Study on Analysis & Synthesis of Bangla Vowel Using Wavelet Transform

A Thesis Submitted to the Department of Electronics and Telecommunication Engineering in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Electronics and Telecommunication Engineering.

By

Ahsan Uddin Khan
ID# 063-19 - 483
Md. Mehedi Hassan
ID# 062 -19 - 463
Md. Jewel Rahman
ID# 063 -19- 529

SUPERVISED BY

MS.SHAHINA HAQUE

SENIOR LECTURER

DEPARTMENT OF ELECTRONICS & TELECOMMUNICATION ENGINEERING

DAFFODIL INTERNATIONAL UNIVERSITY

DEPARTMENT OF ELECTRONICS & TELECOMMUNICATION ENGINEERING

DAFFODIL INTERNATIONAL UNIVERSITY

MARCH 2011

DEPARTMENT OF ELECTRONICS AND TELECOMMUNICATION ENGINEERING

DAFFODIL INTERNATIONAL UNIVERSITY

MARCH 2011

APPROVAL

This thesis entitled "Study on Analysis & Synthesis of Bangla Vowel Using Wavelet Transform" by Ahsan Uddin Khan, Md. Mehedi Hassan, Md. Jewel Rahman has been submitted to the Department of Electronics and Telecommunication Engineering of Daffodil International University in partial fulfillment of the requirements for the Degree of Bachelor of Science in Electronics and Telecommunication Engineering. This thesis has been accepted as satisfactory by the following Honorable member of the Board of Examiners after its presentation that was held on March 03, 2011.

Board of Examiners

1._____

Dr. Md. Golam Mowla Choudhury — Chairman

Professor and Head

Department of Electronics & Telecommunication Engineering

Faculty of Science & Information Technology

Daffodil International University

2. _____

A.K.M. Fozlul Haque — Internal Examiner

Assistant Professor

Department of Electronics & Telecommunication Engineering

Faculty of Science & Information Technology

DaffodilInternationalUniversity

3._____

Mr. Mirza Golam Rashed — Internal Examiner

Assistant Professor

Department of Electronics & Telecommunication Engineering

Faculty of Science & Information Technology

Daffodil International University

4._____

Dr. Subrata Kumar Aditya

Professor & Chairman — External Examine

Department of Applied Physics, Electronics & communication Engineering

University of Dhaka

DECLARATION

We do declare that the work presented in this thesis is done by us under the supervision of MS.SHAHINA HAQUE. We also declare that neither this report nor any part thereof has been submitted elsewhere for the award of any degree or diploma.

_____

A.K.M. Ahsan Uddin Khan

ID# 063-19 - 483

_____

Md. Mehedi Hassan

ID# 062 -19 - 463

_____

Md. Jewel Rahman

ID# 063 -19- 529

COUNTERSIGNED

_____

Ms. Shahina Haque

Senior Lecturer

ACKNOWLEDGEMENT

At the beginning we remember Allah the merciful, the beneficent since, but for His grace it would not be possible on our part to complete this small volume of thesis. Next we would like to thank our respective parents for their unconditional support and care during our whole educational period.

We express our sincere thanks to our supervisor MS.SHAHINA HAQUE for spending her valuable time by guiding us throughout the thesis work. Without her guidance this work would not have been possible. During the working period we have learnt many more things apart from the thesis work from her which will help us in our future life.

We convey our thanks to the Head of the Department,Dr. Md.Golam Mowla Choudhury and our Honorable Dean Dr. S. M. Mahbub-ul-Haque Majumder for their support. We would also like to take the opportunity to show our gratitude to all our teachers.

Lastly we would like to thank all our classmates and friends for their encouragement and also for making our educational period an unforgettable memory.

ABSTRACT

This project deals with the study of Bangla vowel analysis and synthesis which is the basis of Bangla speech processing. The main task is to acquire the Bangla vowel samples/i/, /e/, /æ/, /a/, /ɔ/, /o/, /u/. then study the process of analysis and synthesis using Wavelet Transform (WT). This thesis will explore theory of WT, how to apply WT for Bangla vowel analysis and synthesis, measure the performance of WT for reconstructing the vowels. The performance of WT for synthesizing vowel was measured by calculating Normalized Root Mean Square (NRMSE) between the original and synthesized signal. It is observed from our study that WT with Daubechies 4 (db4) wavelet at decomposition level 5 reproduces the signal with a very small NRMSE in the order of 10^{-11}. Therefore we may say that WT is an appropriate method for Bangla vowel analysis and synthesis.

Contents

Acoustic phonetics classification of speech signal (11-21)

3

Theory and Application of Wavelet Transform To Bangla Speech (22-33)

4

List of Figures

1

INTRODUCTION

Man's primary method of communication is speech. He is unique in his ability to transmit information with his voice. Of the myriad of life sharing our world, only man has developed the vocal means for coding and conveying information beyond a rudimentary stage. At the acoustic level, speech signals consist of rapid and significantly erratic fluctuations in air pressure. These sound pressures are generated and radiated by the vocal apparatus. Speech sounds radiated into the air are detected by the ear and apprehended by the brain. The mechanical motions of the middle and inner ear, and the electrical pulses traversing the auditory nerve, may be thought of as still different coding of speech information.

1.1 Overview

Speech processing is the study of speech signals and the processing methods of these signals. It includes speech analysis, synthesis recognition, coding etc. Nowadays Bangla speech is being analyzed by various speech analysis techniques by many researchers and works on these areas are now being in progress. The signals are usually processed in a digital representation, so speech processing can be regarded as a special case of digital signal processing, applied to speech signal.

1.2 History of previous work

Signal processing, a field which has its roots in the 17^{th} and 18^{th} century mathematics, has become an important modern tool in a multitude of diverse fields of science and

technology. Many of the limitations were overcome with the introduction of digital computer in speech analysis in 1950's. Looking back to the history of this field; we see that in 1961, Bell [1] introduced a method of spectral analysis by synthesis for the reduction of the speech spectra. In this, a spectrum is filled by a synthesis spectrum in terms of poles and zeros. Large number of works has been done successfully in English, Japanese and other prominent languages .But in Bangla, not so much work has been done so far. The process for formant analysis and synthesis of Bangla vowels was first reported in 1977. The formant structure of vowels of Bangla with Japanese and American English were discussed successfully. In 1986, some aspects of automatic generation and recognition of speech is discussed. Software was developed for automatic transliteration of English text into phonetic symbols and reported that the system gives 90 percent correct word transcription.

M.G. Ali [2] made a spectrum analysis of three Bangla vowels / A /, / B /, /G /using short time Fourier analysis. S.A. Hossain [3] carried out work on the power spectrum analysis of a good number of Bangla vowels and consonant. M. R. Talukdar [4] also analyzed the formant frequencies and power spectrum of some Bangla vowels and consonants. M. LutforRahman [5] evaluated the first three format frequencies and Bandwidth of all Bangla vowels for different age group. Also some more nice works in this line have been done by M. K. Hamid [6], M. Jamal Uddin [7], P. Khandakar [8], S. Haque extracted the features of all Bangla phonemes for different age and sex groups and showed the how the pitch and formants vary with age for both males and females. She then performed the synthesis of all voiced phonemes [9],comparative study of extracted parameters by LPC and cepstraltechniques are studied in[10].

1.3 Objective

Speech processing is necessary to make the properties of speech signal clear analytically beforehand. In this connection analytical studies on various languages are being in progress around the world. Bangla, being a language of about 250 million people, there are not so many studies on it. So in order to make effective processing and economic transmission of Bangla speech, it is necessary to study it analytically.

Researchers are still exploring many unexplored areas of Bangla language. To get introduced to the field of speech processing and to give a new flavor to the use of conventional FT method, we have used Wavelet Transform method. Wavelet is being used in speech processing in many languages for the last few years. Wavelet Transform which is recently developed method overcomes many deficiencies of Fourier Transform.

As a first phase of study on Bangla speech processing we selected the Bangla vowels in isolated utterance for the purpose of analysis and synthesis. The object is to obtain better accuracy in speech processing using WT.

1.4 Result

Here we used the seven Bangla vowels [/i/(ই),/e/(এ),/æ/(এ্যা),/a/(আ),/ɔ/(অ), /o/(ও), /u/(ঊ)] for speech analysis and synthesis using WT. Using WT at decomposition level 5 with Daubechies 4 (Db4)wavelet for selected phonemes, reconstructed signal have very small error in the order of 10^{-11} and we have observed in previous works using Fourier transform in Bangla phonemes the reconstructed signal has an error of order around 10^{-2}.

2

The Mechanism of Speech Production

2.1 The Physical Mechanism of Voice Production

There are three requirements for the production of voice (a) a force that puts a vibrating mechanism into action (b) the vibrating mechanism itself (c) a resonator that reinforces certain vibrations. In the instrument that produces the human voice, the force is given by the blast sent through the trachea by the lungs, which act as below – the true vocal cords form the vibrating mechanism and the resonator is made up of all the supraglottic cavities, i.e., the upper part of the larynx, the pharynx, the mouth, and the nose. The laryngeal sound is produced by the passage of air under pressure through the narrow glottis. Vibrations of the vocal cords interrupt the passage of air which becomes intermittent. Thus eddies are formed in the supraglottic cavities.

2.1.1 The speech tract as an Acoustic system

Air is pushed from lungs through vocal tract and out of mouth comes speech. For certain *voiced* sound, vocal cords vibrate (open and close). The rate at which the vocal cords vibrate determines the pitch of voice. Women and young children tend to have high pitch (fast vibration) while adult males tend to have low pitch (slow vibration). For certain fricatives and plosive (or unvoiced) sound, vocal cords do not vibrate but remain constantly opened [11]. The shape of vocal tract determines the sound that you make. As we speak, vocal tract changes its shape producing different sound. The

shape of the vocal tract changes relatively slowly (on the scale of 10 m sec to 100 m sec). The amount of air coming from lung determines the loudness of voice.

.

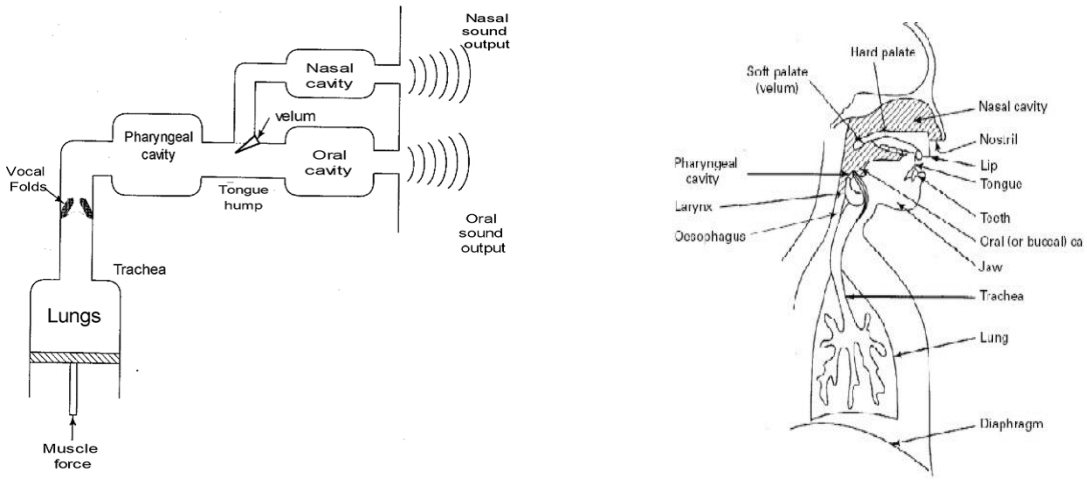

Fig.2.1 Human speech production system along with three components block representation

The natural speech processing technique as perceived by human being can be decomposed in three components- (1) the generation of the excitation source signal. (2) Its modulation by the vocal tract systems. (3) The radiation of speech signal.

The first two requirements are achieved in the normal human mechanism for breathing and the third equisetic is medium, air. Human speech production systems along with its schematic representation and the three components block representation is shown in fig. 2.1 .The lungs and associate respiratory muscles are the vocal supplies for generating the excitation signal.

All the above mechanisms convert the more or less steady pressure (DC power) of lungs into an acoustical signal (AC power) which forms the excitation signal. All sounds generated by the vocal tract apparatus are characterized by properties of (i) the source excitation (ii) the acoustic transmission system. For a given speech sound, the vocal tract represents an acoustic cavity and is characterized by natural frequencies or

formants which correspond to resonant frequencies of the acoustic cavity. Different speech sounds are produced by dynamically changing up the shape of the vocal tract which is affected by the movement of articulators tongue, lips, jaw and velum. These speech sounds are radiated through lips.

In general several major regions as shown in Fig. 2.1 figure prominently in speech production are –

- The relatively long cavity formed at the lower back of the throat in the pharynx region.
- The narrow passage at the place where the tongue is humped
- The variable constriction of the velum and nasal cavity
- The relatively large, forward oral cavity
- The frequency spectrum is shaped by the frequency selectivity of the tube.

In the context of speech production, the resonance frequencies of the vocal tract tube are called formant frequencies which depend upon the shape and dimensions of the vocal tract, each shape is characterized by a set of formant. Different sounds are formed by varying the shape of the vocal tract. Thus; the spectral properties of the speech signal vary with time as vocal tract shape varies.

2.2 Components of the Speech Production System

(a) The Respiratory Organ

Breathing in and out are the common functions of the respiratory system. Production of human speech is not the prime function of organ. It is in fact an incidental activity of this system. When we breathe in, we inhale stream of air from outside into the lungs either through the nose or through the mouth, pharynx and windpipe. When we breathe out a good portion of the air is released out of the lungs through the same

outlet/passage. It is this stream of air released by the lungs that comes to play for human speech.

(b)The Larynx

The larynx is situated at the top of the windpipe. The front part of the larynx protrudes and is popularly known as the "Adam's Apple". Inside the larynx are the vocal cords which have two lips like classic tissues lying opposite to each other across the air passage. It is called the glottis.

