signal calculated using STA/ LTA method and AIC method.

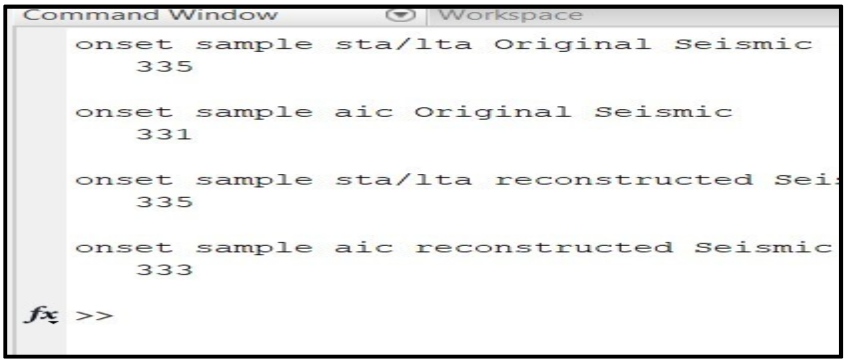

Figure 5.12: Onset time calculations

AIC is applied on a window decided by triggering index of STA/ LTA . The minimum value of AIC array gives the onset. Figure 5.12 shows the minimum of AIC value shown is 331st sample in the original signal and the minimum value shown is reconstructed signal is 333rd which is close to original signal's AIC picker value. The onset values are not exactly same because some randomness was incorporated due to noise addition.

5.4 Magnitude validation

As per the equations of magnitude discussed in section (1.6) magnitudes are based on a logarithmic scale (base 10). This means for each whole number if we go up on the magnitude

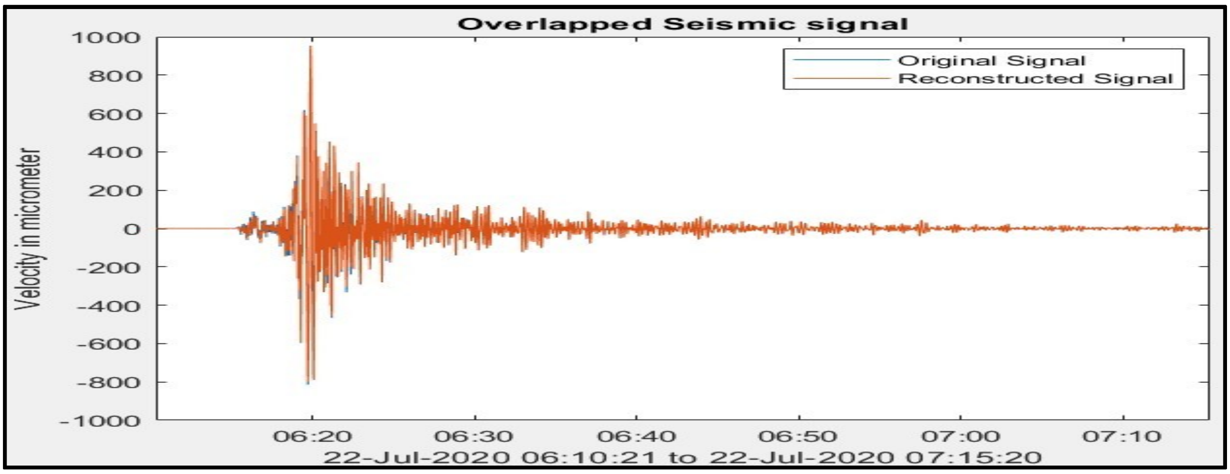

Figure 5.13: Original and reconstructed Overlapped signal

scale, the amplitude of the ground motion recorded by a seismograph goes up ten times. When an earthquake occurs, its magnitude can be given a single numerical value on the Richter magnitude scale. Figure (5.16) shows the reconstructed signal's peak value is within 0.5% of original signal's amplitude (peak value).

