FEATURE EXTRACTION OF BANGLA VOWELS AND CONSONANTS FOR MALE AND FEMALE VOICE

BY

SNIGDHA ISLAM
ID: 062-19-432

AND

SHAKHAWAT HOSSAN

ID: 063-19-472

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Electronics and Telecommunication Engineering

Supervised By

Mrs. Shahina Haque
Assistant Professor
Department of ETE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
AUGUST 2012

Table of Contents

Topic Name

Page No.

APPROVAL

This Project titled Feature Extraction of Bangla vowels and consonants submitted by Snigdha Islam and Shakhawat Hossan to the Department of Electronics and Telecommunication Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Electronics and Telecommunication Engineering and approved as to its style and contents.

BOARD OF EXAMINERS

____

Dr. Md. Fayzur Rahman — Chairman
Professor and Head
Department of ETE
Faculty of Science & Information Technology
Daffodil International University

____

Dr. A. K. M. Fazlul Haque — Internal Examiner
Associate Professor
Department of ETE
Faculty of Science & Information Technology
Daffodil International University

____

Mr. Mirza Golam Rashed — Internal Examiner
Assistant Professor
Department of ETE
Faculty of Science & Information Technology
Daffodil International University

____

Dr. Shubrata Kumar Aditya — External Examiner
Professor and Chairman
Department of Applied Physics, Electronics and Communication Engineering
University of Dhaka

DECLARATION

We hereby declare that, this project has been done under the supervision of **Mrs. Shahina Haque, Assistant Professor, Department of ETE**, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Mrs. Shahina Haque
Assistant Professor
Department of ETE
Daffodil International University

Submitted by:

Snigdha Islam
ID: 062-19-432
Department of ETE
Daffodil International University

Shakhawat Hossan
ID: 063-19-472
Department of ETE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty Allah for His divine blessing makes us possible to complete this project successfully.

We fell grateful to and wish our profound our indebtedness to **Mrs. Shahina Haque, Assistant Professor,** Department of ETE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of wireless network influenced us to carry out this project .His endless patience ,scholarly guidance ,continual encouragement , constant and energetic supervision, constructive criticism , valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to Dr. A. K. M. Faziul Haque, Mr. Golam Rashed, and Head, Department of ETE, for his kind help to finish our project and also to other faculty member and the staff of ETE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

This thesis is on Feature Extraction of Bangla vowels and consonants. This thesis deals with the study of Bangla phoneme analysis which is the basis of Bangla speech processing. The main task is to acquire the Bangla vowels and consonants samples. Then extract the features by Linear Predictive Coding (LPC) method of analysis. The extracted features are pitch, amplitude and first three formants for both male and female voice. This document will describe LPC speech signal analysis technique and apply it to Bangla vowels and consonants to extract features. The speech features obtained by this method have acceptable comparable values.

CHAPTER 1

Introduction

Man’s primary method of communication is speech. He is unique in his ability to transmit with his voice. Of the myriad of life sharing our world, only man has developed the 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 coding of speech information.

1.1 Overview

Speech processing is the study of speech signals and 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 is progress. The signals are usually processed in a digital representation, so speech progress. The signals are usually processed in a digital signal processing, applied to speech signal.

1.2 History of Previous Work

Signal processing, a field which has its roots in the 17th and 18th century mathematics, has become an important modern tool in a multitude of diverse fields of science and technology. Many of the limitations 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 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 197. The formant structure of vowels of Bangla with Japanese and American English were discussed successfully. In 1986, some aspect of automatic generation and recognition of speech is discussed. Software was developed for automatic transliteration of English text phonetic symbols and reported that the system gives 90 percent correct word transcription. M.G Ali made a spectrum analysis of three Bangla vowels / A/,/ B/,/ G/ using short time Fourier analysis. S.A. Hossain carried out work on the power spectrum analysis of a good number of Bangla vowels and consonant. M.R. Talukdar also analyzed the formant frequencies and power spectrum of some Bangla vowels and consnants. M. Lutfor Rahman 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, M Jamal Uddin , P Khandakar , 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 etc.

1.3 Objective

Speech processing is necessary to make the properties of speech signal clear analytically beforehand. In this connection analytically studies on various languages are being in progress around the world. Bangla, being a language of about 250 million people, there are few studies on it. So in order to make effective processing and economic transmission of Bangla speech, it is necessary to study it analytically.

As first phase of study on Bangla speech processing we selected the Bangla vowels consonants in isolated utterance for the purpose of analysis. The object is to get introduced to the vast area of Bangla speech processing

1.4 Result :

There we used oral, nasal vowels and consonants for both male and Female for speech analysis vising Lpc (Linear predictive coding). Here we prove that Bangla Phonemes also maintain the international phonetics rules. The pitch period of male is 0-2000ms and female above 250ms.

Chapter 2
The Mechanism of Voice [Speech] Production
2.1 Introduction

When a person carries on a conversation, certain physical actions must occur within the body or else talking cannot be heard. These actions begin naturally with the crying of a newborn child and are then developed further in the formative years. Since singing has been considered an art form of sustained speech, the mechanism of speech and singing is the same.

2.2 Fundamentals of voice production

The foundation for an effective voice is based on the coordination of three factors:

- Breathing
- Phonation
- Resonance

Breathing air out of the lungs produces the power supply for the voice. This airflow from the lungs makes the vocal folds (or vocal chords) in the larynx (or voice box) vibrate to make the basic sound of the voice; this process is called phonation. Because that sound made by the vocal folds is too weak to be heard, that basic sound is then modified into the sound we recognize as the human voice as it travels up from the larynx through the throat, mouth and nose; this transformation is known as resonance. Production of a natural, effective voice depends on how well we balance or coordinate these three fundamental components of breathing, phonation and resonance. [1]

2.2.1 Breathing

Our intention to produce voice is signaled to the parts of the body involved by impulses from the brain. The first response of the body to these impulses is to breathe in so that there is enough air in the lungs to power the voice.

The breath is taken in through the mouth and nose, passes down the trachea (or windpipe), and is inhaled into the lungs. For air to be inhaled into the lungs, the ribcage needs to expand and the dome-like diaphragm which forms the base of the chest, needs to flatten downwards. When we breathe in effectively, we feel most of this expansion in the area of the lower ribs. Once the air has been inhaled into the lungs and they reach capacity, the elastic tissue of the lung recoils and the air is exhaled or breathed out. The exhaled air then returns up through the trachea and then through the larynx where it encounters the closing vocal folds. [1]

2.2.2 Phonation

When we breathe in and out without speaking, the vocal folds in the larynx are open to allow the air to pass to and from the lungs easily. The impulses sent from the brain when we intend to speak, however, signal to the muscles of the larynx to close the vocal folds. When the air coming up from the lungs encounters the closed vocal folds, the pressure and flow of the air overcomes the resistance of the vocal folds and sets them into a pattern of rapid vibration. That is, the vocal folds open and close repeatedly, around 200 - 220 times per second for women and 100 - 120 times per second for men. This rapid vibration of the vocal folds produces the sound waves in the air which are the basic tones of our voices. The vocal folds are therefore the source of the human voice. The larynx is located on the top of the trachea and is behind the Adam 's apple. The two vocal folds in the larynx are approximately 20 mm in length and are stretched from just behind the Adam's apple in the front of your neck to the back of the larynx. These vocal folds are complex structures made up of four main layers. The outer layer is the mucous membrane (or epithelium). Directly under the mucous membrane is a soft, pliable layer filled with fluid; this layer is known as Ranke's space. The mucous membrane and Reinke's space are together known as the 'cover' of the vocal folds. This cover of the vocal folds must be kept moist and pliable so that it can move freely in a wave-like motion (the 'mucosal wave') over the deeper layers of the folds. If the cover of the Vocal folds becomes dry or stiff, the voice will become rough and the person may experience throat discomfort.