The Vocal cordsvoicing or phonation is the sound that we produce by vibrating the vocal cord. The vocal cords can be made longer, shorter, tenser or more relaxed or be more or less strongly pressed together. The pressure of air below the vocal cords can also be varied. The frequency of vibration of the vocal bands is determined by their length, thickness and degree of tension when they begin to vibrate. When the column of pitch of forced through the narrowed opening between the approximated vocal bands they are literally blown apart then come close together. The job of vibrating the air stream belongs to the vocal cords which are folding of ligaments that stretch from front to back across our windpipe which control the air stream flow like a trap door. This trap door is in a set of cartilages called the larynx located about half way up the front of our neck. The larynx or Adam's apple is so vital to the production of speech that it is sometimes called the voice box.

Using vocal cords men can completely or practically block the air streams flow and control the degree and speed of the blocking action by altering the tension of the ligaments. Changing the tension changes the size of the glottis or opening in the cords. We can open the glottis as much as half an inch or close it off completely. Opening and closing the glottis rapidly, turns the steady stream of air into short bursts. We hear these bursts as a buzzing sound. The faster we open and close the vocal cords, the higher the frequency of the buzz that is produced.

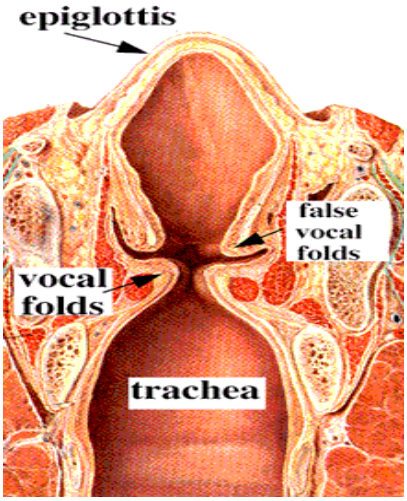

Fig.2.2 The larynx

In normal speech the frequency of the buzz varies from 60 Hz to 350 Hz .The movement of the vocal cords to produce speech is called voicing .But it is not just the frequency of the buzz that distinguishes one voiced sound from another. Speech sounds are more than just a buzz. After it leaves the larynx the air stream passes through the throat, mouth and nose. These organs make up the teeth apparatus known as vocal tract. Thoracic and abdominal musculature serves as a source of energy for speech production. Air is drawn in the lungs by increasing the chest cavity and lowering the diaphragm. The voiced sound is produced by vibrating the vocal cord; the process is also called phonation. As the air stream passes through pharynx and mouth, the natural resonance of these modify the vibration into a variety of sounds. By moving our tongue to various position inside the mouth, opening or closing of teeth, raising or lowering the palate or otherwise changing the shape and site of the vocal tract we change its resonance. This turn cause the air stream to be modified .This modification process is sometimes called modulation.

(c) The Supraglottal Cavities

The supraglottal cavities are located above the glottis. The pharyngeal cavity, the oral cavity and the nasal cavity comprise the supraglottal cavities. A fourth cavity is also

possible. It can be shaped by projection and rounding of lips. This may be called the labial cavity.

The pharyngeal Cavity The cavity formed between the root of the tongue and the back wall of the throat is called the pharyngeal cavity. Air stream passing through the glottis enters this cavity.

The Oral Cavity The entire mouth is called the oral cavity and it has the plate the teeth ridge the teeth, the tongue, the lips and the jaw.

- **Palate** The palate forms the roof of the mouth and separates the oral cavity from the nasal cavity. The velum determines whether a sound will be oral or nasal. If the velum is lowered the air passes through the nasal cavity and the nasal sound in formed. The shape and the volume of the nasal cavity remain always the same. If the velum is raised to close the nasal passage, it helps forming an oral sound. The small tongue at the end of the velum is known as the uvula

- **Teeth ridge or alveolar ridge** the hard palate is divided into two sections-the alveolar ridge and the hard palate. The alveolar ridge is that part of the gums that lies immediately behind the upper front teeth.

- **Teeth (Dental)** Teeth especially the two upper front teeth are used for production of some consonant speech sounds.

- **Tongue** The most important organ in the oral cavity is the tongue. It is extremely flexible and fills almost the entire oral cavity. It is often called the tutor tongue because it has great flexibility for variety of movements for production of human speech.

- **The lips (Bi-labial)** the two lips are also highly mobile and can be used as articulators to produce different kinds of consonant sounds. For production of

sounds lips can have four important positions like the lips may be spread neutral rounded (open) and (close).

- **The jaw** The movement of the jaw specially the lower jaw also influences the position of the lips for production of speech sound

- The Nasal Cavity It is situated at the top of the pharyngeal cavity. The air may pass either through the mouth or through the nose by way of nasal cavity. For production of nasal sound the velum is lowered and the air stream passes through the nasal cavity.

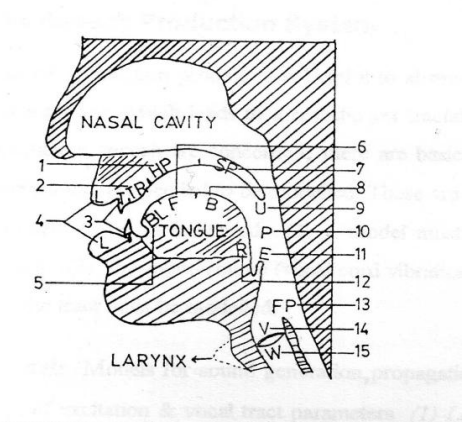

Fig.2.3 Components of vocal tract

1. HP=Hard plate	2. TR=Teeth ridge	3. T=Teeth
4. L=Lips	5. BL=Blade of the tongue	6. F=Front of the tongue
7. SP=Soft plate	8. B=Black of the tongue	9. U=Uvula
10. P=Pharynx	11. E=Epiglottis	12. R=Roots of the tongue
13. FP=Food passage	14. V=Vocal cord	15. W=Windpipe

3

Acoustic Phonetics Classification of Speech Signal

3.1 Introduction

To know how language works one subject must be studied which is sound. The sorts of sound used in speech and how they are produced and detected, this part of linguistics are called phonology or phonemics. Throughout the study of phonology it must be remembered that sounds and differences between them have one and only one function is language to keep utterances apart. Letters of English/Bengali alphabet have got nothing to do with the phonemes/sounds. Letters are written and put together to compose words whereas phonemes/sounds are spoken or used in speech. For example 'P' of English alphabet is not identically similar to /p/ of English phoneme. It must not be confused or wrongly understood. Letters of the English alphabet do not maintain the same identity and quality every time everywhere, whereas each letter of phonetic notation represents a small family of sounds. The quality of the sounds varies to some extent. In ordinary English spelling, it is always not easy to know what sounds the letters stand for.

3.2Phonemes

A language must consist of finite number of distinguishable mutually exclusive sounds, i.e. the language must be constructed of basic linguistic units which have the property that if one replaces another in an utterance in changed. The acoustic manifestation of a basic unit may vary widely. The basic unit for describing how speech conveys linguistic meaning is called phonemes. Its manifold acoustic variations are called allophone [16]. Another way phoneme is a group of sounds that – (a) are felt to be the same by the speaker, (b) cannot be used for distinguishing

between words, (c) differ in ways which are predicable from the context, (d) those stand in contrast with each other in the phonological system.

So the phonemes are code uniquely related to the Articulatory gestures of a given language. The allophone of a given phoneme might be considered representative of the acoustic freedom permissible in specifying a code symbol. This freedom is not only depended upon the phoneme but also upon its position in an utterance. The set of code symbols used in speech and their statistical properties depend upon the language and dialect the communicators. The statistical constraints of a language greatly influence the precision with which a phoneme needs to be articulated. Due to the changing of vocal apparatus in connected speech and the continuous nature of speech wave, human can subjectively segment speech into phonemes.

Phoneticians are able to make written transcriptions of connected speech events and phonetic alphabets have been divided for this purpose. The often accepted standard in modern times is the alphabets of the International Phonetic Association (IPA). This alphabet provides symbols for representing the speech sounds of the most of the major languages of the world. Speech sounds are classified in accordance to their manner and place of articulation. It is very convenient to indicate the gross characteristics of sounds. Speech sounds fall into certain natural division according to the way (process) they are made and the organs with which they are made. Bangla language has thirty six phonemes which are broadly classified into two groups' vowels and consonants. Consonants are, in general, more permanent or stable elements in a language whereas vowels are less and diphthongs least stable or permanent. Consonants, therefore, from the bones and the skeleton of a language and give them their basic structure or shape. Vowels and diphthongs are to speak form the flesh and blood.

3.3 Models for Speech Production

3.3.1 Modeling of the Speech Production System

In studying the speech production process, it is helpful to abstract the important features of the physical system in a manner which leads to a realistic yet tractable

mathematical model. As far as acoustic properties of speech are concerned, there are basically three aspects of the speech production mechanism that are needed to be modeled. These are (1) the geometry of the vocal and nasal tract needs to be parameterized.

(2) A model must be selected to describe wave propagation in the tract. (3)The sound source (vocal cord vibration and turbulent air flow) and their interactions with the tract must be modeled.

3.3.2 Speech Production Models

Models for sound generation, propagation and radiation can be solved with suitable values of excitation & vocal parameters (1) Lossless tube models for speech signal of vocal tract. (2) Digital models for speech signals of vocal tract. (3) Graphical models of vocal tract. (4) Radiation model (5) Excitation model (6) complete model (7) Natural acoustic system (8) Hypothetical equivalent circuit for lossy cylindrical pipe.

3.3.3 Model based upon the acoustic theory (Source-Filter Model)

The important features of the acoustic theory of speech production is that the detailed models for sound generation, propagation, and radiation can in principle be solved with suitable values of the excitation and vocal tract parameters to compute an output speech waveform. Indeed, it can be argued effectively that this may be the best approach to the synthesis of natural sounding synthetic speech [11] Fig 2.4 shows a general block diagram that is representative of numerous models that have been used as the basis for speech processing.

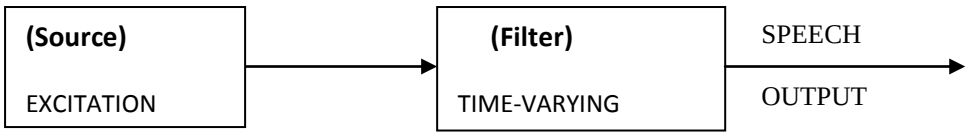

Fig.3.1: Source system model of speech production

These models all have in common that the excitation features are separated from the vocal tract and radiation features. The vocal tract and radiation effects are accounted for by the creates a signal that is either a train of pulses, or randomly

varying(noise).The parameters of the source and system are chosen so that the resulting output has the desired speech-like properties. If this can be done, the model may serve as a useful basis for speech processing [12].

Formant: A peak in the frequency response of an unobstructed vocal tract. One of the simple resonators makes up the complex resonant system of the vocal tract.

A useful analytical model of how speech sounds are produced, which emphasizes the independence of the source of sound in the vocal tract from the filter that shapes that sound. The source-filter model of vowel production states that the frequency content of a vowel may be explained by considering how the spectrum of the sound generated by the larynx is filtered by the vocal tract system. The independence of source and filter explains why vowels of the same timbre can be produced on different pitches, and why vowels of the same pitch can have different timbres. The source filter model also helps to quantify vowels since we can separately measure the contributions of the source and the filter to the final vowel sound. The primary characteristic of the source is its fundamental frequency, while the primary characteristics of the filter can be reduced to the location in frequency of the vocal tract resonances or formants, see figure 3.1.

The source filter model also helps to quantify vowels since we can separately measure the contributions of the source and the filter to the final vowel sound. The primary characteristic of the source is its fundamental frequency, while the primary characteristics of the filter can be reduced to the location in frequency of the vocal tract resonances or formants, see figure 3.1.

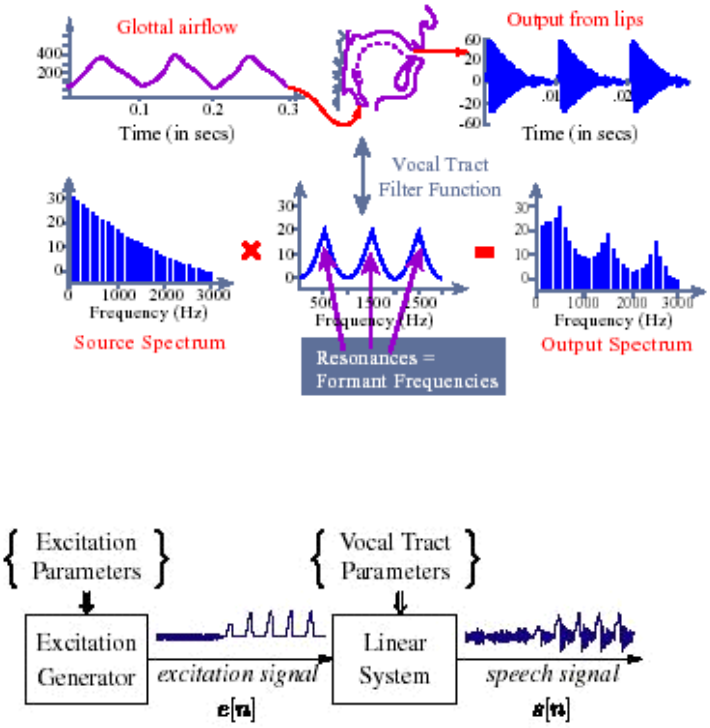

Fig 3.2: Source/System model for speech Signal

The excitation generator on the left simulates the different modes of sound generation in the vocal tract. Samples of a speech signal are assumed to be the output of the time-varying linear system. In general such a model is called a *source/system* model of speech production. The short-time frequency response of the linear system simulates the frequency shaping of the vocal tract system, and since the vocal tract changes shape relatively slowly, it is reasonable to assume that the linear system response does not vary over time intervals on the order of 10 ms or so. Thus, it is common to characterize the discrete time linear system by a system function of the form:

$$H(z) = \frac{\sum_{k=0}^{M} b_k z^{-k}}{1 - \sum_{k=1}^{N} a_k z^{-k}} = \frac{b_0 \prod_{k=1}^{M} (1 - d_k z^{-1})}{\prod_{k=1}^{N} (1 - c_k z^{-1})},$$

Where the filter coefficients $a_k$andb_k (labeled as vocal tract parameters in Figure 3.2) change at a rate on the order of 50~100 times/s. Some of the poles (c_k) of the system function lie close to the unit circle and create resonances to model the formant frequencies. It is sometimes useful to employ zeros (d_k) of the system function to model nasal and fricative

3.4How speech can be modeled as a source signal passing through a filter

The Make-Up of Speech

The components of speech are the words and the voice. Every phrase is a union of these two components - they are the foundations of the spoken language. One or the other does not mean much without its counterpart. Words without voice lack intonation, so they have no meaning. Voice without words is devoid of structure and cannot possibly transfer information. Only the fusion of the two can claim to be such a thing as speech. In biology, the components of speech are produced in different organs. To speak, air is first released over the vocal cords, which expand and contract to give the air column structure. This is the biological concept of words. The words are then passed through the vocal tract where they are shaped, giving them intonation. This shaping of the words is the biological concept of voice. Such a biological process can be easily modeled. So far, we have determined that speech is a collection of words shaped by voice. Here, we present a model of this. In this model, the words are called the source. Since the words are modified by voice, we say the source passes through a filter. This brings us to the source filter model of speech.

