

Estimation of AR Model of Vocal folds for Speaker Identification

by

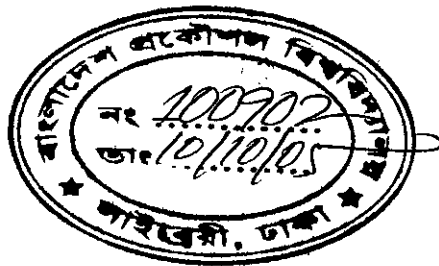
Kazi Zamir Uddin Ahmad

A thesis

Submitted to the Department of Electrical and Electronic Engineering
in partial fulfillment of the requirements for the degree

of

MASTER OF SCIENCE IN ELECTRICAL AND ELECTRONIC
ENGINEERING

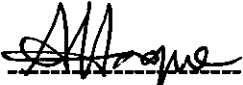


DEPARTMENT OF ELECTRICAL AND ELECTRONIC ENGINEERING
BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY


May 2005

The thesis titled, "Estimation of AR model for Vocal folds for Speaker identification," submitted by Kazi Zamir Uddin Ahmad, Roll No. 100106250P, Session October-2001 has been accepted as satisfactory in partial fulfillment of the requirements for the degree of **Master of Science in Electrical and Electronic Engineering** on 29th May 2005.


BOARD OF EXAMINERS

1. 


Dr. Anisul Haque
Professor
Department of Electrical and Electronic Engineering
BUET, Dhaka-1000.

Chairman
(Supervisor)
2. 


Dr. M. Rezwon Khan
Vice chancellor(Designate)
UIU, Dhaka.

Member
(Co-Supervisor)
3. 

Dr. Mohammad Ali Choudhury
Professor and Head
Department of Electrical and Electronic Engineering
BUET, Dhaka-1000.

Member
(Ex-Officio)
4. 

Dr. Md. Kamrul Hasan
Professor
Department of Electrical and Electronic Engineering
BUET, Dhaka-1000.

Member
5. 


Dr. Md. Abul Kashem Mia
Professor
Department of Computer Science and Engineering
BUET, Dhaka-1000.

Member
(External)

DECLARATION

I hereby declare that this thesis has been completed by me and it or any part of it has not been submitted elsewhere for award of any degree or diploma.

Signature of the Candidate



(Kazi Zamir Uddin Ahmad)

CONTENTS

List of Figures	VII	
List of Tables	IX	
List of Symbols	X	
Abstract	XI	
CHAPTER 1	INTRODUCTION	1
1.1	Background and Present state of the Problem	1
1.2	Objectives of the Thesis	3
1.3	Thesis Outline	3
CHAPTER 2	HUMAN SPEECH MODELING	5
2.1	Introduction	5
2.2	Anatomy and Mechanism of Speech Production	5
2.3	Modeling of Speech Production	12
2.3.1	One-mass Model	14
2.3.2	Two-mass Model	15
2.3.3	Multiple-mass Model	16
2.3.4	Continuum Model	17
2.3.5	Ribbon Model	19

	2.3.6 Body-cover Model	20
	2.3.7 Interactive Model (electrical analog version)	22
	2.3.8 Parametric Model	26
	2.4 Conclusion	30
CHAPTER 3	EXTRACTION OF VOCAL FOLDS PARAMETERS FOR SPEAKER IDENTIFICATION USING INVERSE FILTERING	31
	3.1 Introduction	31
	3.2 Basis for the Proposed Model	31
	3.3 Cepstrum	34
	3.3.1 LPCC	37
	3.4 Estimation of Proposed Model Parameters	39
	3.4.1 Speech Framing	39
	3.4.2 Pre-emphasis	40
	3.4.3 Windowing	41
	3.4.4 Speech Features	42
	3.4.5 Linear Prediction Cepstral Coefficient	42
	3.4.6 AR Model of Vocal Folds	44
	3.5 Speaker Identification	45
	3.5.1 Vector Quantization	46
	3.5.2 Feature Matching	53
	3.6 Conclusion	56
CHAPTER 4	RESULTS	57
	4.1 Introduction	57
	4.2 Data Acquisition	57

	4.2.1 Selected voiced sounds and pitch	57
	4.2.2 Technical Specifications	58
	4.3 Procedural specification	59
	4.4 Results	59
	4.5 Discussion	62
CHAPTER 5	CONCLUSIONS	65
	5.1 Discussions	65
	5.2 Limitations	65
	5.3 Suggestions for Further Work	66
REFERENCES		68
APPENDIX	Typical speech utterance wave forms for different vowels	72

LIST OF FIGURES

Figure 2.1	Human voice production system	6
Figure 2.2	Schematic diagram of voice production	7
Figure 2.3	Anatomy of Larynx and the vocal folds	8
Figure 2.4	Place of Larynx and the vocal folds in the throat.	9
Figure 2.5	Schematic diagram of a normal cycle of vocal folds vibration	10
Figure 2.6	Vocal folds in different states	11
Figure 2.7	Discrete-time model of the voice production	13
Figure 2.8	Schematic diagram of One- mass model	15
Figure 2.9	Schematic diagram of Two-mass model	16
Figure 2.10	Schematic diagram of 16-mass model	17
Figure 2.11	Schematic diagram of Continuum model	18
Figure 2.12	Schematic diagram of ribbon model	19
Figure 2.13	Schematic diagram of the body-cover model	21

Figure 2.14	Glottal area and glottal air flow during phonation	23
Figure 2.15	Electrical analogous of vocal folds	24
Figure 2.16	Electrical analogous of vocal system	25
Figure 2.17	The derivative of glottal flow signal (LF model)	27
Figure 2.18	Glottal flow signal	28
Figure 3.1	Block diagram of the vocal system for voiced sounds	32
Figure 3.2	Component of speech in speech spectrum	35
Figure 3.3	Component of speech in speech cepstrum	36
Figure 3.4	Computation of the cepstrum	37
Figure 3.5	“Short term analysis” of speech utterance	40
Figure 3.6	The frequency response of a pre-emphasis filter	41
Figure 3.7	Schematic of speaker location for three different speakers in a two dimensional feature space	43
Figure 3.8	Inverse filtering to get vocal folds output	45
Figure 3.9	Feature vectors along with code vectors in a two dimensional vector space	47
Figure 3.10	Schematic diagram of LBG technique	52

Figure 3.11	Schematic diagram of matching technique	54
Figure 3.12	Schematic diagram of speaker identification process	55

LIST OF TABLES

Table 4.1	Close set speaker identification through vowels using LPCC	60
Table 4.2	Close set speaker identification through vowels using proposed model	60
Table 4.3	Open set speaker identification through vowels using LPCC	61
Table 4.4	Open set speaker identification through vowels using proposed model	62

LIST OF SYMBOLS

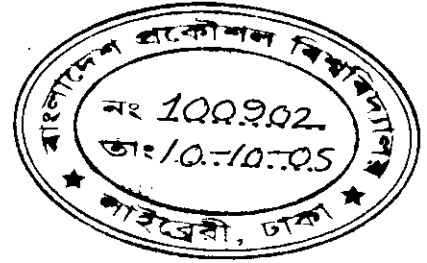
P_s	Pressure of lungs to the vocal folds	14
U_g	volume velocity	14
S_c	coupling stiffness	15
T_m, T_v	longitudinal tensions	17
ψ	displacement vector	17
γ	longitudinal stress	17
Q_a	abduction quotient	20
Q_s	shape quotient	20
Q_b	bulging quotient	20
Q_p	phase quotient	20
F_o	fundamental area frequency	20
Y_g	glottal conductance	22
C_o	oral compliance	23
U'	glottal flow derivative	28
T_o	length of the entire pulse	28
T_p	length of the time that $U > 0$	28
T_e	time of the maximum negative value of U'	28
E_e	value of the maximum negative U'	28
T_a	experimentally determined effective duration of the return phase	28
p	prediction order	38
a_k	prediction coefficients	38
e	prediction order	38
α	Filter coefficient	41

ABSTRACT

In this thesis, speaker identification is done by the AR model parameters of the vocal folds. Speaker identification needs extraction of speaker discriminative features. Mel-frequency cepstral coefficients (MFCC) and Linear predictive cepstral coefficients (LPCC) are well known cepstral techniques which extract speaker discriminative vocal tract properties from speech signal for speaker identification purpose. On the other hand, the vocal folds properties of a speaker can also be used for this purpose as vocal folds vary person to person. But in this case the correct modeling of vocal folds is essential. AR model parameters of the vocal folds is used here as the speaker distinctive features. Vocal folds properties are found by inverse filtering the cepstral of the output speech by the vocal tract properties related LPCC. These model parameters called speaker features are then used to generate the so-called codebook of a speaker by the well established vector quantization technique. Codebooks generated in this way are then used to find the speaker identity using feature matching technique. The result of the proposed model found here is significantly better than that of the previous model for voiced sound.

CHAPTER 1

INTRODUCTION



1.1 Background and Present State of the Problem

The motivation for understanding the mechanism of speech production lies in the fact that speech is the human being's primary means of communication. Through developments in acoustic theory, many aspects of human voice production are now understood. There are areas such as non-linearity of vocal fold vibration, vocal-tract articulator dynamics, knowledge of linguistic rules, and acoustic effects of coupling of the glottal source and vocal tract that continue to be studied. The continued pursuit through this field with the tools of basic speech analysis has provided new and more realistic means of performing speech synthesis, coding, and recognition.

Early attempts for modeling and understanding speech production resulted in mechanical speaking machines. Modern advances have led to the electrical analog devices, and ultimately computer-based systems. One of the earliest documented efforts to produce artificial speech was by C.G. Kratzenstein [1] in 1779 in which he attempted to artificially produce and explain the differences among the five vowels. He constructed acoustic resonators similar in shape to the human vocal tract and excited them with a vibrating reed, which, like the vocal folds, interrupted an air-stream. Further developments in mechanical speech modeling and synthesis continued into the 1800s and early 1900s. One of the first all-electrical networks for modeling speech sounds was developed by J.Q. Stewart [2].

Subsequently, research works on speech modeling and synthesis take a new turn with the development of computer. In many regards, advancement in speech modeling led to the development of better speech coding and synthesis methods - the scientific fields of *Automatic Speech Recognition and Computer Voice Response*. During the latest 20

years, however, several new research areas in computer science have been founded, that focus on solving problems by mimicking the nature. This has in turn led to a number of novel approaches regarding speech understanding and speaking machines. However, the level of expectation from these novel approaches gradually declined over the time, nevertheless, the perpetual crave to unveil new things from this field resulted further new techniques, some of which are very promising [3].

The main problem of speech analysis is to determine the internal structure and movement from speech or other measured data [4]. One aspect of acoustic phonetic, for example, deals with inferring articulatory shapes and movement from the speech waveform or its spectrum. Usually, the success of this inverse problem depends on the ability to accurately model the underlying biophysical processes. Constructing simple predictive models of phonatory acoustics, tissue mechanics, and glottal aerodynamics is formed to be very difficult. By stepwise improvement, some models have been developed that can be used for simulating and thus improving the behavior of the speaker identification tools. One of these early models is the recognized source-filter model [5], which treats the vocal fold as the source of the voice and vocal tract as the filter and they work independently during the voiced sounds. Most of the work in speech modeling and synthesis is based on this model. Now-a-days this model is also used in voice quality enhancement, speaker identification, voice pathology classification, speech coding and synthesis [6-8]. As the model is simple and efficient, scientists working in speech analysis are devoted to improve the source model (vocal folds) and the filter model (vocal tract) separately for the above purposes. For the source (vocal folds), models like: LF model [9], R++model [10], KLGOTT88 model [11], two-mass model [12], multi-mass model [13], continuum model [14-15], four-parameter model [16, 17], body-cover model (three mass model) [18] are used for the above purposes. One of the major issues in speech analysis area is speaker identification, which is widely used around the world and found their appropriate places in the industry through the assistance of opportune research. Even though a lot of efforts has already been made in

this area [19, 20], still there is a lot of room for the improvement and this makes the research in this area up going.

1.2 Objective of the Thesis

In this work, the AR model of vocal fold is studied for speaker identification. So far, the models used to implement the behavior of vocal fold are of two types. Parametric models use some parameters to produce the output similar to that of the vocal folds, which works for synthesis, coding and pathology. Physical models emulate the physical behavior of the vocal folds to understand its dynamic property. Appropriate knowledge about these models gives a proper picture about the vocal folds characteristics and its contribution in speech output waveform. This helps to recognize the contribution of the vocal folds to the speech output. Conventionally, vocal tract properties are used in speaker identification and vocal folds properties are not yet investigated thoroughly for this purpose. In this thesis work, we try to estimate the vocal folds parametric model and investigate it for speaker identification purpose using the available tools. It is to be noted that the AR model parameters of the vocal folds are used here as the speaker distinctive features. This work identifies the speaker in less expensive calculations and produces competitive results.

1.3 Thesis Outline

Certain anatomical properties of vocal folds and vocal tract carry the information of speaker identity. These properties are very difficult to extract through the size, shape, tension, dimension of these limbs. But it can be realized through the output of the vocal system. By using some establish theory; it can be extracted in some parametric form. This thesis proposes a model for vocal folds, using well established *Linear Predictive Coding* and *inverse filtering* which conveys the speaker discriminating properties in an explicit form.

In the following chapter (chapter 2) the anatomy of the voice production system has been reviewed first. The vocal system is then ramified into three subsystems according to the contribution they make in the output of the system and then each one is studied in brief. The first subsystem- the vocal folds, is then studied thoroughly through its existing models.

In the next chapter (chapter 3) we try to establish the process to extract the vocal folds parametric model from the speech output. These model parameters are then used to generate the so-called *codebook* of the speaker using *vector quantization* method. Finally, the test speech is matched with the codebook to identify the speaker using *quantization distortion*.

In the chapter 4, the data acquisition for this system is presented and specifications related to the data acquisition, as well as the identification process, is stated. Speaker identification rate for proposed model as well as conventional vocal tract model are shown later. Finally, a discussion is made about the results found for both models.

The concluding chapter (chapter 5) provides a comprehensive summary of the whole work followed by a brief discussion on the limitations of this work and some suggestions for future works.

CHAPTER 2

HUMAN SPEECH MODELING

2.1 Introduction

In the early days speech modeling was mainly related to the construction of speech machines, which emulate the human speech. With the development of computer, continuous research in this field has chronologically discovered various aspects of speech modeling. However, the procedure to estimate a speech model has not been much changed over the years. The essential and primary aspect for this is to know thoroughly the physical process of human speech production (Anatomy of speech production) and then investigate it through its existing models. This will give the necessary working knowledge and objective to further analyze and develop human speech models. In this chapter, the anatomy of the human speech production is discussed at the beginning. Later some of the existing models, specially the model of vocal folds are examined to some extent. This knowledge gives a good understanding of speech modeling techniques and helps to extract the correct AR model parameters of vocal folds (proposed model) for speaker identification purpose.

2.2 Anatomies and Mechanism of Speech Production

The speech waveform is an acoustic sound pressure wave that originates from voluntary movements of anatomical structures, which make up the human speech production system. The components of this anatomical structure are the lungs, trachea (windpipe), larynx (where the vocal fold resides), pharyngeal cavity (throat), oral cavity (mouth), and the nasal cavity (nose). Some finer components of this, obviously critical for speech production, are the vocal folds, velum, tongue, teeth and lips [21] as shown in the Fig-2.1. All these components, called *articulators* by the speech scientists, move to different

positions to produce various sounds. Based on number of parts involved in the production, speech sounds can be divided into voiced and unvoiced speech [22, 23].

Acoustic speech output in human and many nonhuman species are commonly considered to result from a combination of the *source* of sound energy (e.g. the vocal folds) modulated by the *filter* (vocal tract) function determined by the shape of the supralaryngeal vocal tract. This combination results in a shaped spectrum with broadband energy peaks. This model is often referred to as the “source-filter theory of speech production” and stems from the

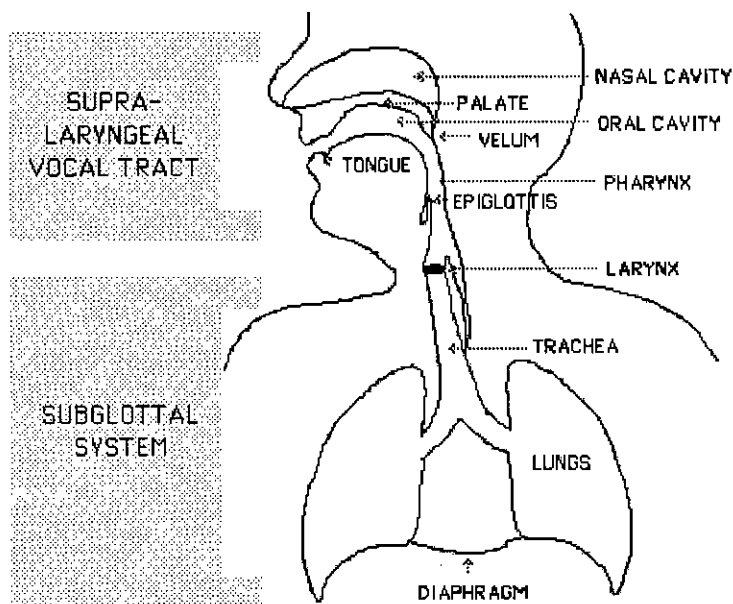


Figure 2.1: Human voice production system.

experiments of Johannes Muller [22] in which a functional theory of phonation was tested by blowing air through larynges excised from human cadavers. In this model the

source of acoustic energy is at the larynx (vocal folds) and the supralaryngeal vocal tract serves as a variable acoustic filter whose shape determines the phonetic quality of the sound.

With minor modification to the source-filter model, the speech production system can be divided into three stages: first stage is the sound source production, second stage is the articulation by vocal tract, and third stage is the sound radiation or propagation from the lips and /or nostrils [14]. A *voiced sound* is generated by vibratory motion of the vocal folds powered by the airflow generated by expiration. The main acoustic filter (pharyngeal cavity, vocal cavity and nasal cavity) is then excited and loaded at its main output by radiation impedance due to the lips. Another type of sound called *unvoiced sound* is produced by the turbulent of airflow passing through a narrow constriction in the vocal tract [22, 24]. A simplified acoustic model illustrating these ideas (source filter model) is shown in Fig. 2.2.

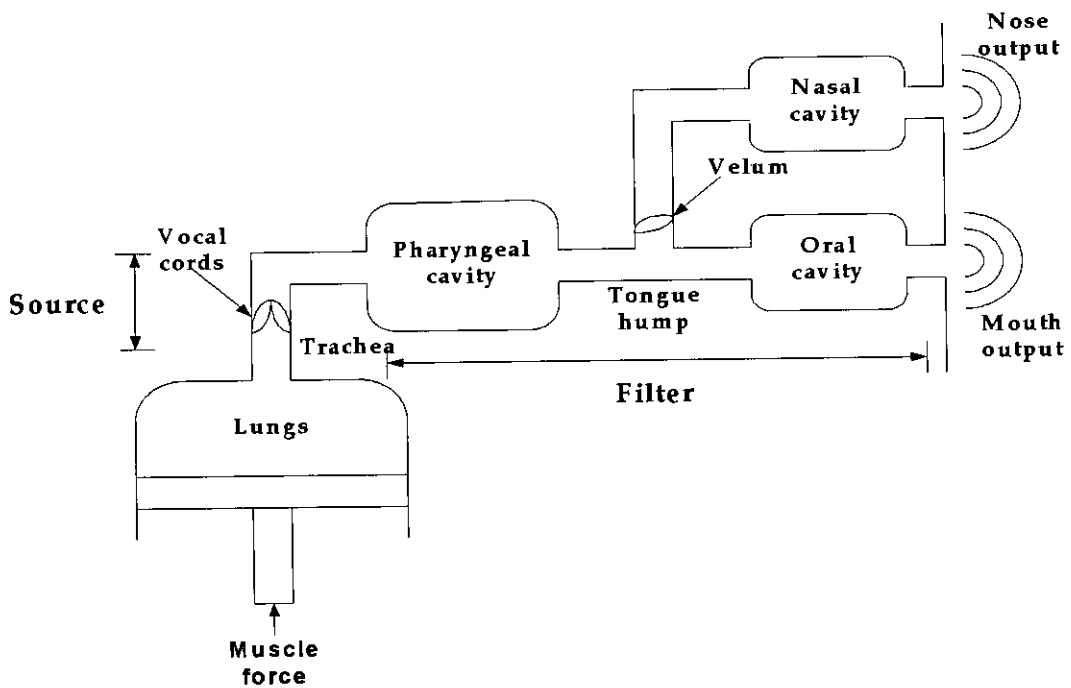


Figure 2.2: Schematic diagram of voice production.

When the larynx serves as a source of sound energy, voiced sounds are produced by a repeating sequence of events. First, the vocal folds are brought together (adduction), temporarily blocking the flow of air from the lungs and leading to a increased subglottal

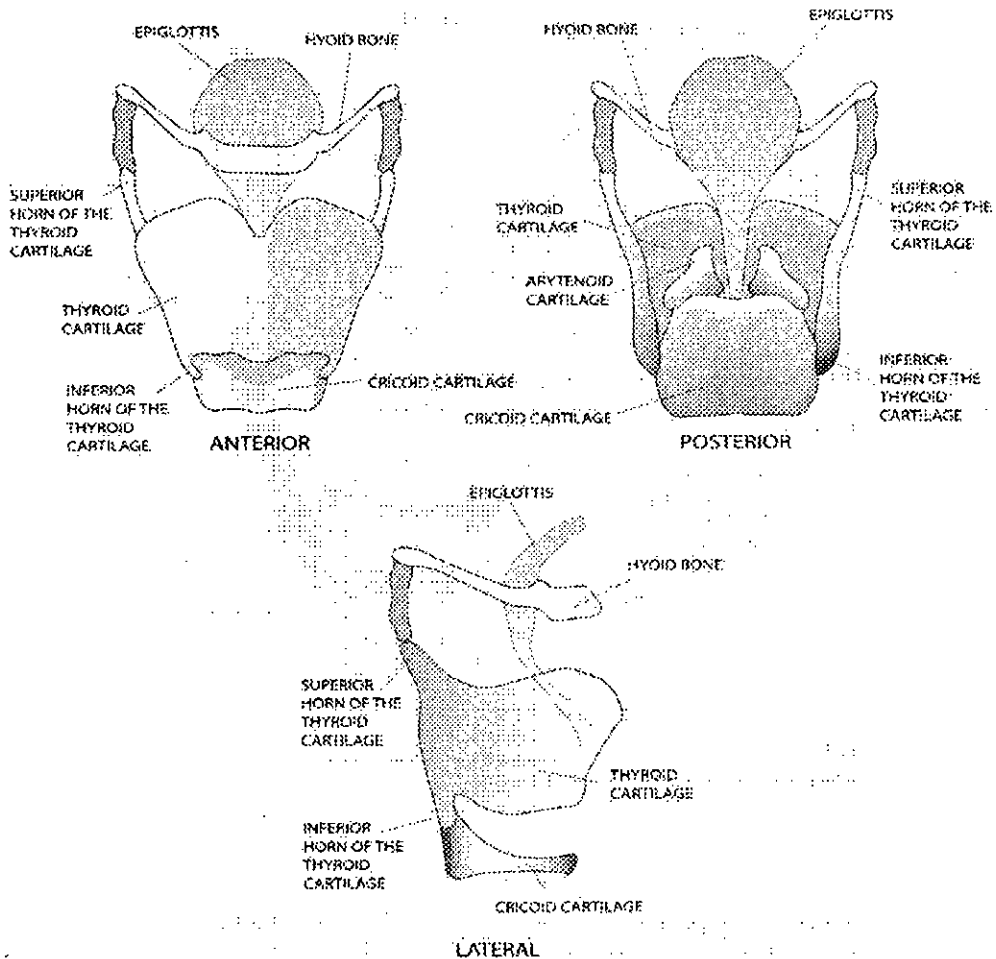


Figure 2.3: Anatomy of Larynx and the vocal folds

pressure. When the subglottal pressure becomes greater than the resistance offered by the vocal folds, they open again. The folds are then closed rapidly due to a combination of factors, including their elasticity, laryngeal muscle tension, and the Bernoulli Effect. If the process is maintained by a steady supply of pressurized air, the vocal folds will continue to open and close in a quasiperiodic fashion. As they open and close, puffs of

air flow through the glottal opening. The frequency of these pulses determines the fundamental frequency of the laryngeal source and contributes to the perceived pitch of the produced sound.

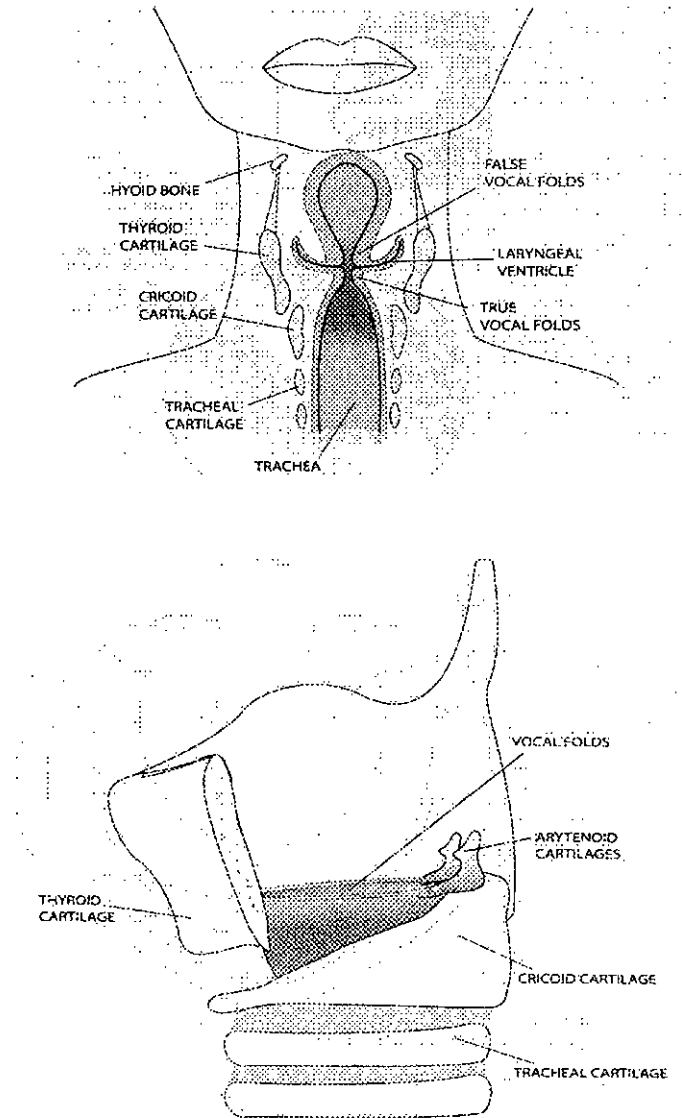
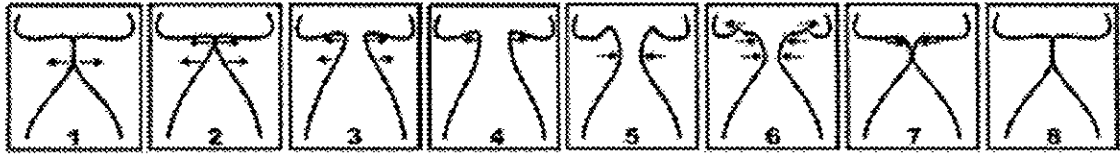
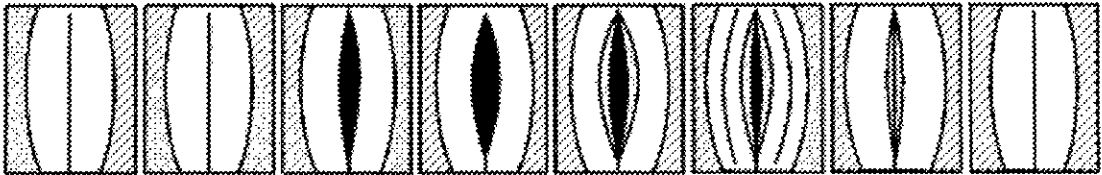


Figure 2.4: Place of Larynx and the vocal folds in the throat.