Under the cover of the vocal folds is the vocal ligament. This ligament is made up of elastic tissue that allows the vocal folds to change shape easily when the deepest and least pliable layer of the vocal folds, the muscle, changes shape. The basic tone of the voice can be varied in many different ways, depending on the way in which we use the vocal folds and other parts of the voice mechanism. [1]

The main aspects of the voice that can be varied are:

- Pitch
- Loudness
- Quality

Pitch

Refers to how high or low the voice sounds. It is determined mainly by the speed of vibration of the vocal folds, the thickness of the edge of the folds, and the length of the folds. The higher the voice, the faster is the rate of vibration of the vocal folds. The more elongated and thinner the edges of the vocal folds become, the higher the pitch will be. On the other hand, if the vibrating edges of the vocal folds become thicker and shorter, and the vocal folds vibrate at a slower rate, the pitch will be lowered. We use variations in pitch during speech to signal meaning and emotion and this is referred to as intonation. [1]

Loudness

It refers to how loud or soft a voice is. It is dependent on the amount of air pressure from the lungs and the muscle tension in the vocal folds. The greater the air pressure and the more tense the vocal folds, the louder the sound will be. The lower the air pressure from the lungs is and the slacker the vocal folds are, the softer the voice will be. We also use variations in loudness during speech to signal meaning and emotion and this is referred to as stress. To emphasis the importance of a particular word, for example, we increase the loudness of voice on that word. [1]

Quality

It refers to how clear the voice sounds. Voice quality is determined by many complex factors including how relaxed the muscles of the larynx are, how moist the cover of the vocal folds is, how smoothly the vocal folds vibrate, and whether or not the vocal folds are able to close sufficiently during phonation. If the muscles of the larynx are excessively tense, the cover is dry, the folds move in an irregular way, and/or the folds cannot close together, the voice quality will sound rough, strained and/or breathy. [1]

2.2.3 Resonance

The sound waves produced by the vocal folds in the larynx are too weak to be recognized as voice and so this basic tone must be amplified or resonated as it travels up through the spaces of the throat, mouth and nose. The shape, size and muscle tension of these spaces will determine the eventual sound of the voice we will hear. Because every person is built differently in the throat, mouth and nose, the basic voice tone is modified differently in each of us so that we will all have a recognizably unique timbre of voice. This process of resonance in our voices is similar to the way in which the shape and size of a musical instrument such as a trumpet gives the basic tone produced by the reed its unique sound. Just as the resonance process in a trumpet makes the sound of the trumpet carry throughout a concert hall, resonance in the human voice gives us the ability to control its carrying power or projection. [1]

2.3 Components of the Speech Production System

(a) The Respiratory Organ: The respiratory organs include all of those organs involved with breathing, so your nose, mouth, trachea, lungs, and other organs are all included in this

Category. Another respiratory organ to consider is the larynx... that is because the only reason that we can speak is that we can control breathing to force air through the larynx. All of the respiratory organs can be considered to be either upper or lower respiratory tract organs. So, the nose and mouth fit into the upper respiratory tract category, while the larynx and lungs fit into the lower respiratory tract category.

This figure is pointing out all of the major respiratory organs from both the upper and lower respiratory tracts. Please note that the mouth isn't considered a "respiratory organ" specifically... that is because it is also a digestive organ.

(b) The Larynx:

The pharynx carries the air into the larynx. The entire function of the larynx is to produce sounds. The larynx is commonly known as the voice box. The larynx contains vocal cords within it, which are laryngeal folds that project into the space within the larynx (you'll learn more about these in lab). You see, as air travels out under pressure through the larynx, the vocal cords in the larynx vibrate. This vibration causes the air running through it to vibrate, which can also be thought of as forcing the air to move in waves. When you learned about sound and hearing, you learned that sound is simply air traveling in waves. Therefore, the general idea of the larynx is to force air into waves to produce sounds. If the vocal cords vibrate quickly, the sound waves will be fast, and a high-pitched sound will be made. If the vocal cords vibrate slowly, the sound waves will be slow, and a low-pitched sound will be made.

Finally, if the air is not forced through the larynx under pressure, the vocal cords will not vibrate at all, and no sound will be made (normal exhaling). The larynx is highly structured. You see, in order to have vocal cords vibrate at different frequencies (speeds), we have to be able to adjust the tension on the vocal cords. We do that by having muscle attached to the cords, regulating their tension. When we want to make a high-pitched sound, we increase the tension on the vocal cords to force them to vibrate faster. Mainly because of this need for regulation of the vocal cords with muscle, we need a strong, supportive structure for the larynx. This structure

is provided by a number of pieces of cartilage, as seen in this picture below.

The thyroid, cricoid, corniculate, and arytenoids cartilage are all supportive cartilaginous structures for the larynx. The entire larynx is also supported by attachment to the hyoid bone.

One other item in this picture is the "epiglottis cartilage." This piece of cartilage does not support the larynx. Instead, the epiglottis functions to block off the air passageways of the lower respiratory tract when swallowing. You see, since both food and air are in the pharynx, and the pharynx leads to the larynx, there has to be a way to prevent the food in the pharynx from entering the larynx, yet still a way to allow air to enter the larynx. The epiglottis is a movable structure, performing this function. We will focus more on the epiglottis.

(c) Supra glottal cavities:

The Supra glottal cavities are located above the glottis .The air stream coming from the larynx will be subject to further modifications by the action of the articulators, which will shape the different supra glottal cavities and, in doing so, they will alter the resonating effects of such cavities. We will distinguish 4 main resonators in the vocal tract: pharyngeal, oral, nasal and labial. The pharyngeal resonator is the result of the shape of the pharynx at the moment the air passes through.

The oral resonator will change in shape and volume depending on the position of the different articulators inside.

The nasal resonator has a fixed influence on the quality of the sound and it appears only when a sound is nasal. The labial resonator determines a special characteristic in those sounds where the lips do not adopt a relaxed position, but rather tense (spread or rounded) as we will see in the following sections. Only the nasal cavity and the oral cavity will be used in the classification of English sounds.

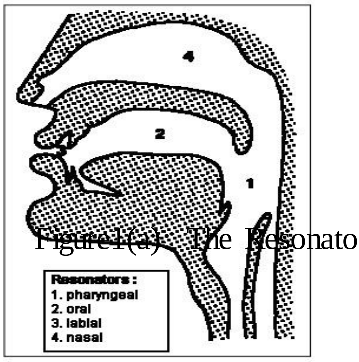


Figure1(a): The Resonators

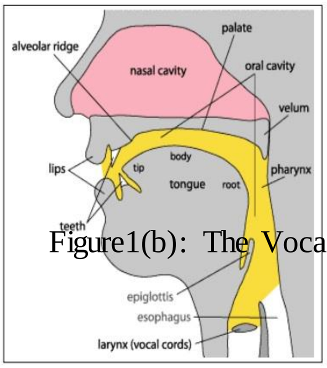


Figure1(b): The Vocal tract

The pharynx extends from the trachea and esophagus, past the epiglottis and the root of the tongue, to the rear region of the soft palate. The raising of the larynx and the action of the different muscles around it will alter the shape of this cavity and, therefore, the resonance effect it can produce.