Signal Processing Considerations

The source filter model can easily be extended to signal processing. The source is simply a signal x (t). This signal is the input to the filter and is called the excitation signal since it excites the vocal tract. The vocal tract is a filter similar to all filters we have studied so far: it is a linear time-invariant system with impulse response h (t). This is sometimes called the transfer function of speech since it is what transfers the excitation signal to speech - it adds voice to words. Speech is the output y (t) of the

source signal x (t) passed through the filter with impulse response h (t). Thus, the output is given by y (t) = x (t) * h (t). This is depicted below:

Signal Processing Representation of the Source Filter Model

From the equation: An input x (t) to a filter with impulse response h (t) yields the convolution of the two. Since speech is simply a convolution of a source signal x (t) with a filter's input response h (t), we can analyze these signals to determine the characteristics of a speech signal y (t). However, we must first de-convene these signals so that they can be processed individually.

Properties of vowel sounds

We can observe a number of properties of vowel sounds which tell us a great deal about how they must be generated: (i) they have pitch, so they are periodic signals, (ii) different vowels have different timbres, so they must have different harmonic amplitudes in their spectra, (iii) the same vowel can be spoken on different pitches, so the pitch must be set independently from the vowel identity, (iv) the same vowel can be spoken on different voice qualities, so the voice quality must be set independently from the vowel identity, (v) different vowel qualities can be produced on the same pitch, so that vowel quality doesn't affect pitch, (vi) vowel quality seems to depend mostly on tongue position: front-back and open-close, and (vii) vowel quality is also affected by the position of other articulators, the jaw, lips and velum.

Source-filter model

All of these characteristics of vowels can be explained by the source filter model of sound production in the vocal tract. This model of sound production assumes a source of sound and a filter that shapes that sound, organized so that the source and the filter are independent. This independence allows us to measure and quantify the source separately from the filter. For vowel sounds, the source of sound is the regular vibration of the vocal folds in the larynx and the filter is the whole vocal tract tube between the larynx and the lips. For fricative sounds, the source of sound is the turbulence generated by passing air through a constriction, and the filter is the vocal tract tube anterior to the constriction.

Vowel Source

Vibration in the larynx is caused by blowing air between two tensed and approximated membranes: the vocal folds. The periodic buzz produced by the vibrating folds has a large number of harmonics up to 5000Hz or so, although the energy drops off with increasing frequency. Fundamental frequencies for men are typically in the 100-200Hz range, while for women are in the 150-300Hz range.

Vowel Filter

The frequency response of the vocal tract filter for vowels shows a small number of resonant peaks called formants. In a formant model of the vocal tract frequency response, each peak is considered to be a separate simple resonator; thus we tend to think of formants as individual resonances of the vocal tract (even though they are not really independent of one another) see figure 3.2. Studies of formant frequencies for different phonetic vowel qualities show a rough relation between the frequencies of the first two formants (F1, F2) and the position of the vowel on the vowel quadrilateral. This leads to the rule of thumb that F1 is associated with increasing open-ness of vowel articulation, while F2 is related to increasing front-ness of vowel articulation (see figure 3.2).

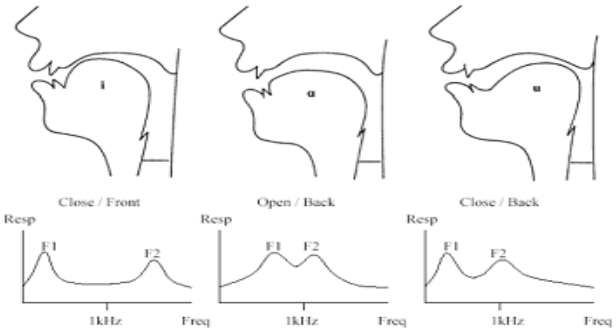

Figure 3.3: Vocal tract shapes for different vowels lead to different frequency Responses Close / Front Open / Back Close / Back

3.5 Fundamental properties of Speech Signal

Three main differences are found that are the fundamental properties of speech signal loudness, pitch and quality.

Variation of intensity of loudness: Loudness is the sensory response to the amplitude of the sound waves. This movement results from volume velocity of the glottal air pulse which is created by the interaction of glottal resistance and the force of the breath. Consequently loudness in controlled primarily by the vibrator but directly with air pressure.

Variation in frequencies of source or pitch

Pitch is the result of change of vibration or frequency changes of the source (glottis). Perceptually pitch is what the auditory system perceives. The 'highness" or "lowness" of a sound depends on the cycles per second of its variations. The number of tones can be emitted by the human voice and their positions in the musical scale vary with age and sex. The frequencies range of newborn is limited to approximately three semitones. As the child grows the frequencies range increases mainly by the addition of higher tones, but also of low tones (Fig: 3.2).

Variation in sound quality or Formants

In the context of speech production the resonance frequencies of the vocal tract tube are called formant frequencies or simply formants. The formants frequencies depend upon the shape and dimensions of the vocal tract; each shape is characterized by a set of formant frequencies. Different formats are formed by varying the shape of the vocal tract. Thus spectral properties of the speech signal vary with times as the vocal tract varies. We can determine the frequencies of the formants of vocal tracts specified as in Fig: 3.1 in terms of – (1) the size of the minimum cross-sectional area $A_{min,}$ (2) the distance /L/ of this minimum area from the glottis, (3) the lip opening.

According to mode of excitation speech sounds can be classified into three distinct classes

- Voiced sounds are produced by forcing air through the glottis with the tension of the vocal cords adjusted so that the vibrate in a relaxation oscillation, thereby producing quasi-periodic pulses of air which excite the vocal tract and

is the excitation signal for producing voiced sound. All vowels are voiced sounds.

- Fricative or unvoiced sounds are generated by forming a constriction at some point in the vocal tract (usually toward the mouth end) and forcing air through the constriction at a high enough velocity to produce turbulence. For unvoiced sound production, the vocal tract is excited by random white noise and the shape of the vocal tract uniquely determines the sound that is produced. So a brief transient excitation occurs.

- Plosive sounds result from making a complete closure (usually toward the front of vocal tract) building up pressure behind the closure and abruptly releasing it. Plosive sounds are /p/, /b/, so for plosive sound generation, the lungs and associated respiratory muscles power is converted into short burst of noisy signal by the sudden release of pressure which is build up by completely closing the vocal tract for short duration.

3.6 Articulatory Phonetics

All the sounds we speak are the result of muscle contracting. The respiratory muscles in the larynx which are the power source for speech production produce many different modifications in the flow of air from the chest to the mouth. After passing through the larynx and vibrating vocal cords the air goes through the vocal tract which ends at the mouth and nostrils. Here the air from the lungs escapes into the atmosphere. We have a large and complex set of muscles that can produce changes in the shape of the vocal tract and in order to learn how the sounds of speech are produced it is necessary to become familiar with the different parts of the vocal tract which is discussed in the chapter II. This different part which is flexible speech organs such as tongue, palate, teeth, lips etc. called articulators and the study of them is called Articulatory phonetics. Articulation is the process of changing the shape of the mouth to control the production of sounds.

Positions of Articulation the articulators which it is convenient to differentiate are the dorsum, the center and the ballad of tongue, the tip of the tongue, and the lower lip. The points of articulation are: the velum (sometimes requiring subdivision into front

and back), the dome, the alveolar ridge, the backs of the upper teeth approximately at the edge of the gum, the cutting edge of the gum, the cutting edges of the upper teeth, and the upper lip. Occasionally the last two function together. A combination of articulator and points of articulation constitutes a position of articulation. Positions of articulation are labeled by a compound term, the first part designating the articulator, the second part, the point of articulation. Thus we have dorso-velar, front and back dorso-velar, centro-doma, lamino-domal, lamino-alveolar, apice-domal, apico-alveolar, apico-dental, apico-interdental, apico-labial, labio-dental and labio-labial for the last of these the term bilabial is usually substituted.

3.6.1 Acoustics Phonetics of Bangla Vowels

Vowel phonemes are voiced sounds which are produced by forcing air through the glottis with the tension of the vocal cords adjusted so that they vibrate in a relaxation oscillation thereby producing quasi-periodic pulses of air which excite the vocal tract. During normal articulation, the tract is maintained in a relatively stable configuration during most of the sound and negligible nasal coupling occurs, so radiation only from the mouth takes place. If the nasal tract is effectively coupled to the vocal tract during the production of a vowel, the vowel becomes nasalized. Each time the vocal cords opens and close there is a pulse of air from the lungs which act like sharp tubs on the air in the vocal tract which is accordingly set into vibration in a way that is determined by its shape and size. The air in the vocal tract vibrates at three or four frequencies irrespective of the fundamental frequency which are frequencies of that particular vocal tract shape and rate of vibration. There are eleven vowel alphabets for writing purpose whereas for speech we have seven vowel phonemes in Bangla such as / A /, During the production of vowel the tongue position remains stationary throughout the time it takes to say the vowel i.e. the vowel remains exactly the same both at the end and the beginning, so it remains pure and chaste as tongue position does not undergo any change, so the quality of purity remains. Bangla vowels sounds take different abridge forms when used in combination with Bangla consonant sounds. E.g. Bangla vowel phonemes are naturally short.

4

Theory and Application of Wavelet Transform to Bangla Speech

4.1 Speech Analysis and Synthesis

Analysis: Analysis is the process of breaking a complex topic or substance into smaller parts to gain a better understanding of it. The technique has been applied in the study of mathematics and logic since before Aristotle (384–322 B.C.), though analysisas a formal concept is a relatively recent development. [15]

Synthesis: In general, the noun synthesis refers to a combination of two or more entities those together forms something new[14].Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer, and can be implemented in software or hardware. A text-to-speech system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech [16,17].

Voice analysis is the study of speech sounds for purposes other than linguistic content, such as in speech recognition. Such studies include mostly medical analysis of the voice i.e. phonetics, but also speaker identification. More controversially, some believe that the truthfulness or emotional state of speakers can be determined using Voice Stress Analysis or Layered Voice Analysis. [18]

4.2 Digital Signal Processing

This section will give a brief introduction to, and a formal definition of, the basic concepts and methods of the theory of digital signal processing used in this project. They form the basis for the methods presented in the subsequent chapters. Necessarily, this introduction will be very brief and restricted. In particular,we will consider the theory of digital signal processing apart from the more general theory of analog signalprocessing, and will pass over the various problems and mathematical prerequisites one has to take care of(stability, convergence) [20,21,22] .

4.3 Signal Analysis Techniques

There are various kinds of Speech analysis techniques. They are described shortly below.

- Cepstral[23]
- Linear Predictive Coding(LPC)[23]

Cepstral Analysis Techniques: The observed speech signal is the result of convolution of excitation source signal and the linear system impulse response in the time domain or a product of the excitation source and the system spectra in frequency domain. The Cepstral analysis technique is a method of separating the two components by transforming the product of two components into their sum and exploiting the fact that the two convolved signals have quite different spectral characteristics [14]. These forms are useful in the form of Fourier transform platform.

LPC analysis technique: LPC starts with the assumption that a speech signal is produced by a buzzer at the end of a tube (voiced sounds), with occasional added hissing and popping sounds (sibilants and plosive sounds). Although apparently crude, this model is actually a close approximation to the reality of speech production. The glottis (the space between the vocal folds) produces the buzz, which is characterized by its intensity (loudness) and frequency (pitch). Like Cepstrum these form is useful in the form of Fourier transform platform.

4.3.1 Disadvantages of these techniques based on Fourier Transform and How to overcome

Bangla Vowelanalysis using conventional methods is primarily based on inspection of Fourier transform (FT) which has resolution problem. In order to produce better accuracy, we attempted Wavelet Transform for Bangla vowel analysis and used Daubechies wavelet for analyzing and synthesizing the seven Bangla vowels. The performance of thesynthesized speech signal is measured by calculation. The normalized root mean square error (NRMSE) of the reconstructed vowel waveform is calculated for all the seven vowel of Bangla and is found to be in the order of 10^{-11}. It is observed from our study that Db4 wavelet at decomposition level 5 reproduces the signal with a very small NRMSE.

4.4 Wavelet Transform

The main purpose of the procedure that has been called wavelet transform is to decompose arbitrary signals into localized contributions that can be labeled by a 'scale parameter'.

Now that we know some situations when wavelet analysis is useful, it is worthwhile asking "What is wavelet analysis?" and even more fundamentally,
"What is a wavelet?" A wavelet is a waveform of effectively limited duration that has an average value of zero.
In 1982 Jean Morlet a French geophysicist, introduced the concept of a `wavelet'. The wavelet means small wave and the study of wavelet transform is a new tool for seismic signal analysis.Immediately, Alex Grossmann theoretical physicists studied inverse formula for the wavelet transform. The joint collaboration of Morletand Grossmann [29] yielded a detailed mathematical study of the continuous wavelet transforms and their various applications, of course without the realization that similar results had already been obtained in 1950's by Calderon, little wood,Paley and Franklin. However, the rediscovery of the old concepts provided a new method for decomposing a function or a signal. Some other works on WT are found in [30,31].