(a)



(b)

Figure 2.5: Schematic diagram of a normal cycle of vocal folds vibration.

(a) side view (b) top view

The rate at which the vocal folds open and close during phonation can be varied in a number of ways and is determined by the tension of the laryngeal muscle and the air pressure generated by the lungs. The shape of the spectrum is determined by the details of the opening and closing movement, and is partly independent of fundamental frequency.

Myoelastic-aerodynamic theory of vocal folds has construed the changes of vocal folds frequency better. It has two parts. The myoelastic part arises from the fact that changes in the frequency of vibration of the vocal folds occur as the muscles (myo) of the vocal folds change the elasticity and tension. The mass of the vocal folds also affects the vocal fold vibratory frequency. The frequency of vibration is lowered as the vocal folds become shorter and thicker. Again, when folds are stretched to make them tenser they vibrate at a higher frequency as they become longer and thinner. Elastic folds vibrate

faster because they are able to “bounce” back at a more rapid rate. In a nutshell tense folds vibrate faster than slack folds.

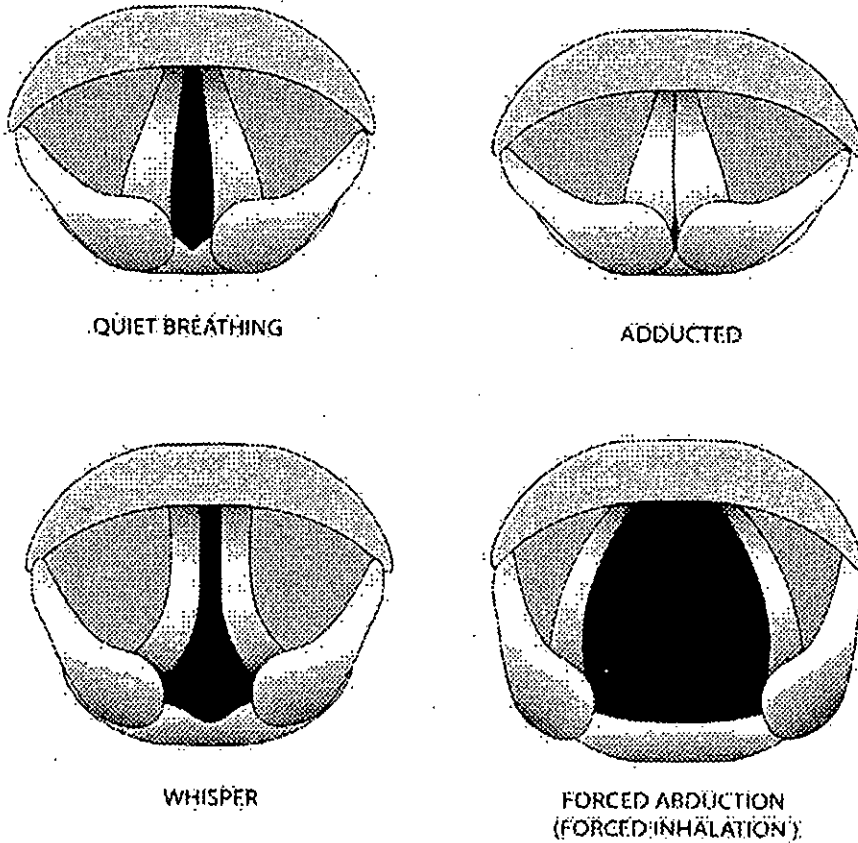


Figure 2.6: Vocal folds in different states.

The muscles regulate the thickness and tension of the vocal folds. The aerodynamic part of the theory says that the driving force for vocal folds vibration is airflow. The air expelled from the lungs activates the vibration of the vocal folds. The Bernoulli effect is one factor that affects the vibration of the vocal folds; and another is the recoil force of the vocal folds.

The supralaryngeal vocal tract, consisting of both the oral and nasal airways, can serve as a time-varying acoustic filter that suppresses the passage of sound energy at certain frequencies while allowing its passage at other frequencies at which local energy maxima are sustained by the supralaryngeal vocal tract. These local maxima are determined, in part, by the overall shape, length and volume of the vocal tract. The detailed shape of the filter (transfer function) is determined by the entire vocal tract serving as an acoustically resonant system combined with losses including those due to radiation at the lips. The formant frequencies, corresponding to the peaks in the function, represent the center points of the main bands of energy that are passed by a particular shape of the vocal tract. The flexibility of the human vocal tract, in which the articulators can easily adjust to form a variety of shapes, results in the potential to produce a wide range of sounds.

Each voiced vowel sound has its own higher characteristic frequency components of harmonics (frequency spectrum) due to the pharynx and oral cavity acting as resonators to reinforce and absorb different frequencies [25]. These frequencies are always higher than that of the fundamental one and are called *formants*. Each vowel sound has its own characteristic formants just as each musical instrument has, and hence the human voice is recognized as a human voice because of its special characteristics. Therefore, we can conclude that each individual voice has its own frequency spectrum so that an analysis of an individual's voice could be used for person identification.

2.3 Modeling of Speech Production

The above discussion of human speech production reveals three separate areas for modeling. These include the source excitation, vocal-tract shaping, and the effect of speech radiation. For example, a single phoneme such as a vowel, modeled over finite time, can be represented as the product of the following three functions:

$$S(\omega) = U(\omega) \cdot H(\omega) \cdot R(\omega) \quad (2.1)$$

Where $S(\omega)$, is the Fourier transform of the vowel sound (speech output), $U(\omega)$, representing the voice waveform (source excitation), $H(\omega)$, representing the dynamics of the vocal tract, and $R(\omega)$, radiation effects. where the input of the system is considered as the impulse train. For unvoiced excitation the voice source will be replaced by noise source.

As implied by the representation in equation (2.1), the majority of modern speech modeling techniques assume that these components are linear and separable. Accordingly, the speech production system is assumed to be the concatenation of subsystems: vocal folds (source), vocal tract (filter) and radiation from lips and muscles, with no dependency between two adjacent subsystems. A discrete-time model of this concept is shown in the Fig-2.7.

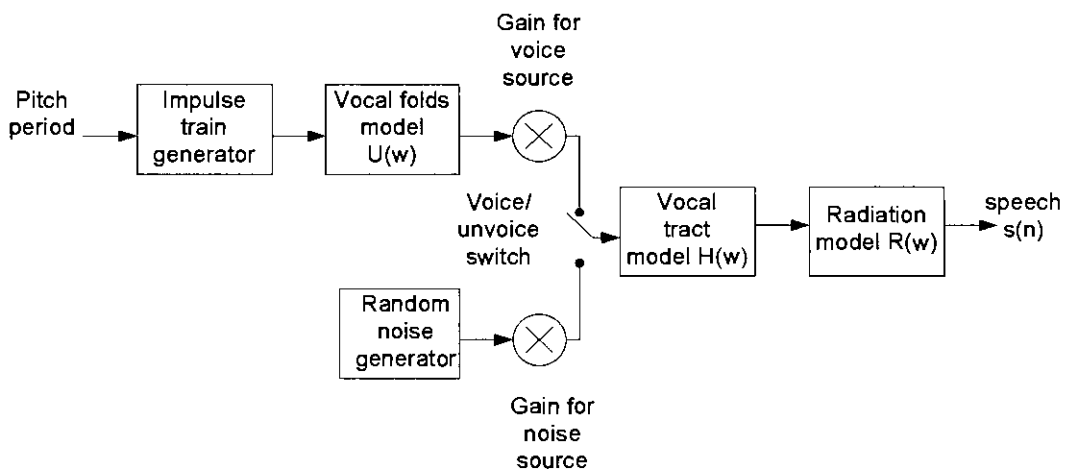


Figure 2.7: Discrete-time model of the voice production

The vocal-tract model $H(\omega)$ and the radiation model $R(\omega)$ are excited by a discrete-time glottal excitation signal $u(n)$. During unvoiced speech activity, the excitation source is a flat spectrum noise source embodied by a random noise generator. During periods of

voiced speech activity, the excitation uses an estimation of the local pitch period to set an impulse train generator that drives the vocal folds model $U(\omega)$.

Throughout the history of speech modeling, voiced excitation has always received more research attention than its unvoiced counterpart. This is due to the fact that studies in speech perception suggest that accurate modeling of voiced speech is crucial for natural-sounding speech both in coding and in synthesis application. Hence, the development of this area is still open. The objective of this thesis is to explore it even further – to identify a speaker with the help of the voice source (vocal folds) model so that its characteristic for different speakers will be quite vivid. So far, two types of models are used to implement the vocal folds: physical model and parametric model [26]. Parametric models fit the glottal signal with piecewise analytical functions, using a small number of parameters, such as LF model [9] characterizes one cycle of the flow derivative using as few as four parameters. Physical models describe the glottal system in terms of physiological quantities. These models capture the basic non-linear mechanisms that initiate self-sustained oscillations, and can simulate subtle features; however they involve many parameters and are not suitable for identification purposes [27]. Some of these existing models are briefly discuss in the subsequent article.

2.3.1 One-mass Model

This model represents the glottal source as lumped oscillator and the sub-glottal system as an air reservoir with pressure P_s that provides air flow with the volume velocity (U_g) [8]. The lumped oscillator, model by a single mass-spring, is driven by airflow from the lungs. This model is simple and it has low computational burden. In this model source-tract (vocal folds and vocal tract) interaction is taken into account whereas phase-difference between the motions of folds edges is ignored.

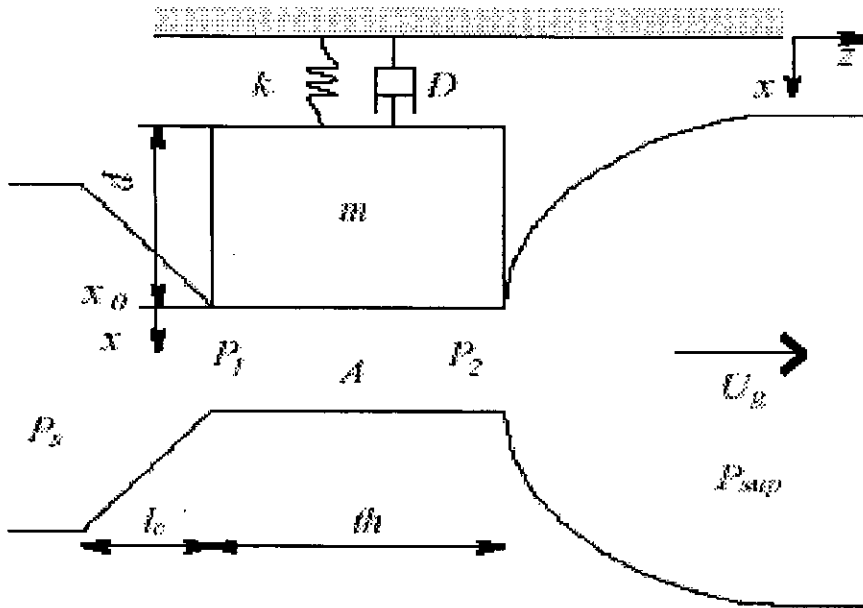


Figure 2.8: Schematic diagram of One- mass model.

The glottal area and volume velocity can be simulated from this model. It is shown in the figure 2.8.

2.3.2 Two-mass Model

In the two-mass model, the vocal folds are divided into an upper and a lower mass [28]. This is because of the anatomic and functional division between the mucosa and the vocalis. Each part consists of a simple mechanical oscillator having mass, spring, and damper (m , s , and r) as shown in the fig-2.9. The springs and dampers represent the elastic properties of the folds and dissipative forces such as viscosity and friction respectively. The coupling stiffness S_c represents the interaction between the two masses. The coupling stiffness represents the fact that as one of the masses is displaced relative to the other, there is a force tending to restore the masses to their equilibrium position relative to one another. The two-mass model considers the phase-difference

between the motions of folds edges so that the simulation of glottal properties is more realistic. With a reasonable computational burden natural speech can be produced.

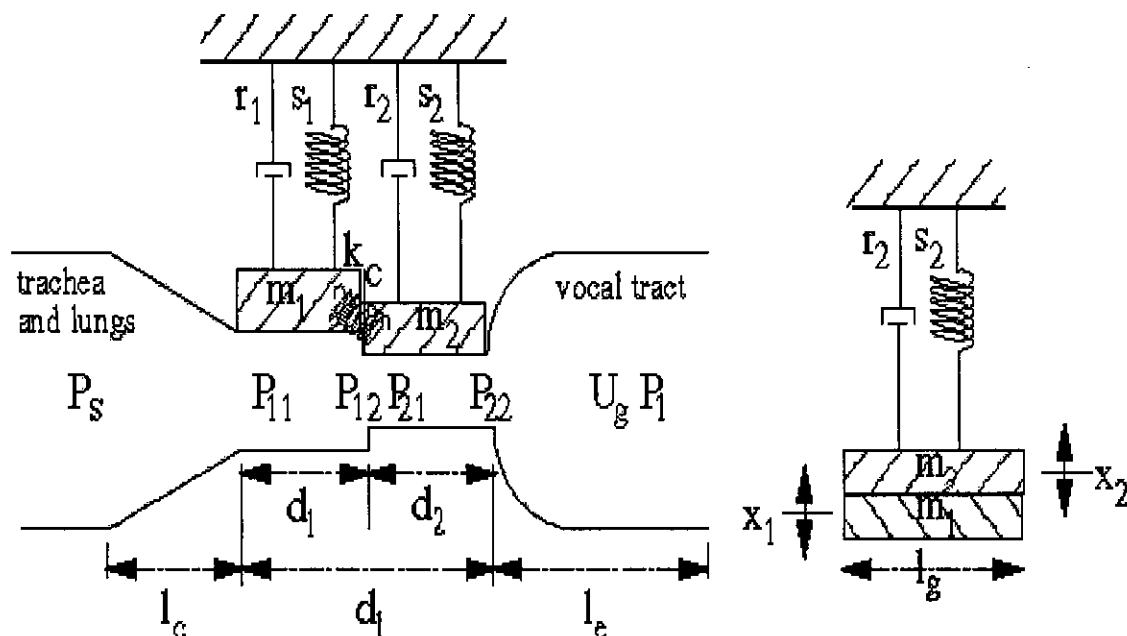


Figure 2.9: Schematic diagram of Two-mass model.

2.3.3 Multiple-mass Model

The two-mass model is considered as a milestone in qualifying vocal folds vibration but it models only the vocal folds as a minimal mechanical structure capable of responding to aerodynamic forces and sustaining oscillation. It is not capable of exhibiting various longitudinal vibratory modes observed in human phonation. Titze [17], made an attempt to enlarge the horizontal degree of freedom, proposed a 16-mass model, which is composed of two rows with eight masses each as shown in the figure-2.10.

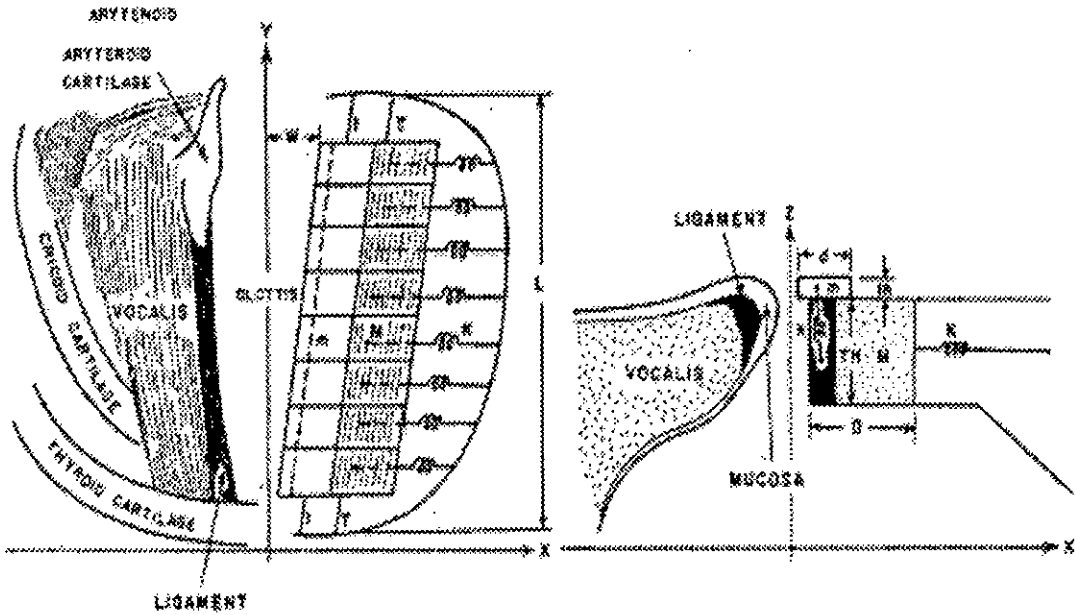


Figure 2.10: Schematic diagram of 16-mass model.

The one row of masses represents the mucosa and the other row represents primarily the vocal ligament and the vocalis muscle. The forces T_m and T_v represent the longitudinal tensions. Specially, the spring constants for the upper and lower rows increase nonlinearly with elongation of the vocal folds. The 16-mass model is complex and has high computational burden.

2.3.4 Continuum Model

The vocal folds are represented as a continuous deformable medium in continuum model [15, 16] as shown in the figure-2.11 where the origin of the co-ordinate system is centered at the vocal processes. The rectangular parallelepiped represents the vocal fold part of the vocal folds. Surface 1, 2 and 3 are fixed, and others are free. ψ represents the displacement vector of the differential element and Y is the longitudinal stress of the

differential element. A coupling between the horizontal and vertical motion exists, which is understood by the

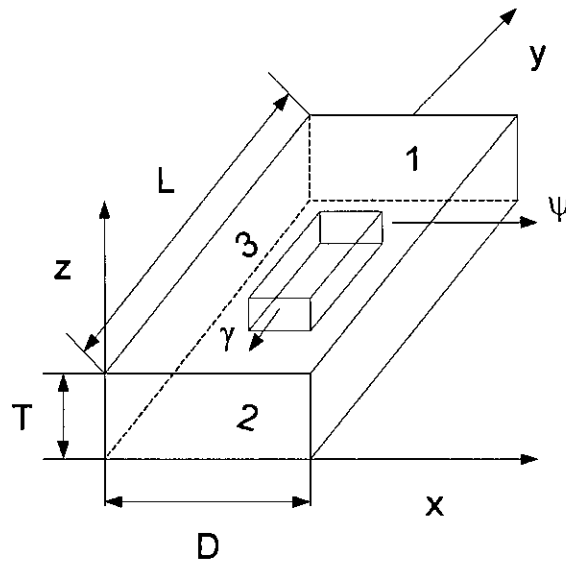


Figure 2.11: Schematic diagram of Continuum model.

incompressibility of the vocal folds. An important consequence of the incompressibility of the vocal folds is that the most easily excited vibratory mode appears to involve vertical phase differences, since this mode tends to preserve the volume of the vocal folds. It was also shown that the layered structure of the vocal folds is ideally adapted to support vocal folds vibration. The longitudinal fibrous structure is looser in the vertical direction than in the longitudinal direction. This allows vertical phase differences to occur.

The continuum model gives clear information about the relationship between the vocal folds structure and the vocal folds vibratory modes. But it has the limitation that the shape of the vocal folds in this model is restricted to a rectangular form. The tissue properties are uniform in the plane normal to the longitudinal direction for ease of manipulation. It has another limitation that the model lacks a complete representation of

the interaction between the aerodynamic airflow and the elastic vocal folds tissue because the normal modes of the vocal folds vibration are derived based on an eigenvalue analysis of the fold tissue.

2.3.5 Ribbon Model

Vocal folds vibration occurs mainly in a thin layer of the non-muscular tissue at the vocal folds surface. It is estimated that the effective depth of vibration into the vocal folds is on the order of 1mm. Hence, one can think of the vibrating portion as a stretched ribbon that is fixed at the horizontal endpoints ($Y=0$ at the posterior arytenoids part, $Y=L$ at the anterior part) but is free to bend and flex in the vertical dimension between those endpoints. So the motion of the ribbon can be described by a wave equation with appropriate boundary conditions, and its eigen-function will give the approximate vibration patterns of the vocal folds. Using this concept, a kinematics four-parameter

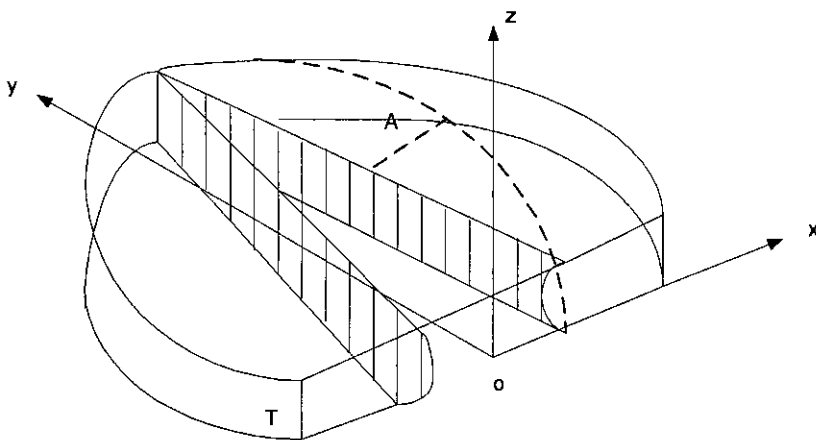


Figure 2.12: Schematic diagram of ribbon model.

model for three-dimensional glottis was presented by Titze [17-18] as shown in the figure-2.12. The four-parameter model can provide the glottal flow, glottal area, and

vocal folds contact area waveforms. The static glottis is controlled by the abduction quotient (Q_a) and the shape quotient (Q_s) and the bulging quotient (Q_b). The phase quotient (Q_p) and fundamental area frequency (F_o) control the dynamic glottis. The displacement function is sinusoidal and is used to calculate the glottal area.

2.3.6 Body-cover Model

In body-cover model the vocal folds is divided into two mass tissue layers with different mechanical properties. The body layer consists of muscle fibers and some tightly connected collagen fibers of the vocal ligament [29]. The cover layer consists of pliable non-contractile tissue that acts as a flexible sheath around the body layer. The layer typically is loosely connected to the body during vibration. The motion of the cover layer is usually observed as a surface wave. This wave propagates from the bottom of the vocal folds to the top and so experiences movement in both the lateral and vertical directions. Self-sustained vocal folds oscillation is highly dependant on this surface-wave behavior and is the primary mechanism for transferring energy from the glottal flow to the tissue to fuel the vibration. The body layer is primarily involved in lateral motion. Based on his findings, Hirano in 1974 suggested that the vocal folds should be treated as a double structured vibration with stiffness parameters that should be based on the relative actions of the thyroarytenoid and cricothyroid muscles. Thus, the resultant vibration of the vocal folds is composed of the coupled oscillations of the body and cover layers. In the two-mass model, the lower mass is made thicker (vertical dimension in the coronal plane) and more massive than the upper element in an attempt to include the effects of the body layer. But, because a provision does not exist for coupled oscillation of both layers, the two-mass model is essentially a “cover” model rather than a “body-cover” model. In order to present more realistically the body-cover vocal folds structure, Story and Titze in 1995 extended the two-mass model to the body cover model as shown in the figure-2.13.

The three-mass model consists of two “cover” masses coupled laterally to a “body” mass by nonlinear springs and viscous damping element. In this model body mass represents muscle tissue which is further coupled laterally to a rigid wall (assumed to represent the thyroid cartilage) by a nonlinear spring and a damping element. The two cover springs are intended to represent the elastic properties of the epithelium and the lamina propria, while the body spring simulates the tension produced by contraction of the thyroarytenoid muscle. Thus, contractions of the cricothyroid and thyroarytenoid muscles are incorporated in the values used for the stiffness parameters of the body and cover springs. The two cover masses are coupled to each other through a linear spring, which can represent vertical mucosal wave propagation.