One important activity going on at this point is the position of the soft palate (or velum) in the pharynx. Depending on the position, the sound will be nasal (as in the first sound of nap) or oral.

a) The soft palate may be lowered, as in normal breathing, thus allowing for the air to escape through the mouth and the nose. When that happens, we get a nasal sound.

b) The soft palate may be raised, so that the escape of the air is only possible through the mouth. When that happens, we get an oral sound.

The mouth is possibly the "busiest" cavity in the production of speech mechanism. This is due to the existence of different elements which can provoke many different alterations to the shape of this cavity. The originators of these alterations are called articulators. That is, the resonance effect of the mouth will be altered by the participation of the articulators. A detailed description of these articulators follows below.The lips articulate with each other to make bilabial sounds. Another possible articulation occurs when the lower lip articulates with the upper lip, then forming labiodental sounds. The teeth usually articulate with the tongue, forming

Dental sounds. Another common articulation is that of the teeth and the lower lip, making labiodental sounds.

Just behind the upper teeth, we find the alveolar ridge. The alveolar ridge is a bumpy area which articulates with the tongue for the articulation of alveolar sounds. Other sounds are possible: post alveolar sounds, which are made with the blade of the tongue articulating at the back of the alveolar ridge. Retroflex sounds are produced when the tip of the tongue is curled back to articulate within the area of the alveolar ridge.

What comes after the alveolar ridge is the bony area known as the hard palate (or simply, palate). The tongue articulates with the palate to form palatal sounds.

The rear portion of the palate is known as the soft palate or velum. The back of the tongue articulates at this area to form velar sounds.

At the rear of the soft palate, we find the uvula. The tongue articulates with the uvula to make uvular sounds. But there are not uvular sounds in English.

The tongue is, obviously, the most active articulator in the mouth. It is present in the articulation of many types of sounds, including: dental, alveolar, post-alveolar, retroflex, palatal, velar and uvular; that is, everything but those articulated at the labial area. Being a large muscle, it is usually divided into different areas to identify which part specifically participates in each articulation.

Now that we have revised every single element in the oral cavity or mouth, we must go on with the last resonator: the nasal cavity (nose). This cavity is important, as we saw before, according to the position of the velum. Depending on such position, the sound will be nasal (as in the first sound of nap) or oral.

a) The soft palate may be lowered, as in normal breathing, thus allowing for the air to escape through the mouth and the nose. When that happens, we get a nasalized sound.

b) The soft palate may be raised, so that the escape of the air is only possible through the mouth. When that happens, we get an oral sound.

c) The soft palate may be lowered with a complete obstruction at some point in the mouth. When that happens, we get a nasal sound.

Chapter 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 sounds 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 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 and sounds are spoken or used in speech. For example ‘P’ of English alphabet is not identically similar to /p/ of English phonemes. It must not be confused or wrongly understood. Letters of English alphabet do not maintain the same identity and equality every time everywhere, whereas each letters of phonetic notation represents a small family of sounds. The equality of sounds varies to some extent. In ordinary English spelling, it not easy to know what sounds the letters stands for.

3.2 Phonemes

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 is an utterance in changed. The acoustic manifestation of basic unit may vary widely. The basic unit for describing how speech conveys linguistic meaning is called phonemes. Its manifolds acoustic variations are called allophone. 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 is predicable from the context, (d) those sounds 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 dependent upon the phonemes 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)

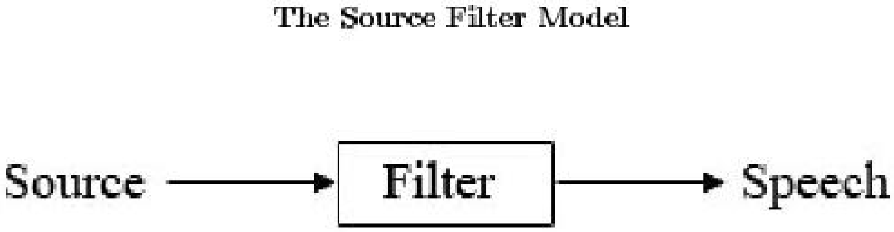


Fig.2: Source system model of speech production

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 fig. 2 shows a general block diagram that is representative of numerous models that have been used as the basis for speech processing.

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 Study on Analysis & Synthesis of Bangla Vowel 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.

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. 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

The source filter model also helps to qualify vowels since we can separately measure the contribution of the source and the filter to the final vowel sound. The primary characteristics of the source is its fundamental frequency while the primary characteristics of the filter can be reduce to the location in frequency of the vocal tract resonance or formants, see figure 3.

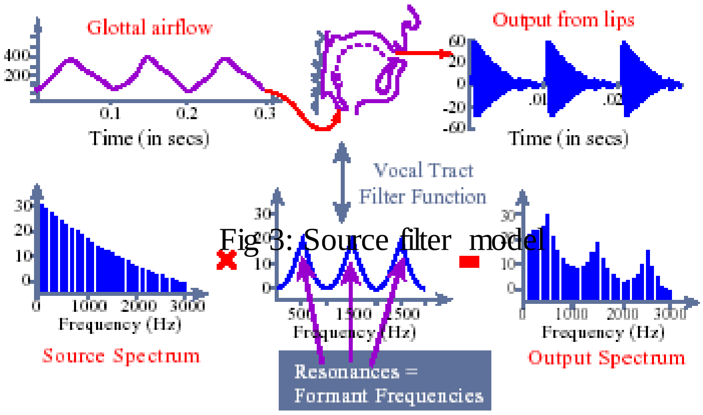


Fig 3: Source filter model

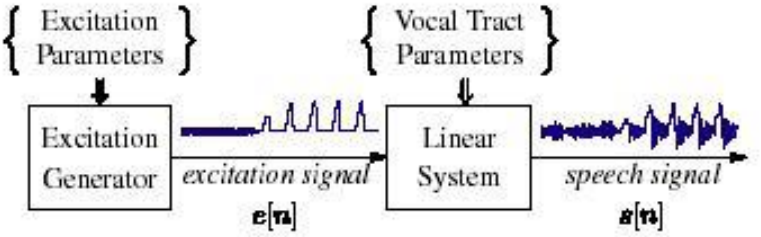


Fig 4: Source/System model for speech signal

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 system changes shape relatively solely, it is reasonable to assume that the linear system does not 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 that 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 *ak* and *bk* (labeled as vocal tract parameters in Figure 4) change at a rate on the order of 50~100 times/s. Some of the poles (*ck*) 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 (*dk*) of the system function to model nasal and fricative.

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

3.4.1 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.

3.4.2 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:

3.4.3 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.

3.4.4 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.[4]

3.4.5 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. [2]

3.4.6 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. [2]

3.4.7 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 5. 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 5).[2]

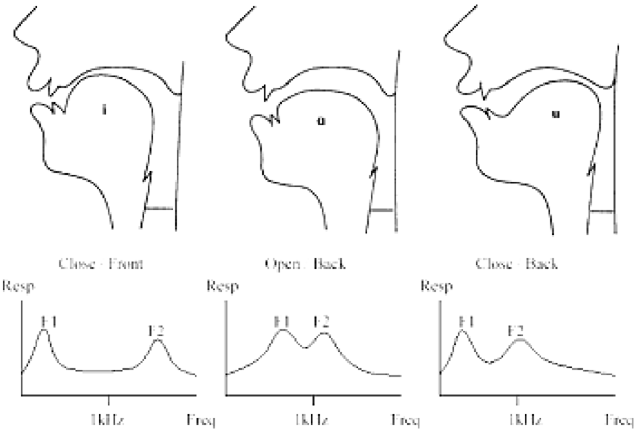


Figure 5: Vocal tract shapes for different vowels lead to different frequency Responses Close / Front Open / Back Close / Back

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.