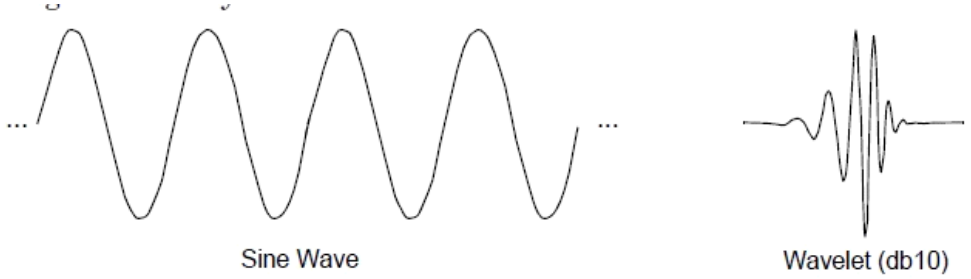

Figure 4.1: Analysis of a simple sine wave in Wavelet transform.

If we compare wavelets with sine waves, which are the basis of Fourier analysis. Sinusoids do not have limited duration — they extend from minus to plus infinity. And where sinusoids are smooth and predictable, wavelets tend to be irregular and asymmetric.

Fourier analysis consists of breaking up a signal into sine waves of various frequencies. Similarly, wavelet analysis is the breaking up of a signal into shifted and scaled versions of the original (or *mother*) wavelet.

From the pictures of wavelets and sine waves, it can be observed that signals with sharp changes might be better analyzed with an irregular wavelet than with a smooth sinusoid, just as some foods are better handled with a fork than a spoon.

It also makes sense that local features can be described better with wavelets that have local extent.

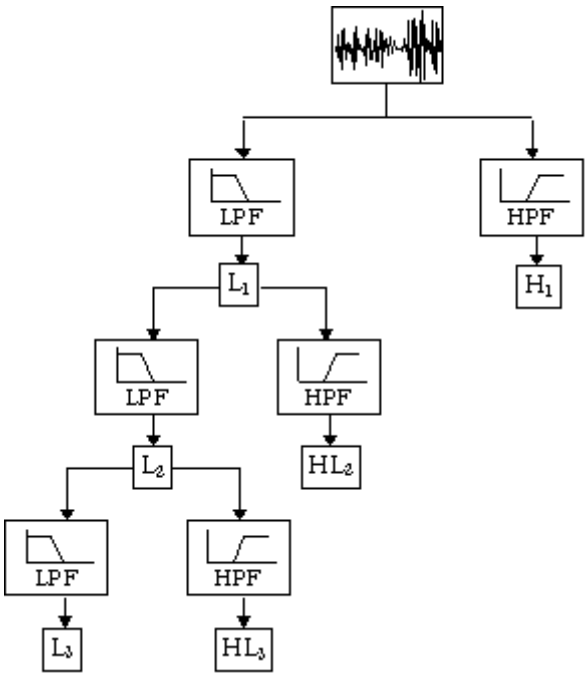

Figure 4.2: Filter bank representation of the DWT dilations.

The fundamental idea behind wavelets is to analyzes according to scale. The wavelet analysis procedure is to adopt a wavelet prototype function called an analyzing wavelet or mother wavelet. Any signal can then be represented by translated and scaled versions of the mother wavelet. Wavelet analysis is capable of revealing aspects of date that other signal analysis techniques such as Fourier analysis miss aspects like trends, breakdown points, discontinuities in higher derivatives, and self-similarity. Furthermore, because it affords a different view of data than those presented by traditional techniques, it can compress or de-noise a signal without appreciable degradation.

Wavelet transform can be viewed as transforming the signal from the time domain to the wavelet domain. This new domain contains more complicated basis function called wavelets, mother wavelets or analyzing wavelets. A wavelet prototype function at a scale s and a spatial displacement u is defined by

$$\Psi s, \qquad u(x) = \sqrt{s}\Psi\left[\frac{(x-u)}{s}\right] \qquad \text{.....(2.1)}$$

The continuous wavelet transforms (CWT) is given mathematically by Eq.2.2

C(s,u)

$$= \int_{-\infty}^{\infty} f(t)\sqrt{s}\Psi\left[\frac{(x-u)}{s}\right] dt \qquad \text{.....(2.2)}$$

Which is the sum over all time of the signal multiplied by scaled and shifted versions of the wavelet function Ψ . The results of the CWT are many wavelet coefficients C, which are a function of scale and position. Multiplying each coefficient by the appropriately scaled and shifted wavelet yields the constituent wavelets of the original signal.

The wavelet coefficients are calculated for each wavelet segment, giving a time-scale function relating the wavelets correlation to the signal. This process of translation and dilation of the mother wavelet is depicted in Figure 4.3[24].

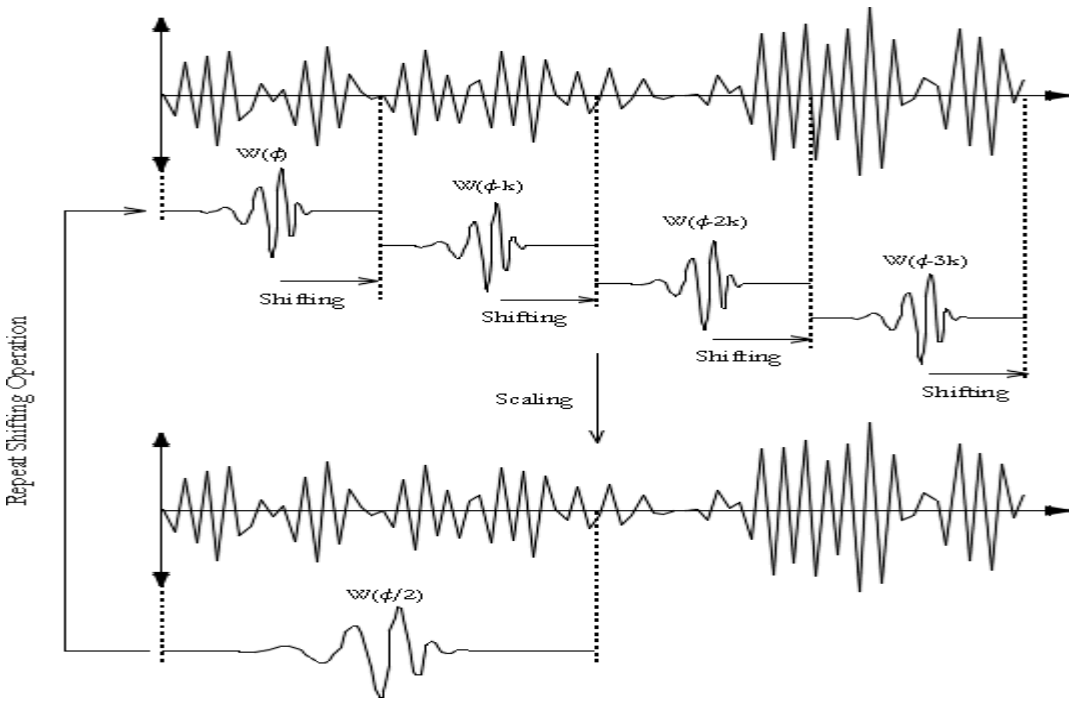

Figure 4.3: The scaling and shifting process of the DWT.

4.5Advantages of WT

- One of the main advantages of wavelets is that they offer a simultaneous localization in time and frequency domain.
- The second main advantage of wavelets is that, using fast wavelet transform, it is computationally very fast.
- Wavelets have the great advantage of being able to separate the fine details in a signal. Very small wavelets can be used to isolate very fine details in a signal, while very large wavelets can identify coarse details.
- A wavelet transform can be used to decompose a signal into component wavelets.
- In wavelet theory, it is often possible to obtain a good approximation of the given function f by using only a few coefficients which is the great achievement in compare to Fourier transform.
- Wavelet theory is capable of revealing aspects of data that other signal analysis techniques miss the aspects like trends, breakdown points, and discontinuities in higher derivatives and self-similarity.
- It can often compress or de-noise a signal without appreciable degradation.

4.6 Vowel signal processing using Wavelet Transform

Traditional techniques for speech signal analysis use Fourier methods for signal processing. Fourier analysis, however, only details the spectral content of a signal in the frequency domain. The time domain information for a particular event is lost during Fourier transformations because preservation of time instances is not considered. This condition can be overlooked if the signal is stationary. However, for non-stationary signals, like speech, time *and* frequency domain information is necessary to avoid any loss of significant information in the signal. Wavelet analysis provides an alternative method to Fourier analysis for signal processing. Wavelets apply the concept of multiresolutionanalysis (i.e., time and frequency scale representations) to produce precise decompositions of signals for accurate signal representation. They can reveal detailed characteristics, like small discontinuities, self-similarities, and even higher order derivatives that may be hidden by the conventional Fourier analysis.

4.7 Types of wavelet transform

Haar:

Any discussion of wavelets begins with Haar wavelet, the first and simplest. Haar wavelet is discontinuous, and resembles a step function. It represents the same wavelet as Daubechiesdb1.

Daubechies:

Ingrid Daubechies, one of the brightest stars in the world of wavelet research, invented what are called compactly supported orthonormal wavelets — thus making discrete wavelet analysis practicable. The names of the Daubechies family wavelets are written dbN, where N is the order, and dbthe "surname" of the wavelet. The db1 wavelet, as mentioned above, is the same as Haarwavelet. Here is the wavelet functions psi of the next nine members of the family:

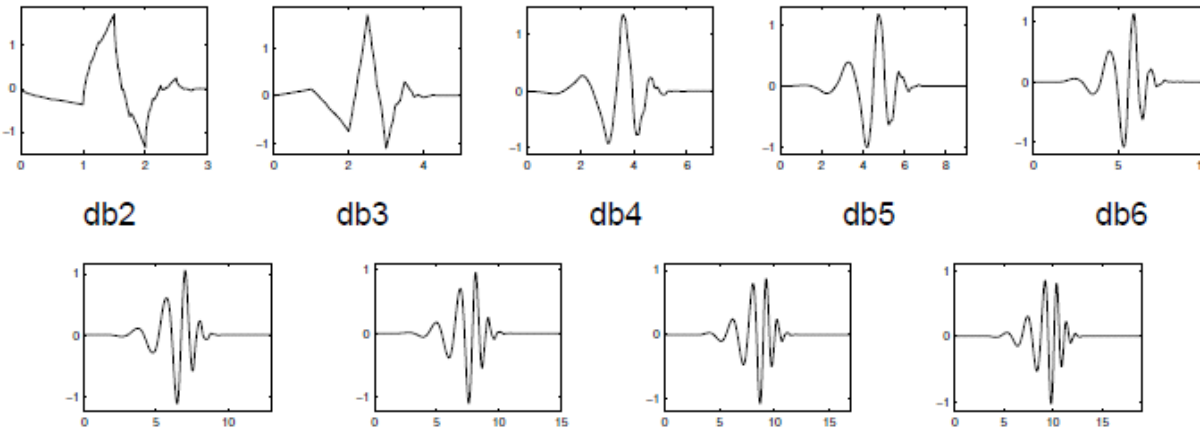

Figure 4.4: Sample diagram of Daubechies Wavelet Transform.

Biorthogonal:

This family of wavelets exhibits the property of linear phase, which is needed for signal and image reconstruction. By using two wavelets, one for decomposition (on the left side) and the other for reconstruction (on the right side) instead of the same single one, interesting properties are derived.

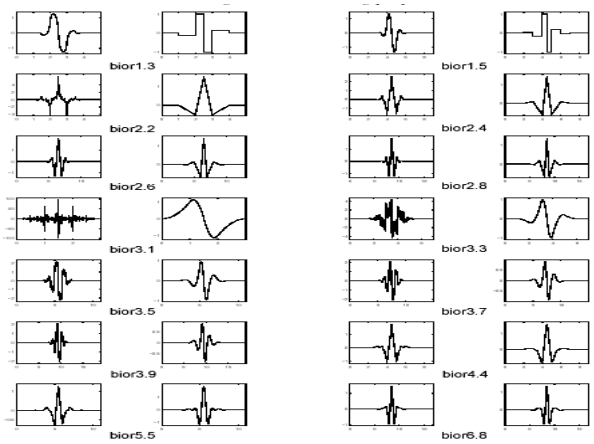

Figure 4.5: Sample figure of Biorthogonal Wavelet Transform.

Coiflets :

Built by I. Daubechies at the request of R. Coifman. The wavelet function has 2*N* moments equal to 0 and the scaling function has 2*N*-1 moments equal to 0.The two functions have a support of length 6*N*-1. You can obtain a survey of the main properties of this family by typing wave info ('coif') from the MATLAB command line.

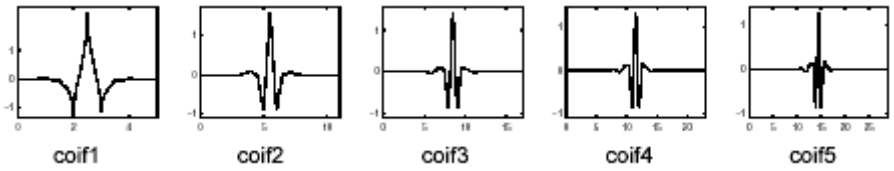

Figure 4.6: Sample figure of Coiflets Wavelet Transform.

4.8Application of Wavelet Transform

4.8.1 Detecting Discontinuities and Breakdown Points I

The purpose of this example is to show how analysis by wavelets can detect the exact instant when a signal changes. The discontinuous signal consists of a slow sine wave abruptly followed by a medium sine wave.

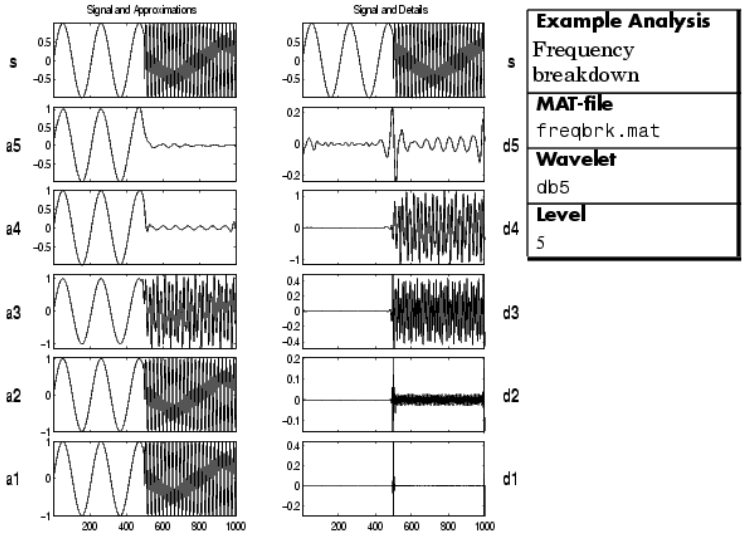

Figure 4.7: Detecting Discontinuities and Breakdown Points (I).