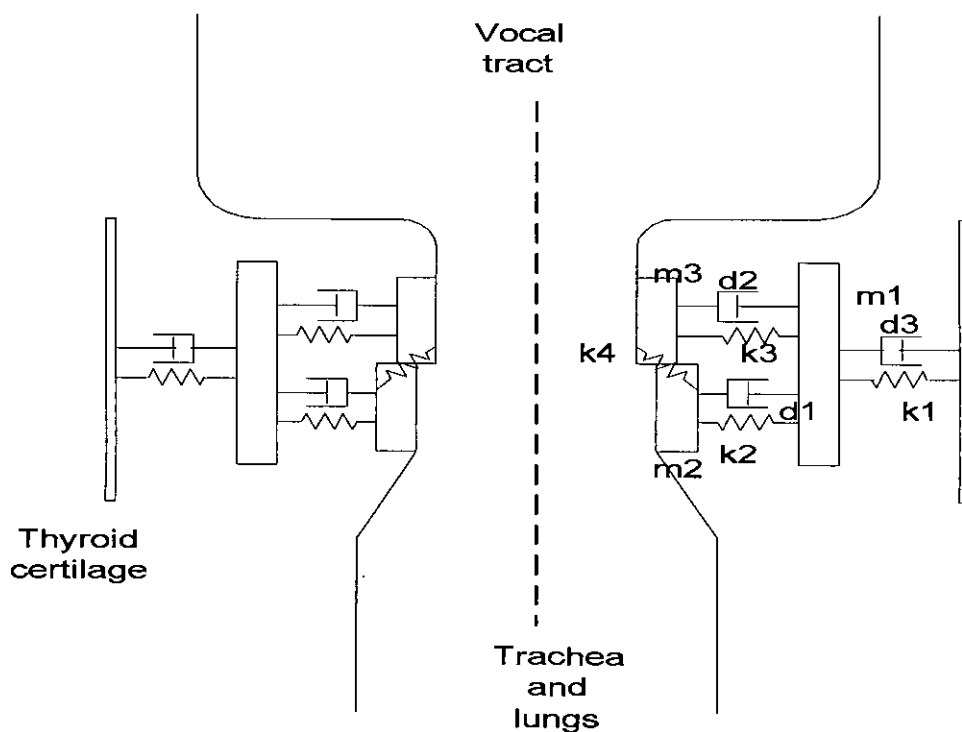


Figure 2.13: Schematic diagram of the body-cover model.

2.3.7 Interactive Model (electrical analog version)

In this model the glottal constriction can be thought of as a purely dissipative flow resistance which is inversely proportional to the glottal area [30]. In addition, the acoustic impedance of the supraglottal and subglottal system can be approximated by an inertive reactance at F_0 (fundamental frequency) and those glottal harmonics falling below F_1 (first formant) and below the lowest subglottal acoustic resonance (for the subglottal system). This is because that the supraglottal acoustic impedance as seen by the glottis is inertive for frequencies more than a few percent less than F_1 and the subglottal acoustic impedance as seen by the glottis also tends to be inertive for frequencies between the highest respiratory tissue resonance, which of the order of magnitude of 10 Hz in adults [31], and the lowest acoustic resonance, which roughly 300 to 400 Hz in adults.

Since the subglottal and supraglottal air masses can be considered to be more inertive (mass-like) than compliant (compressible), if the vocal folds open after remaining closed a long time, there will be a delay or lag in the build-up of air flow relative to the increase in area, as the lungs pressure acts to overcome the inertia of the combined air mass. This lag is shown fig-2.14 by the left-most horizontal arrow of the sketch of the glottal area and flow waveforms. Fig-2.15 shows the solution of the nonlinear differential equation that results when glottis is represented by a time-varying resistance, and the subglottal and the supraglottal acoustic system by a single constant inertance [31]. The system is shown in the figure 2.15 in its analogous electrical circuit form, where

$Y_g = 1/R =$ the glottal conductance;

$P_L =$ the average alveolar pressure in the lungs;

$L_f =$ the sum of subglottal and the supraglottal inertance near F_0

$U_g =$ the glottal volume velocity.

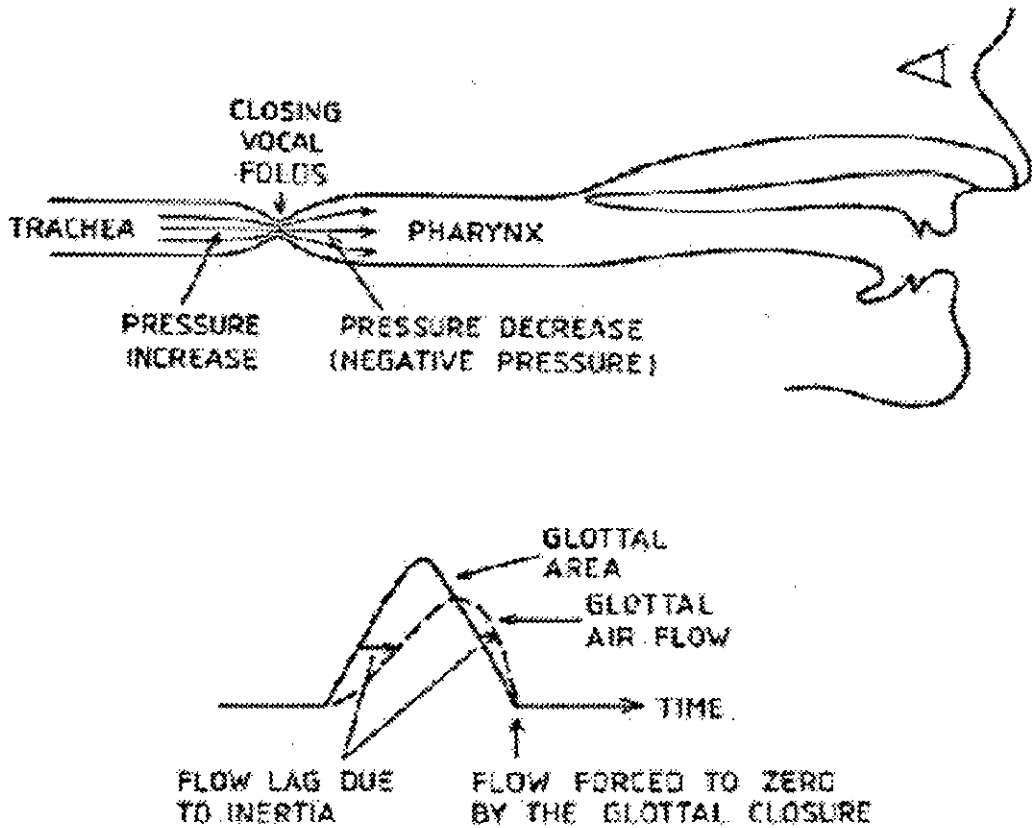


Figure 2.14: Glottal area and glottal air flow during phonation.

The R, L model in the figure 2.15 does not include the interaction with the first formant. To include a first-order approximation of the first formant, the model can be modified by adding an oral compliance, C_o . This oral compliance can be considered a lumped approximation to the compressibility of the supraglottal air and at lower values of F_1 , a small component due to the effective compliance of the walls of supraglottal tract. In this model, the supraglottal inertance is split into two parts, one on the either side of the oral compliance.

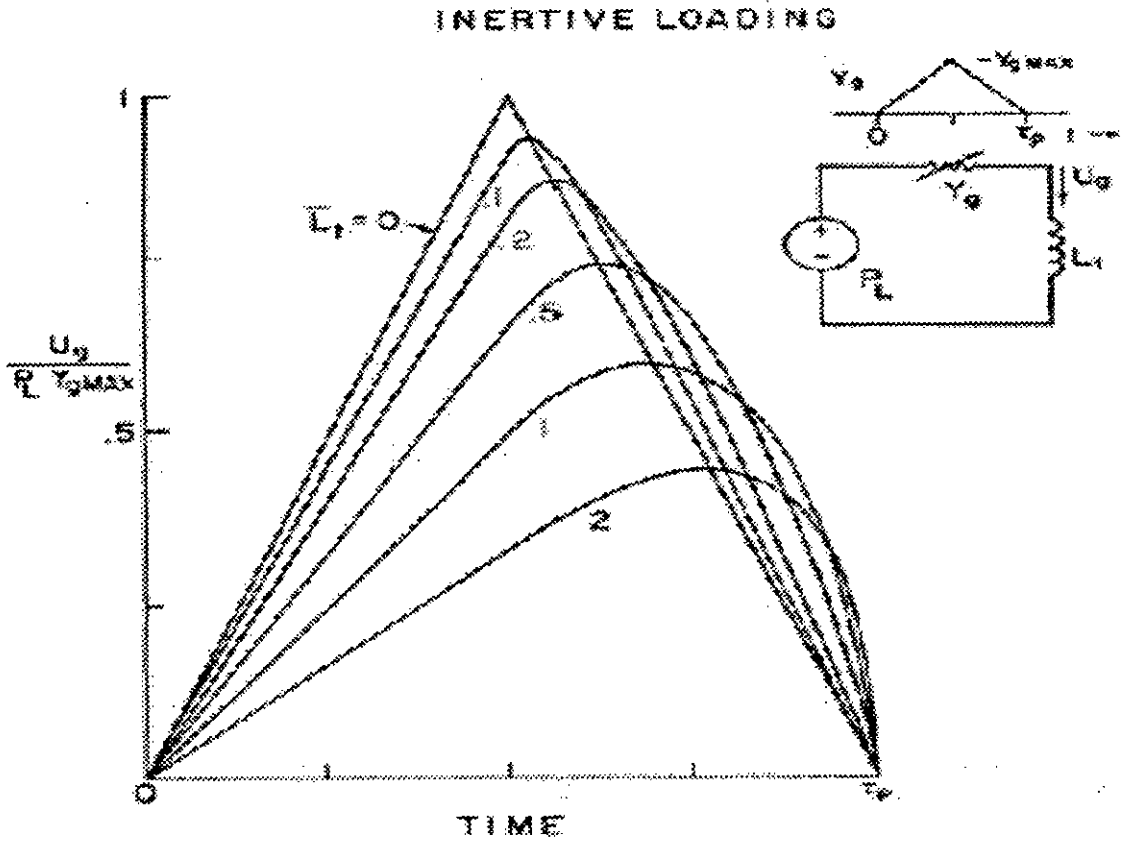


Figure 2.15: Electrical analogous of vocal folds.

In this model, a back vowel such as “a” would have a high value for the pharyngeal inertance and a low value for the oral component, while the reverse would hold for a front vowel such as “i”.

The dissipative elements associated with the vocal tract, R_{oc} , R_{oL} , and R_{oN} in the figure-2.16 are shown dashed, since not all may be needed in a simple model. R_{oc} primarily

represents the dissipation associated with the compressibility of the air flow and the compliance of the cavity walls. R_{oL} represents the dissipations associated with the

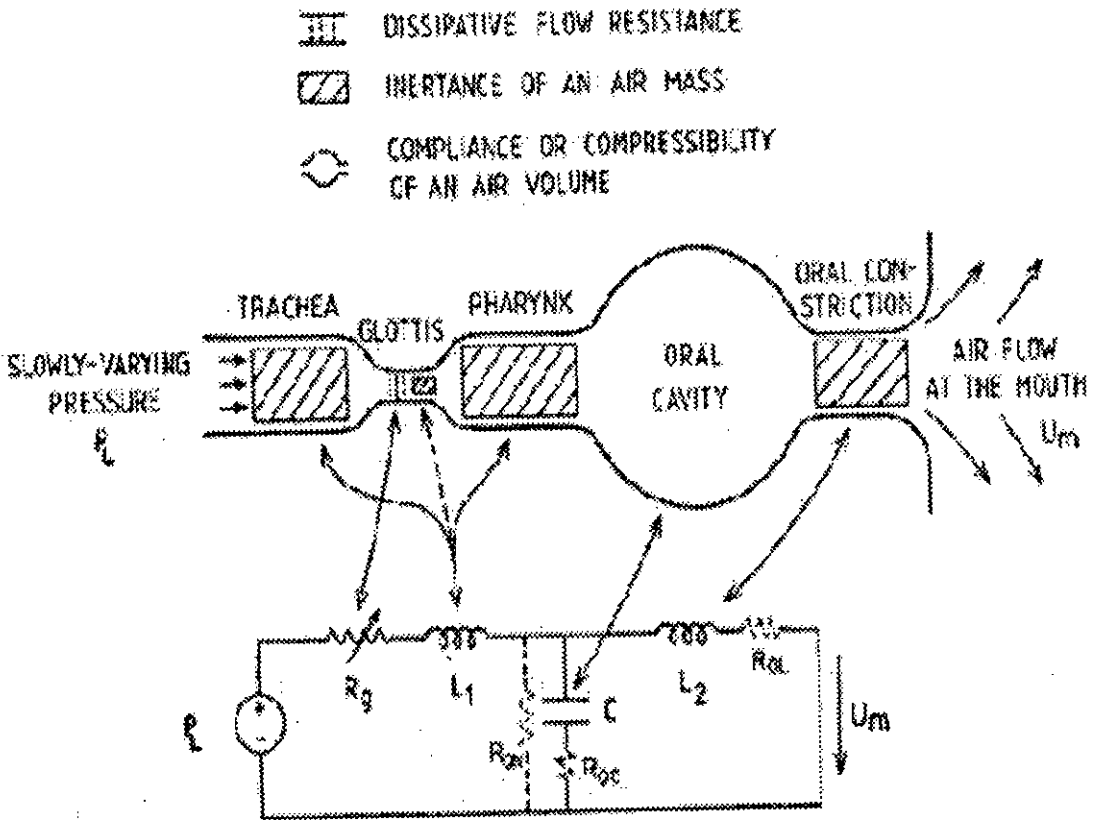


Figure 2.16: Electrical analogous of vocal system.

velocity of the air flow (boundary layer effects, etc.), and R_{oN} represents any shunting effects, such as a small velopharyngeal leakage. For non-nasal vowels with a high value of F_1 , the main effect of oral dissipation is to determine the damping of F_1 during the period of glottal closer. Since the total dissipative loss is generally very small in this case, anyone of these components can be used. However, for low value of F_1 or

nasalized vowels, the placement and the distribution of the dissipative loss elements should be reconsidered.

2.3.8 Parametric Model

Parametric models use parameters to show the vocal folds properties. Though there are various types of parametric models, all of them utilize almost the same method for estimating the parameters. Among the parametric models, LF-model is considered as the best [32]. LF-model is a 5-parameter model and the estimation of this parameters are done by fitting the LF-model to the glottal flow signal which is procured by means of inverse filtering [32,33]. Inverse filtering, in one way, is similar to deconvolution. If the transfer function of the vocal tract is known, glottal volume velocity can be determined through the deconvolution of the sound pressure waveform and this is called the inverse filtering of speech waveform. The vocal tract transfer function is generally estimated over the closed phase interval (during the shut of the vocal folds) of the pitch period to obtain an all-pole model. Since the estimated function is all-pole, vocal tract configurations with dominant spectral zeros (such as nasals) are poorly represented. Additionally, the inverse filtering paradigm assumes a time-invariant vocal tract over each pitch-period.

Electroglottography (EGG) is used to determine the instant of closer and the information about opening location. Inverse filtering is performed based on this information and then the LF-model is fit to the inverse filtered waveform. The LF-model of Fant, Liljencrants and Lin describes the smooth derivative of the glottal pulse waveform, referred to as differentiated glottal volume velocity (DGVV) waveform in terms of an exponential growing sinusoid in the open phase and a decaying exponential in the closed phase. The estimated DGVV obtained through inverse filtering will be used to resemble the shape of the LF-model waveform.

There are some techniques to find the parameters from the inverse filtered output. Prony [34] proposed a method for solving the parameters with the help of the equation of the form

$$X(n) = \sum_{i=0}^N \alpha_i e^{\beta_i n}. \quad (2.2)$$

The technique, known as Prony's method, decomposes the problem into two sets of linear equations. First, the modes β_i of the signal are calculated, and then the residues α_i are calculated. Both α_i and β_i are complex. The equation (2.2) is in the form of a sum of complex exponentials in both the closed and open phases. Therefore, Prony techniques can be applied in each phase to fit these equations to the inverse filtered output.

Another method, called Gradient Descent Technique, a very common iterative search technique, can be used to estimate the open phase LF-model parameters that can best fit the inverse filtered waveform in a least square sense.

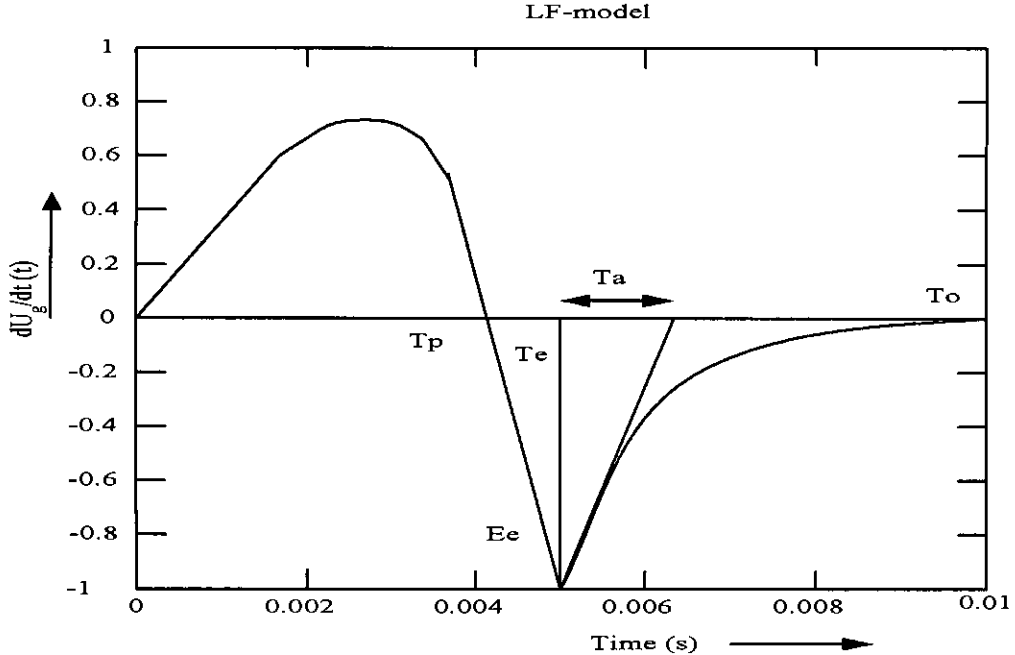


Figure 2.17: The derivative of glottal flow signal (LF model).

In the figure-2.17 a five-parameter glottal flow model (LF-model) is given. Several points on the voice source pulse serve as parameter for minimizing errors between the model and the source. These points are usually the following major features of the glottal flow derivative (U'):

T_0 = the length of the entire pulse;

T_P = the length of the time that $U > 0$;

T_e = the time of the maximum negative value of U' ;

E_e = the value of the maximum negative U' ;

T_a = the experimentally determined effective duration of the return phase.

The glottal signal flow waveform found from DGVV is shown in the figure 2.18.

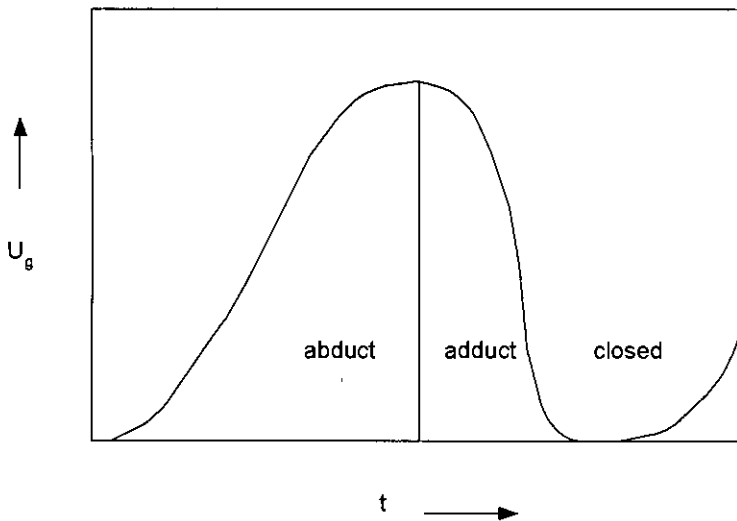


Figure 2.18: Glottal flow signal.

In the figure 2.18 the three phases of the glottal cycle are shown. When the pressure difference becomes sufficiently large, the vocal folds are forced apart and air begins to

flow through the glottis; this is the abduction phase. When the pressure difference between the sub-glottal and supraglottal passages is sufficiently reduced, airflow begins to reduce and the glottis begins to close; this is adduction phase. Adduction occurs more rapidly than the abduction. Then glottis quickly closes, resulting in the closed phase of the glottal cycle.

2.4. Conclusion

In this chapter the anatomy of the voice production system is discussed at the beginning. Then three major subsystems of the voice system -vocal folds, vocal tract and lips and nose are explained briefly with their particular role in the voice production. Finally the major emphasis was given to the vocal folds, starting with its anatomy to understand its contribution to the voice output followed by some of its existing models. Physical models of the vocal folds describe the glottal system in terms of physiological quantities. Parametric models fit the glottal signal with piecewise analytical functions, using a small number of parameters.

CHAPTER 3

EXTRACTION OF VOCAL FOLDS PARAMETERS FOR SPEAKER IDENTIFICATION USING INVERSE FILTERING

3.1 Introduction

The basic of the speaker identification is to extract the speaker discriminative properties from the speech. There are some methods used to extract these properties such as Linear predictive cepstral coefficients, mel-frequency cepstral coefficients [22]. Every method has its advantages as well as shortcomings. In this thesis work, a new model (vocal folds model) is proposed here to investigate the speaker identification process. The whole process is divided into two parts: finding the vocal folds model parameters through cepstral coefficients and inverse filtering, identifying the speaker by *vector quantization* (VQ) and feature matching technique. Generally, cepstral analysis separates vocal tract properties from the speech output and cepstral coefficients carry the information of the vocal tract. These coefficients are used with the help of inverse filtering to get the vocal folds model parameters. On the other hand, VQ process is used to produce the *code* of a speaker from the vocal folds model parameters and matching the test speech with this *code* does the identification. In this chapter, the process for obtaining the cepstral coefficients and finding the vocal folds model parameters from these is explained. Then the VQ method that produces the speaker code is described. Finally, the feature matching technique is presented to identify the speaker.

3.2 Basis for the Proposed Model

Vocal folds act as the source of voice production. Its oscillation produces a quasi-static air pressure, which is further modulated by the vocal tract, mouth and nasal cavities. It is interesting to observe that only the change of size and shape of the vocal tract cavity can

produce different sounds although the vibration of the vocal folds remains unchanged. Hence, it is quite clear that the variation in sound is caused mostly by the variation of vocal cavities (mainly vocal tract) but properties of vocal folds do not vary much during that time. Therefore, there is a strong that the properties of vocal folds can be investigated for speaker identification.

Speaker identification needs to extract speaker discriminating features from the speech signal. *Mel-frequency cepstral coefficients* (MFCC) and *Linear predictive cepstral coefficients* (LPCC) are well known cepstral coefficients derived from two different techniques which extract speaker discriminative vocal tract properties from speech signal used for speaker identification purpose [22]. But according to the above discussion vocal folds properties (proposed model parameters) are more likely to give better speaker discriminative features.

To extract the vocal folds model parameters from the speech output it is necessary to know the over all model for the vocal system. Early but still recognized model of human speech production is the source-filter model, which treats the vocal folds as the source of the voice and vocal cavities (mainly vocal tract) as the filter and these two parts work independently during the utterance especially for the voiced sounds. This implies that these components are linear and separable as shown in the figure 3.1.

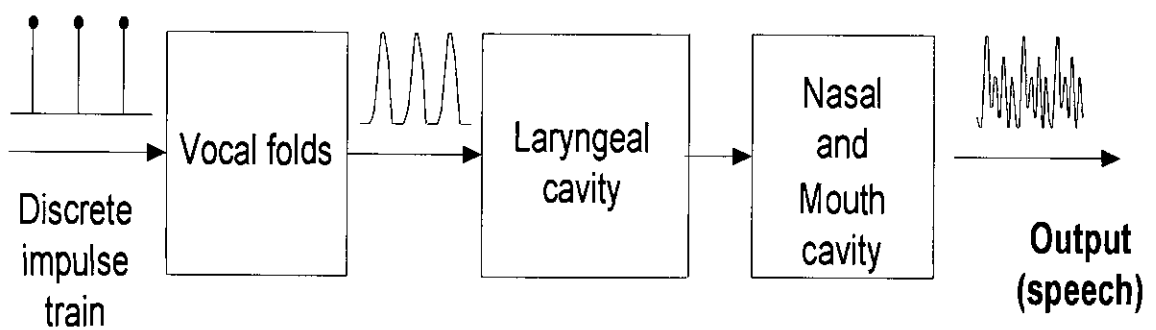


Figure 3.1: Block diagram of the vocal system for voiced sounds.

According to the above block diagram if $s(n)$ represents the speech output with a predetermined sampling rate satisfying the Nyquist criteria for sampling, the relation of the components of the vocal system with the output utterance in the frequency domain will be

$$S(\omega) = G(\omega).H_v(\omega).H_c(\omega).H_r(\omega) \quad (3.1)$$

Where the $G(\omega)$ is the Fourier transform of the input, $H_v(\omega), H_c(\omega)$ and $H_r(\omega)$ are the transfer functions of vocal folds, vocal cavity (vocal tract) and nasal and mouth cavities respectively. But in the sample voice data only the $s(n)$ is known. Primarily the $G(\omega)$ is taken as a discrete-time impulse train (i.e. $G(\omega) = 1$). Again, the overall system can be considered as the *Autoregressive Moving Average* (ARMA) model where the *zeros* of this model are basically produced by mostly the nasal and mouth cavities. Generally, zero of a model corresponds to the phase of the output waveform. But for recognition purpose the magnitude of the Fourier transform will correctly resemble the output of the voice system. Hence taking the AR model of overall system considering the dominant poles only ignores the most part of the contribution of nasal and mouth cavity. Then the system can be represented in a reduced form as in the equation (3.2)

$$S(\omega) = H_v(\omega).H_c(\omega) \quad (3.2)$$

That is the output is the multiplication of the two major parts of the vocal system: vocal folds and vocal tract in the frequency domain.

The salient feature of the $H_c(\omega)$ is found in the Fourier transform of the output where the resonance of the sound gives the indication of the poles of the vocal cavity (vocal tract). The above equation (3.2) is more precise for sounds that come through the vibration of vocal folds. This type of sounds is called *voiced sound*. Hence, all vowels are *voiced sounds*. *Unvoiced sounds* are generated by forming a constriction at some point along the vocal tract, and forcing air through the constriction to produce turbulence. As for

example the /s/ sound in “six” is an *unvoiced sound*. In the model for *unvoiced sounds*, the $G(\omega)$ and $H_v(\omega)$ are replaced by flat spectrum white noise source [22].

For speaker identification purpose Cepstral method is used to get the speaker distinctive features from some parts (vocal tract) of the vocal system. Conventionally the cepstral coefficients generated by this cepstral method are used to get these features where the cepstral coefficients carry the information of speaker discriminative vocal tract properties. Again, according to the equation (3.2) if the speech output is subjected to inverse filtering by the cepstral coefficients in the cepstral domain the outcome will be the properties of the vocal folds especially for *voiced sounds*. In this research work, only the vowels (*voiced sounds*) are used. Hence, the filtered output produced in the way just mentioned, is the correspondence of the output of the vocal folds. AR model parameters of this vocal folds output are used as the features of the speakers and according to above discussion it can serve a very useful element for speaker identification purpose.