3.5.1 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: 5).

3.5.2 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 Amin, (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

i. 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.

ii. 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.

iii. 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.[5]

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.

Chapter 4

Mathematical Tools Used

Speech processing is the study of speech signals and the processing methods of these signals.

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.

It is also closely tied to natural language processing (NLP), as its input can come from / output can go to NLP applications. E.g. text-to-speech synthesis may use a syntactic parser on its input text and speech recognition's output may be used by e.g. information extraction techniques.

Speech processing can be divided into the following categories:

- Speech recognition, which deals with analysis of the linguistic content of a speech signal.
- Speaker recognition, where the aim is to recognize the identity of the speaker.
- Speech coding, a specialized form of data compression, is important in the telecommunication area.
- Voice analysis for medical purposes, such as analysis of vocal loading and dysfunction of the vocal cords.
- Speech synthesis: the artificial synthesis of speech, which usually means computer-generated speech.
- Speech enhancement: enhancing the intelligibility and/or perceptual quality of a speech signal, like audio noise reduction for audio signals.

4.1 Discrete Time Signal

Discrete time is the discontinuity of a function's time domain that results from sampling a variable at a finite interval. For example, consider a newspaper that reports the price of crude oil once every day at 6:00AM. The newspaper is described as sampling the cost at a frequency of 24 hours, and each number that's published is called a sample. The price is not defined by the newspaper in between the times that the numbers were published. Suppose it is necessary to know the price of the oil at 12:00PM on one particular day in the past; one must base the estimate on any number of samples that were obtained on the days before and after the event. Such a process is known as interpolation. In general, the sampling period in discrete-time systems is constant, but in some cases non uniform sampling is also used.[6]

Discrete-time signals are typically written as a function of an index n (for example, x (n) or x_n may represent a discretisation of x (t) sampled every T seconds). In contrast to Continuous signal systems, where the behavior of a system is often described by a set of linear differential equations, discrete-time systems are described in terms of difference equations. Most Monte Carlo simulations utilize a discrete-timing method, either because the system cannot be efficiently represented by a set of equations, or because no such set of equations exists. Transform-domain analysis of discrete-time systems often makes use of the Z transform.

If the independent variable (t) takes on only discrete values, for example t = ±1, ±2, ±3 ...

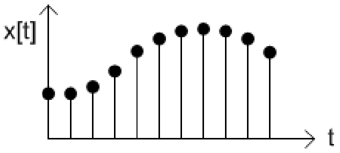


Fig 6: Discrete time signal

4.2 Sampling Theorem

In signal processing, **sampling** is the reduction of a continuous signal to a discrete signal. A common example is the conversion of a sound wave (a continuous signal) to a sequence of samples (a discrete-time signal). A **sample** refers to a value or set of values at a point in time and/or space. A **sampler** is a subsystem or operation that extracts samples from a continuous signal.

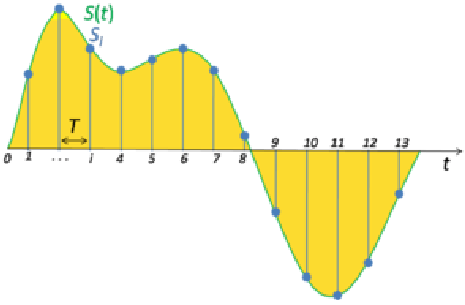


Fig 4.2: Sampling Theorem

A theoretical **ideal sampler** produces samples equivalent to the instantaneous value of the continuous signal at the desired points.

Sampling can be done for functions varying in space, time, or any other dimension, and similar results are obtained in two or more dimensions.[7]

For functions that vary with time, let s(t) be a continuous function (or "signal") to be sampled, and let sampling be performed by measuring the value of the continuous function every T seconds, which is called the sampling interval. Thus, the sampled function is given by the sequence**:**

s(nT), for integer values of n.

The sampling frequency or sampling rate $\mathbf{f_s}$ is defined as the number of samples obtained in one second (samples per second), thus $\mathbf{f_s = 1/T}$.

Reconstructing a continuous function from samples is done by interpolation algorithms. The Whittaker–Shannon interpolation formula is mathematically equivalent to an ideal low pass filter whose input is a sequence of Dirac delta functions that are modulated (multiplied) by the sample values. When the time interval between adjacent samples is a constant (T), the sequence of delta functions is called a Dirac comb. Mathematically, the modulated Dirac comb is equivalent to the product of the comb function with s (t). That purely mathematical function is often loosely referred to as the sampled signal.

The Nyquist–Shannon sampling theorem is a fundamental result in the field of information theory, in particular telecommunications and signal processing. Sampling is the process of converting a signal (for example, a function of continuous time or space) into a numeric sequence (a function of discrete time or space). Shannon's version of the theorem states:

If a function x (t) contains no frequencies higher than B hertz, it is completely determined by giving its ordinates at a series of points spaced 1/ (2B) seconds apart. The theorem is commonly called the Nyquist sampling theorem.

4.3 The Fourier Transforms

The **Fourier transform** is a mathematical transform with many applications in physics and engineering. Very commonly, it expresses a mathematical function of time as a function of frequency, known as its frequency spectrum. The Fourier inversion theorem details this relationship.

For instance, the transform of a musical chord made up of pure notes (without overtones) expressed as amplitude as a function of time, is a mathematical representation of the amplitudes and phases of the individual notes that make it up. The function of time is often called the time domain representation, and the frequency spectrum the frequency domain representation. The inverse Fourier transform expresses a frequency domain function in the time domain. Each value of the function is usually expressed as a complex number (called complex amplitude) that can be interpreted as a magnitude and a phase component. The term "Fourier transform" refers to both the transform operation and to the complex-valued function it produces.[8]

In the case of a periodic function, such as a continuous, but not necessarily sinusoidal, musical tone, the Fourier transform can be simplified to the calculation of a discrete set of complex amplitudes, called Fourier series coefficients. Also, when a time-domain function is sampled to facilitate storage or computer-processing, it is still possible to recreate a version of the original Fourier transform according to the Poisson summation formula, also known as discrete-time Fourier transform. These topics are addressed in separate articles. For an overview of those and other related operations, refer to Fourier analysis or List of Fourier-related transforms.

The following images provide a visual illustration of how the Fourier transforms measures whether a frequency is present in a particular function. The function depicted f (t) = cos (6πt) $e^{-\pi t2}$ oscillates at 3 hertz (if t measures seconds) and tends quickly to 0. (The second factor in this equation is an envelope function that shapes the continuous sinusoid into a short pulse. Its general form is a Gaussian function). This function was specially chosen to have a real Fourier transform which can easily be plotted. The first image contains its graph. In order to calculate $\hat{f}(3)$ we must integrate $e^{-2\pi i(3t)}f(t)$. The second image shows the plot of the real and imaginary parts of this function. The real part of the integrand is almost always positive, because when f (t) is negative, the real part of $e^{-2\pi i(3t)}$ is negative as well. Because they oscillate at the same rate, when f (t) is positive, so is the real part of $e^{-2\pi i(3t)}$. The result is that when you integrate the real part of the integrand you get a relatively large number (in this case 0.5). On the other hand, when you try to measure a frequency that is not present, as in the case when we look at $\hat{f}$ (5), the integrand oscillates enough so that the integral is very small.

The general situation may be a bit more complicated than this, but this in spirit is how the Fourier transform measures how much of an individual frequency is present in a function f (t).