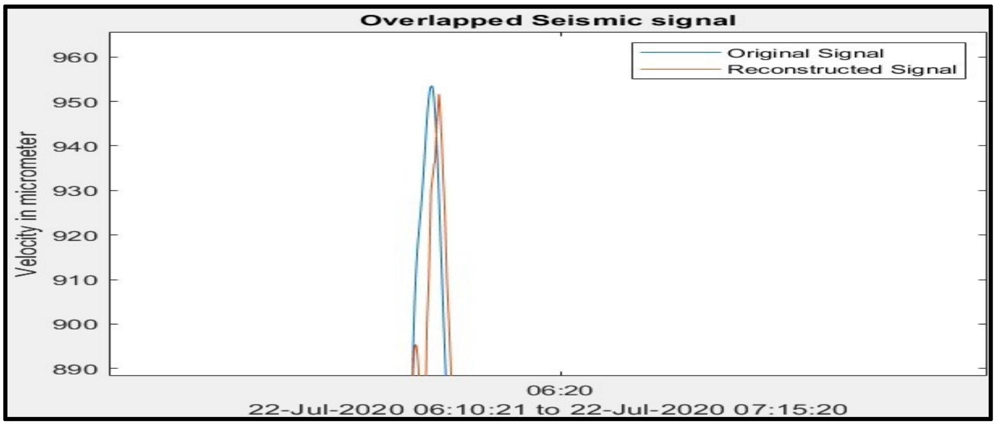

Figure5.14: Zoomed image of figure 45

5.5 Quantitative analysis of onset time difference

A number of seismic signals were feed into the code to verify the compression algorithm. STA and LTA windows of overlapping mode are selected with duration, STA window:1 second , LTA window : 10 seconds

Table 5.2							
Origi- -nal size	**Percentage compression**	**%Energy Retained**	**Percentage correlation**	**STA/LTA Trigger point difference**	**AIC onset trigger difference**	**Sample Rate**	**Time differenc e (seconds)**
312 kB	95.9	97.7	98.8	0	2	20	0.1
624 kB	88.7	99.9	88.7	13	14	40	0.35
624 kB	96.7	96.2	98.1	6	12	40	0.3
312 kB	95.9	86.9	73.2	31	34	40	0.85
624 kB	89.1	98.5	91.5	13	11	40	0.275
624 kB	89.6	99.7	74.9	13	12	20	0.6
624 kB	88.4	98	71.9	12	12	40	0.3
312 kB	88.9	98.6	70.7	14	13	20	0.6
624 kB	95.8	98.9	99.4	6	6	40	0.15
312 kB	96.5	99.3	99.6	2	8	20	0.2

Table 5.2: Time showing onset time difference between Original and reconstructed Signal

Table 5.2 shows original seismic file size, the percentage compression obtained, energy retained, percentage correlation, STA/ LTA difference of samples, AIC difference in samples, the sample rate of the file and onset difference in seconds between original and reconstructed. It is observed that cross correlation between original and reconstructed signal is more than 98%, PER is more than 97% and have less difference in onset time (0.1 – 0.3 seconds) and the signals having low value of cross correlation and PER are having more time difference (up to 0.8 seconds). On an average we get 0.4 seconds as the difference in onset time between original and reconstructed signal.

5.6 Error plot

The onset time difference in seconds of original and reconstructed wave is plotted in histogram. It can be clearly seen the average time difference is 0.2 to 0.4 seconds for the signals having high PER and high cross correlation value, which is satisfactory as explained in section (1.5).