3.3 Cepstrum

The human speech signal $s(n)$ can be represented as a “*quickly varying*” source (vocal folds) signal $e(n)$ convolved with the “*slowly varying*” impulse response $h(n)$ of the vocal tract represented as a linear filter [22] as shown in the figure 3.2 conforming to the equation (3.2). It is often desirable to eliminate one of the components though only the output speech signal can be accessed. Separation of the source (vocal folds) and the filter (vocal tract) from the mixed output is in general difficult problem when these components are combined nonlinearly [22]. The cepstrum is the representation of the signal where these two components are resolved into two additive parts [22] as shown in the figure 2.3-a and figure 2.3-b.

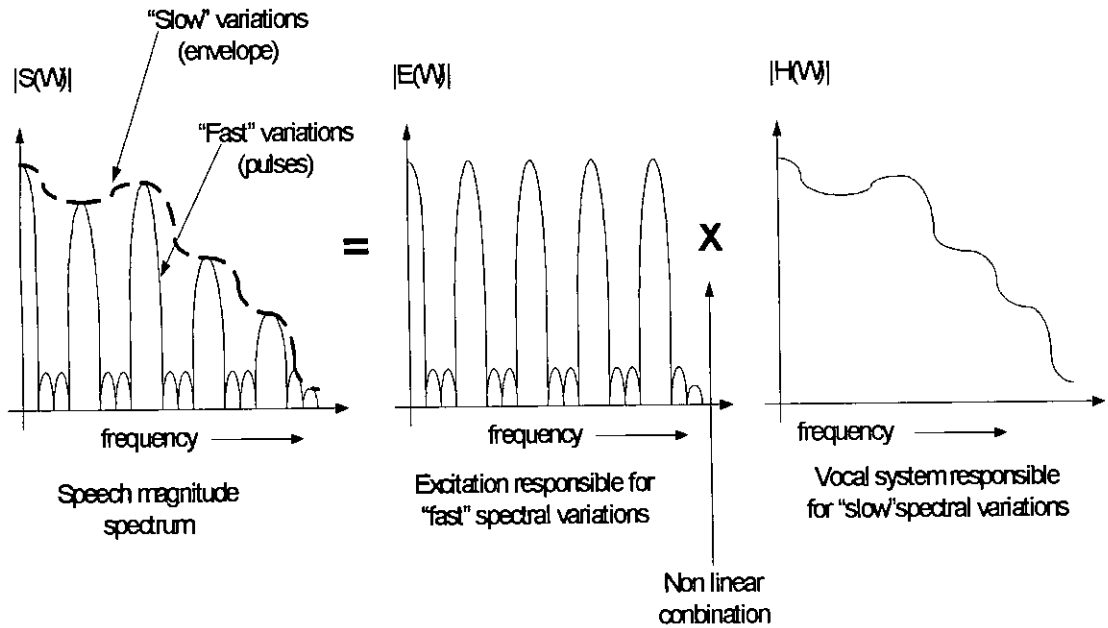


Figure 3.2: Component of speech in speech spectrum.

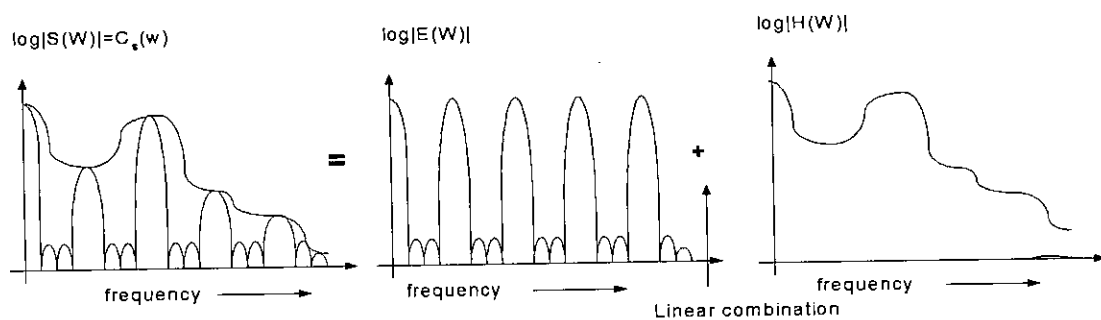
Mathematically, according to the discussion above:

$$|S(\omega)| = |E(\omega)| \cdot |H(\omega)| \quad (3.3)$$

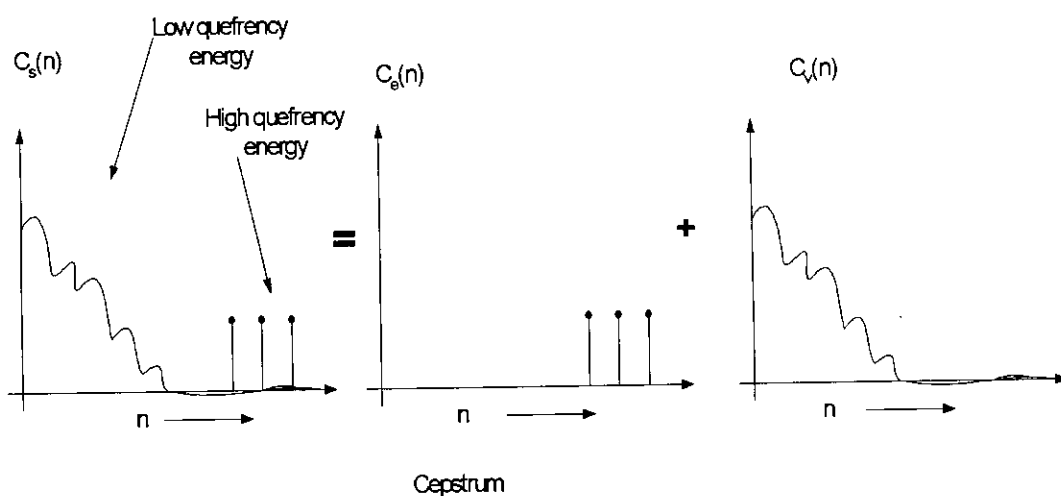
Where, $E(\omega) = G(\omega) \cdot H_v(\omega) = H_v(\omega)$ and $H(\omega) = H_c(\omega)$ according to the equation (3.1) and equation (3.2). Taking log on both sides of equation (3.3) will convert the product in the right hand side of the equation (3.3) to additive form.

$$\log |S(\omega)| = C_s(\omega) = \log |E(\omega)| + \log |H(\omega)| \quad (3.4)$$

Hence the two components of the vocal system are now linearly combined. It is shown in the figure 3.3-a. These additive components will then be converted to the cepstral coefficients in the cepstrum. It is computed by taking the inverse *Discrete-Time Fourier Transform* (DTFT) of the logarithm of the magnitude spectrum of a frame of a speech.



(a)



(b)

Figure 3.3: Components of speech in speech cepstrum
(a) conforming to equation (3.4) (b) conforming to equation (3.6)

This is represented in the equation (3.5).

$$\text{cepstrum}(\text{frame}) = \text{IDTFT}(\log(|\text{DTFT}(\text{frame})|)) \quad (3.5)$$

Taking the IDTFT in both sides of the equation (3.4) according to the equation (3.5) will convert the equation (3.4) to its cepstrum counterpart. It is given in the equation (3.6).

$$c_s(n) = c_e(n) + c_v(n) \quad (3.6)$$

Where $c_s(n)$, is the IDTFT of $C_s(\omega)$, $c_e(n)$ is the IDTFT of $\log|E(\omega)|$ and $c_v(n)$ is the IDTFT of $\log|H(\omega)|$.

The block diagram to get the cepstral coefficients from the speech output is shown schematically in the figure 3.4

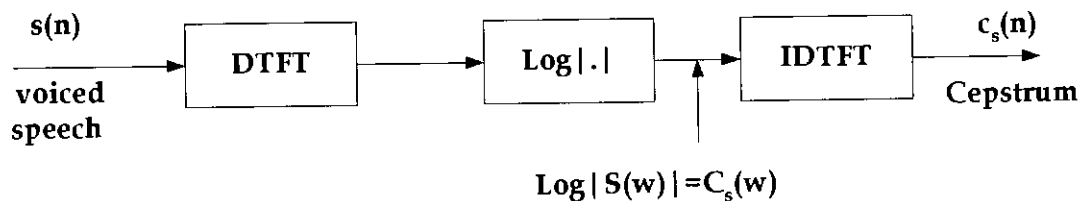


Figure 3.4: Computation of the cepstrum

In the speech magnitude spectrum $|S(w)|$, two components $|E(w)|$ and $|H(w)|$ is multiplied. Once the logarithm of the spectral magnitude is taken, the two components have their additive correlates in the new "signal," $C_s(w)$. When the IDTFT is taken, two parts of the voice system are found clearly distinctive as shown in the figure 3.3-b. Conventionally, the cepstral coefficients found from the cepstral analysis carry the characteristic of vocal tract. LPCC and MFCC are generally used as the cepstral coefficients for this purpose.

3.3.1 LPCC

LPCC are the cepstral coefficients derived from the *linear predictive coding* (LPC) technique where the system is taken as all-pole model (AR model). For speech

DP

production system LPC is based on the source-filter model conforming to the equation (3.2). The main idea behind LPC is that the given speech sample can be approximated as a linear combination of the past speech samples as given in the equation (3.7).

$$\hat{s}(n) = -\sum_{k=1}^p a_k \cdot s(n-k) \quad (3.7)$$

where $\hat{s}(n)$ is an approximation of the present output, $s(n-k)$ are past outputs, p is the *prediction order*, and a_k are the model parameters called the *prediction coefficients*. *Prediction error* e is defined as the difference between real and predicted output. Autocorrelation and covariance methods are usually applied to procure these coefficients.

In speaker recognition task, we can use LPC based on the short-term analysis approach. Because of the quasi-stationary nature of speech, we can compute a set of prediction coefficients from every frame. The prediction order depends on the sampling rate and frame size.

The preoccupation with the AR model (all-pole model) of the speech production system, however arises from the fact that the very powerful and simple technique, LPC is used to derive the AR model parameters from a given speech utterance. The main problem with this model is that it can exactly preserve the magnitude spectrum but it may not retain the phase characteristics. Nevertheless, a waveform with correct spectral magnitude is frequently sufficient for coding, recognition and synthesis [22].

However, in speaker identification the major disadvantage of LPC is that it does not resolve the vocal-tract characteristics from the vocal folds dynamics. In LPC derived cepstral coefficients, the cepstral coefficients are not computed directly from the speech utterance, rather it is computed from the impulse response of the LP model (AR model).

Therefore, the cepstral coefficients are computed from a sequence that has already been “smoothed” in the sense that the excitation has been removed [22].

It has already been mentioned in the forgoing article that the MFCC and LPCC are derived from two well-known techniques used in speaker identification to describe signal characteristics, relative to the speaker discriminating vocal tract properties. There is no general agreement in the literature about which method is better. However, it is generally considered that LPCC are computationally less expensive while MFCC provide more precise result [35] at the cost of higher computational complexity. However, it is well known that certain phoneme classes, most notably the vowels, involve vocal tract configurations that are acoustically resonant, and are therefore appropriately modeled by all-pole AR model structure [22]. Hence, for the *voiced sound* this LPCC derived from the *all-pole* model LPC will provide better result and can be a strong competitor for MFCC [36].

3.4 Estimation of Proposed Model Parameters

Before the estimation of the vocal folds parameters, the speech output needs to be preprocessed. Speech framing and pre-emphasis and windowing are the preprocessing steps. Then the preprocessed data will be used to find the vocal folds parameters as describe in the following sub-articles.

3.4.1 Speech Framing

When the speech is examined over a sufficiently short period of time (20-30 milliseconds) it has quite stable acoustic characteristics [22]. It leads to the useful concept of describing human speech signal, called “*short-term analysis*,” where only a portion of the signal is used to extract signal features at one time. It works in the following way: the speech is broken into fixed length frames typically between 20 ms and 40 ms in duration with an overlapping (usually 30-50% of the window length)

between the adjacent frames. Overlapping is generally needed to avoid losing of information. It is shown schematically in the figure 3.5.

3.4.2 Pre-emphasis

In many cases pre-emphasis is applied to the input signal. This is done mainly because the recording device attenuates the higher frequencies more than the lower ones. Higher frequencies get attenuated while propagating through air and the human ear also emphasizes the higher frequencies. Pre-emphasis is typically done by a simple first order high pass filter that increases the relative energy of the higher frequency spectrum [37].

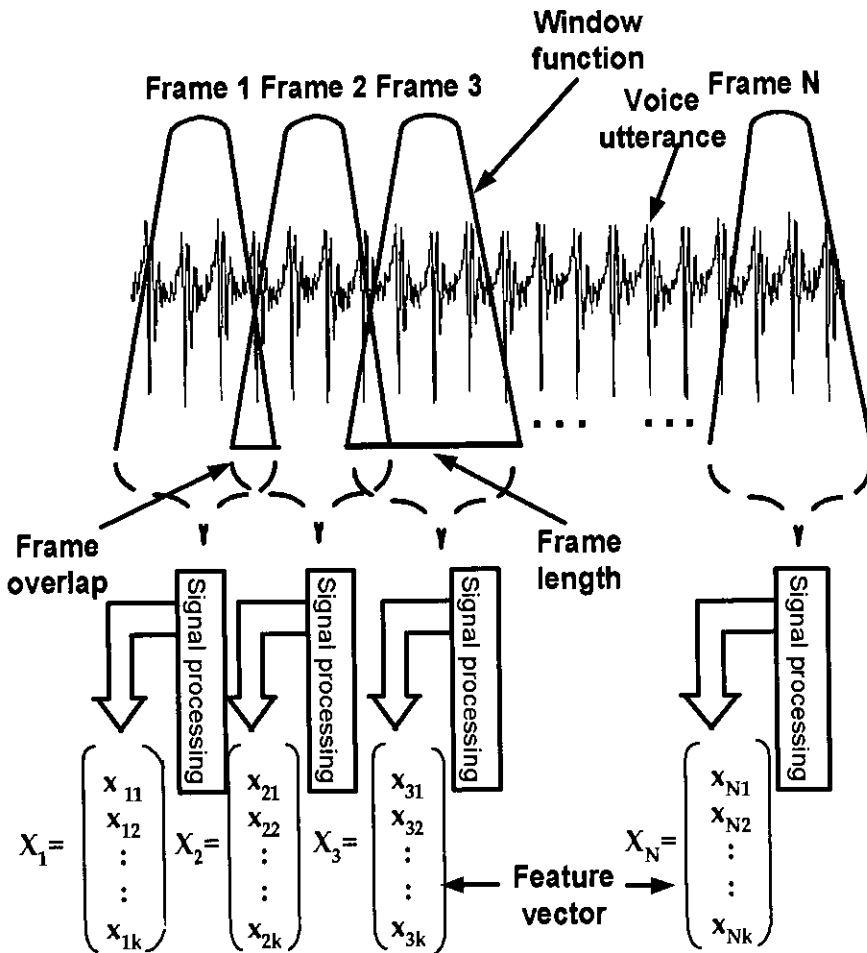


Figure 3.5: "Short term analysis" of speech utterance

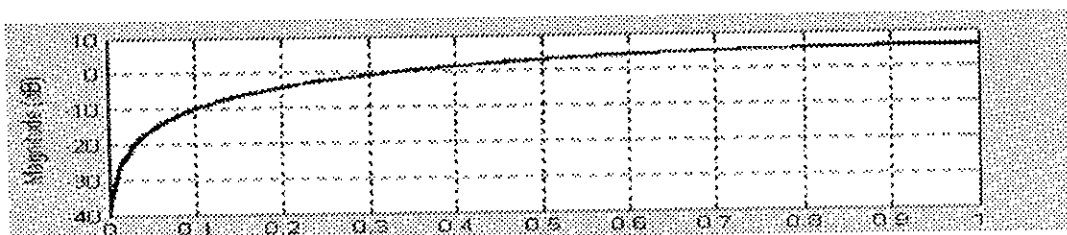
The Z-transform of this high pass filter can be given by the equation (3.8),

$$H(z)=1-\alpha z^{-1} \quad (3.8)$$

Where α is around 0.96 to 0.99. The frequency response of this filter with $\alpha =0.99$ is shown in the figure 3.6 where the scale is normalized with respect to sampling frequency.

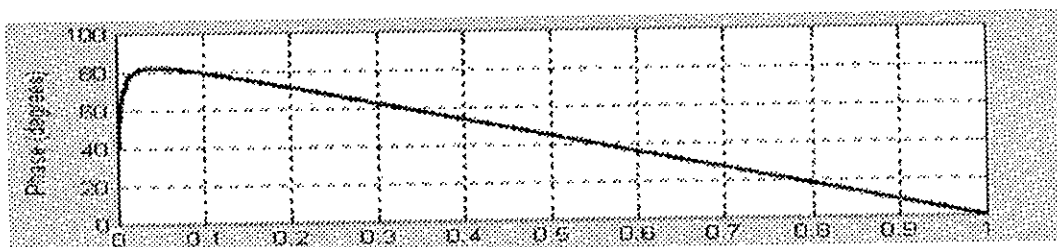
3.4.3 Windowing

In order to prevent an abrupt change at the end points of a frame, it is usually multiplied by a window function. These processed frames are called *windowed frames*. There are several window functions used in speaker recognition area, but the most popular one is the *Hamming window*.



Normalized frequency

(a)



Normalized frequency

(b)

Figure 3.6: The frequency response of a pre-emphasis filter;
(a) magnitude response (b) phase response.

The equation of the *Hamming window* is stated as in the equation (3.9)

$$W(n) = 0.54 - 0.46\cos\left(\frac{2n\pi}{N-1}\right) \quad (3.9)$$

Where $n = 0, 1, \dots, N-1$

3.4.4 Speech Features

A set of features extracted from a frame is called *speech feature* (or *speech vector*). Speech information is usually conveyed in the spectrum of the speech. The logical choice of a speech feature should represent the spectrum of the speech in a compact way. It is clear that any feature will contain information about both the speech and the speaker. Over short period of time the features represent the sounds and in a lesser way the speaker. Over a longer period many sounds are uttered and the accumulation of the features represent the speaker more clearly. Figure 3.7 illustrates how features arrange themselves in two-dimensional feature-space for three different speakers. As can be seen from the figure, the features of each speaker are concentrated at a specific location in the feature-space. Extracting speech features with the proper model makes *Speaker Identification* possible. Speech features should be extracted in such a way that it will reduce the data while retaining the speaker discriminative information.

3.4.5 Linear Prediction Cepstral Coefficient (LPCC)

The LPCC carry the information of the vocal tract properties of a speaker and the calculation to find the LPCC is less expensive compared to MFCC. These LPCC are procured from the speech signal through some steps. First the data is preprocessed as described in the framing, pre-emphasis and windowing sub-articles.

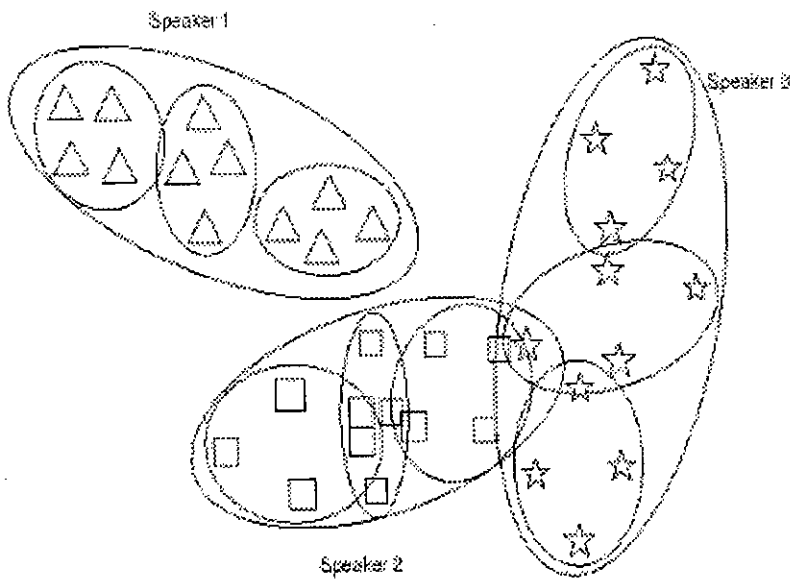


Figure 3.7: Schematic of speaker location for three different speakers in a two dimensional feature space.

Then *Linear Predictive Coding* (LPC) of each *windowed frame* is calculated according to the Equation (3.10).

$$H(z; m) = \frac{1}{A(z; m)} = \frac{G}{1 + \sum_{i=1}^P a_i(m)z^{-i}} \quad (3.10)$$

Where $H(\omega) = H_v(\omega).H_c(\omega)$ (according to equation (3.2)), m is the frame index, P is the order of the LPC, $a_i(m)$ is the set of model parameters of the m^{th} frame and $A(z; m)$ is the Z-transform of the inverse filter. Hence the parameters found from this equation are the AR model parameters of the speech system (entire vocal system).

Now the cepstral coefficients (LPCC) related to the vocal tract properties of the voice system can be found recursively from the model parameters derived by the equation (3.11) as follows,

$$c_{lp}(n) = a_n + \sum_{i=1}^{n-1} \left(\frac{i}{n}\right) c_{lp}(i) a_{n-i} \quad (3.11)$$

With $c_{lp}(0) = \ln(e)$ and $c_{lp}(1) = -a_1$, where e is the minimum *prediction error* describe in art: 3.3.1.

3.4.6 AR Model of Vocal Folds

In our proposed model the AR model parameters are used to find the speaker distinctive features for speaker identification. The speech signal can be represented as a “quickly varying” source signal convolved with the “slowly varying” impulse response of the vocal tract represented as a linear filter. The separation of the source (vocal folds) and the filter (vocal tract) parameters from the mixed output is generally difficult. In this case cepstrum technique can be a useful tool as representative of the component signals will be separated is the cepstrum [22] LPCC is one of the techniques which represent the vocal tract properties of the speaker [22]. It is already mentioned through the equation (3.2) that the voice system can be represented as the multiplication of the two major parts of the vocal system: vocal folds and vocal tract. Therefore, if the speech output is subjected to inverse filtering in the cepstral domain by these vocal tract related LPCC, the information about the vocal folds of the speaker can be found. This is illustrated in the figure 3.8.

Well-established cepstral coefficients carry the vocal tract information as the identification tool for speaker. As both vocal folds and vocal tract vary from person to person, therefore, our vocal folds model can also be used for speaker identification purpose. As in the voiced signal the vocal folds property remained almost unchanged, it

is possible that vocal folds related properties can give even better results for speaker identification than the conventional vocal tract related LPCC.

The output of the vocal folds in the cepstral domain found from the inverse filtering by LPCC is used to find the AR model parameters of the vocal folds using Yule-Walker equations, which are solved by the Levinson-Durbin recursion [38]. These AR model parameters will be considered as the speech features, which carry the information of the speaker.

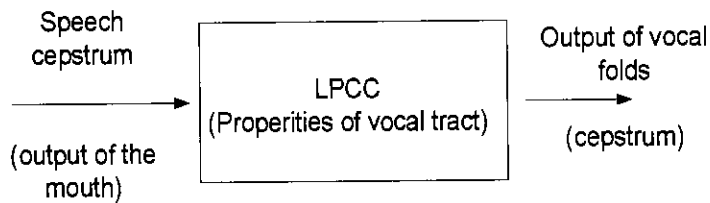


Figure 3.8: Inverse filtering to get vocal folds output

3.5 Speaker Identification

In this phase the AR model parameters will be used for speaker identification purpose. Speaker identification is a decision making process of determining the author of a given speech signal based on the previously stored or learned information [39]. This step is usually divided into two parts, namely *coding* and *matching*. The coding is a process of enrolling a speaker to the identification system by constructing a so-called codebook of that speaker, based on the features extracted from that speaker's speech sample. The matching is a process of computing a *matching score*, which is a measure of the similarity of the features extracted from the unknown speech sample and the speaker codebook [24]. For text independent speaker identification scheme coding is done by first averaging the features over a relatively long period of time. This average features,

which are less than the total number of features extracted from an unknown speech, will be the codebook of the speaker. The process of averaging of the features is known as *vector quantization*, which is the part of the coding in speaker identification. Then this codebook will be used to test the unknown speech for speaker identification.

There are mainly two kinds of speaker identifications: *close set* and *open set*. In *close set* identification, the speaker to be tested is already enrolled in the identification system i.e. his/her codebook is constructed in the system before. In *open set* identification, the speaker may not be enrolled in the system.

3.5.1 Vector Quantization

Vector quantization (VQ) is a process of mapping vectors from a vector space to a finite number of regions in that space. These regions are called *clusters* and represented by their *central vectors* (centroid) or *code vectors*. A set of *code vectors* which represents the whole vector space, is called *codebook*. In speaker identification, VQ is applied on the set of *feature vectors* extracted from the speech sample and as a result, the speaker's *codebook* is generated. Such *codebook* has a significantly smaller size than extracted vector set as shown in the figure 3.9.

This VQ creates clusters with rigid boundaries in a sense that every vector belongs to one and only one cluster [35] and all the vectors within a cluster is represented by the *code vector*. The generated *codebook* in this process represents the speaker i.e. each speaker has its own *codebook*. In this connection it is worth mentioning that the feature should be extracted in a way so that it better represents the speaker than the speech for the purpose of speaker identification.

To get a proper *codebook* for a speaker it needs a sufficiently long speech so that all the statistical properties of the speaker are captured in the speech feature. Each frame produces one set of speech features called *feature vector*, which corresponds to a point in

the vector space. The dimension of the vector space will be equal to the number of elements (coefficients) in a set of speech features, which of course depends on the technique used to extract these features.