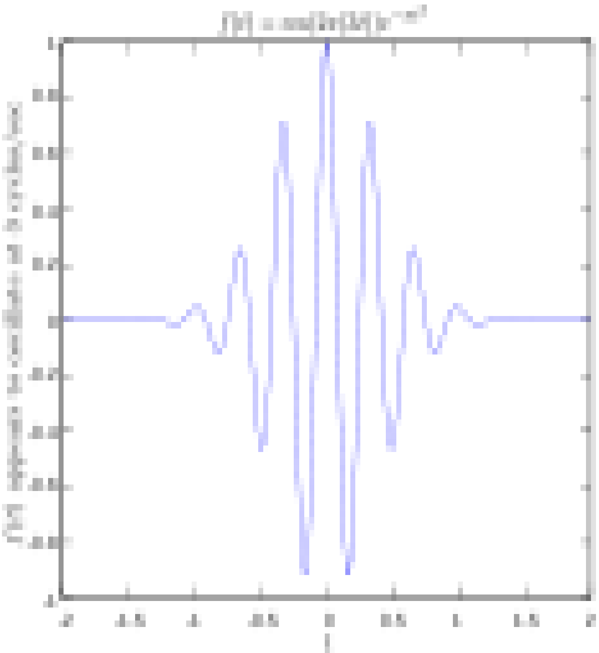


Fig 8: Original function showing oscillation 3 hertz.

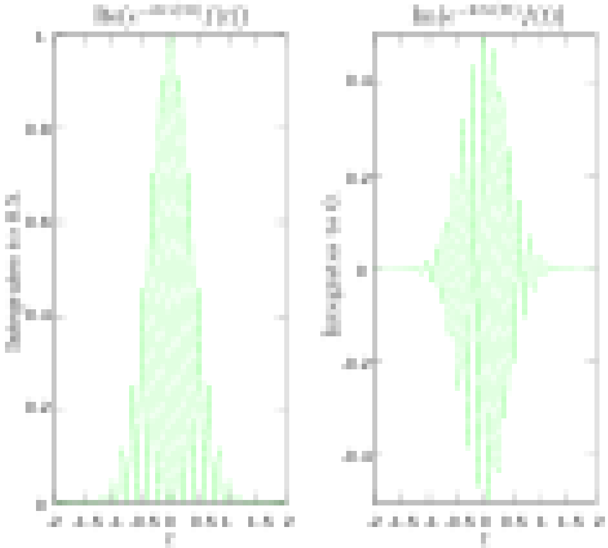


Fig 9: Real and imaginary parts of integrand for Fourier transform at 3 hertz

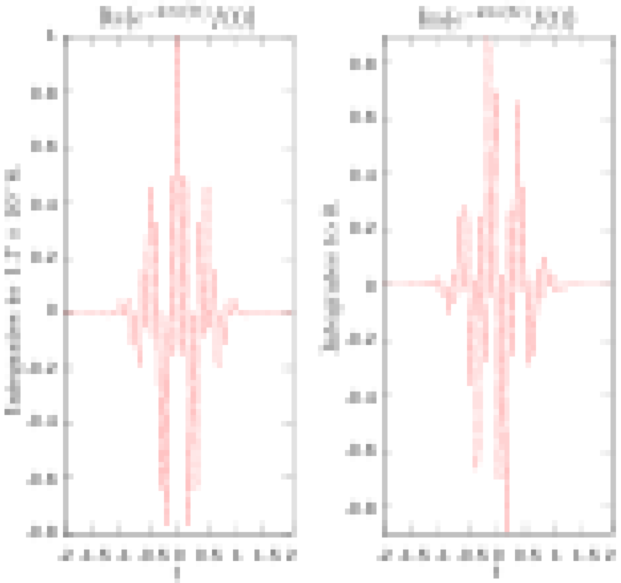


Fig 10: Real and imaginary parts of integrand for Fourier transform at 5 hertz

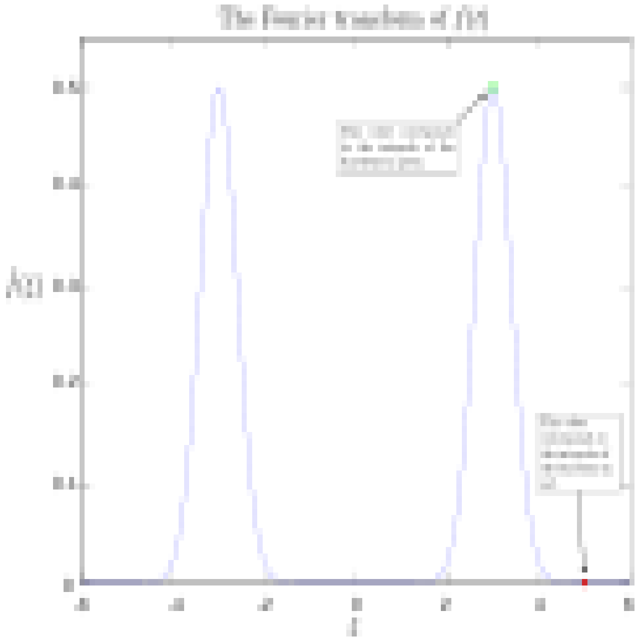


Fig 11: Fourier transform with 3 and 5 hertz labeled.

4.3.1 The Discrete Fourier Transform (DFT)

The **discrete Fourier transform (DFT)** is a specific kind of discrete transform, used in Fourier analysis. It transforms one function into another, which is called the frequency domain representation, or simply the DFT, of the original function (which is often a function in the time domain). The DFT requires an input function that is discrete. Such inputs are often created by sampling a continuous function, such as a person's voice.

The discrete input function must also have a limited (finite) duration, such as one period of a periodic sequence or a windowed segment of a longer sequence. Unlike the discrete-time Fourier transform (DTFT), the DFT only evaluates enough frequency components to reconstruct the finite segment that was analyzed. The inverse DFT cannot reproduce the entire time domain, unless the input happens to be periodic. Therefore it is often said that the DFT is a transform for Fourier analysis of finite-domain discrete-time functions.[9]

The input to the DFT is a finite sequence of real or complex numbers (with more abstract generalizations discussed below), making the DFT ideal for processing information stored in computers. In particular, the DFT is widely employed in signal processing and related fields to analyze the frequencies contained in a sampled signal, to solve partial differential equations, and to perform other operations such as convolutions or multiplying large integers. A key enabling factor for these applications is the fact that the DFT can be computed efficiently in practice using a fast Fourier transform (FFT) algorithm.

4.3.2 Fast Fourier Transform (FFT)

A **fast Fourier transform** (**FFT**) is an efficient algorithm to compute the discrete Fourier transform (DFT) and it is inverse. There are many distinct FFT algorithms involving a wide range of mathematics, from simple complex-number arithmetic to group theory and number theory; this article gives an overview of the available techniques and some of their general properties, while the specific algorithms are described in subsidiary articles linked below.

A DFT decomposes a sequence of values into components of different frequencies. This operation is useful in many fields (see discrete Fourier transform for properties and applications of the transform) but computing it directly from the definition is often too slow to be practical. An FFT is a way to compute the same result more quickly: computing a DFT of N points in the naive way, using the definition, takes O(N^2) arithmetical operations, while an FFT can compute the same result in only O(N log N) operations. The difference in speed can be substantial, especially for long data sets where N may be in the thousands or millions—in practice, the computation time can be reduced by several orders of magnitude in such cases, and the improvement is roughly proportional to N / log (N). This huge improvement made many DFT-based algorithms practical; FFTs are of great importance to a wide variety of applications, from digital signal processing and solving partial differential equations to algorithms for quick multiplication of large integers.