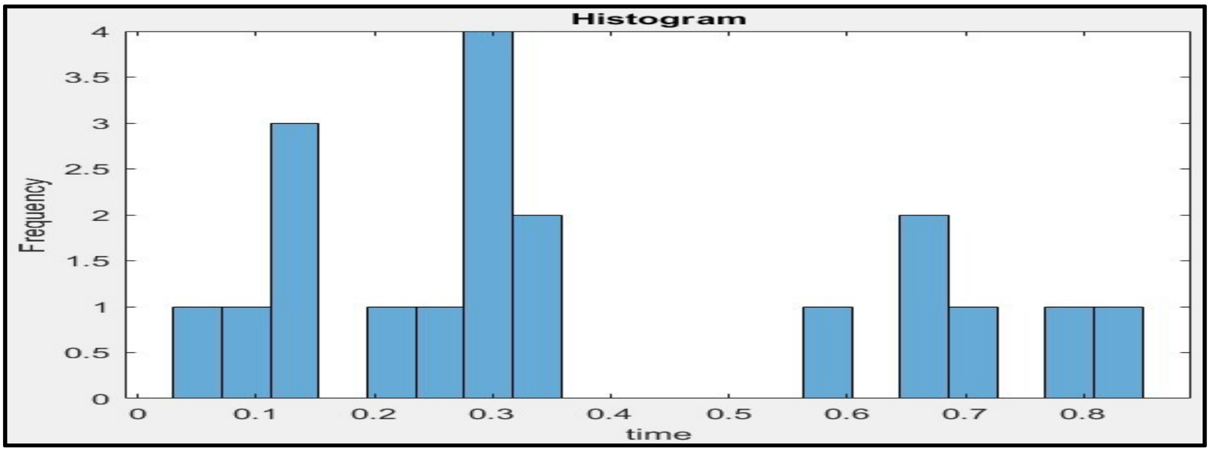

Figure 5.15: Histogram showing onset time difference of original and reconstructed signal

This chapter presented results of the project, tabulated them for easy comparison and understanding. The validation of onset time reveals that the other parameters of earthquake like origin time and location of focus will also be same for original and reconstructed signal. The frequency filtration of signal, STA, LTA window duration, threshold are important parameters for correct onset detection time hence, a noise survey of the seismic station location should be carried out to conclude proper values of these parameters.

6 Conclusion

6.1 Conclusion Remarks

- The goal of this project was to bring out the suitable algorithm for seismic data compression based on multi- scale resolution.
- Wide variety of seismic signals obtained from IRIS website which includes earthquake events, chemical explosions etc. are analyzed in the software developed for compression and it gives satisfactory performance.
- Choosing or designing the right wavelet is crucial for a successful wavelet transform application but choosing the right wavelet for a specific application has been an unanswered question. The MSE and PER proved to be the good criteria for selecting mother wavelet and number of levels also. As per observations (shown in results) the levels varies from 4 to 7 in general depending on the magnitude/ SNR of the signal, hence 6 was selected as the optimum level.
- This compression algorithm can be extended to other formats of seismic data files.
- Variation in magnitude comes out to be negligible and onset time difference (average of 0.4 seconds) between original and reconstructed file is acceptable for seismic applications.
- On comparing with commercial software (ZIP) we get very high compression percentage (90%) with less loss in quality.

6.2 Future Scope

Future work can include implementation of this algorithm for online seismic data compression. As the earthquake event can occur at any time so the instruments needs to be in power on state throughout the day, compiling huge amount of data. The online implementation will compress the seismic file at station itself before transmission thus saving bandwidth and ensuring timely reporting of events.

Further, during wavelet decomposition after every filtering stage, a downsampler removes half of the computed coefficients, since they are redundant. Thus computational efficiency is wasted by unnecessarily calculating these coefficients. A better scheme would be possible if the down sampling stage is some how introduced before the filtering operations. This leds to more efficient way of implementing DWT through lifting scheme. The lifting scheme is a technique for both designing wavelets and performing discrete wavelet transform. Actually it is useful to merge these steps and design the wavelet filters while performing the wavelet transform. This is then called the 2nd generation wavelet transform.