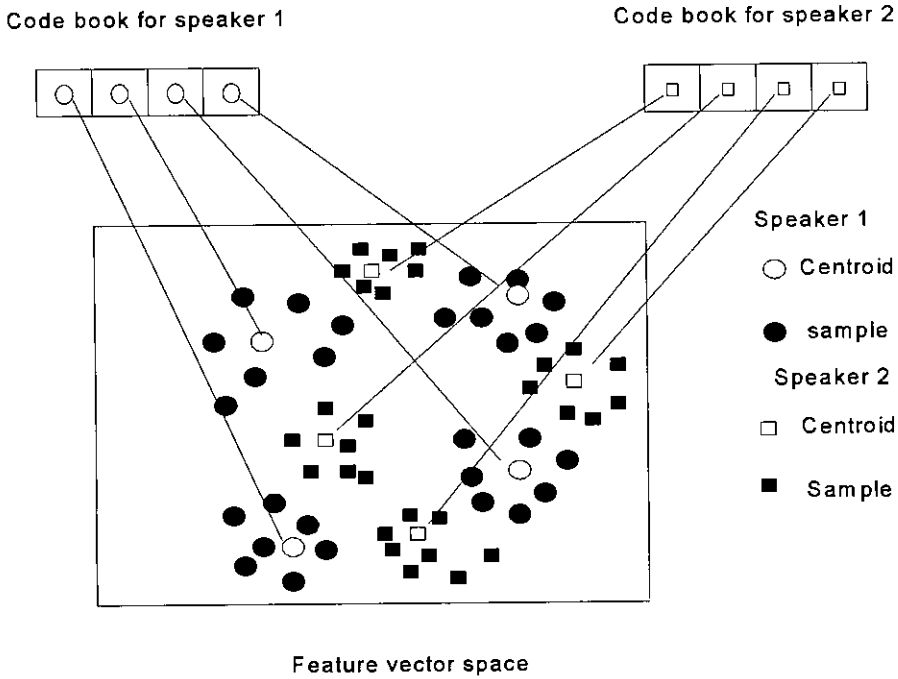


Figure 3.9: Feature vectors along with code vectors in a two dimensional vector space

One feature vector thus found can be represented by the equation (3.12)

$$X_i = (x_{i,1}, x_{i,2}, \dots, x_{i,K}), \quad i = 1, 2, \dots, N. \tag{3.12}$$

Where X_i is the *feature vector* of the i^{th} number of frame, K is the number of elements $x_{i,k}$ (coefficients) in the *feature vector*. In this work, AR model parameters of one frame is the set X_i , where the elements $x_{i,k}$ are the poles of that frame. Hence each frame will produce a set of AR model parameters which will be a point in K dimensional vector space. If the total number of frames is N for a speaker, then N number of *feature vectors*

are procured, each of which has K number of elements. It can be expressed as in the equation (3.13)

$$T = \{X_1, X_2, \dots, X_N\} \quad (3.13)$$

Hence, T represents the entire set of *feature vectors* of a speaker, which can be visualized as N number of *feature vector* points in the K dimensional vector space. The *feature vectors* are clustered in some particular regions in the vector space that corresponds to the characteristics of that speaker. Different speakers would produce different clustered regions.

Now, VQ is applied on these *feature vectors* to find the *code vectors* of the speaker, which are the representatives of the clusters. The set of all *code vectors* of a speaker is called the *codebook* of that speaker and is given in the equation (3.14)

$$B = \{C_1, C_2, \dots, C_M\} \quad (3.14)$$

Where, B is the *codebook* of the speaker, $C_j, j = 1, 2, \dots, M$, are the *code vectors* generated by VQ from T , and M is the total number of *code vectors* for that speaker. The value of M is a low as 4 to as high as 128, and of course depends on the process of the extraction of the speaker features. Generally, the higher is the number of speakers, the higher will be the value of M . Each *code vector* has the same number of elements as that of the *feature vector* as given in the equation (3.15).

$$C_n = (C_{n,1}, C_{n,2}, \dots, C_{n,K}), \quad n=1, 2, \dots, M \quad (3.15)$$

When these *code vectors* are plotted in the vector space along with the *feature vectors* these will represent the centroids of the clustered *feature vectors* as depicted in the figure 3.9.

Let S_n be the encoding region in the K dimensional vector space associated with *code vector* C_n and R be the set of these encoding regions, then the set R can be written as in the equation (3.16).

$$R = \{ S_1, S_2, \dots, S_M \} \quad (3.16)$$

The set denotes the partition of the vector space and if the *feature vector* X_i is in the encoding region S_n , then VQ quantizes X_i (denoted by $Q(X_i)$) in the region S_n as C_n :

$$Q(X_i) = C_n, \text{ if } X_i \text{ is in } S_n \quad (3.17)$$

Hence, C_n represents all those X_i that fall within S_n .

In other way, every C_n will fulfill two criteria:

1. The encoding region S_n should consist of all those *feature vectors* that are closer to C_n than any of the other *code vectors*. For those *feature vectors* lying on the boundary, any tie-breaking procedure will do.
2. The *code vector* C_n should be the average of all those *feature vectors* that are in the encoding region S_n . And at least one *feature vector* belongs to each encoding region.

If the speech features are extracted in a proper way, then these *code vectors* of different speaker will fall in different encoded regions, which can be very useful and efficient speaker discriminative tool for speaker identification.

There are several algorithms used to generate a codebook from the extracted features. The LBG design algorithm is the most efficient method [22]. This algorithm is an

iterative one, which alternatively solves the above two optimality criteria. The algorithm requires an initial code-vector C and it is set as the average of the entire feature vectors. This code-vector is then split into two. The iterative algorithm is run with these two vectors as the initial codebook and each of them is split into two in the same way and the process is repeated until the desired number of code-vectors is obtained.

This is summarized below in the following steps:

1. Find the average of the entire *feature vectors* of a speaker,

$$C = \frac{1}{N} \sum_{i=1}^N X_i \quad (3.18)$$

Where C is the initial *code vector*, X_i is the *feature vector* of the i^{th} frame, N , is the total number of frames or *feature vectors*.

2. Calculate the average distance of *feature vectors* with respect to the initial *code vector*,

$$D_{ave} = \frac{1}{N*K} \sum_{i=1}^N |X_i - C|^2 \quad (3.19)$$

Where $N*K$ is the total number of elements in the whole vector space of that speaker and C is the initial *code vector*.

3. Split the initial *code vector* into two,

$$\begin{aligned} C_1 &= (1 + \varepsilon) \cdot C \\ C_2 &= (1 - \varepsilon) \cdot C \end{aligned} \quad (3.20)$$

where ε is a small positive number.

4. For $i = 1, 2, \dots, N$, find the minimum value of

$$|X_i - C_{t'}|^2 \quad (3.21)$$

where, $t' = 1, 2$. Let t be the index which achieves the minimum set, then

$$Q(X_i) = C_t \quad (3.22)$$

where Q is an operator indicates that X_i is quantized into C_t . That is, C_t is the centroid of all the vectors who produce minimum distance with C_t . In this way the *feature vectors* will be clustered into two sets.

5. Find the average of the two clustered *feature vectors* to get two new *code vectors* C_1 and C_2 .
6. Calculate the average distance of each set of clustered *feature vectors* with respect to its *code vector* and then find the total average distance D_{at} by averaging these two distances.
7. Then

$$\text{if } \frac{D_{ave} - D_{at}}{D_{ave}} > \epsilon ;$$

$$D_{ave} = D_{at}$$

Go to step-4.

In summery the whole process can be illustrated as in the figure 3.10

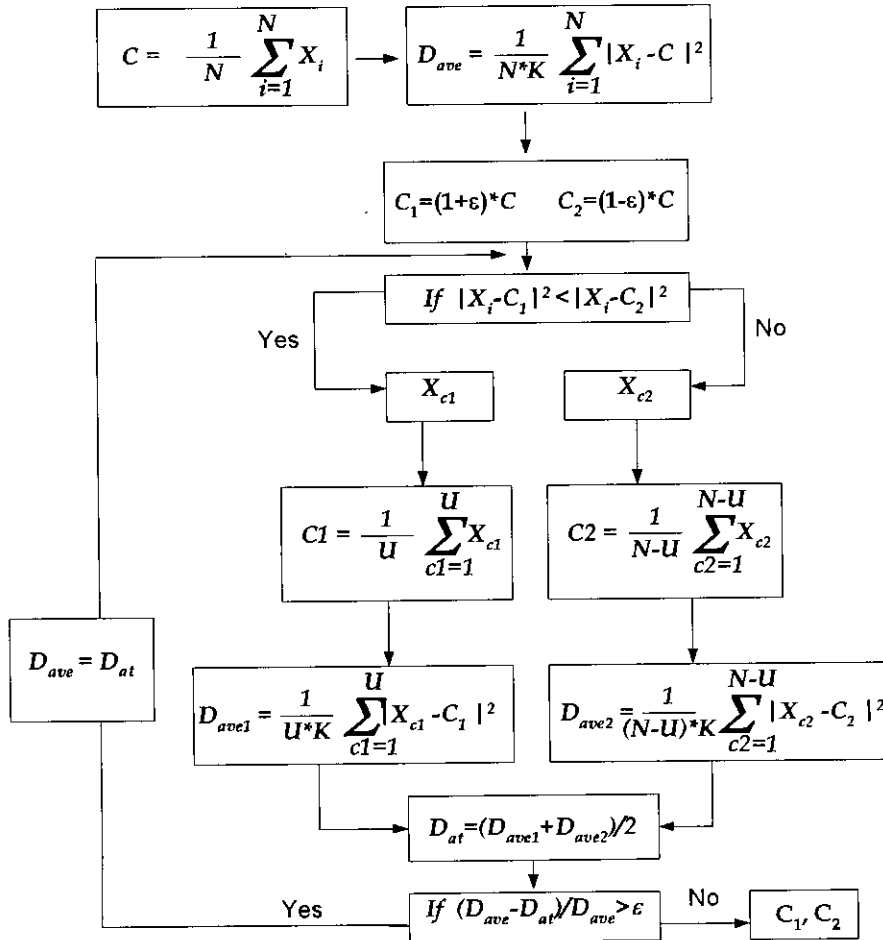


Figure 3.10: Schematic diagram of LBG technique

In this way the initial *code vector* C is split into two vectors C_1 and C_2 , then each of them is split into two (total four) and the whole process is repeated until desired number of *code-vectors* are obtained for the *code book* of the speaker. Generally, the number of code vectors depends on the number of speakers enrolled, and the number of frames taken for processing the code vectors.

This same process is applied for generation of *codebook* for each and every speaker. These *codebooks* are then be compared to identify the speaker from an unknown speech.

3.5.2 Feature Matching

Matching step in VQ consists of computing the *quantization distortion* between *feature vectors* produced from the unknown speech and *code vectors* of the *codebooks* that are already enrolled. Usually the nearest distance between *feature vectors* and the *code vectors* is used as the *quantization distortion*. Commonly it is done by partitioning the extracted *feature vectors*, using the minimum distance between the *feature vectors* and *code vectors* from the speaker *codebook*, and calculating the *quantization distortion*. Another choice for matching score is *mean squared error* (MSE), which is computed as the sum of the squared distances between the *feature vectors* and the *code vectors*, divided by the number of *feature vectors* extracted from the speech sample.

Let us assume that after feature extraction from an unknown speech, we have N number of *feature vectors*, and V number of speakers is enrolled in the system and each speaker has P number of *code vectors*. The number of operations needed for matching step is equal to $N*V*P$ as we need to calculate the distance between that *feature vector* (test vector) and every *code vector* of a *codebook* and select the *code vector* with the smallest distance. This process should be repeated for *codebook* of every speaker. This process is illustrated schematically in the figure 3.11.

Hence, each and every *feature vector* will select a *code vector* from every *codebook* according to smallest distance. In this way, there will be N number of smallest distances between N number of *feature vectors* and the selected *code vectors* for a *codebook*. Average of these distances will be the *quantization distortion* between the *feature vectors* of the unknown speech and the *codebook* of that speaker. Therefore, the number of *quantization distortions* will be equal to the number of speaker enrolled. Finally, the speaker that produces the minimum *quantization distortion* will be the identified speaker.

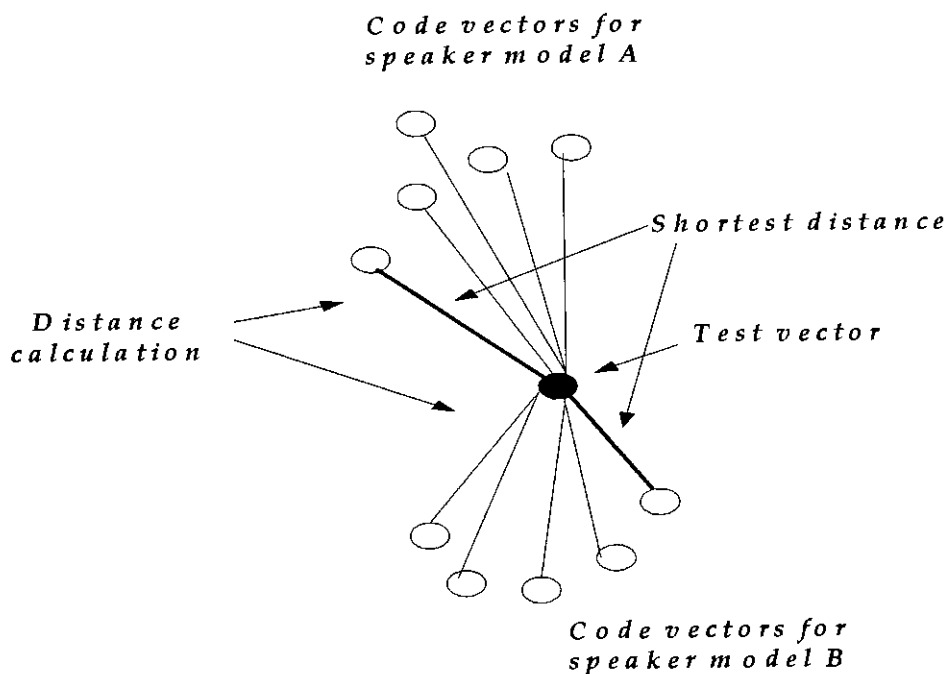


Figure 3.11: Schematic diagram of matching technique

According to the figure: 3.11, the *shortest distance* of each codebook will be measured for each and every codebook, then the codebook that produces the smallest *shortest distance* will be considered the speaker of the test speech.

Hence, each and every *feature vector* will select a *code vector* from every *codebook* according to smallest distance. In this way, there will be N number of smallest distances between N number of *feature vectors* and the selected *code vectors* for a *codebook*. Average of these distances will be the *quantization distortion* between the *feature vectors* of the unknown speech and the *codebook* of that speaker. Therefore, the number of *quantization distortions* will be equal to the number of speaker enrolled. Finally, the speaker that produces the minimum *quantization distortion* will be the identified speaker.

For *close-set* speaker identification, the unknown speech will belong to one of the speakers who are already enrolled. Therefore, the minimum quantization distortion will be sufficient to find the speaker identity. But for *open-set* speaker identification, unknown speech may not belong to the set of speakers who are enrolled in the system. In that case the *threshold quantization distortion* will be used to identify the speaker from its unknown speech. It can be found from the average of the all *quantization distortions* as stated in the equation (3.23):

$$D_n = f \cdot D_{av} \quad (3.23)$$

Where D_n is the *threshold quantization distortion*, D_{av} , average of all *quantization distortions* produce by all the speakers to the unknown speech except the minimum quantization distortion. The value of f can be determined empirically, which is a part of training the speaker identification system.

In summery the whole process, starting from the extraction of feature and identification of the speaker, is schematically depicted as in the figure 3.12.

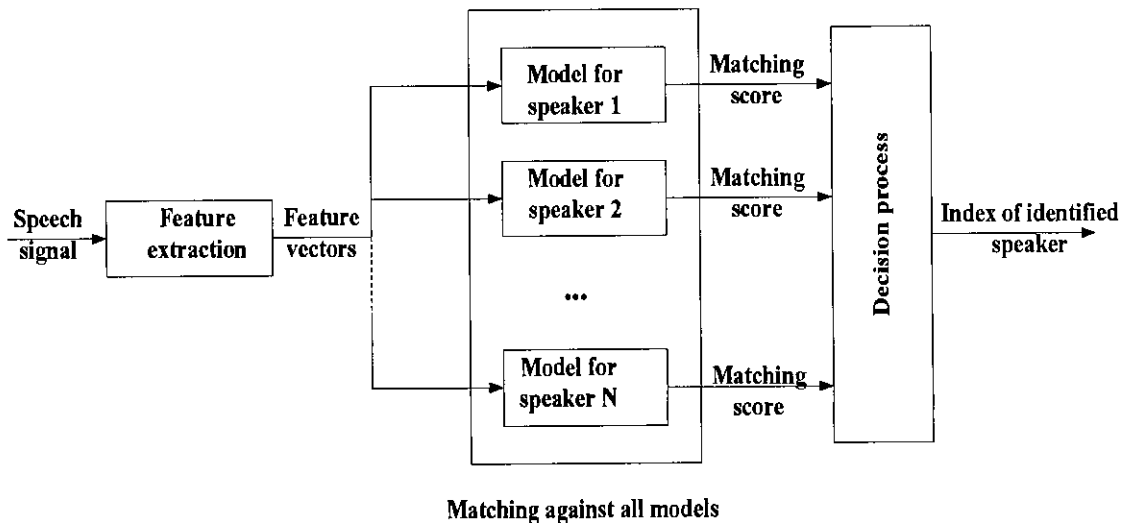


Figure 3.12: Schematic diagram of speaker identification process

3.6 Conclusion

In this chapter, the basis for the proposed model is given at the beginning. Then one of the two well-recognized techniques for cepstral coefficients derived from LPC (LPCC) is discussed to extract the properties of vocal tract. Later, the process of procuring the proposed model parameters is put forth thoroughly. This procured model parameters are used as the speaker discriminating features and with the help of VQ it is converted into the *codebook* of a speaker. Finally, the *codebook* that is used to identify speaker with the help of feature matching technique is discussed. This chapter gives a detailed description of extracting the vocal folds model (proposed model) parameters and identifying the speaker from these parameters.

CHAPTER 4

RESULTS

4.1 Introduction

At the beginning of this chapter the data acquisition process along with some of its specification for this research work is given. Following the discussion of the previous chapter, the speech data is processed to make the codebook of each individual speaker and test speech is matched for identification. For open set identification a test speech of a speaker is chosen whose codebook is not generated in the scheme. Then the results are made on both open-set and close-set for proposed method and for conventional method. Finally, a discussion is made upon the results as well as the comparison with previous method.

4.2 Data Acquisition

The experiment was carried over 4 individuals (2 males and 2 females). Each person performed monotonic speech of utterance in several voiced sounds in different pitch. The types of voiced sounds and the pitch are discussed below.

4.2.1 Selected voiced sounds and pitch:

Six different vowel-like sounds are chosen for this purpose:

1. /a/, as pronounced in the word 'father'.
2. /au/, as pronounced in the word 'autumn'.
3. /o/, as pronounced in the word 'low'.
4. /i/, as pronounced in the word 'fish'.
5. /u/, as pronounced in the word 'bull'.

6. /ae/, as pronounced in the word 'cat'.

Each speaker has to utter one vowel-like sound in eight different pitches. Pitch is increased from lower to higher in a fashion that the last pitch is one octave higher than the first one. As each speaker has to utter the same set of pitch sounds like every other speaker we have chosen the eight common pitch used in the musical domain. They are in Bengali, 'sa'(low), 're', 'ga', 'ma', 'pa', 'dha', 'ni', 'sa'(high) where the scale is fixed in the B-flat.

All 48 (6 sound*8 pitches) utterances are concatenated to produce the long speech of a speaker. Then, this long speech is subjected to *framing*, *pre-emphasis*, and *windowing*. Next, this windowed frame is used to find LPCC and from LPCC AR model parameters are found for that speaker. Each frame produces one set of AR model parameters or *feature vector*. Finally, these *feature vectors* are subjected to VQ and feature matching technique for speaker identification.

4.2.2 Technical Specification:

The sounds were recorded with a microphone, which is kept at a certain constant distance from all the speakers' mouth. Careful attention was made on the invariability of the mouth cavity. Other important notes are-

Sampling Frequency	44.1 KHz, CD quality
Quantization	16 bit
Channel	Mono
Recorder software	the default windows sound recorder
File Format	'*.wav' format.

Typical speech utterance (wave form) is attached in the *appendix B*

4.3 Procedural specification

The procedure to find the proposed model and identify the speaker with these model parameters is already discussed in the previous chapter (chapter 3). Here the specification is given related to the procedure:

Speech Framing	1000 samples, 23 ms (approximate), 50% overlapping
High pass filter coefficient	$\alpha = 0.99$
Prediction Order	24
Code Vectors	16
LBG specification	$\epsilon=0.01$
Threshold coefficient	$f=0.65$ (chosen empirically)
Program software	MATLAB 6.5

4.4 Results

Here the AR model parameters of vocal folds are used for speaker recognition through vowels using the VQ and feature matching technique, which is discussed in the foregoing chapter. This is also compared with results found by standard LPCC of vocal tract using the same VQ and feature matching technique. For the generation of the codebook, all 48 utterances (six vowels and eight pitches) of a speaker are used. And each utterance takes more than 60 frames. The results are given in the tabular form. In the following tables [4.1-4.4], the column indicates the pitch of an utterance and the row indicates how many vowels are taken for that particular pitch for identification purpose. For close set speaker identification, four individuals are used, but for open set speaker identification, another speaker's speech is taken, whose codebook was not in the system.

Table 4.1: Close set speaker identification through vowels using LPCC

Number of vowel	Pitch							
	sa(low)	re	ga	ma	pa	dha	ni	sa(hi)
One	63%	71%	75%	67%	71%	67%	71%	63%
Two	70%	78%	68%	70%	79%	84%	88%	91%
Three	81%	79%	85%	86%	92%	87%	84%	80%
Four	89%	78%	84%	89%	93%	95%	91%	87%
Five	94%	96%	97%	95%	89%	93%	96%	92%
Six	97%	98%	97%	96%	94%	91%	89%	93%

The results of the conventional LPCC and vector quantization based close-set speaker identification shown in the table: 4.1 using voiced sound. It is found from the table that the percentage matching is as low as 63% and the highest is 98%.

Table 4.2: Close set speaker identification through vowels using proposed model

Number of vowel	Pitch							
	sa(low)	re	ga	ma	pa	dha	ni	sa(hi)
One	100%	100%	100%	100%	100%	100%	100%	100%
Two	100%	100%	100%	100%	100%	100%	100%	100%
Three	100%	100%	100%	100%	100%	100%	100%	100%
Four	100%	100%	100%	100%	100%	100%	100%	100%
Five	100%	100%	100%	100%	100%	100%	100%	100%
Six	100%	100%	100%	100%	100%	100%	100%	100%

The results of the proposed vocal folds AR model parameters and vector quantization based close-set speaker identification shown in the table: 4.2 using voiced sound.

Table 4.3: Open set speaker identification through vowels using LPCC

Number of vowel	Pitch							
	sa(low)	re	ga	ma	pa	dha	ni	sa(hi)
One	42%	45%	47%	41%	46%	45%	44%	48%
Two	54%	49%	56%	48%	55%	53%	49%	52%
Three	58%	59%	57%	61%	60%	59%	58%	62%
Four	65%	63%	61%	62%	67%	65%	64%	63%
Five	67%	68%	64%	69%	65%	67%	69%	70%
Six	80%	81%	79%	76%	78%	77%	76%	73%

Results for open-set are presented in Tables 4.3 and 4.4 for the existing LPCC and vector quantization based identification and the proposed model respectively. The results in Table 4.3 show that it is really difficult to identify a speaker from the conventional for an open set. The percentage matching is as low as 41% and the highest is only 81%. The open-set identification is done by threshold quantization distortion. This is done in this way. The average of the quantization distortions (D_{av}) made by the test speech to all codebooks are taken apart from the minimum distortion. It is found in the experiment that wrong speaker produce quite larger quantization distortion compare to that of the right speaker. The minimum quantization distortion will be quite smaller to the D_{av} if this is produced by the right speaker. It is also found that the right speaker produce distortion 0.65 or less than the D_{av} .

Table 4.4: Open set speaker identification through vowels using proposed model

Number of vowel	Pitch							
	sa(low)	re	ga	ma	pa	dha	ni	sa(hi)
One	91%	93%	90%	92%	95%	94%	95%	96%
Two	99%	97%	98%	97%	99%	98%	98%	97%
Three	100%	100%	100%	100%	100%	100%	100%	100%
Four	100%	100%	100%	100%	100%	100%	100%	100%
Five	100%	100%	100%	100%	100%	100%	100%	100%
Six	100%	100%	100%	100%	100%	100%	100%	100%

It is seen in the table-4.4 that the proposed model can identify the speaker if the combination of vowels is at least three. This is for open set identification where the speaker identification is done if the unknown speech produces *quantization distortion* with any of the *codebooks* less than the *threshold quantization distortion*.

4.5 Discussion

For close set speaker identification the result is positive for the proposed model and not so convincing (can not identify speaker thoroughly) for the previous LPCC method. The reason why the LPCC did not do well in this test is that, the number of various utterances taken for generation of the codebook is very small (six), which cannot produce the average vocal tract property of the speaker. Instead it produces more the information of the speech uttered by the speaker than the speaker distinctive properties.

On the other hand, the new model gives better result (identify speaker thoroughly) for the same data sequence as LPCC. It may be concluded that vocal folds carry detectable signature of the speaker whose properties are more consistent for speaker identification. In the vowel output, the vocal folds property is found to be more pronounced. When the inverse filtering of the LPCC was taken, the properties of vocal tract were filtered out

from the overall properties and the properties of vocal folds are exposed. As vocal folds vary from person to person, it carries the information of a person (speaker). From the experimental results two important decisions can be made.

1. The vowels carry little information of the speaker as a property of vocal tract. Vocal tract only holds the property of the vowel i.e. the information of the speech.
2. Vocal folds properties vary very little when vowels are produced. Variation of vowels and pitch has little effect in the variation of the vocal folds properties.

The first decision is almost inevitable. But the second one is very interesting. These methods serve as a basis for future investigations in two ways: researchers can investigate the vocal folds properties which helps them for speech synthesis, voice pathology, speech coding and voice quality enhancement. Secondly, it can serve as a strong tool for speaker identification purpose especially for voiced sound.

For open set speaker identification, it is seen that the speaker identification is done if three vowels are taken at a time. This is because of vowels less than three, the property of the speaker is not very pronounced. When more than two vowels are taken the speaker's characteristics are strongly found in the codebook so that a *threshold quantization distortion* can be set to find whether the speaker of the unknown speech is in the system or not.

Another important aspect of this model is that, once the model (*code book*) is produced for a speaker it can identify the speaker if he/she speaks only a single vowel for close-set identification. For open-set it takes at least three vowels.