The first- and second-level details (D1 and D2) show the discontinuity most clearly, because the rupture contains the high-frequency part. Note that if we were only interested in identifying the discontinuity, db1 would be a more useful wavelet to use for the analysis than db5.

The discontinuity is localized very precisely: only a small domain around time = 500 contain any large first- or second-level details.

Here is a noteworthy example of an important advantage of wavelet analysis over Fourier. If the same signal had been analyzed by the Fourier transform, we would not have been able to detect the instant when the signal's frequency changed, whereas it is clearly observable here.

Details D3 and D4 contain the medium sine wave. The slow sine is clearly isolated in approximation A5, from which the higher-frequency information has been filtered.[19]

4.8.2 Detecting Discontinuities and Breakdown Points II

The purpose of this example is to show how analysis by wavelets can detect a discontinuity in one of a signal's derivatives. The signal, while apparently a single smooth curve, is actually composed of two separate exponentials that are connected at time = 500. The discontinuity occurs only in the second derivative, at time = 500.

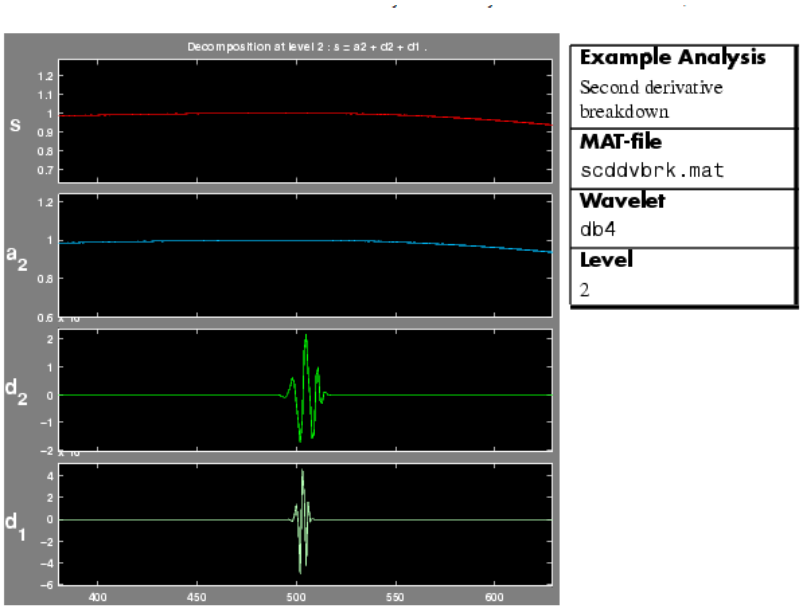

Figure 4.8: Detecting Discontinuities and Breakdown Points (II).

We have zoomed in on the middle part of the signal to show more clearly what happens around time = 500. The details are high only in the middle of the signal and are negligible elsewhere. This suggests the presence of high-frequency information -- a sudden change or discontinuity -- around time = 500[19].

4.8.3 Detecting Long-Term Evolution

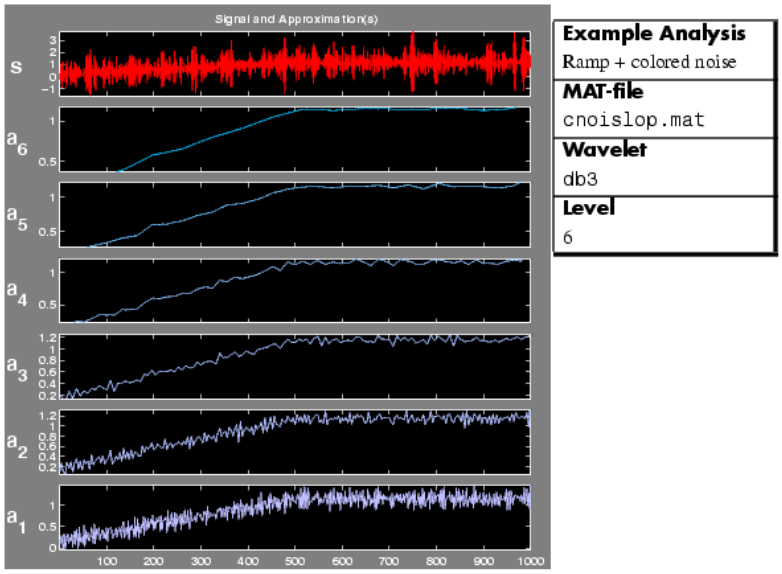

Figure4.9: Detecting Long-Term Evolution.

The purpose of this example is to show how analysis by wavelets can detect the overall trend of a signal. The signal in this case is a ramp obscured by "colored" (limited-spectrum) noise. (We have zoomed in along the x-axis to avoid showing edge effects.)There is so much noise in the original signal, s, that its overall shape is not apparent upon visual inspection. In this level-6 analysis, we note that the trend becomes more and clearer with each approximation, A1 to A6.

The trend represents the slowest part of the signal. In wavelet analysis terms, this corresponds to the greatest scale value. As the scale increases, the resolution decreases, producing a better estimate of the unknown trend.

Another way to think of this is in terms of frequency. Successive approximations possess progressively less high-frequency information. With the higher frequencies removed, what's left is the overall trend of the signal.[19]

More applicationsof WT

- Detecting Self-Similarity
- Identifying Pure Frequencies
- De-Noising Signals
- Compressing Image

5

Apply Wavelet Transform to Bangla Speech

The aim of this section is to acquire the speech samples and process them using WT. The experimental part consists of recording each of the isolate Bangla oral[/i/, /e/, /æ/, /a/, /ɔ/, /o/, /u/]Vowels at a normal speaking rate three times in a quiet room by three male native Bangla speakers (age around 27 years) is a DAT tape at a sampling rate of 48 kHz and 16 bit value. The best one of these three speakers' voice and best speech sample was chosen for our work these digitized speech sound are then down sampled to 10 kHz mono and then normalized for the purpose of analysis.

5.1 Recording process of voice signal

There are various techniques and software we can use to record speech.

We can devide them depending on operating system.

Windows

Windows sound recorder tool, Jet audio, Cybercorder 2000,Vox Recorder, DART, Scancorder, WEaudio recorder, Autorec.

Linux

- Alliance Arcane
- Open Unified Recording

We have used Windows sound recorder (we get it from start>All programs> Accessories> Entertainment> sound recorder)

We record each Bangla vowels and analysis them. But there occur a lot of noises.

Then we collect previously recorded data which are of good quality (Digital Audio Sound) which is recorded in studio.

5.2Work we have done

- In the speech analysis scenario, first we need to record voice speech.
- We record it in a noise free quiet place at a frequency rate of 48 KHz in a stereo sound system.
- Down sample it to 10 KHz in mono system.
- Filtered it.
- After getting the data we windowing it in 25.6 ms length per window.

MATLAB toolbox

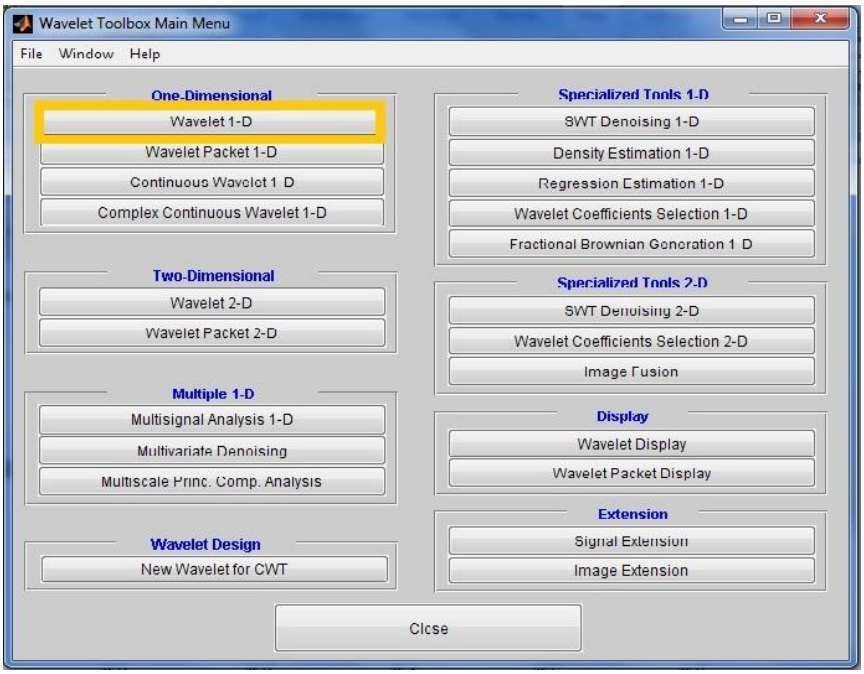

Figure 5.1: MATLAB toolbox.

>>Then we choose wavelet 1D

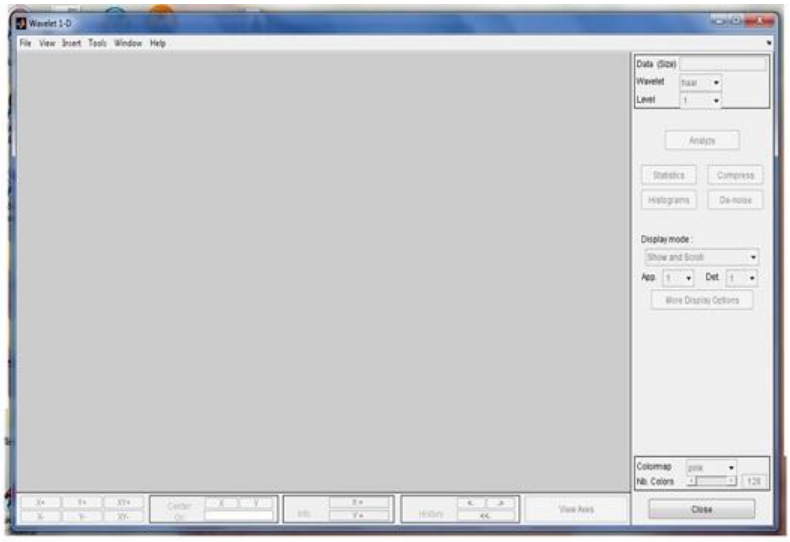

Figure 5.2: Wavelet 1D platform.

5.3Processing the data using WT

We have collected these data:

/ɔ/(অ),/a/(আ),/æ/(এা),/e/(এ),/i/(ই),/o/(ও), /u/(ঊ)

Then we process the data using db4 at decomposition level 5

Figure (5.3 a to 5.9 b) shows the processing

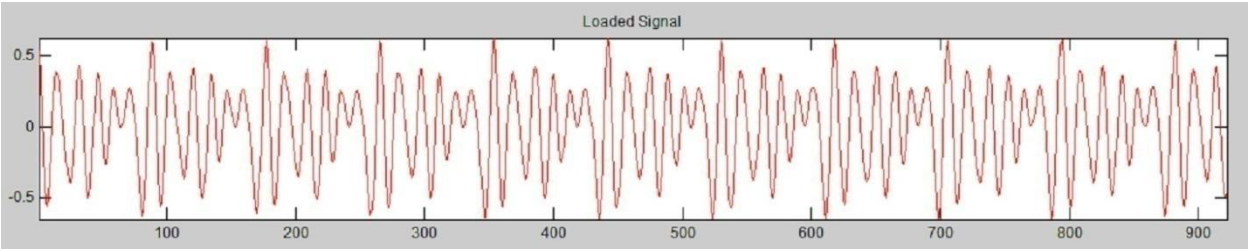

Figure 5.3.a: Original signal for (/ɔ/(অ)).

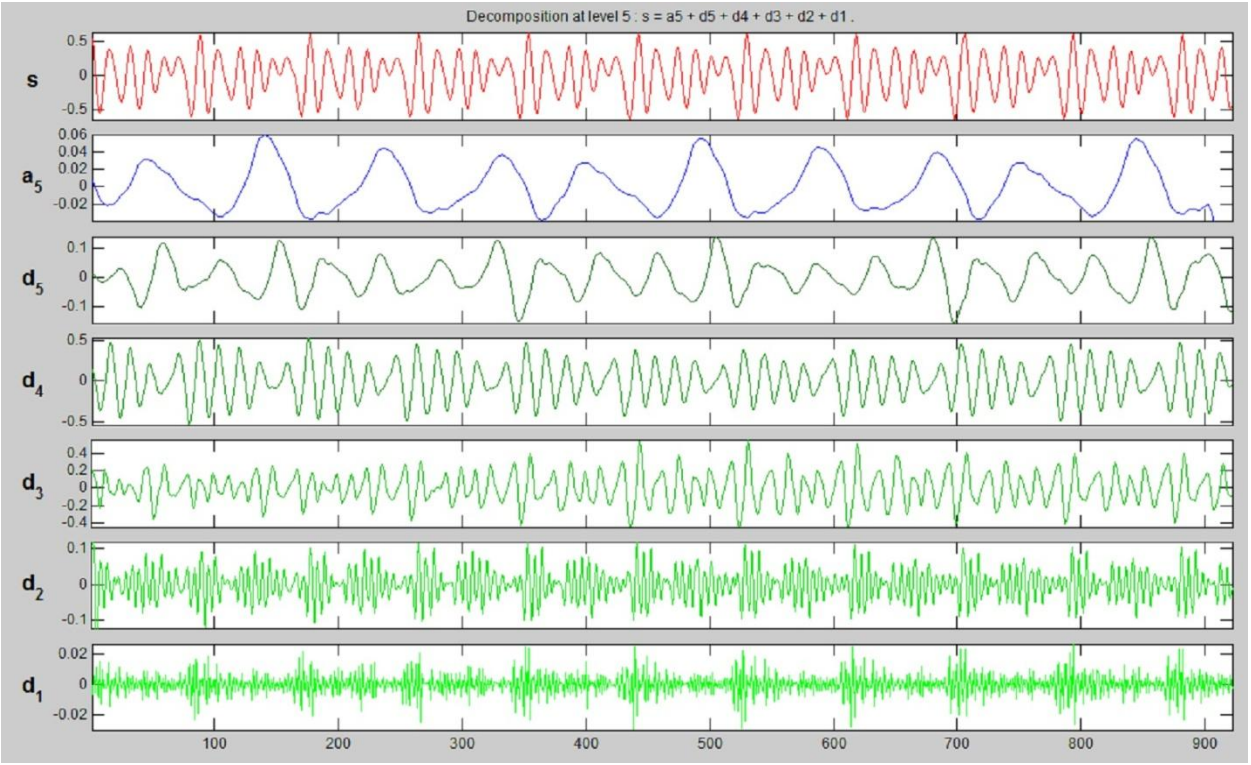

Figure5.3.b: Decomposed data for(/ɔ/(অ)).