7 References

1. M. Sifuzzaman, M.R. Islam and M.Z. Ali “Application of Wavelet Transform and its Advantages Compared to Fourier Transform”, Journal of Physical Sciences, Vol. 13, *121- 134,*2009
2. Carlos Fajardo, Oscar M. Reyes and Ana Ramirez “Seismic Data Compression Using 2D Lifting Wavelet Algorithms” ,vol. 11, no. 21, pp. 221–238,2015
3. Pankaj Hedaoo and Swati S Godbole, “Wavelet thresholding approach for denoising”, International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.4, July 2011
4. Guomin Luo and Daming Zhang, “Advances in Wavelet Theory and Their Applications in Engineering, Physics and Technology (Wavelet Denoising ,chapter 4) “ , Nanyang Technological University Singapore, 4 April2012
5. Paul L. Donoho, Raymond A. Ergas, Chevron Petroleum Technology Co. and John D. Villasenor, “High-performance seismic trace compression”, University of California, Los Angeles
6. Mohammed A. AI-Mooldmeed', “Towards an Efficient Compression Algorithm for Seismic Data”, College of Computer & Information Sciences, KSU P.O.Box ,51178,Riyadh, SaudiArabia
7. Vidosav Stojanovid, Miomir StankoviC and Igor Radovanovid, “Application of Wavelet Analysis to Seismic Signals” 13-1 5. October 1999, NiS,Yugoslavia
8. A.I. Megahed, Member, IEEE, A. Monem Moussa, H. B. Elrefaie and Y.M. Marghany, “Selection of a Suitable Mother Wavelet for Analyzing Power System FaultTransients”,2008
9. PallaAakreeth Reddy1, PulgamLokanadh Reddy 2, Thota Vinay Kumar, “ 5/3 Lifting Scheme Approach for Image Interpolation” , Volume 12, Issue 1, Ver. I (Jan.-Feb. 2017), pp 08-13
10. VidosavStojanovid, MiomirStankoviC and Igor Radovanovid, “ Application of Wavelet Analysis to Seismic Signals” 13-1 5. October1999
11. Maeda N,1985, “A method for reading and checking phase times in autoprocessing system of seismic wave data”, Jishin,vol. 38, pp. 365–379

12. Kurtis Gurley and Ahsan Kareem, “ Applications of Wavelet Transforms in Earthquake, Wind and Ocean Engineering”, University of Notre Dame, Notre Dame, IN,46556
13. Roman Lara-Cueva, Paul Bernal, Deigo Benitez, “Time and Frequency Feature Selection for Seismic Events from Cotopaxi Volcano”, July2015
14. Julie Zelenski with minor edits by Keith Schwarz , “Huffman Encoding and Data Compression”, May 23,2012
15. Ian n. witten, radford m. neal, and john g. Cleary, “Arithmetic encoding for data compression”, une 1987 Volume 30 Number6
16. Website :Geophysics.eas.gatech.edu/GTEQ/Scream4.4/GCF_Specification.htm
17. Prof. V.MGadre, “Advanced Digital Processing-Multirate and Wavelets-a video course”,www.nptel.ac.in
18. Robipolikar, “The Wavelet tutorial” Second edition,Part1
19. AmadejTrnkoczy (formerly Kinemetrics SA), “Introduction to STA/LTA: Understanding and parameter setting of STA/LTA trigger algorithm”, September1999
20. Pankaj Hedaoo1 and Swati S Godbole,” Wavelet thresholding approach for Wavelet denoising ” , International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.4, July 2011
21. Praveen Kumar, EfiFoufoula-Georgiou, ”Wavelet Analysis for geophysical applications” , Reviews of Geophysics, 35, 4 / November 1997 pages385-412
22. Hui CAO And Ming LAI, “the appliance of wavelet transform in the analysis of seismic structure dynamic reliability”, 12WCEE,2000
23. Kanwaldip Singh Anant and Farid U.Dowla, “Wavelet transform methods for phase identification in 3 component Seismogram”, Bulletin of Seismological society of America Vol 87, No. 6, pp. 1598-1612, December.
24. David L. Donoho, Iain M.Johnston, “Adapting to unknown smoothness via wavelet shrinkage” Department of statistics, Stanford university, July 20, 1994
25. Vanderkulk, W.F.Rosen and Lorenz, “Large aperture seismic array signal processing study,” in IBM Final Report, ARPA Contract Number SD-296, 1965.
26. Houliston D.J, Waugh. G and Laughlin.J, “Automatic Real-time event detection for Seismic Networks,” in computers & Geosciences, Vol.10, pp.413-436, 1984.
27. Sleeman.R and T.van Eck, “Robust automatic P-phase picking: an on-line implementation in the analysis of broadband seismogram recordings” in Phys. Earth

Planet. Interiors 113, pp.265-275, 1999.

28. H.kobayashi, L.R.Bahl, "Image data compression by predictive coding 1: prediction algorithm", IBM Journal of research and development, Vol.18, pp 164-67, March 1974.
29. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F, editors. Second international symposium on information theory; 1973. p. 267–81.