In this proposed model we use four speakers for identification purpose and it needs further investigation with larger number of speakers to establish the robustness of the proposed method. However, our objective was to investigate whether the vocal tract or

100902

vocal folds carry the better speaker discriminative information. From our results, it is clear that not the conventional vocal tract but the vocal folds (proposed model) carry the better speaker distinctive information.

CHAPTER 5

CONCLUSIONS

5.1 Discussions

In this thesis the AR model parameters of vocal folds are studied for speaker identification. Conventionally, speaker identification process works with the property of vocal tract. But throughout establishment of our method we try to make a point that the vocal folds properties can also be used for this purpose. The reason behind is that both vocal folds and vocal tract vary from person to person. LPCC and MFCC methods are base on the known evidence that the information carried by the low-frequency component of the speech signal is phonetically more important for human than that of the high-frequency components [22] which is related to vocal tract. LPCC separates the low-frequency information of speech from its higher one using LPC and convert it to its cepstrum where as MFCC warp the frequency to place more emphasis on the low frequency. In the proposed model LPCC are used to inverse filter the speech output so that the information of the high frequency is found which corresponds to vocal folds. This is tested for speaker identification purpose and gets very promising results for voiced sound. So it is evident that the high frequency information corresponding to vocal folds carry the information of the speaker especially for voiced sound.

5.2 Limitations

In the analysis presented here, the number of persons is taken four (two male and two female). As we have to acquire the data using the musical domain pitch, it was not very easy to get sufficient number of persons. However, if the number of person increases, some modification should be needed to identify a speaker.

For close set speaker identification it is thoroughly identify the speaker using only a single vowel, but for open set it needs at least three different vowels to identify a speaker.

Here we use only vowels for identification purpose, but consonant can be included here for increasing the robustness of the identification system.

5.3 Suggestions for Further Work

Based on the limitations discussed above, following works can be carried out in future:

In the future work, the number of persons will be increased for this purpose. In that case the number of *code vectors* should be increase to accommodate higher number of speakers. Frame size and number of prediction order may need to be increased in this purpose.

Again, this model can be tested for some other method like inverse filtering of MFCC or inverse filtering of first-order derivatives of cepstrum called delta features. There is another method based on MFCC found on [40] that can also be tested for this purpose.

Further, any kind of speech can be tested with this model. In that case, first the voiced part of the speech will be extracted from the speech. Then this will be used for procuring AR model parameters. Finally, identification will be done by the same way as it is done here.

Though the proposed model produces better results for vowels uttered at the right pitch, the same model may also be suited for normal speech. In that case, at first we have to extract the voiced part of the normal speech, and then the rest of the process will take exactly the same way. And that will also produce the similar results as the voiced sound carry the information of the speaker, whereas the uncharacteristic

constriction produced by the unvoiced sound is independent of the speaker. As for example, /s/ sound of 'six' never carry the information of the speaker.

This new system can be tested for the *Gaussian Mixture Model* (GMM) [41] instead of VQ model.

The vocal folds model so derived can be utilized for pathology, speech synthesis or speech enhancement for the future work.

REFERENCES:

- [1] Flanagan, J. L. "Voices of men and machines," *Journal of the Acoustical Society of America*, vol. 51, pp. 1375-1387, Mar. 1972.
- [2] Stewart, J. Q. "An electrical analogue of the vocal cords," *Nature*, vol.110 p.311,1922.
- [3] Proakis, J.G., and Manolakis, D.G., "Digital Signal Processing Principles, Algorithm and Applications," 3rd edition, *Prentice-Hall,Inc.*, Sep. 2002.
- [4] Ishizaka, K., and Flanagan, J.L., "Synthesis of Voiced Sounds from a Two-Mass Model of the Vocal Cords," *Bell Syst.Tech.J.*, vol.56, no.6, pp.889-918, 1972.
- [5] Fant, C.G.M., "Acoustic Theory of Speech Production," *The Hague, The Netherlands, Mounton*, 1960.
- [6] Childers, D.G., and Lee, C.K., "Vocal quality factors: Analysis, synthesis and perception," *Journal of the Acoustic Society of America*, vol.90, no.5, pp.2394-2410, November 1991.
- [7] Lalwani, A.L., and Childers, D.G., "Modeling vocal disorders via formant synthesis," *Proceedings of the IEEE*, pp.505-508,1991.
- [8] Yanguas, L.R., Quatieri, T.F. and Goodman, F., "Implications of glottal excitation for Speaker and dialect identification," *IEEE Int. conf. Acoustics , Speech and Signal Processing, Phoenix, Arizona*, March 1999.
- [9] Fant, G.Liljencrants, J. and Lin,Q. " A four-parameter model of glottal flow," *STL-QPSR*, vol.85 no.2 , pp.1-13,1985.
- [10] Veldhuis, R.," A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation," *Journal of the Acoustic Society of America*, vol.103, pp.566-571, 1998.
- [11] Klatt D. and Klatt L., "Analysis, synthesis and perception of voice quality ariations among female and male talkers," *J. Acoust. Soc. Am.* vol.87, pp.820-857, 1990.
- [12] Ishizaka, K., and Flanagan, J.L., "Synthesis of Voiced Sounds from a Two-Mass Model of the Vocal Cords," *Bell Syst.Tech.J.*, vol.51, pp.1233-1268, 1972.

- [13] Kob, M., Alhauser, N., and Reiter, U., "Time-domain model of the singing voice," *Proc. of DAFx99 Workshop*, pp.143-146, Norway, Dec. 1999.
- [14] Furui, S., "Digital Speech Processing, Synthesis and Recognition," *New York, Marcel Dekker*, 2001.
- [15] Titze, I.R., "The human vocal cords: A mathematical model," *Part 1*, vol. 28: 129-170, 1973.
- [16] Titze, I.R., "The human vocal cords: A mathematical model," *Part 2*, vol.29, pp. 1-12, 1974.
- [17] Titze, I.R., "Parameterization of the glottal area, glottal flow and vocal fold contact area," *J. Acoust. Soc. Am.*, vol.75, No.2, pp.520-580, 1984.
- [18] Titze, I.R., "A four-parameter model of the glottis and vocal fold contact area," *Speech Comm.*, Vol.8, pp.191-201, 1989.
- [19] Do, M., and Wagner, M., "Speaker Recognition with Small Training Requirements Using a Combination of VQ and DHMM," *Proc. of Speaker Recognition and Its Commercial and Forensic Application*, pp.169-172, Avignon, France, April 1998.
- [20] Karpov, E., "Real-Time Speaker Identification," *University of Joensuu, Department of Computer Science*, Jan. 2003.
- [21] Sherwood, L., "Human physiology," *St Paul, Minn. West Publishing*, 1989.
- [22] Deller, J.R., Hansen, J.H.L., and Proakis, J.G., "Discrete-Time Processing of Speech Signals," *Piscataway (N.J), IEEE Press*, 2000.
- [23] Huang, X., Acero, A., and Hon, H. W., "Spoken language processing," *Upper Saddle River, New Jersey, Prentice Hall PTR*, 2001.
- [24] Campbell, J.P., "Speaker Recognition: A Tutorial," *Proc. of the IEEE*, vol.85 no.9, pp.1437-1462, Sept. 1997.
- [25] Flanagan, J.L., and Landgraf, I.L., "Self-oscillating source for vocal tract synthesizer," *IEEE Trans. Audio And Electro-acoustics*, Vol.16, pp57-64, 1968.
- [26] Carlo D., and Federico A., "Model-Based Synthesis and Transformation of Voiced Sounds," *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00), Verona, Italy*, Dec. 2000.

- [27] Federico A., Carlo D., and Paavo A., "Synthesis of the Voice Source using a Physically- Informed Model of the Glottis," *Helsinki University of Technology, Lab. of Acoustic, Isma2001-Revised Version*, p. 1, 2001.
- [28] Mammone, R.J., Zhang, X., and Ranachandra, R.P., "Robust speaker recognition: a feature based approach," *IEEE Sig. proc. mag.*, Sept. 1996.
- [29] Story, B. H., and Titze, I.R., " Voice simulation with a body-cover model of the vocal folds," *J. Acoust. Soc. Am.*, vol. 97, no.2 pp. 1249-1260, 1995.
- [30] Berg, V.D., and Jew, "An electrical analogue of the trachea, lungs and tissues," *Acta Physiol. Pharnacol. Neer landica*, vol. 9,pp. 361-385, 1960.
- [31] Martin R., "An Interactive Model for the Voice Source," *Proceedings of the Vocal Fold Physiology Conference , Madison Wisconsin*, May, 1981.
- [32] Wong, D. J., Markel, J.D., and Gray, A. H., "Least squares glottal inverse filtering from the acoustic speech wave," *IEEE Transanctions on Acoustics, Speech, and Signal Processing* , vol. ASSP-27, pp.350-355, August, 1979.
- [33] De Veth, J., Cranen, B., and Strik, H., " Extraction of control parameters for the voice source in a text-to-speech system," *Proceeding of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 301-304, April, 1990.
- [34] Kumaresan, R., Tufts, D.W., and Scharf, L.L., "A prony method for noisy data: Choosing the signal components and selecting the order in exponential signal models," *Proceedings of the IEEE*, vol.72, pp.230-233, Feb.1984.
- [35] Gish, H., and Schmidt, M., "Text independent Speaker Identification," *IEEE Signal Processing Magazine*, vol.11, no.4, pp.18-32,1994.
- [36] Rabiner, L., and Juang, B.H., "Fundamentals of speech recognition," *Englewood Cliffs(N.J.) Prentice Hall Signal Processing Series*,1993.
- [37] Atal, B., "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification ," *J. Acoust. Soc. Am.* vol.55, pp.1304-1312, 1974.

- [38] Ljung, L., "System Identification: Theory for the User," *Englewood Cliffs, (NJ) Prentice Hall*, Pgs. 278-280, 1987.
- [39] Atal, B.S., "Automatic Recognition of Speakers from their Voices," *Proceeding of the IEEE*, vol.64, pp.460-475, 1976.
- [40] Ezzaidi, H., Rouat, J., and O'Shaughnessy, D., "Towards Combining Pitch and MFCC for Speaker Identification Systems," *Aalborg, Eurospeech, Scandevia*, 2001.
- [41] Reynolds, D., and Rose, R., "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models," *AIIEE transaction on speech and audio processing*, vol.3, no.1, pp.72-8, 1995.

Appendix

Typical speech utterance wave-form for different vowels.

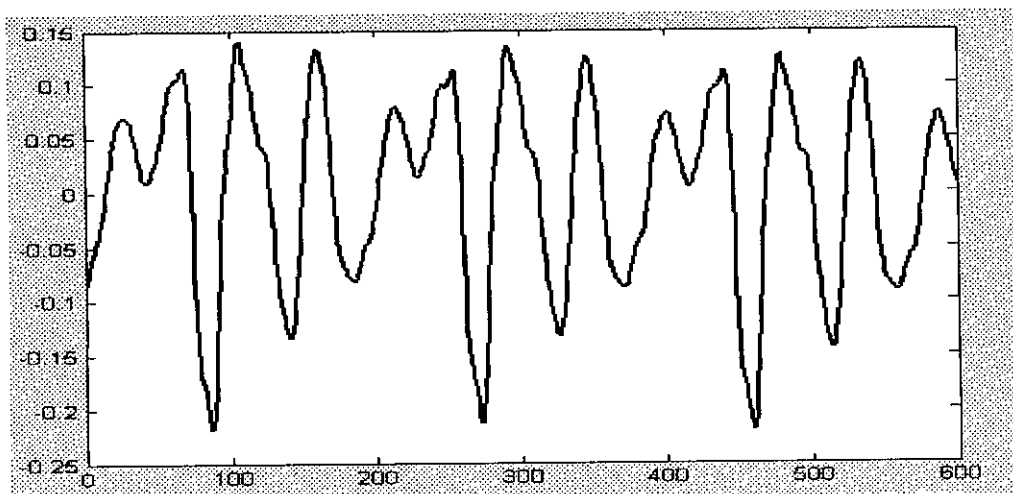


Figure 1: voice : "aa" pitch : sal(low)

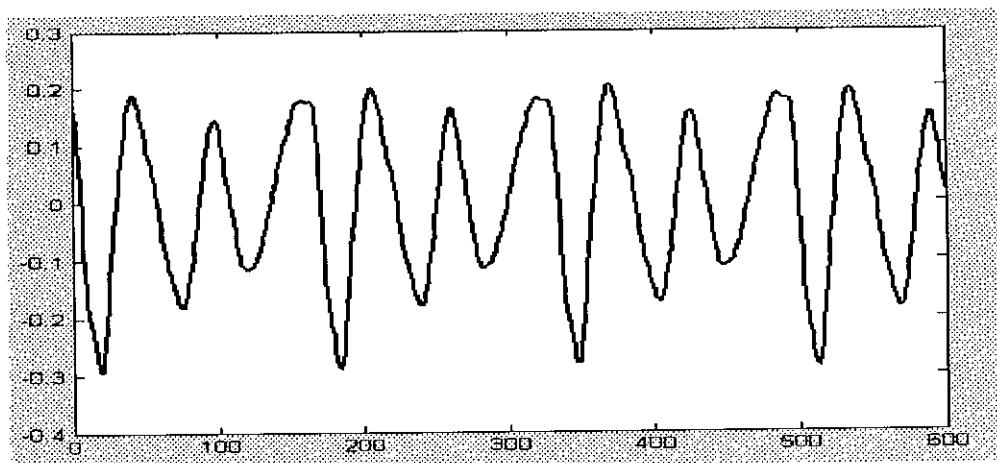


Figure 2: voice : "aa" pitch : re

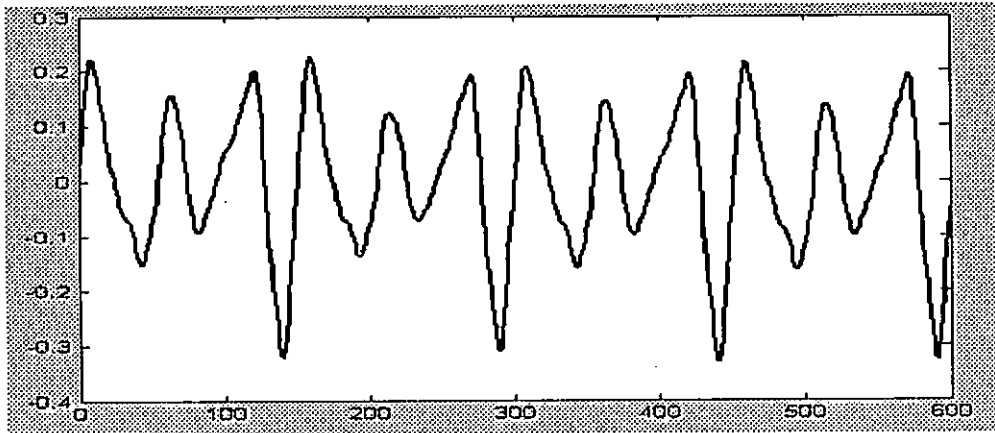


Figure 3: voice : "aa" pitch : ga

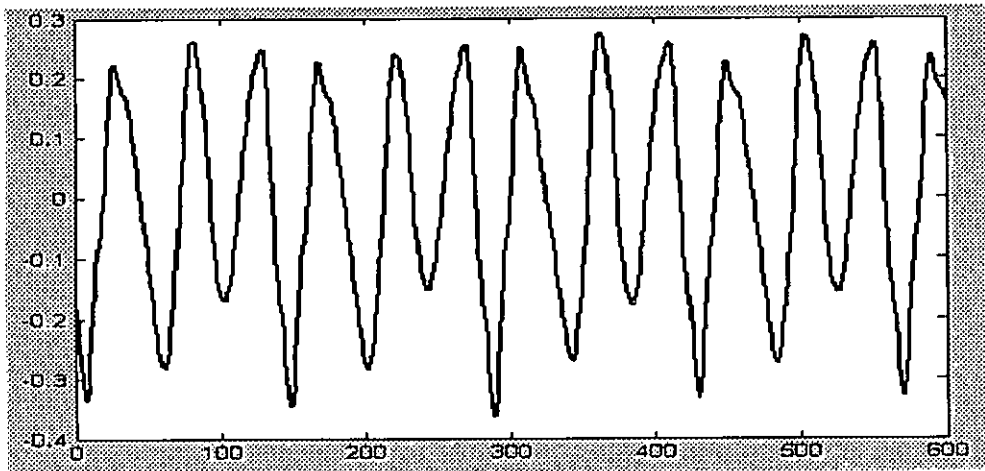


Figure 4: voice : "aa" pitch : ma

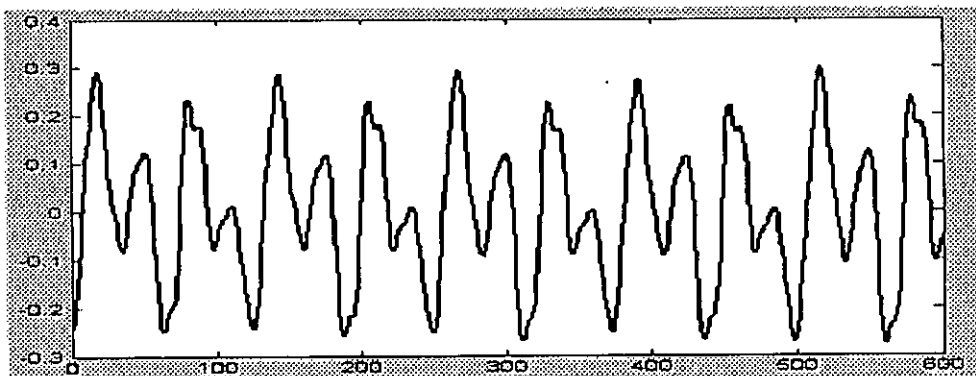


Figure 5: voice : "aa" pitch : pa

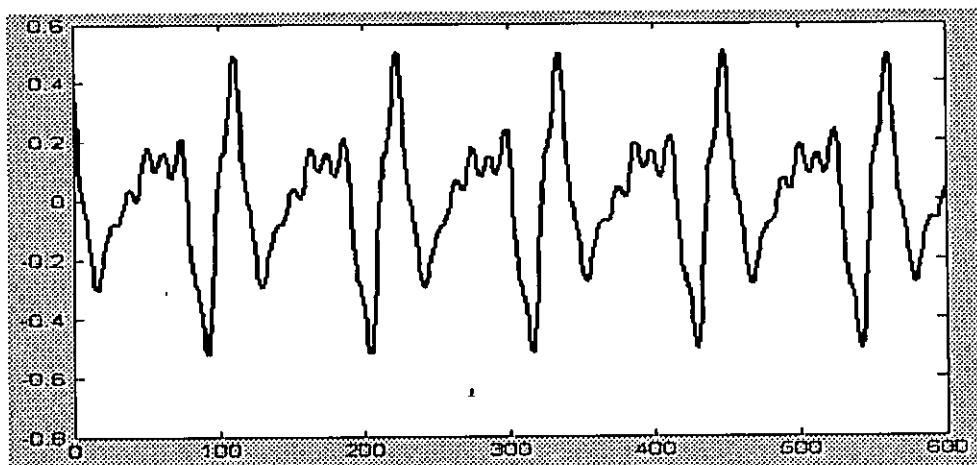


Figure 6: voice : "aa" pitch : dha

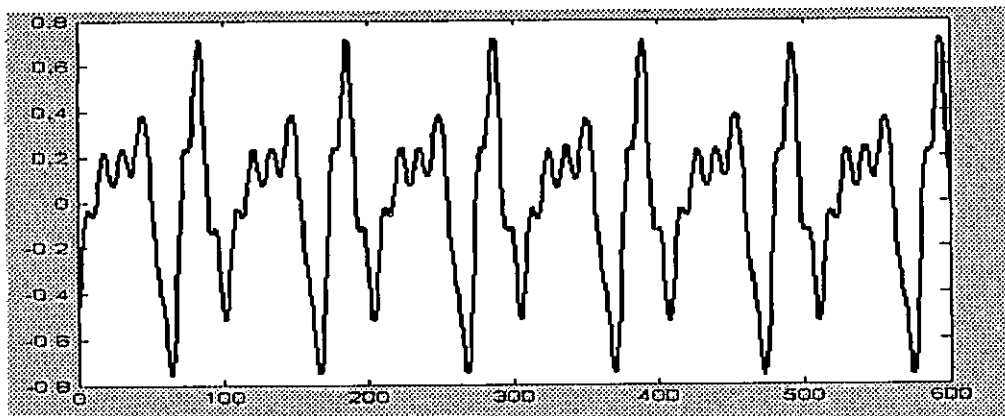


Figure 7: voice : "aa" pitch : ni

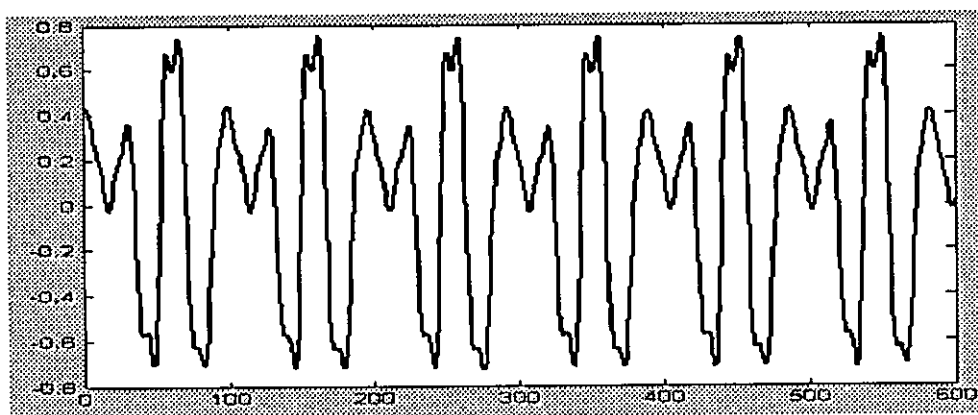


Figure 8: voice : "aa" pitch : sa(hi)

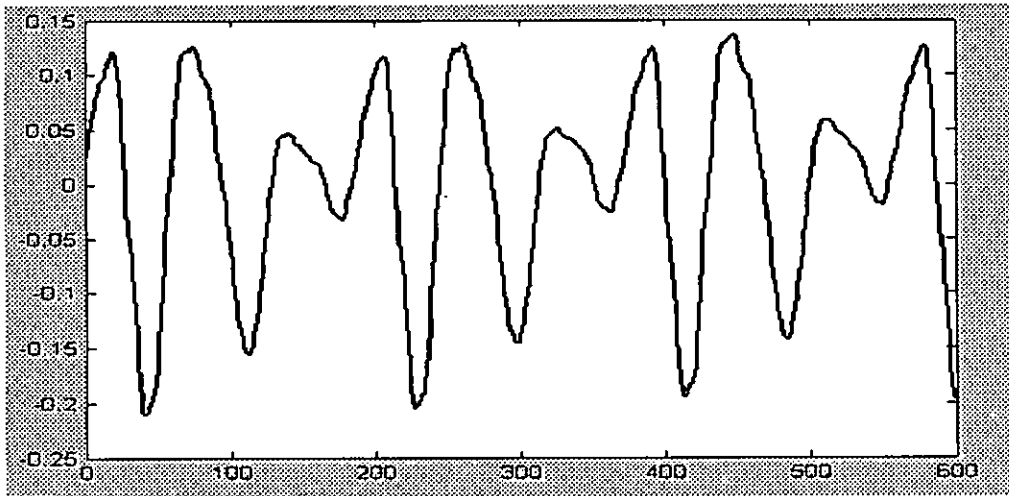


Figure 9: voice : "au" pitch : sa(low)