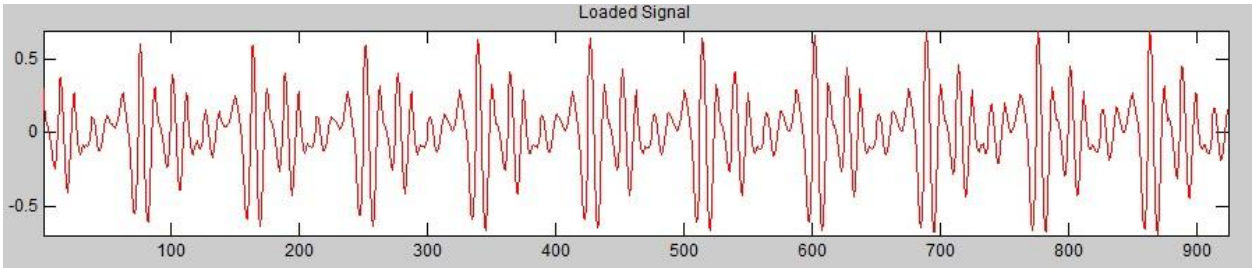

Figure 5.4.a: Original signal for (/a/(আ)).

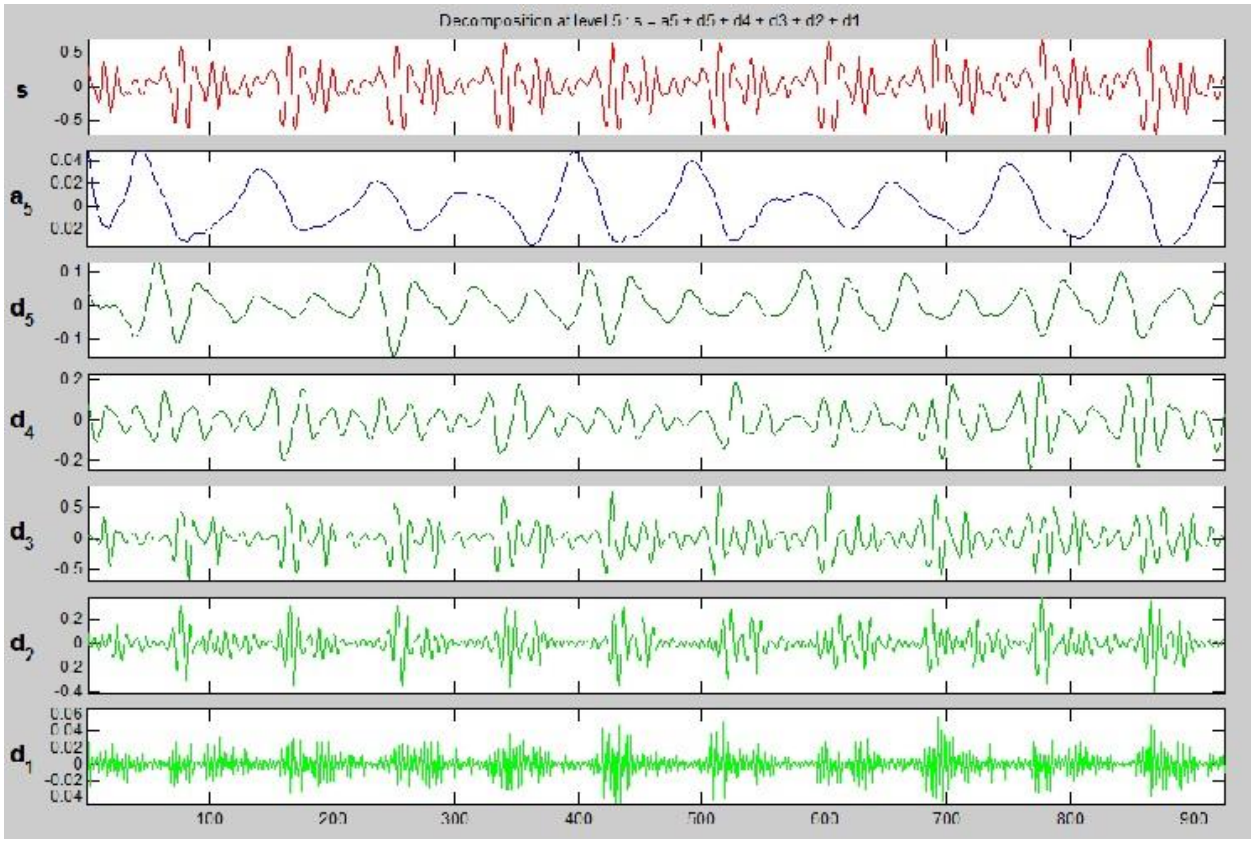

Figure5.4.b: Decomposed data for (/a/(আ)).

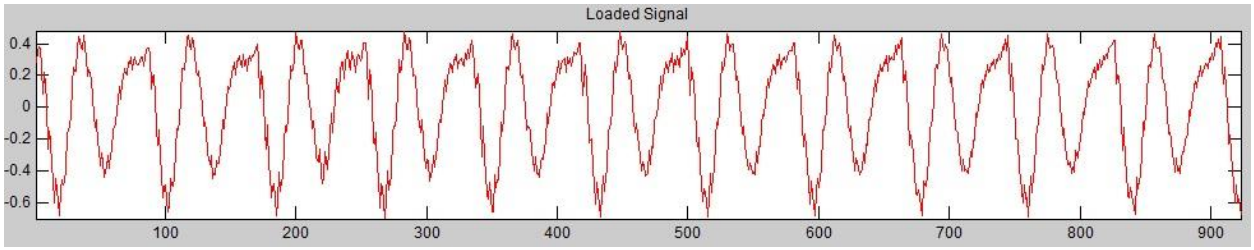

Figure5.5.a: Originalsignal for (/æ/(এ্যা)).

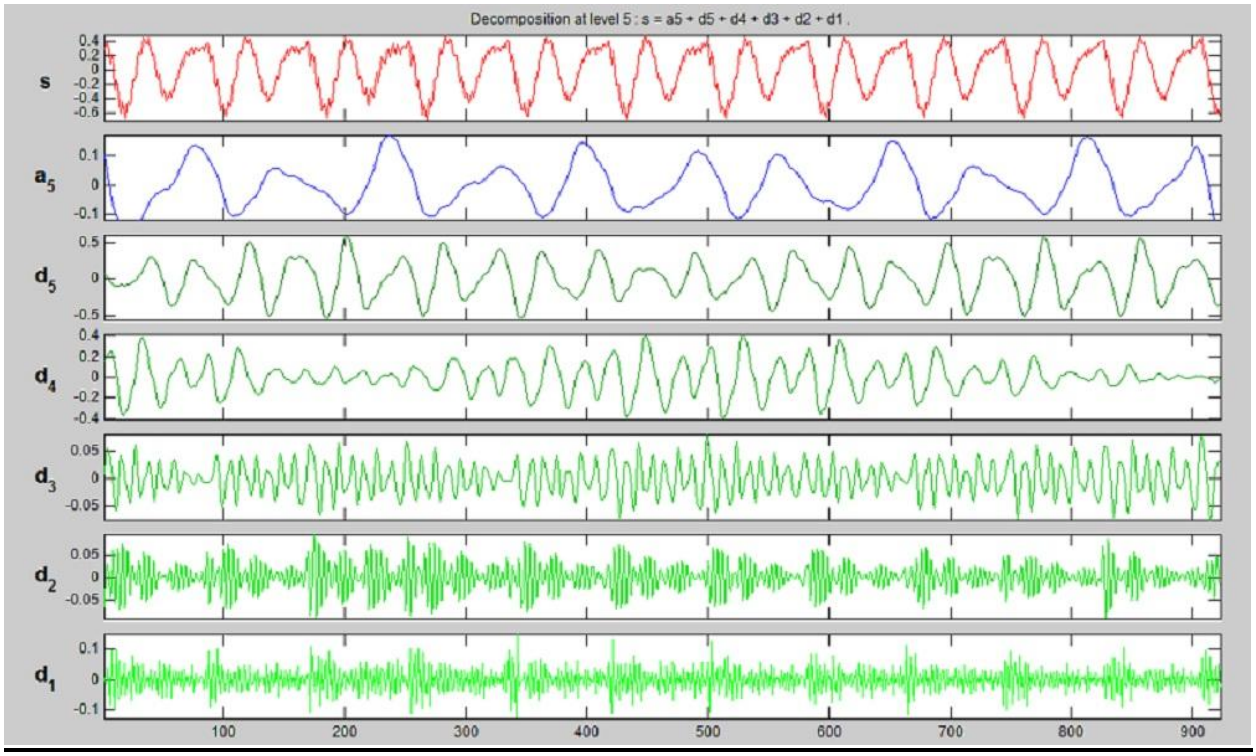

Figure5.5.b: Decomposed data for(/æ/ (অ্যা)).

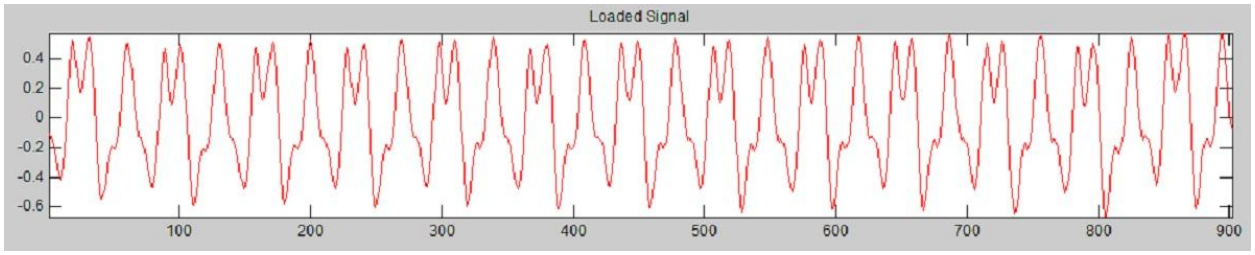

Figure5.6.a: Original signal for (/e/ (এ)).

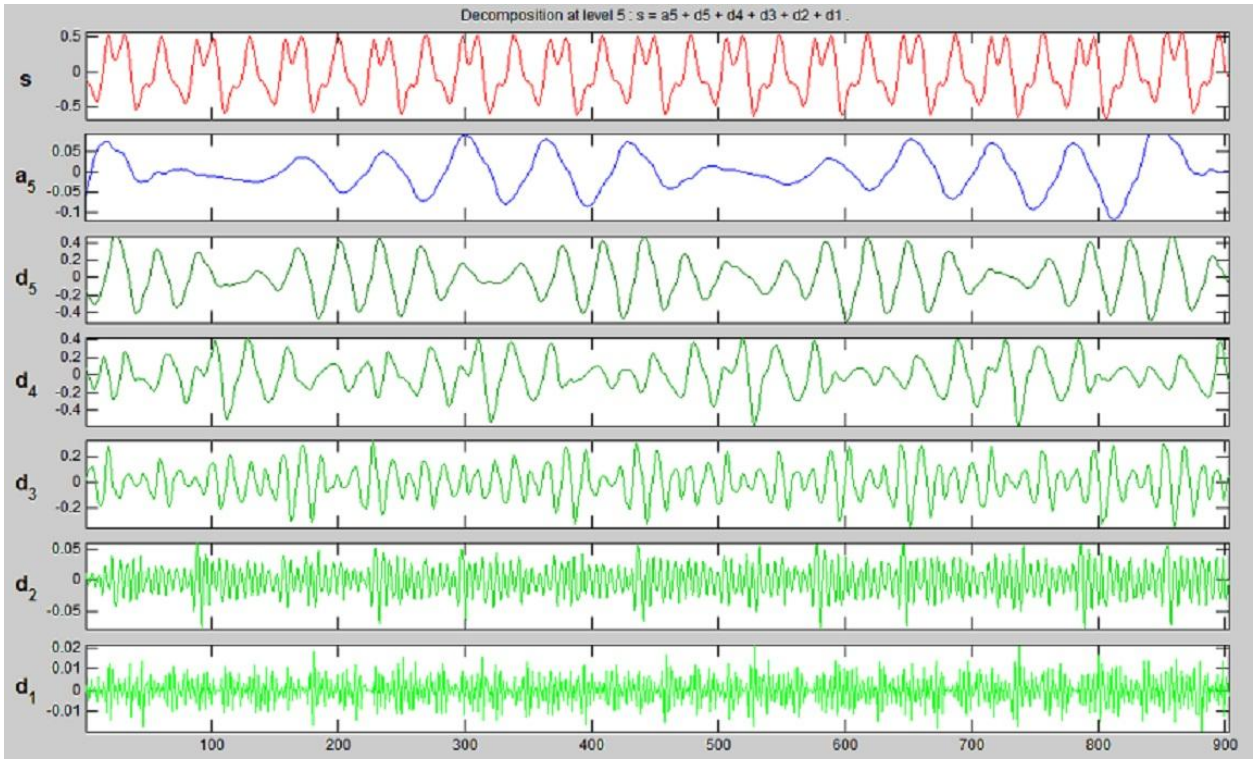

Figure5.6.b: Decomposed data for (/e/(এ)).