Chapter 5

Experimental Procedure of Bangla Consonants and Oral, Nasal Vowel of Male and Female Voice Analysis

This chapter involves discussion of analysis of speech sounds taking into consideration their method of production. The level of processing is between the digitized acoustic waveform and the acoustic feature vectors. The extraction of interesting information is an acoustic vector.

Speech analysis techniques provides a brief scientific overview of the speech signal analysis techniques involved with a particular focus on variable resolution spectral analysis, i.e.-emphasis ,variable resolution spectral analysis, Filter-bank analysis (Filter-bank speech analysis), Linear predictive analysis (Linear prediction speech analysis), LPC analysis, Deltas and normalization (Delta acceleration and feature normalization).

Our goal in processing the speech signal is to obtain a more convenient or more useful representation of the information carried by the speech signal. The required precision of this representation is dictated by the particular information in the speech signal that is to be preserved. For example, the purpose of the digital processing may be to facilitate the determination of whether a particular wave-form corresponds to speech or not. In a similar but somewhat more complicated vain' we may wish to make a 3-way classification as is whatever a section the signal .u voiced speech, unvoiced speech, or silence background not-r". The endeavor of speech analysis technique is to analyze the speech signal and estimate the parameters used for the given speech processing application. Since the parameter used in most of the speech processing application that are derived from frequency-domain representation, the main task is to compute the speech spectrum.

Before performing any type of digital processing on speech signal, it is first necessary to digitize the analog signal. For this the speech signal is filtered by a low pass with a cutoff frequency of W Hz" to avoid aliasing effect, It is then digitized by an analog –to digital converter at a sampling frequency higher than the Nyquist rate of 2w Hz. It is preferable to select the cutoff frequency w, to be high enough to get more information in the digitized speech signal which might be useful in a given speech processing application.

The value of cutoff frequency W, depend on the speech processing application and is typically in the range of 3- l0 kHz. The speech signal is non-stationary in nature, but it can be assumed to be stationary over short duration for the purpose of analysis. For the sake of validity of stationary assumption, it is necessary to choose as short an analysis segments as possible. Thus in practice, the speech signal is analyzed frame-wise, with a frame-rate of 50-100 frame V sec, and for each frame the duration of speech segment is taken to be 20-50 m-sec.

The short-time Fourier transform (STFT) of a speech signal has two components: the short-time magnitude. Spectrum and the short-time phase spectrum. It is traditionally believed that the short-time magnitude spectrum plays the dominant role for speech perception at small window durations (20-40 ms). However, recent perceptual studies have shown that the short-time phase spectrum can contribute as much to speech intelligibility as the short-time magnitude spectrum. In order to study spectral properties of speech signals, we shall find it convenient to formally introduce the concept of a time-varying Fourier representation of a signal. Frequency domain representation of speech information appears advantageous from two stand points. First, acoustic analysis of the vocal mechanism shows that the normal made or natural frequency concept permits concise description of speech sounds. Second, clear evidence exists that the ear makes a crude frequency analysis at an early stage in its processing. Presumably than, features salient in frequency analysis are important in production and perception, and consequently hold promise for efficient coding. Further, the vocal mechanism is a quasi-stationary source of sound. Its excitation and portal modes change with time.

5.1 Speech Materials

The experimental part consists of recording each of the consonants and oral,nasal vowels uttered three times at a normal speaking rate by true native Bangla male and female speakers. The recording was done in a quite room using a microphone interfaced with computer at a sampling rate of 48 kHz and 8 bit resolution. These digitized speech sounds are then down-sampled to l0 kHz for the purpose of analysis. The best recorded of these sounds is chosen for our work.

5.2 Linear Predictive Coding (LPC)

Linear predictive coding (LPC) is a tool used mostly in audio signal processing and speech processing for representing the spectral envelope of a digital signal of speech in compressed form, using the information of a linear predictive model. It is one of the most powerful speech analysis techniques, and one of the most useful methods for encoding good quality speech at a low bit rate and provides extremely accurate estimates of speech parameters.

Linear predictive coding (LPC) is defined as a digital method for encoding an analog signal in which a particular value is predicted by a linear function of the past values of the signal. It was first proposed as a method for encoding human speech by the United States Department of Defence in federal standard 1015, published in 1984. Human speech is produced in the vocal tract which can be approximated as a variable diameter tube. The linear predictive coding (LPC) model is based on a mathematical approximation of the vocal tract represented by this tube of a varying diameter. At a particular time, t, the speech sample s(t) is represented as a linear sum of the p previous samples. The most important aspect of LPC is the linear predictive filter which allows the value of the next sample to be determined by a linear combination of previous samples. Under normal circumstances, speech is sampled at 8000 samples/second with 8 bits used to represent each sample. This provides a rate of64000 bits/second. Linear predictive coding reduces this to 2400 bits/second.

5.3 LPC Model

The particular source-filter model used in LPC is known as the linear predictive coding model. It has two key components: analysis or encoding and synthesis or decoding. The analysis part of LPC involves examining the speech signal and breaking it down into segments or blocks. Each segment is than examined further to find the answers to several key questions:

- Is the segment voiced or unvoiced?

- What is the pitch of the segment?
- What parameters are needed to build a filter that models the vocal tract for the current segment?

LPC analysis is usually conducted by a sender who answers these questions and usually transmits these answers onto a receiver. The receiver performs LPC synthesis by using the answers received to build a filter that when provided the correct input source will be able to accurately reproduce the original speech signal. Essentially, LPC synthesis tries to imitate human speech production. Figure demonstrates what parts of the receiver correspond to what parts in the human anatomy. This diagram is for a general voice or speech coder and is not specific to linear predictive coding. All voice coders tend to model two things: excitation and articulation. Excitation is the type of sound that is passed into the filter or vocal tract and articulation is the transformation of the excitation signal into speech.

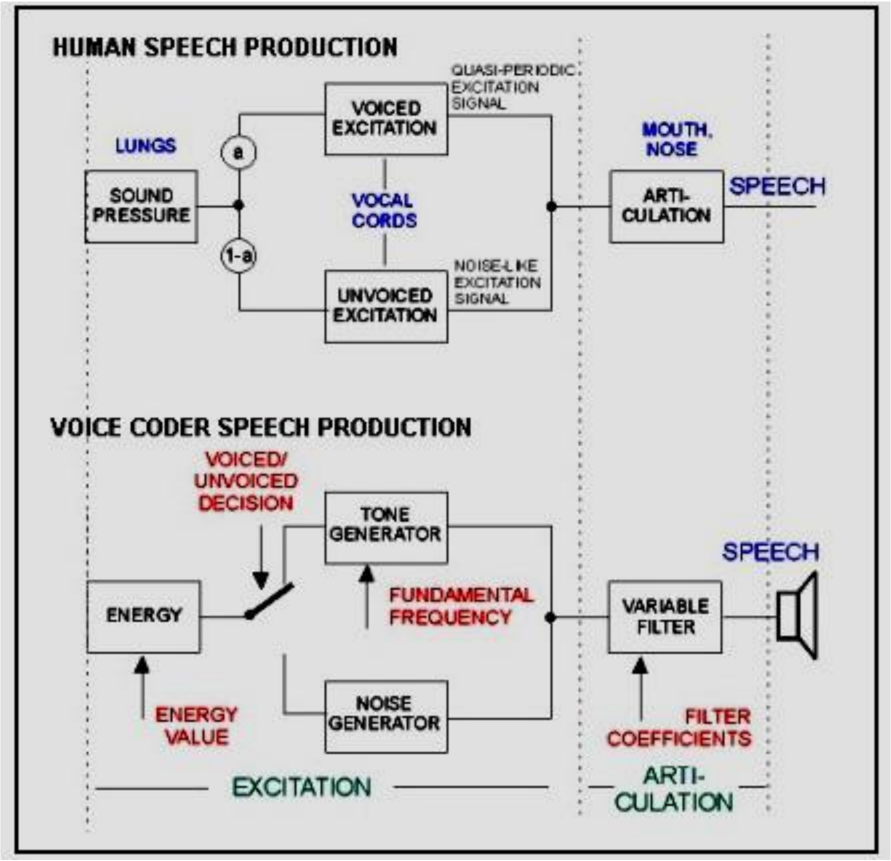


Fig 12: Human vs. Voice Coder Speech Production

Chapter 6

Result, Discussion and Conclusion

Result

Here table represents the numerical values of Consonants and oral, nasal vowels for both male and female of the extracted feature of LPC. The features include pitch period, power content and the first three formant frequencies.