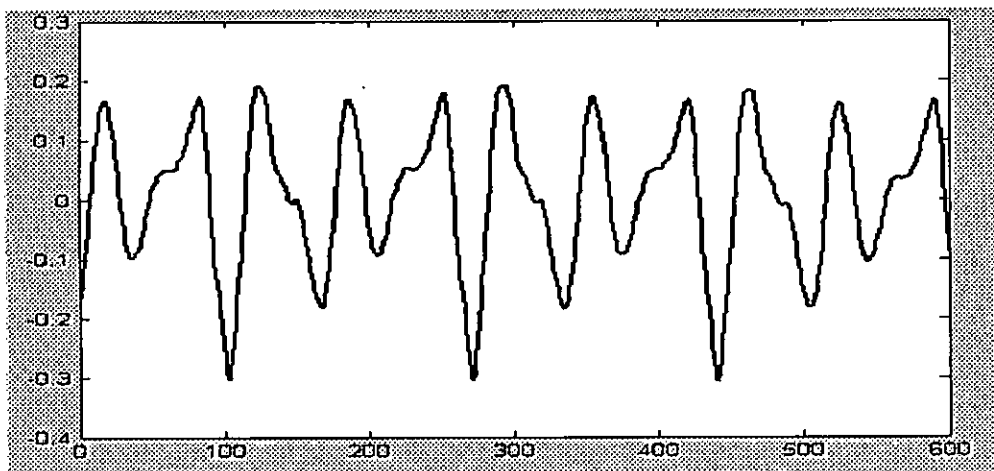


Figure10: voice : "au" pitch : re

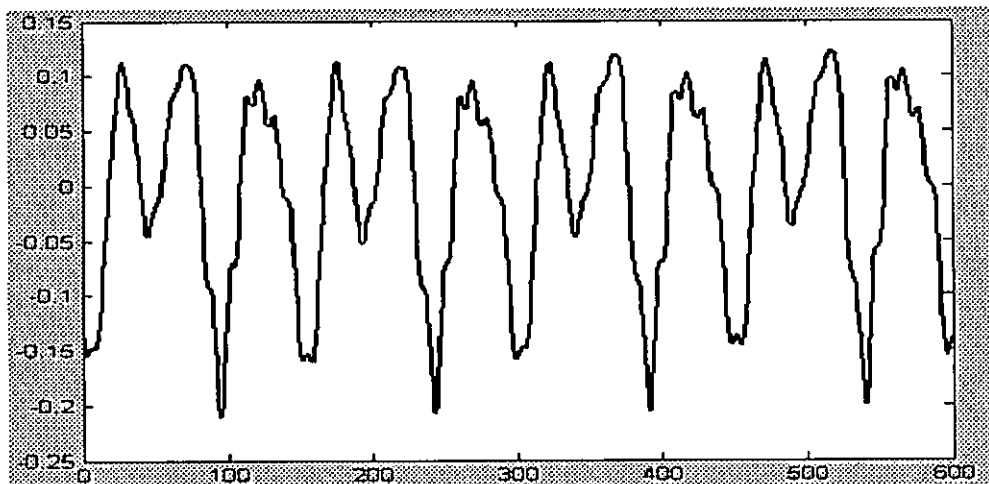


Figure11: voice : "au" pitch : ga

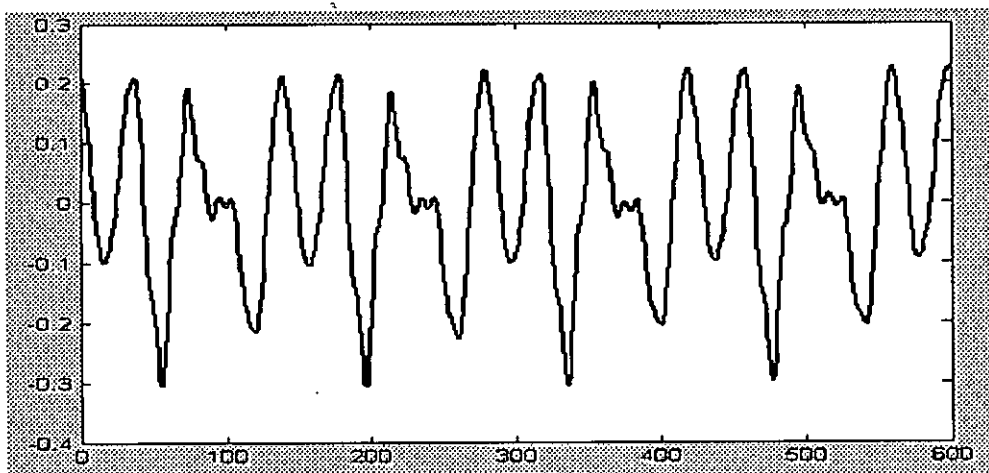


Figure12: voice : "au" pitch : ma

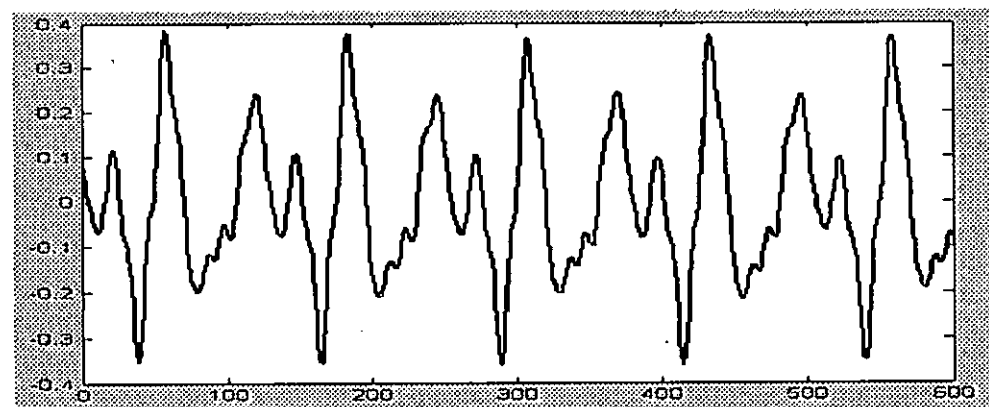


Figure13: voice : "au" pitch : pa

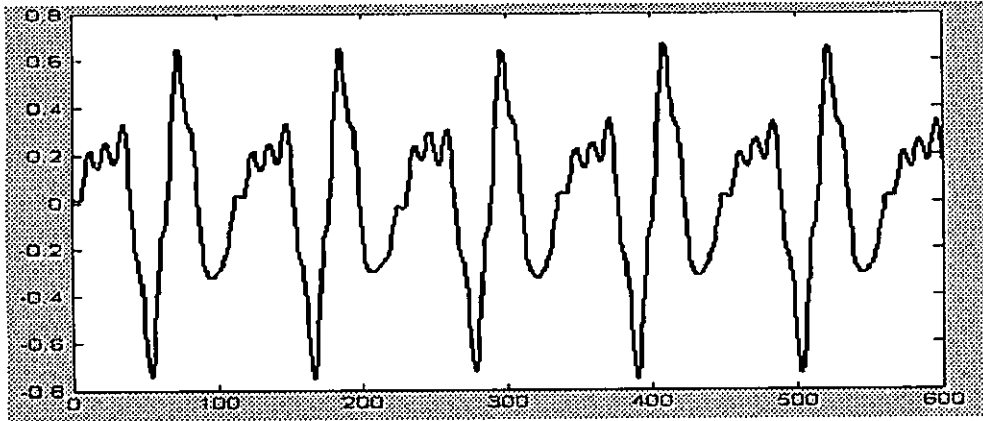


Figure14: voice : "au" pitch : dha

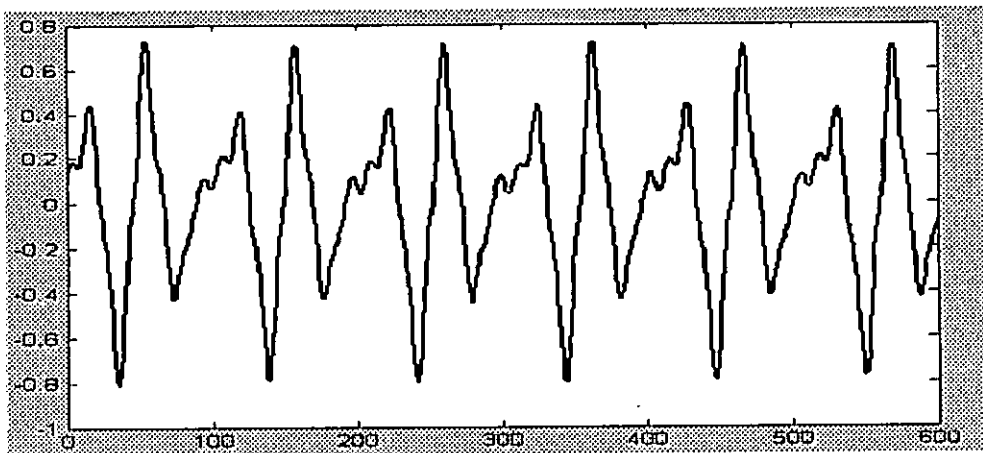


Figure15: voice : "au" pitch : ni

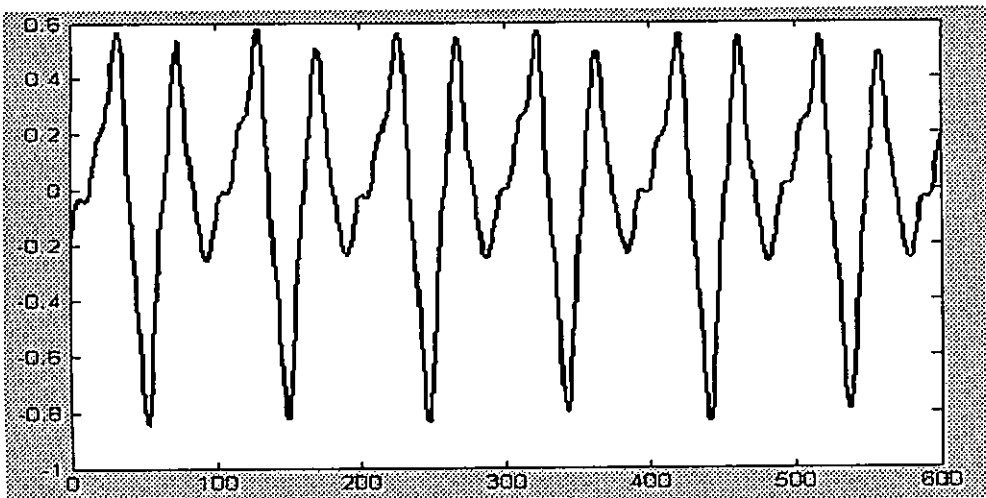


Figure16: voice : "au" pitch : sa(hi)

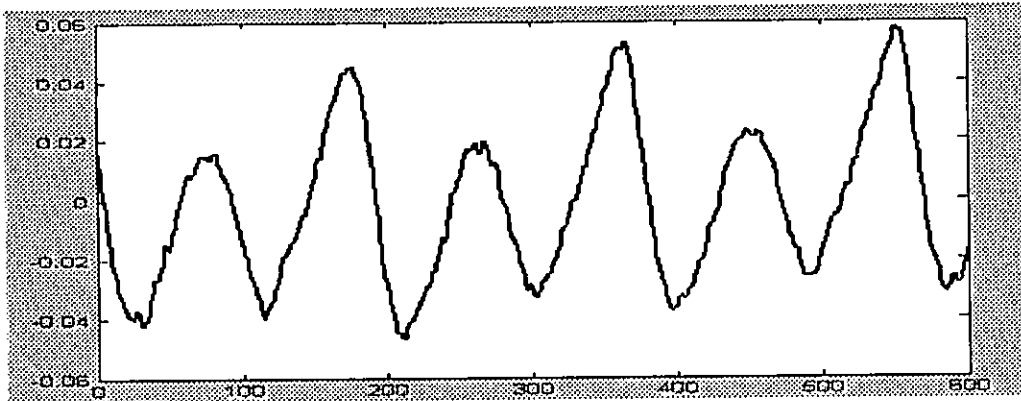


Figure10: voice : "uu" pitch : sa(low)

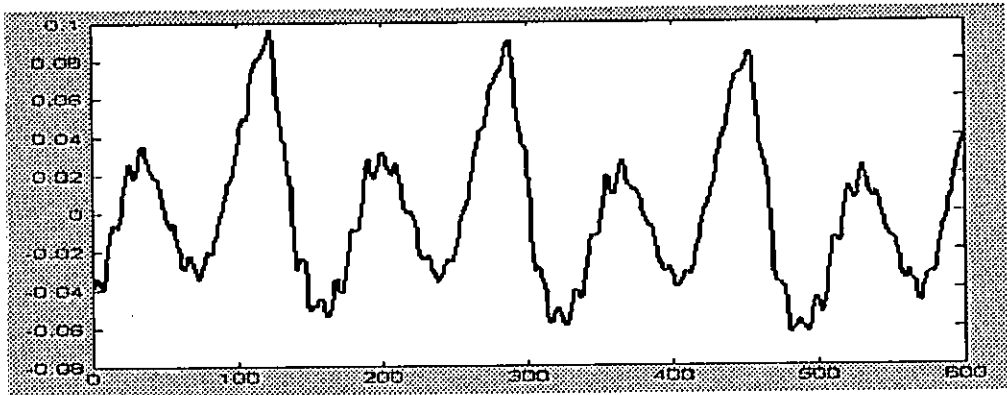


Figure18: voice : "uu" pitch : re

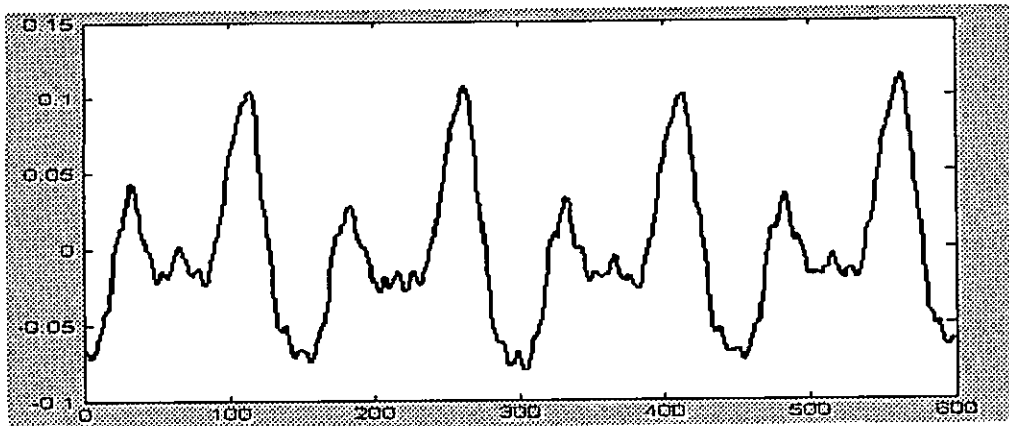


Figure19: voice : "uu" pitch : ga

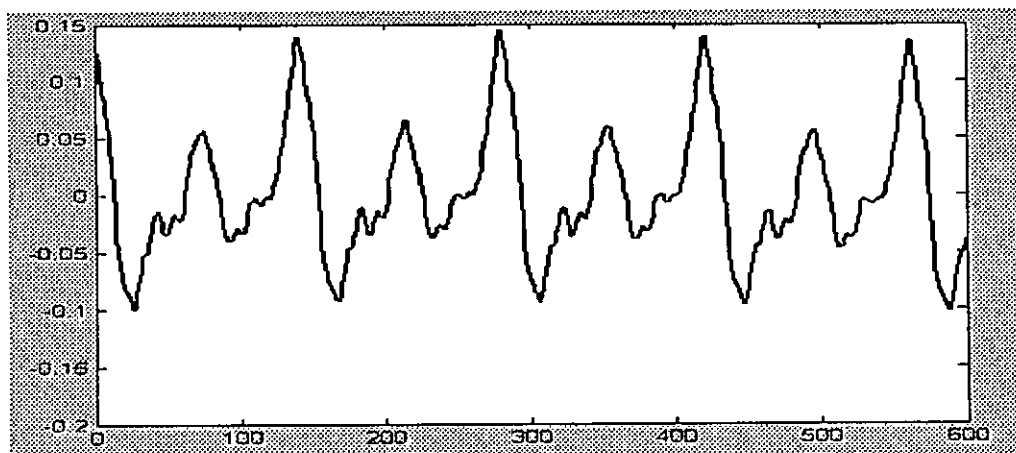


Figure20: voice : "uu" pitch : ma

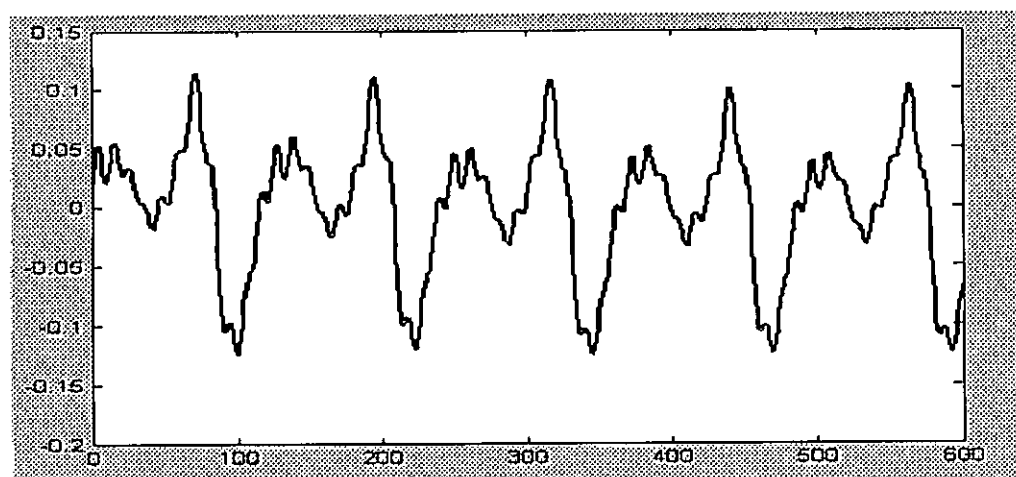


Figure21: voice : "uu" pitch : pa

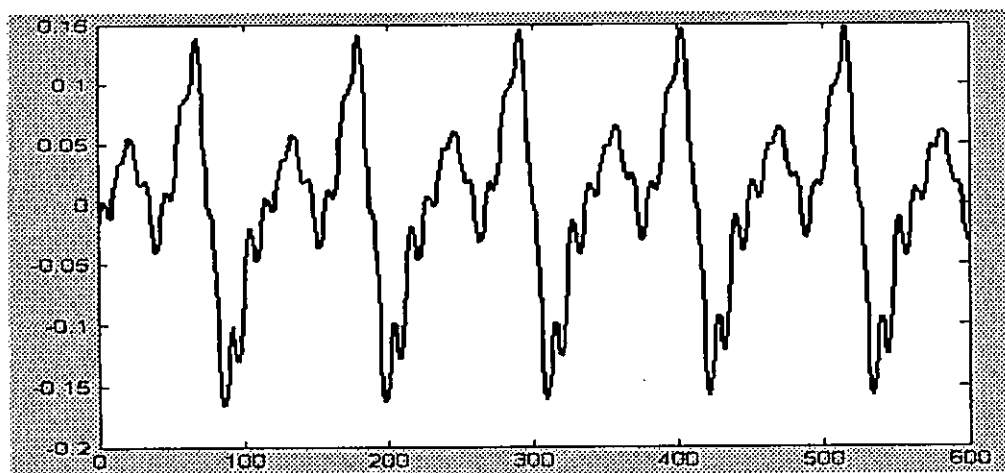


Figure22: voice : "uu" pitch : dha

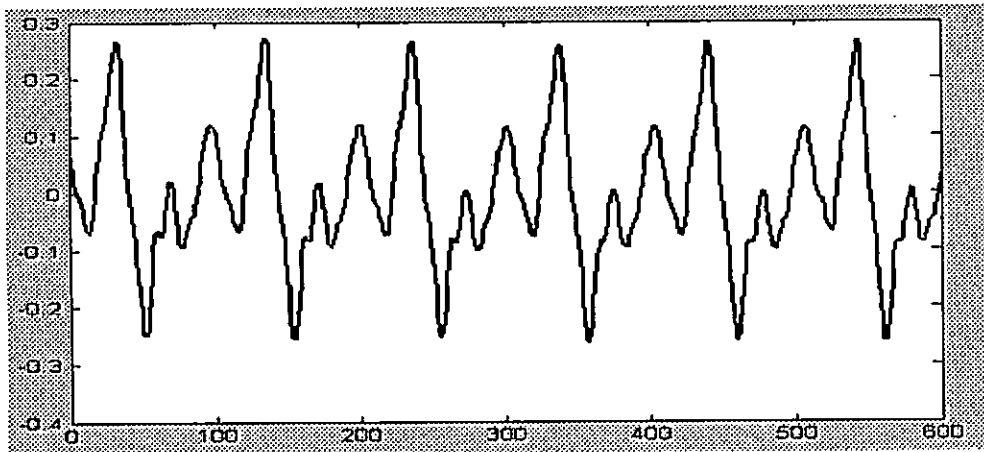


Figure23: voice : "uu" pitch : ni

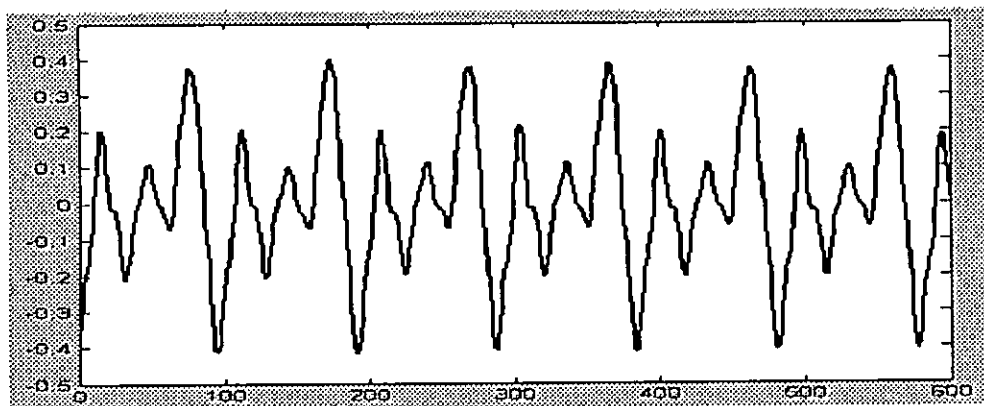


Figure24: voice : "uu" pitch : sa(hi)

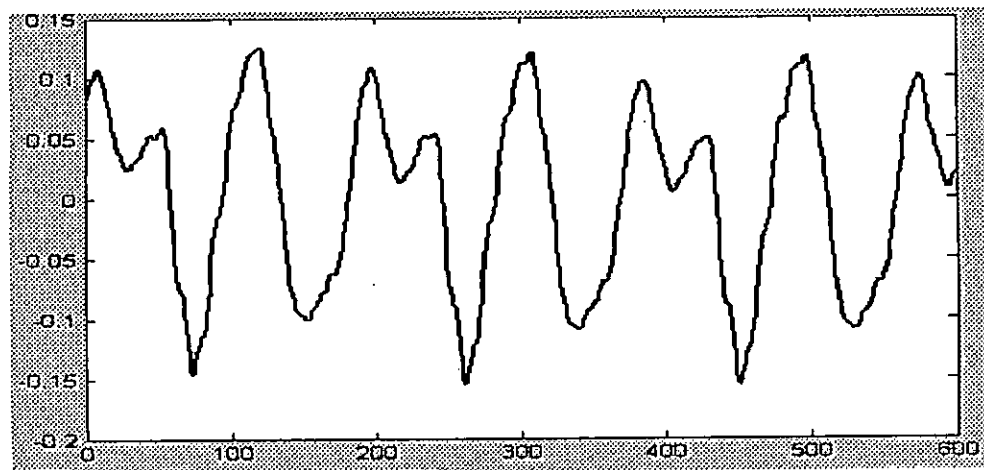


Figure25: voice : "oo" pitch : sa(low)