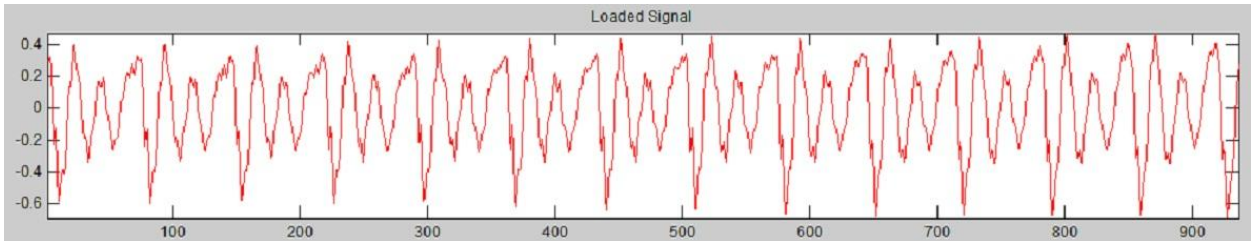

Figure 5.7.a: Original signal for (/i/(ই)).

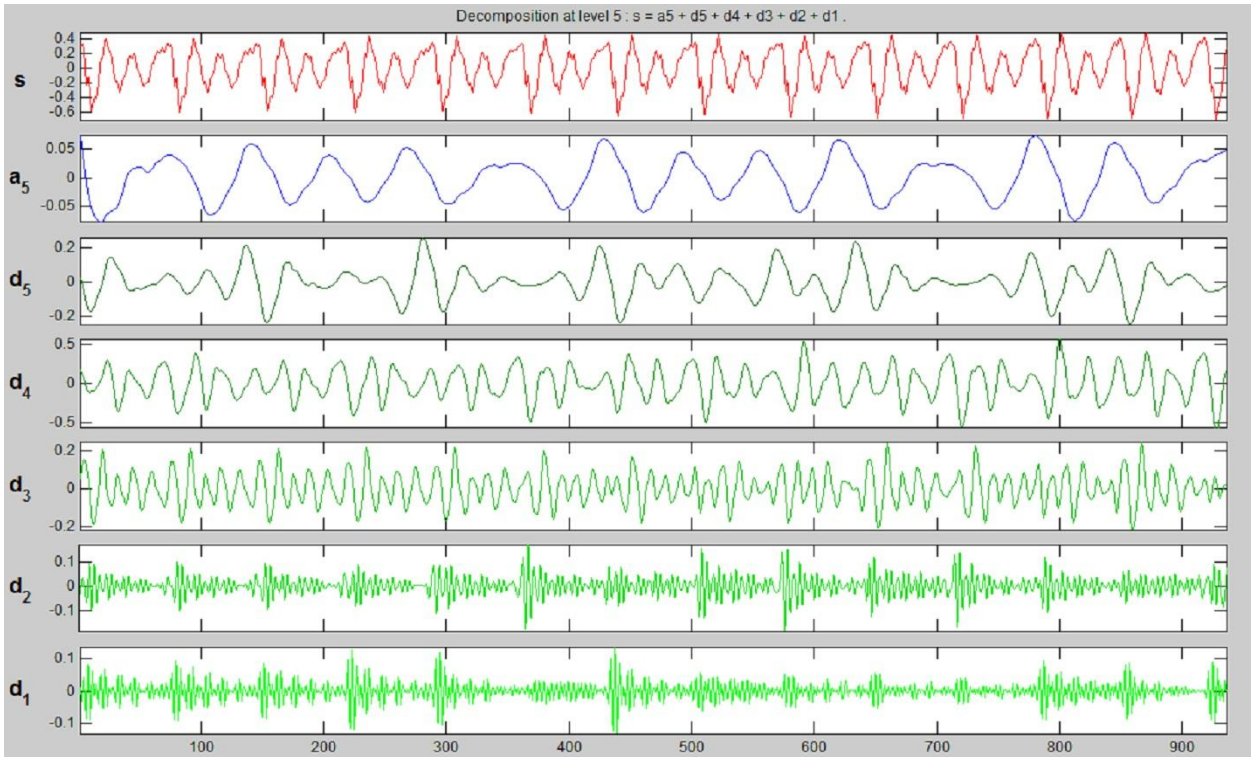

Figure5.7.b: Decomposed data for (/i/(ই)).

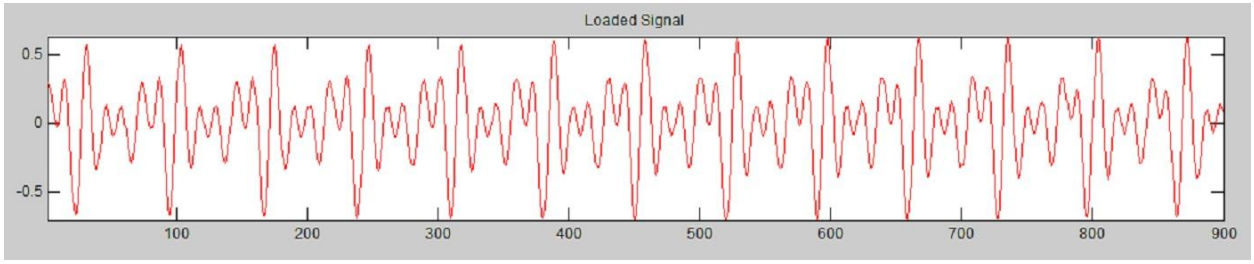

Figure5.8.a: Original signal for (/o/(ও)).

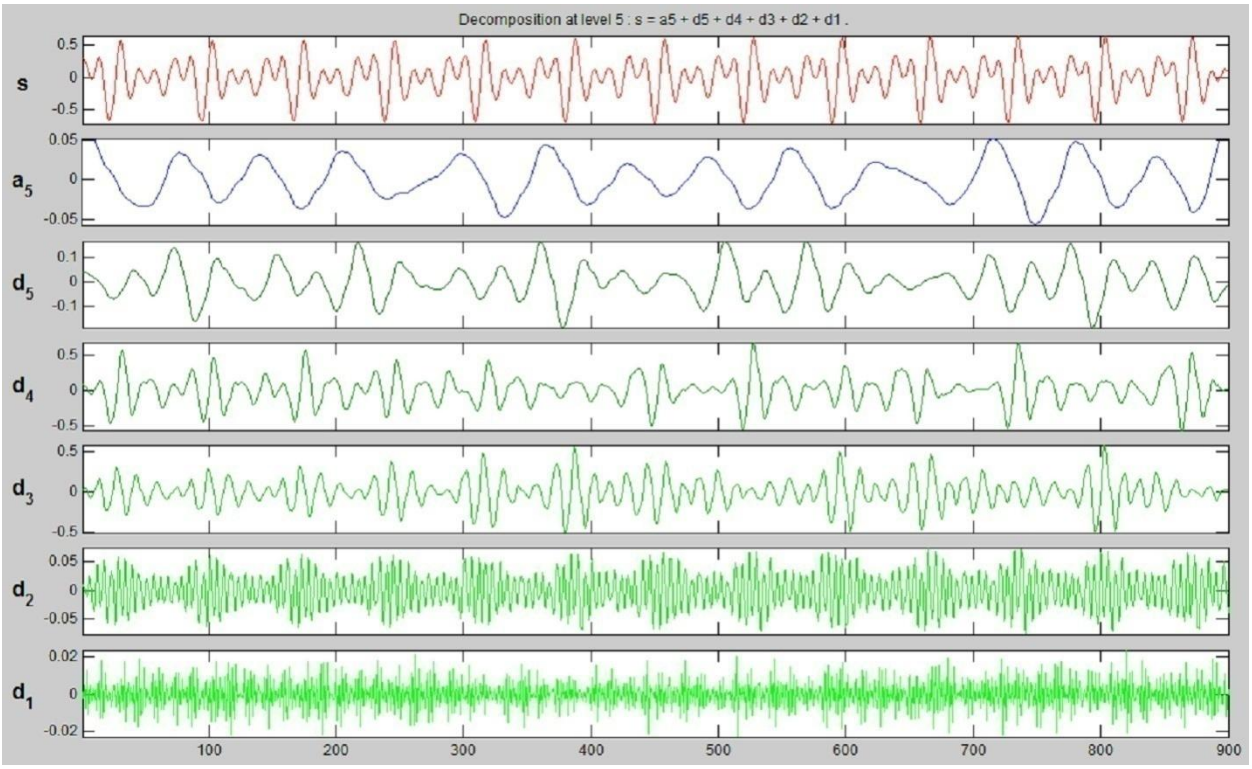

Figure5.8.b: Decomposed data for (/o/(ও)).

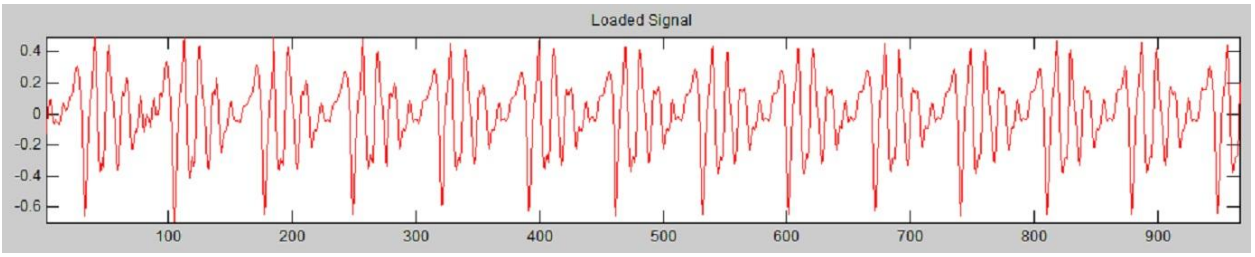

Figure5.9.a: Original signal for (/u/(ঊ)).

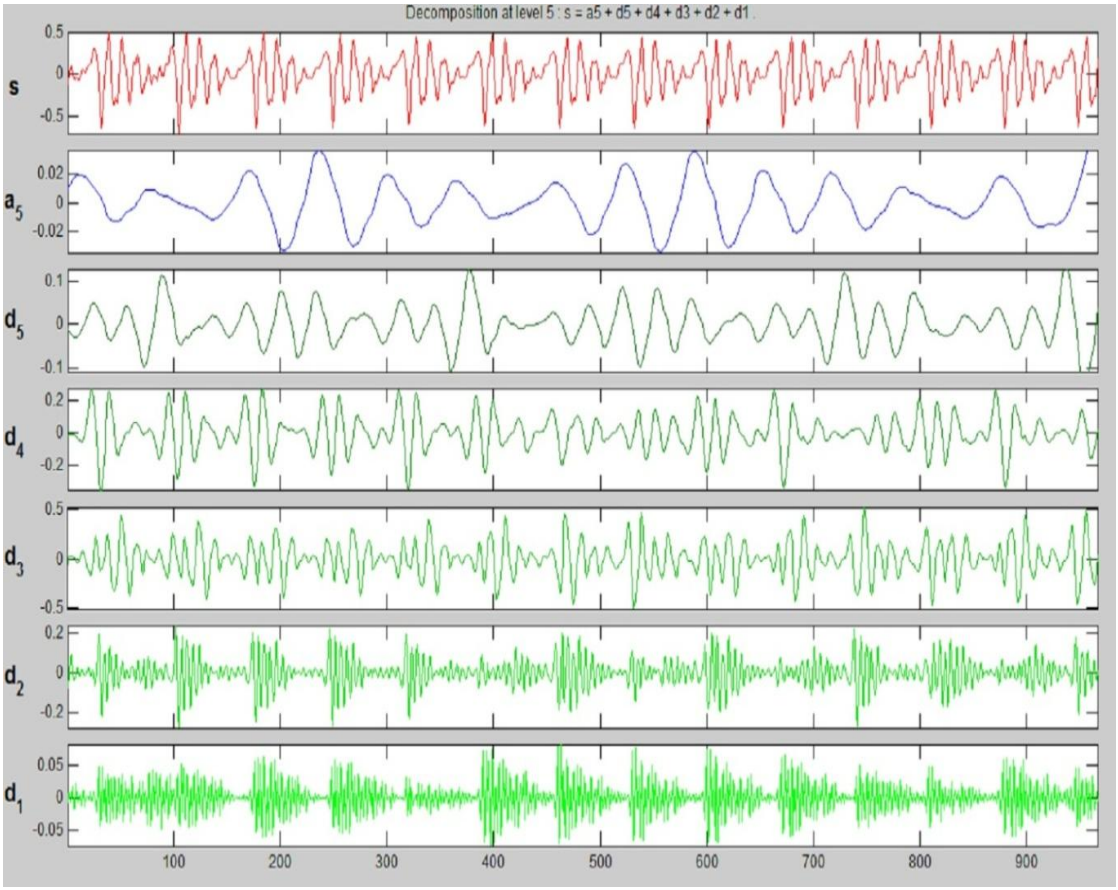

Figure5.9.b: Decomposed data for(/u/ (ঊ)).

Description of the above Figures

In figure [5.3b to 5.9b (all b figure)], first one(s) is the original signal and second one (a_5) is the approximation of the original signal and other plots (d_5, d_4, d_3, d_2, d_1) indicates the decomposition of the signal.

S=a5+d5+d4+d3+d2+d1

5.4 Block Diagram of working procedure

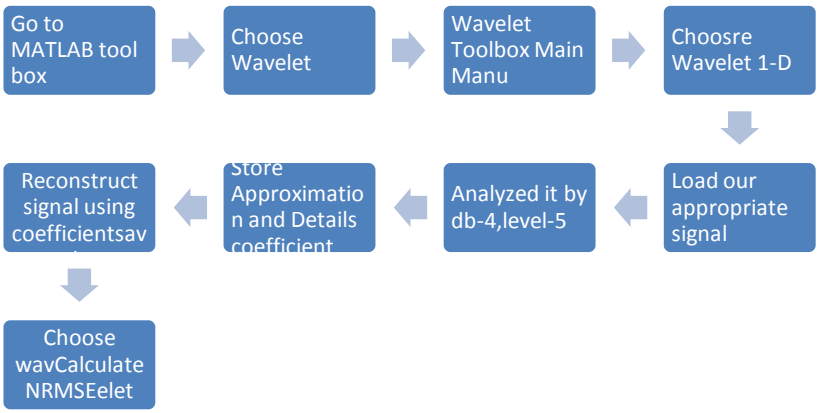

Figure 5.11:Block diagram of working procedure of analysis and synthesis of our selected vowel phonemes.