Result for Bangla vowel (oral):

Speech Parameter	Bangla Vowel (oral)	Pitch Period (m/s)	Pitch Period (m/s)	Power (db)	Power (db)	Formants Freq.			Formants Freq.		
		Male	Female	Male	Female	Male			Female		
						F1	F2	F3	F1	F2	F3
LPC	অ	100	250.5	42	52.5	600	800	1000	990	1500	3101
	আ	105	250	45	50.01	700	1300	1600	999.1	1750	3500
	ই	140	270.5	45	55.1	350	1050	1400	950	1700	3250
	উ	115	280.5	39	53.5	700	850	1700	550	1500	3200
	এ	120	260	41	50.5	700	1200	1850	540	2110	3251
	ও	110	260	43	52.2	650	940	1700	500	850	3400
	অ্যা	110	260.5	49	50.5	700	900	1600	850	2451	2800

Fig13: Speech parameter of Bangla Vowels (oral) both male and female

Result for Bangla vowel (nasal):

Speech Parameter	Bangla Vowel (nasal)	Pitch Period (m/s)	Pitch Period (m/s)	Power (db)	Power (db)	Formants□Freq.			Formants Freq.		
		Male	Female	Male	Female	Male			Female		
						F1	F2	F3	F1	F2	F3
LPC	অ	130	270	45	56.2	550	800	950	560.5	1800	3300
	আ	120	270.5	43	53.5	700	1000	1250	570.5	1400.5	2050
	ই	125	300	36	54	800	1450	1800	560	1500	3250
	উ	127	310	42.5	55.5	800	1050	1650	990	1500.5	3240
	এ	120	270	40	52	600	1550	2080	900	1200	2250
	ও	125	270.5	45.5	53	800	1250	1670	760	2500	3500
	অ্যা	105	275.5	49	53.5	750	1350	1950	900	2850	3340.5

Fig14: Speech parameter of Bangla Vowels (nasal) both male and female

Result for Bangla consonant:

Speech Parameter	Bangla Consonant	Pitch Period (m/s)	Pitch Period (m/s)	Power (db)	Power (db)	Formants Freq.			Formants Freq.		
		Male	Female	Male	Female	Male			Female		
						F1	F2	F3	F1	F2	F3
LPC	ক	110	270	83	99	650	1050	1750	980	1201	3361
	খ	110	240.5	85	94	550	950	1100	900.5	1301	3250
	গ	90	240.5	84	94	530	1180	1630	260	1901	2450
	ঘ	100	240	78	93	640	1100	1650	250	1500	2900
	ঙ	105	325	82	98	700	1050	1750	900	2400	3801
	চ	100	290	75	99	500	950	1480	750	2250	3350
	ছ	100	245	77	91	500	900	1400	750	1900	3150
	জ	96	240	83	95	570	1330	1900	250	2600	3600
	ঝ	95	225	78	90	550	960	1500	250	2400	3501
	ঞ	110	260	77	98	840	1350	2050	270	3400	3851
	ট	100	270	82	96	690	1400	1850	1700	3400	3600
	ঠ	100	260	88	93	510	1200	1650	550	1851	3490
	ড	90	225	81	94.5	750	1300	1640	250	2000	3250
	ঢ	92	240	82	92	490	790	1500	250	1600	3050
	ণ	99	250	81	99	690	1110	1420	280	2010	2420
	ত	105	270	84	94	490	700	900	1000	3010	3800
	থ	102	280	85	92	610	1040	1440	900	2900	3500
	দ	103	250	83	95	700	1110	1350	250	1900	3400
	ধ	98	260	82	94.5	800	1430	1620	250	1500	3100
	প	99	310	85	95	830	1425	1800	500	1800	3400
	ফ	92	260	80	92	500	600	860	900	2100	3400
	ব	89	240	86.5	95	700	450	1750	350	1701	3150
	ভ	98	240	79	94	600	1290	1700	260	2500	3400
	ম	96	250	83	98	740	960	1550	250	2200	3400
	য	95	250	78	95	500	970	1595	260	2501	3800
	র	96	240.56	84	91	480	550	990	510	2000	3500
	ল	94	248.5	87	95	360	420	1020	250	1801	3000
	স	180	260	77	92	550	1250	1650	550	2000	2951
	হ	92	250	80	93	700	1150	1800	1050	2600	3500

Fig15: Speech parameter of Bangla Consonant for both male and female

Discussion

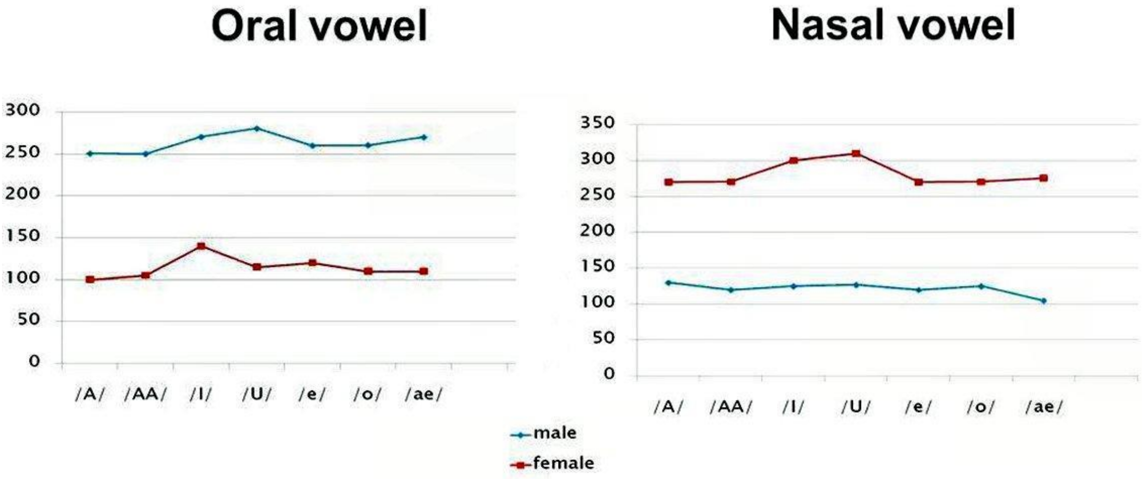
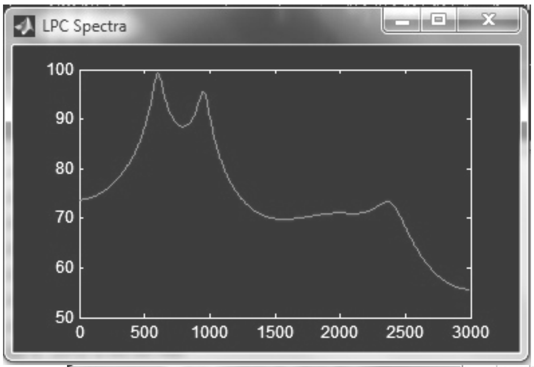