Description of the block diagram

- First, we open MATLAB and select the current directory to where our original data is stored (in our experiment we have "L:\Thesis\Data\data-with-power")
- Second, we go to start at the MATLAB user interface and found Toolbox. Select it and go to Wavelet and choose Wavelet Toolbox.
- There we get Wavelet Toolbox Main menu from where we select 1D Wavelet.
- Another prompt open (See Fig: 5.2). There we go to file > load signal > select our desired signal (for example: 1) from the current directory which we select at step 1.
- After loading the signal we choose db 4 at level 5 and analyze it.
- Save the data.
- Like this way all the data (we have seven data) we analyze and saved.
- Then we go to our main MATLAB prompt and all the saved data we got- import them to workspace. Select all and copy them to MS Excel worksheet.
- After that we calculate the NRMSE which described at the error analysis portion 6.1.

5.5 Description of the process

5.5.1 Choosing the Decomposition Level

The WT of a given signal, the decomposition level can reach up to level L =2K, where K is the length of the discrete signal. Thus we can apply the transform at any of these levels. But in fact, the decomposition level depends on the type of signal being analyzed. For the processing of speech signals, decomposition up to scale 7 is adequate [28]. In this paper, level5 DWT is obtained for every signal.

5.5.2 Choosing Appropriate Daubechies Wavelets

The type of wavelet is of high importance for such experiments. It directly affects the Signal to Noise Ratio (SNR) of the output signal. Choosing the appropriate wavelet will maximize the SNR and minimizes the relative error. As mentioned earlier Daubechies wavelets have good compression property for wavelet coefficients [27], giving better SNR ratios. Wavelets with more vanishing moments provide better reconstruction quality. Daubechies wavelets are developed with maximum regularity; the number of zero moments is maximized, leading to the best wavelet family for compression. The selected members of this orthogonal Daubechies family are db4, db8, db10 and db20. In this work, we have used db4.

5.5.3The Daubechies D4 Wavelet Transform

The Daubechies wavelet transform is named after its inventor (or would it be discoverer?), the mathematician Ingrid Daubechies. The Daubechies D4 transform has four wavelet and scaling function coefficients. The scaling function coefficients are

$$h_0 = \frac{1+\sqrt{3}}{4\sqrt{2}}$$
$$h_1 = \frac{3+\sqrt{3}}{4\sqrt{2}}$$
$$h_2 = \frac{3-\sqrt{3}}{4\sqrt{2}}$$
$$h_3 = \frac{1-\sqrt{3}}{4\sqrt{2}}$$

Each step of the WT applies the scaling function to the the data input. If the original data set has N values, the scaling function will be applied in the wavelet transform step to calculate N/2 smoothed values. In the ordered wavelet transform the smoothed values are stored in the lower half of the N element input vector.

The wavelet function coefficient values are:

$g_0 = h_3$
$g_1 = -h_2$
$g_2 = h_1$
$g_3 = -h_0$

Each step of the wavelet transform applies the wavelet function to the input data. If the original data set has N values, the wavelet function will be applied to calculate N/2 differences (reflecting change in the data). In the ordered wavelet transform the wavelet values are stored in the upper half of each N element input vector.

The scaling and wavelet functions are calculated by taking the inner product of the coefficients and four data values. The equations are shown below:

Daubechies D4 scaling function:

$$a_i = h_0 s_{2i} + h_1 s_{2i+1} + h_2 s_{2i+2} + h_3 s_{2i+3}$$
$$a[i] = h_0 s[2i] + h_1 s[2i+1] + h_2 s[2i+2] + h_3 s[2i+3];$$

Daubechies D4 wavelet function:

$$c_i = g_0 s_{2i} + g_1 s_{2i+1} + g_2 s_{2i+2} + g_3 s_{2i+3}$$
$$c[i] = g_0 s[2i] + g_1 s[2i+1] + g_2 s[2i+2] + g_3 s[2i+3];$$

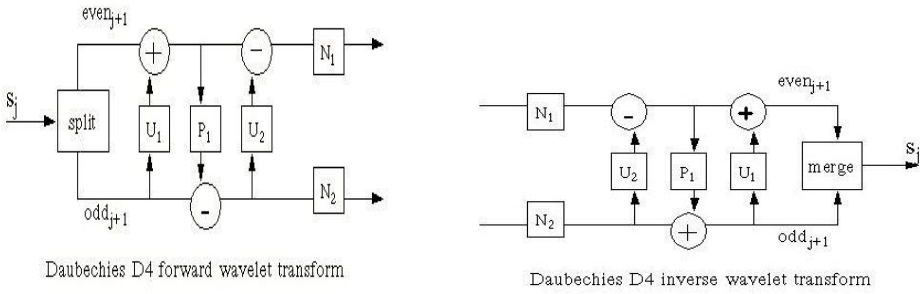

Fig 5.11: Daubechies D4 forward and reverse WT.

6

Results and Discussion

In this section we discuss the performance of the synthesized signal. We calculate the NRMSE between the original and the reconstructed vowel at decomposition levels 5.

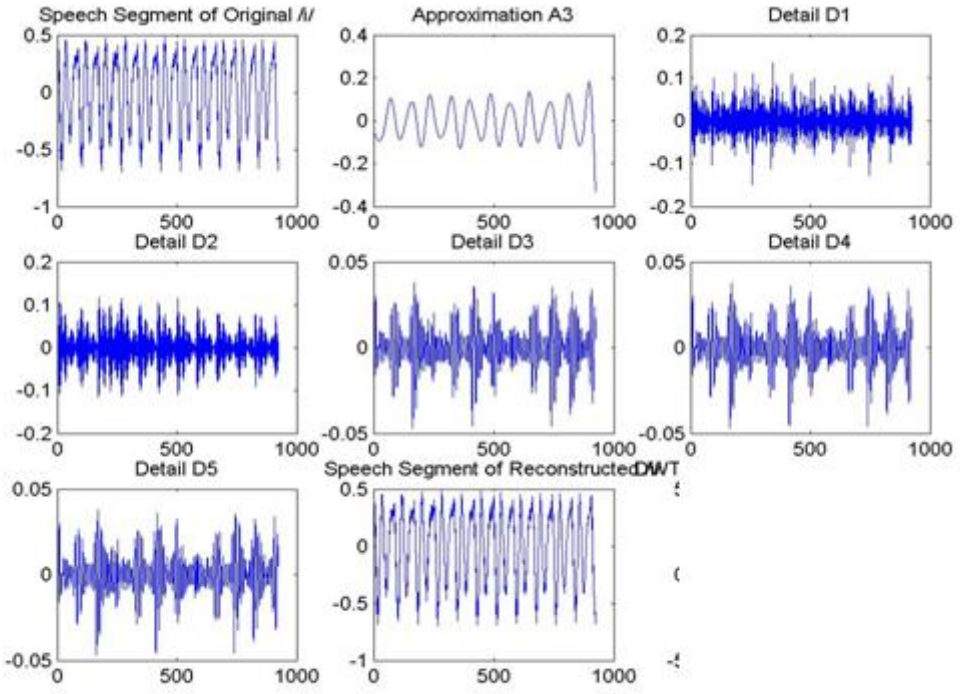

Figure 6.1: Waveform of Original, Reconstructed, Approximations, Details (at Different Scales) for vowel */i/*

Fig 6.1 shows a sample speech signal /i/ and approximations of the signal, at five different scales. These approximations are reconstructed from the coarse low frequency coefficients in the wavelet transform vector. This figure shows that the original speech data is still well represented by the level 5 approximation. The

NRMSE between the original and the reconstructed waveform for all the seven Bangla vowels is given in Fig 6.2 which is very small being the order of 10^{-11}.

We examined the performance of the synthesized signal. The NRMSE of the reconstructed vowel waveform is calculated for all the seven vowels of Bangla and is found to be in the order of 10^{-11}. It may be said that the reconstructed vowel waveform obtained by WT is almost similar to the original waveform. Therefore, we may say that WT preserves the important speech information with few parameters.

The process of NRMSE calculation is done in the following manner.

Error Analysis

Mathematically,

$$NRMSE = \sqrt{(A/B)}$$

Where

$$A = \text{Mean (Difference Square)}$$

$$= \text{Mean } (O - R)^2$$

$$B = \text{Mean } (O - \text{Mean } (O))^2$$

Here,

O = Original signal

R = Reconstructed signal.

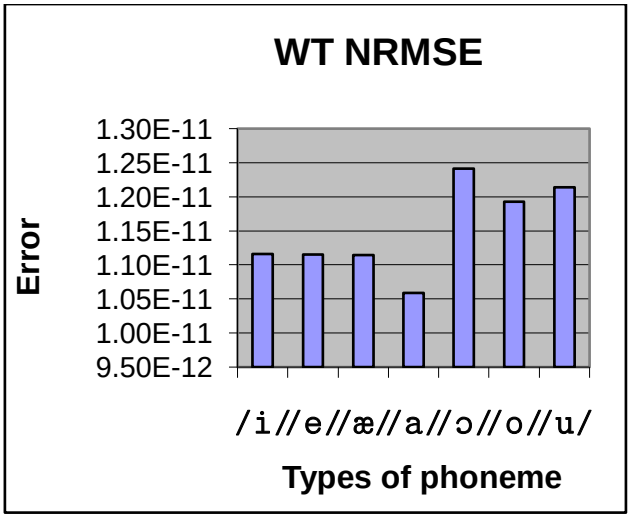

Figure 6.2: Root mean square error NRMSE.

Discussion

We have presented Wavelet transform techniques and details of how to use them in Bangla vowel phoneme analysis and synthesis. Analyzing a signal by db4 at decomposition level 5 and reconstructing the signal make a scheme of analysis and synthesis as if an artificial signal has been made virtually using a model, which is then compared to the original signal and thus we found a very small value of error (nearly 10^{-11}).

Future work

Future work will investigate other wavelet techniques that can be used to overcome some of the deficiencies in the methods presented. For example, the derivation of a unique mother wavelet may provide good accuracy and speaker variability and capture the essence of speech. Also, hardware implementations will also be studied to implement practical wavelet based speech recognition system.

REFERENCES

1. **I.Bell,** C.G. et. al, J. Acoust. Soc.Ampr, 33 (1725-1736), 1961.
2. **M.G. Ali,**"Processing of Short Duration", M. Sc. Thesis, Department of Applied Physic and Electronics, R.U. 1990.
3. **S.A.Hossain.**, "Experimental and Computer aided studies on active filter and analog and digital processing of Music and Bangla Speech.", M.Sc. Thesis, Dept. of Applied Physics and Electronics, R.U., 1991.
4. **M.M.R. Talukdar**, "Spectral and Formant analysis of Bangla Speech", M.Sc. Thesis, Dept. of Applied Physics and Electronics, R.U., 1992.
5. **L.Rahman,**"Power Spectrum and Formant Analysis of Bangla Speech", M.Sc. Thesis, Dept. of Applied Physics and Electronics, R.U., 1994.
6. **M.K. Hamid,**"Software Development for Computer Processing of Bangla Speech", M.Sc. Thesis, Dept. of Applied Physics and Electronics, R.U., 1994.
7. **M.Jamal Uddin,** "Computer aided Spectral Analysis of Bangla Phonemes", M.Sc. Thesis, Dept. of Applied Physics and Electronics, R.U., 1994.
8. **P.Khandaker,** "Computer Based Modeling of Bangla Phonemes and Software development for Speech synthesis", M.Sc. Thesis, Dept. of Applied Physics and Electronics, R.U., 1995.
9. **S.Haque,** "Comparative study of extracted features of phonemes with different age and sex groups and synthesis of voiced phonemes by developed software", M.Sc. Thesis, Dept. of Applied Physics and Electronics, R.U., 1997.
10. **M.S.Sabuj et.al.,** "Study on Bangla vowel analysis", B.Sc. Thesis (P00949), DIU, 2009
11. **L.R.Rabiner and R.W.Schafer**., "Digital Processing of Speech Signal", Engle Wood Cliffs, NJ, Prentice-Hall.Inc, 1978.
12. **W. Koeing,** et.al. J. Acoust. Soc. An. 17(19-49), 1946.
13. **W.Koeing,H.K. Dunn and L.Y .Lacy,** "The Sound Spectrograph", J. Acoust. Soc. Am. Vol. 17, pp. 19-49, Jul 1946.
14. **R.H.Stetson,** Motor Phonetics 2nd Ed., Berlin College, Ohio, 1951.

15. http://en.wikipedia.org/wiki/Synthesis
16. http://plato.stanford.edu/entries/analysis/
17. http://en.wikipedia.org/wiki/Speech_synthesis#cite_ref-0
18. **Jonathan Allen, M. Sharon Hunnicutt, Dennis Klatt,**From Text to Speech: The MITalk system. Cambridge University Press: 1987. ISBN 0-521-30641-8
19. http://en.wikipedia.org/wiki/Voice_analysis
20. http://www.mathworks.com/help/toolbox/wavelet/ug
21. **Stuart Rosen and Peter Howell**. Signals and Systems for Speech and Hearing. Academic Press,London, 1991.
22. **Tony Robinson**. Speech Analysis, 1998.Online tutorial.
23. **Alan V. Oppenheim and Ronald W. Schafer**. Digital Signal Processing. Prentice–Hall, 1975.
24. **SadaokiFurui**. Digital Speech Processing, synthesis and recognition—Marcel Dekker,2001.
25. http://www.tradeways.org/wave_1.php
26. **Shahina Haque**."Using Wavelet Transform for Bangla Phoneme Synthesis",
27. **Y. T. Chan** "Wavelet Basics", Kluwer Academic Publishers, 1995.
28. **J. I. Agbinya,** "Discrete Wavelet Transform Techniques in Speech Processing,"IEEE Digital Signal Processing Applications Proceedings, IEEE, New York, pp: 514 – 519, 1996.
29. **A. Grossmannand,J.Morlet, "**Decomposition of Hardy functions into square integrable wavelets of constant shape". SIAM Journal of Analysis,15: 723-736, 1984.
30. **J.Morlet,G.Arens, E. Fourgeau and D.Giard,** Wave propagation and sampling theory, Part1: Complex signal land scattering in multilayer media. Journal of Geophysics, 47: 203-221, 1982.
31. **L.Debnath,** Wavelet Transformation and their Applications. Birkhäuser Boston, 2002.