Fig.16 : Pitch Period



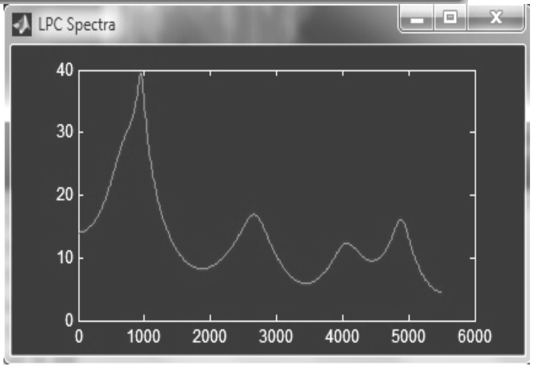
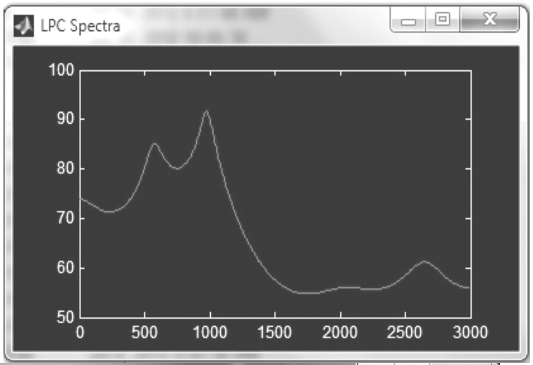


Fig17(a): LPC spectrum of ‘ক’ (male) Fig17(b): LPC spectrum of ‘ক’ (female)



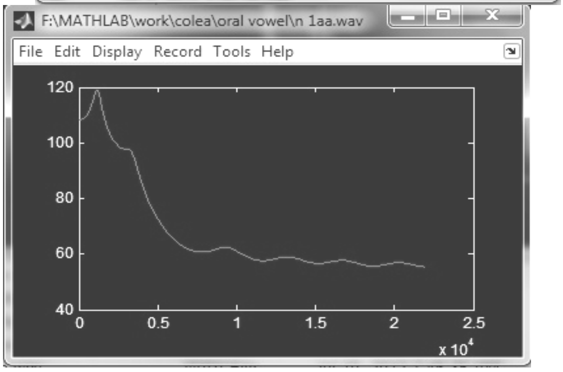
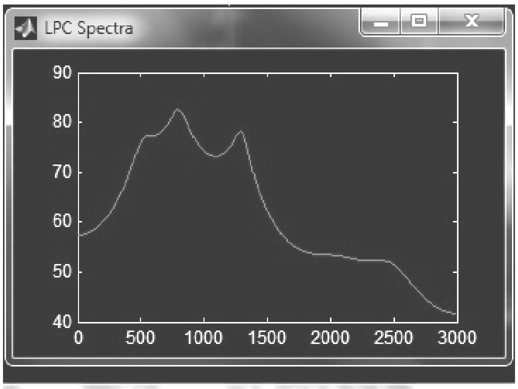


Fig18(a): LPC spectrum of ‘অ oral vowel (male) Fig18(b): LPC spectrum of ‘অ oral vowel (female)



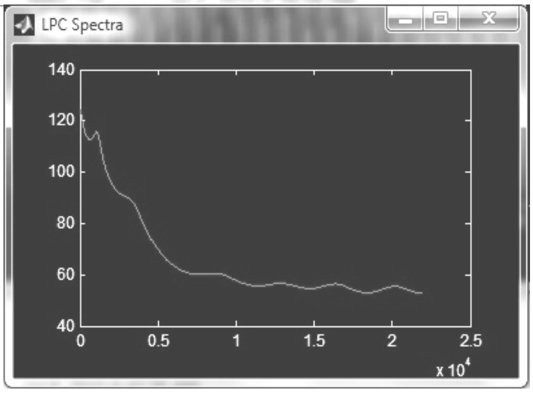


Fig19(a): LPC spectrum of ‘অ’ nasal vowel (male) Fig19(b): LPC spectrum of ‘অ’ nasal vowel (female)

Speech processing is the extraction of speech parameters form the speech signal for convenient representation. The ultimate aim of study Bangla vowels and consonants are to provide a complete Bangla based computer speech processing.

In this work, Bangla vowels and consonants in isolated utterance have been analyzed using a speech analysis method and the parameters obtained from the method have been compared. We analyzed Bangla vowels and consonant according to LPC (linear predictive coding).

The result obtained from the study

Pitch: pitch obtained by the LPC has acceptable comparable value for each vowels and consonants

Power: power obtained has acceptable comparable value

F1, F2, & F3: acceptable comparable value

The variation in result may be due to order of LPC, suppression of formants which are near to each other; LPC has a greater flexibility of formant extraction.

The result obtained from the study may be improved by

Recording voice in a noise free environment

Using of speech data from a phonetically trained person

This work can be extended to future research on other phoneme analysis or other speech unit analysis. These extracted parameters can be stored and later used for speech synthesis or recognition.

Conclusion

Speech processing is the extraction of speech parameters from the speech signal for convenient representation. The ultimate aim of study Bangla vowels and consonants are to provide a complete Bangla based computer speech processing. Now a days speech processing is very necessary. It is used for speech detection. It is also used in detective associations. By using pitch the detectives can find out the terrorist.

References

[1]. Department of Education and Early Childhood Development – Voice Care for Teacher Program

[2]. www.phon.ucl.ac.uk/courses/spsciaccoustics

[3]. Alku, P., (1991). "Glottal Wave Analysis With Pitch Synchronous Iterative Adaptive Inverse Filtering", in Proceedings of Second European Conference on Speech Communication and Technology, Genova, Italy.

Baken, R.J. and Orlikoff, R.F. (2000). Clinical Measurement of Speech and Voice. 2nd ed. Singular Publishing Group, San Diego, California.

[4]. Gillian Brown and George Yule, Discourse Analysis (New York: Cambridge University Press, 1983).

[5]. Ashby, P. (2005). *Speech sounds*. London: Routledge, Davenport, M., & Hannahs, J. (2006). *Introducing phonetics and phonology*. London: Arnold, Garn-Nunn, P., & Lynn, J. (2004). *Calvert's descriptive phonetics* (3rd ed.). New York: Thieme.

[6]. Gershenfeld, Neil A. (1999). *The Nature of mathematical Modeling*. Cambridge University Press. Wagner, Thomas Charles Gordon (1959). *Analytical transients*. Wiley.

[7]. Matt Pharr and Greg Humphreys, *Physically Based Rendering: From Theory to Implementation*, Morgan Kaufmann, July 2004

[8]. Boashash, B., ed. (2003), *Time-Frequency Signal Analysis and Processing: A Comprehensive Reference*, Oxford: Elsevier Science.

[9]. Brigham, E. Oran (1988). *The fast Fourier transform and its applications*. Englewood Cliffs, N.J.: Prentice Hall.

[10]. Cooley, James W.; Tukey, John W. (1965). "An algorithm for the machine calculation of complex Fourier series". *Math. Comput.* **19** (90): 297–301.

[11]. V. Hardman and O. Hodson. Internet/Mbone Audio (2000) 5-7.

Scott C. Douglas. *Introduction to Adaptive Filters*, Digital Signal Processing Handbook (1999) 7-12.