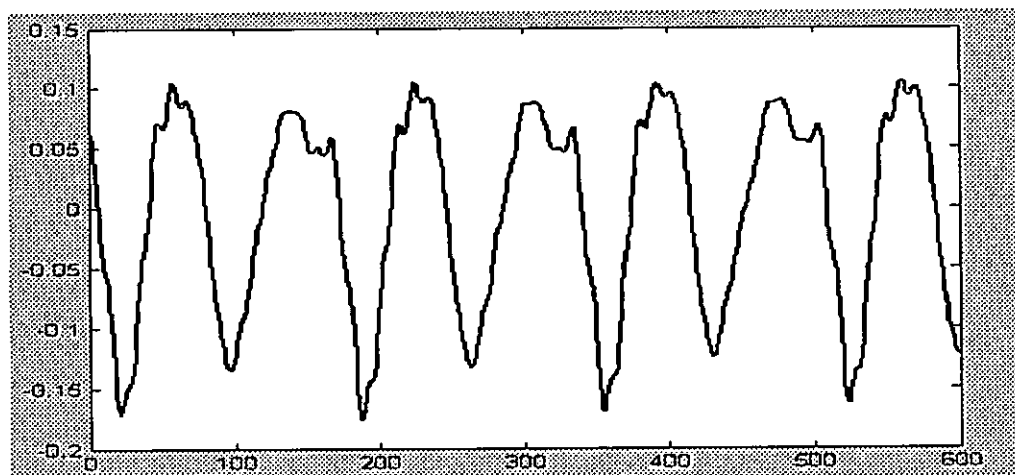


Figure26: voice : "oo" pitch : re

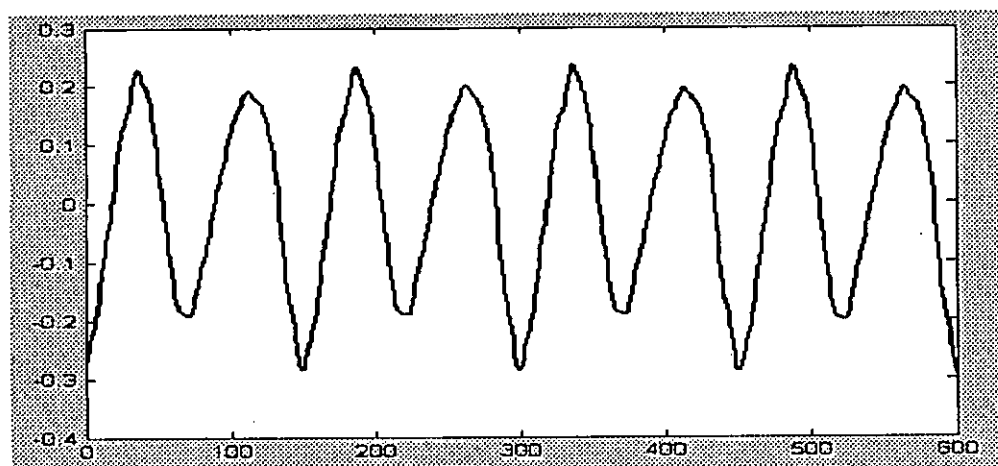


Figure27: voice : "oo" pitch : ga

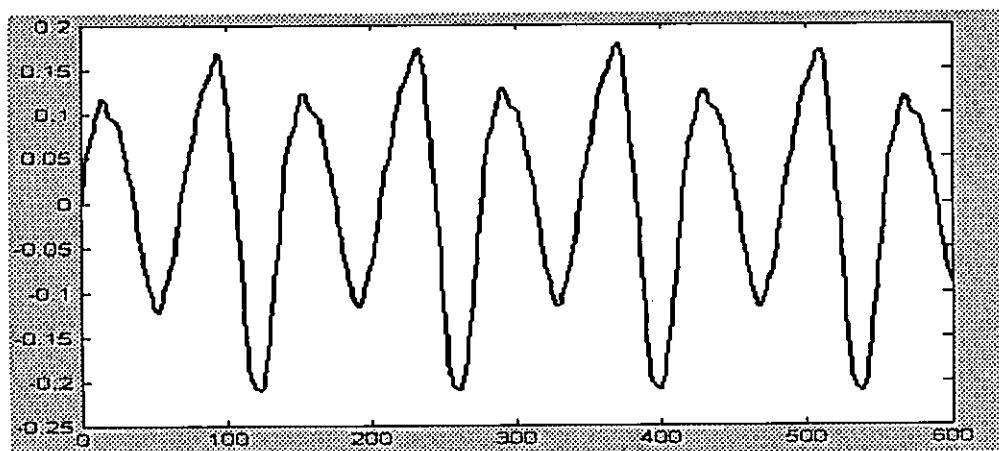


Figure25: voice : "oo" pitch : ma

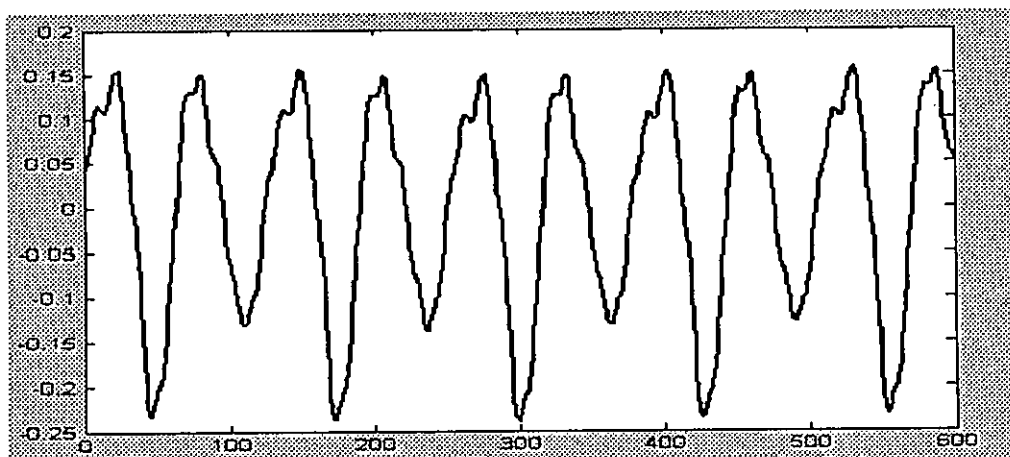


Figure29: voice : "oo" pitch : pa

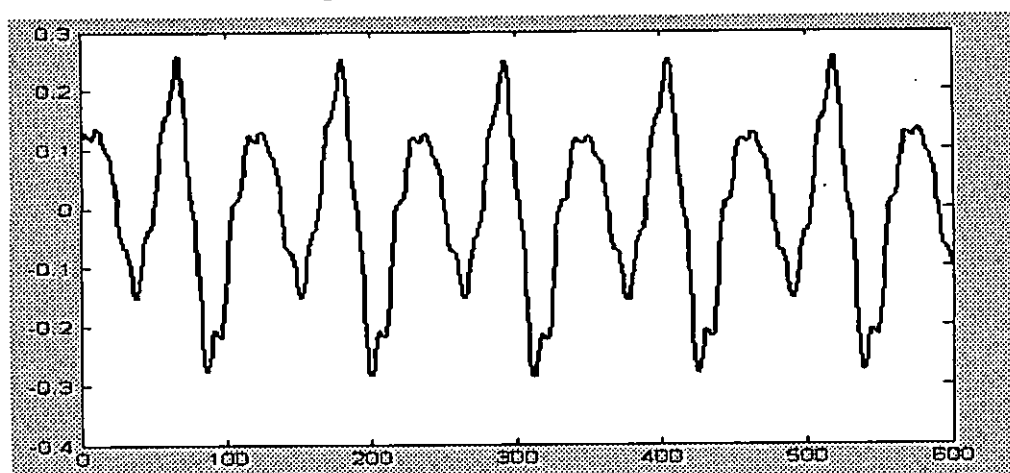


Figure30: voice : "oo" pitch : dha

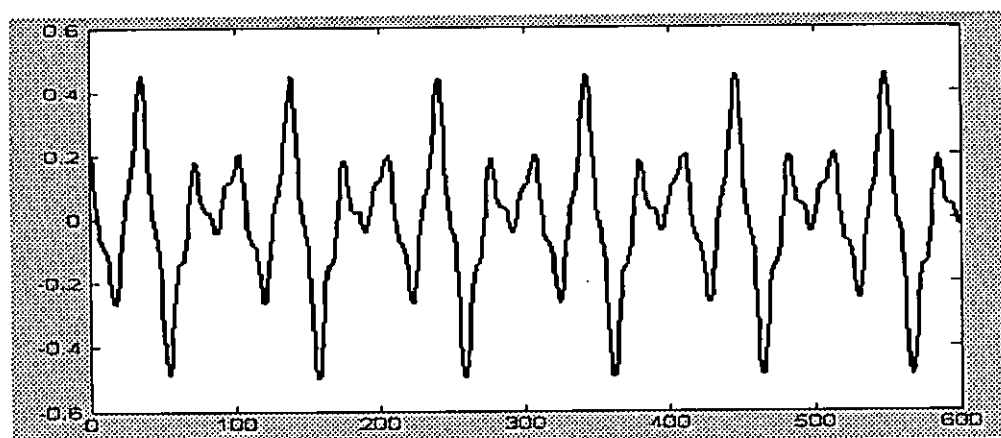


Figure31: voice : "oo" pitch : ni

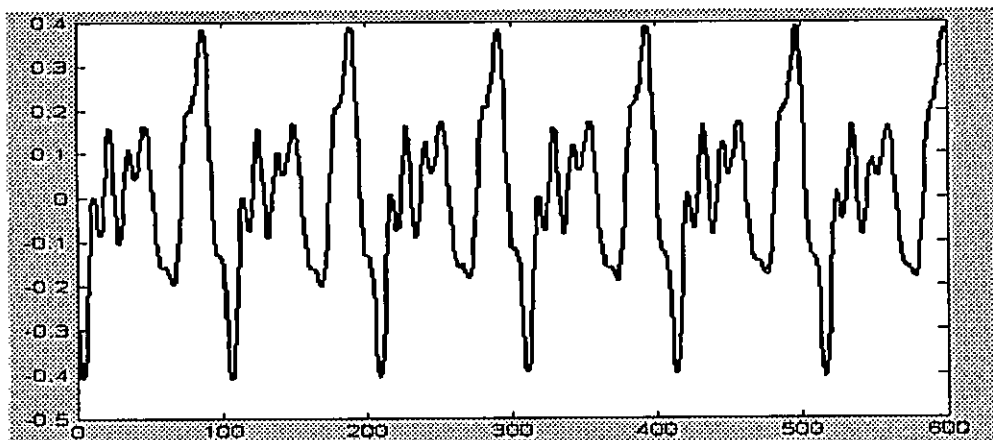


Figure32: voice : "oo" pitch : sa(hi)

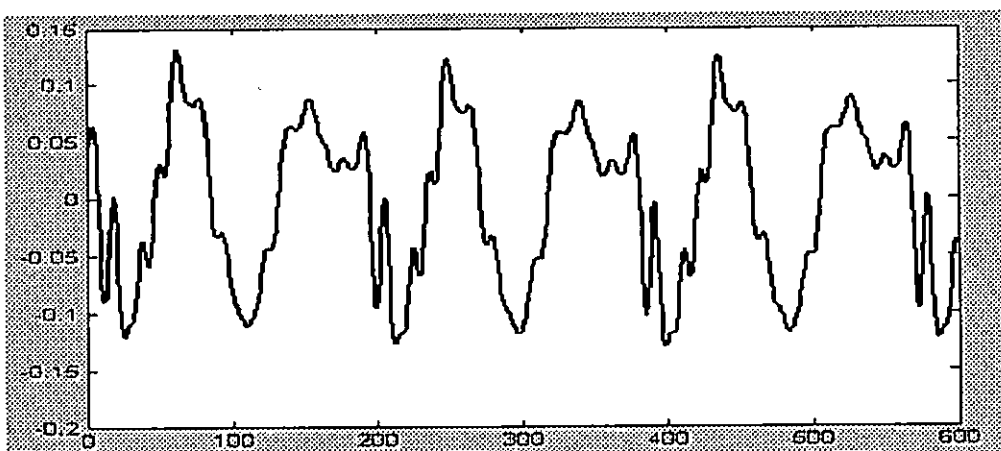


Figure33: voice : "ee" pitch : sa(low)

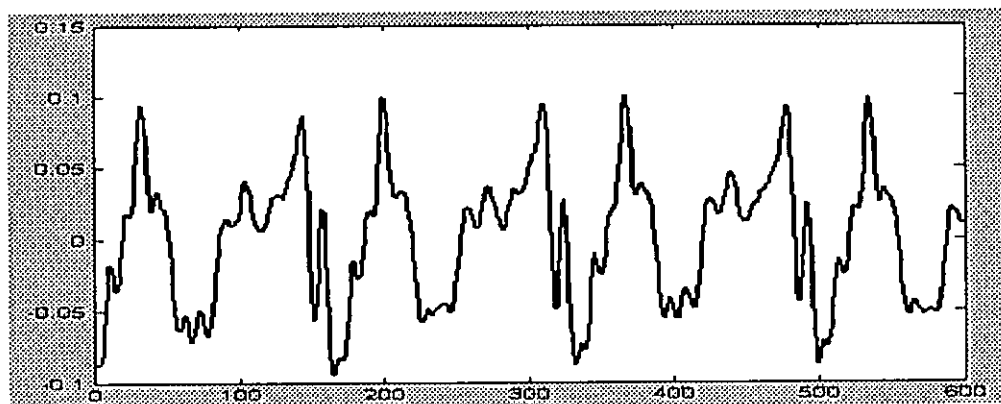


Figure34: voice : "ee" pitch : re

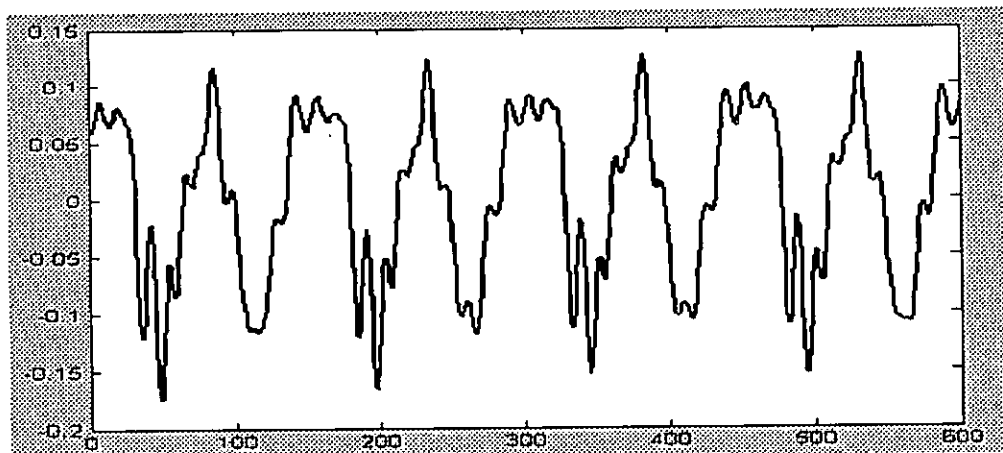


Figure35: voice : "ee" pitch : ga

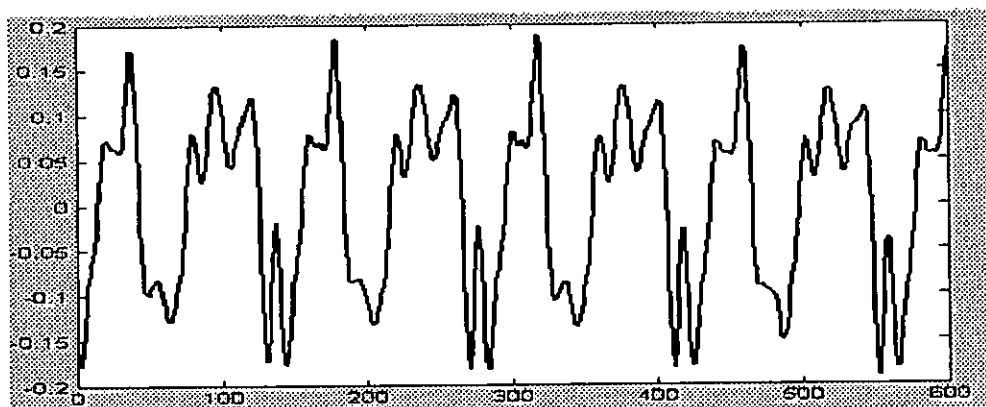


Figure36: voice : "ee" pitch : ma

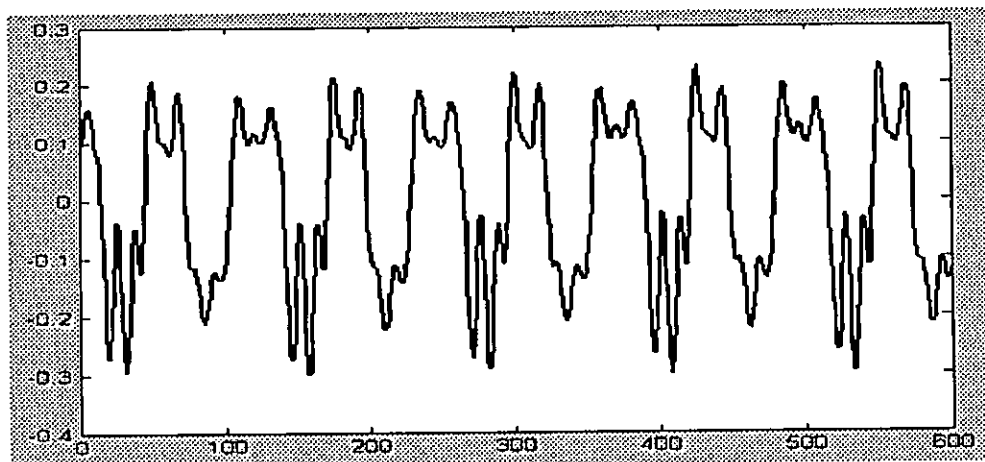


Figure37: voice : "ee" pitch : pa

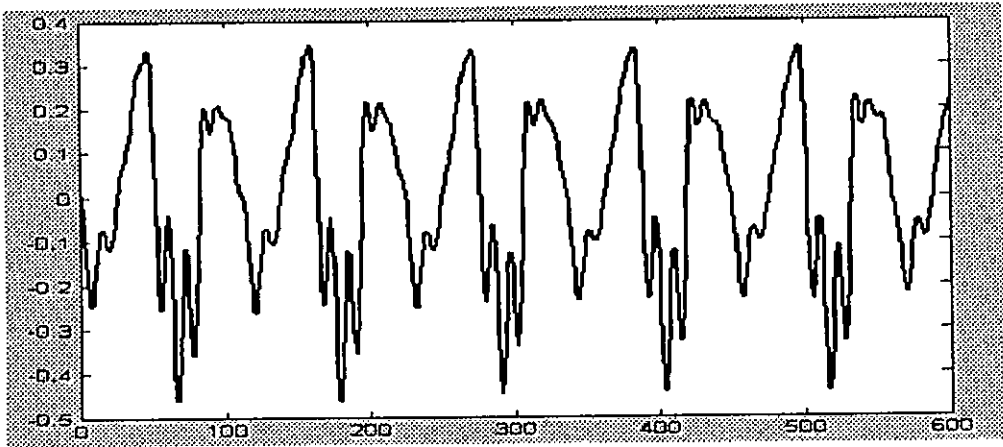


Figure38: voice : "ee" pitch : dha

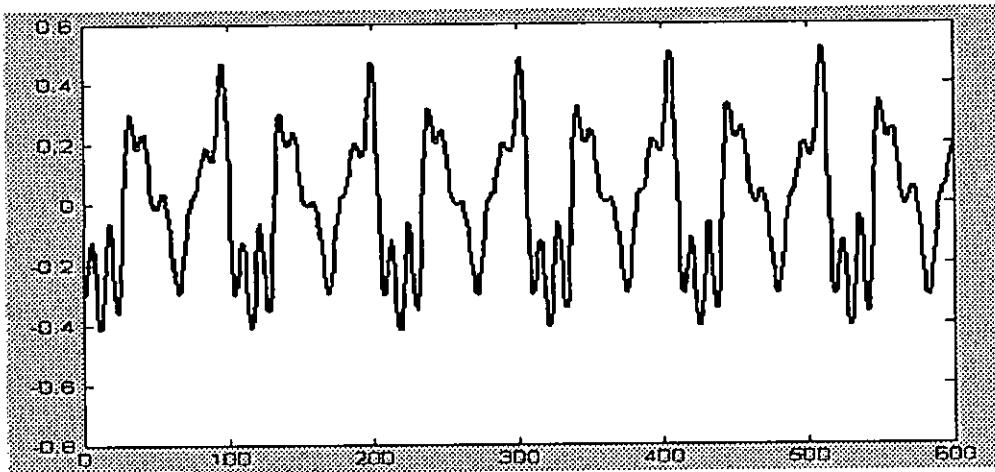


Figure39: voice : "ee" pitch : ni

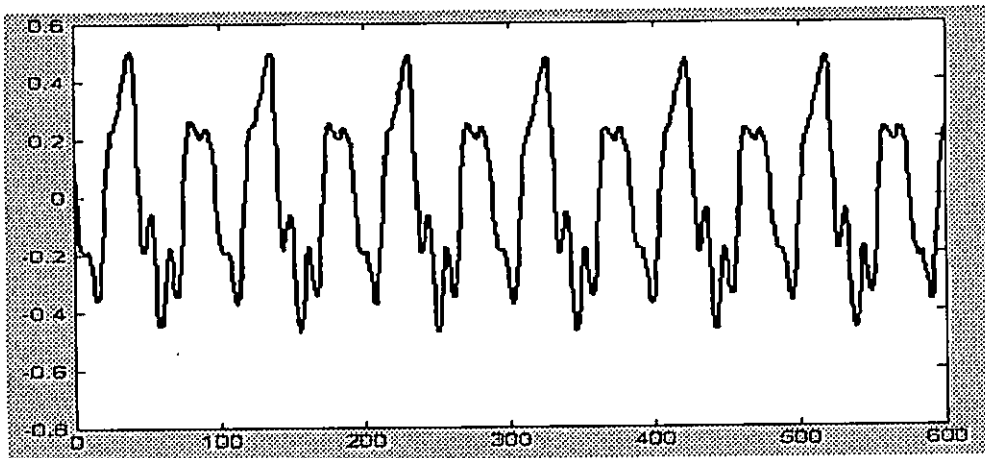


Figure40: voice : "ee" pitch : sa(hi)

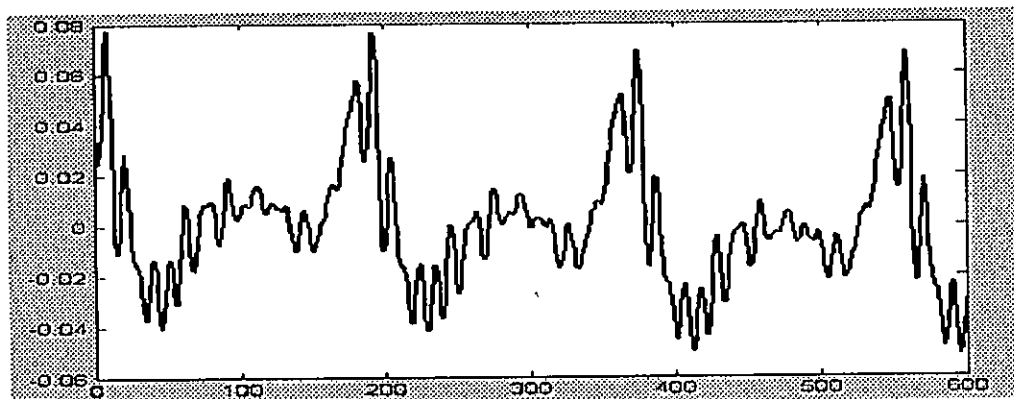


Figure41: voice : "ii" pitch : sa(low)

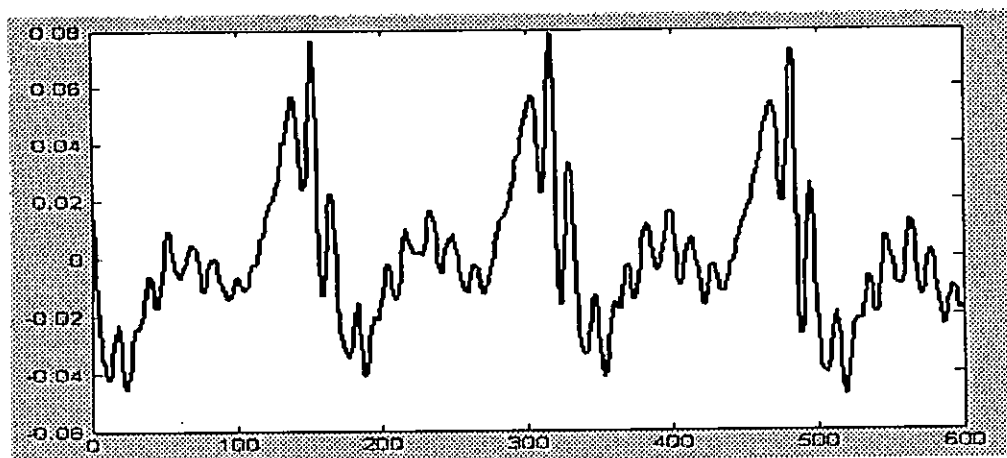


Figure42: voice : "ii" pitch : re

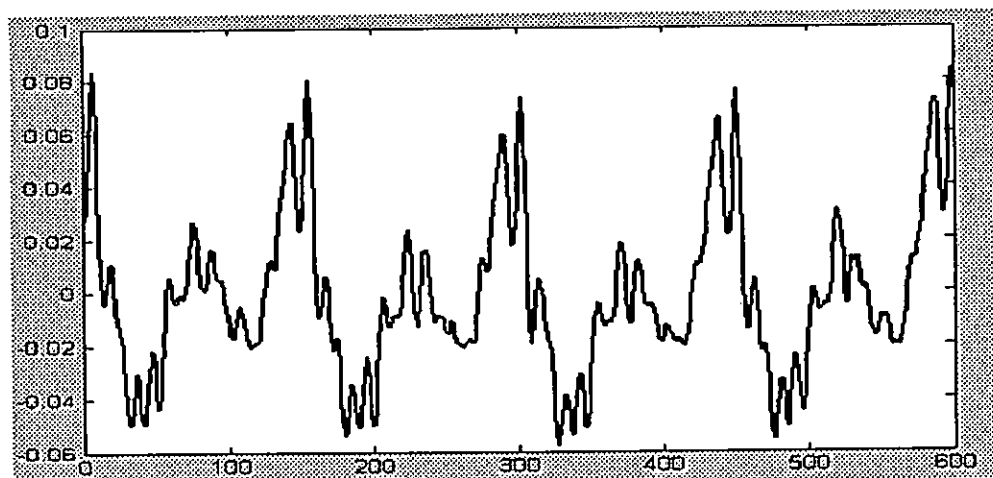


Figure43: voice : "ii" pitch : ga

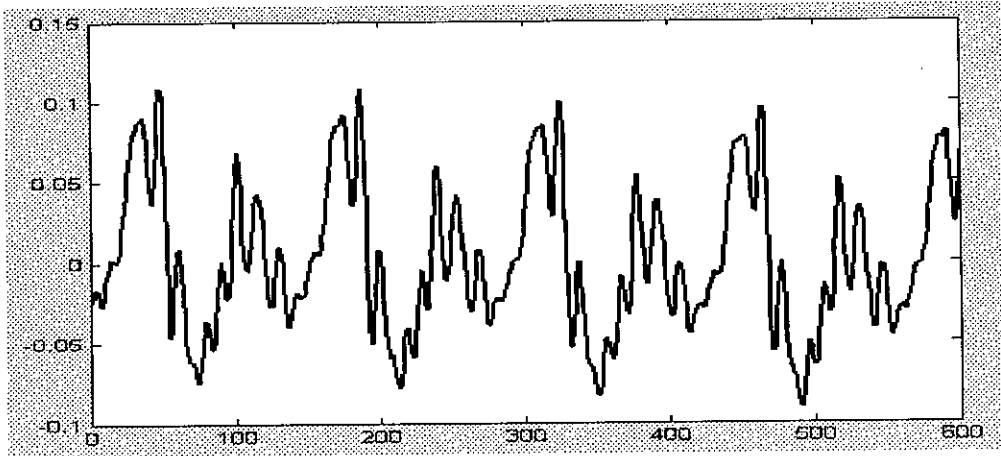


Figure44: voice : "ii" pitch : ma

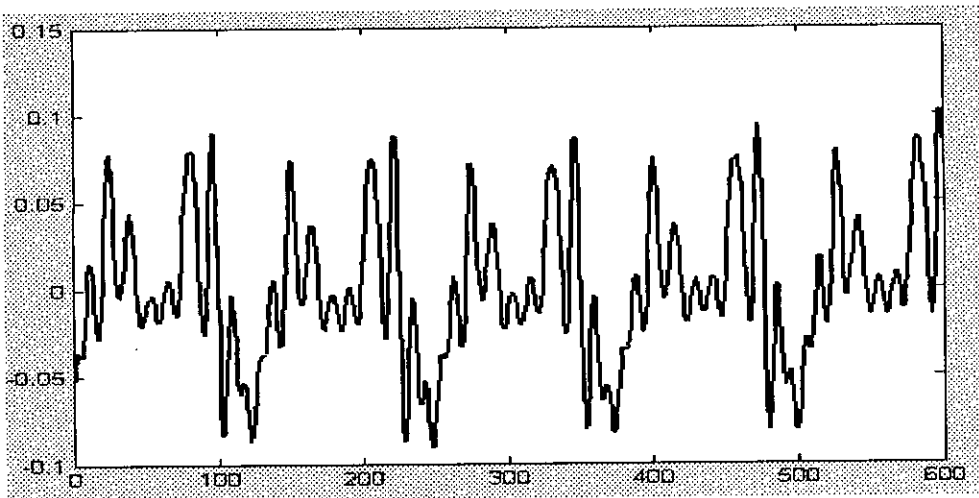


Figure45: voice : "ii" pitch : pa

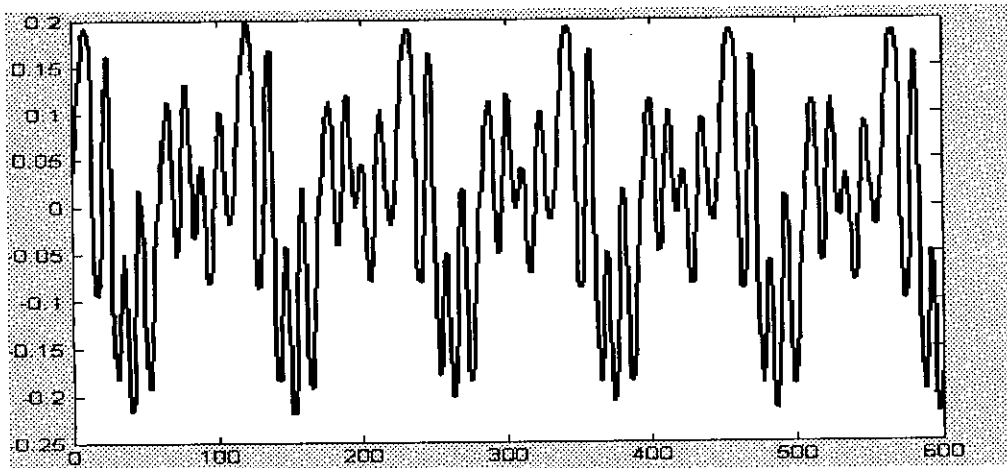


Figure46: voice : "ii" pitch : dha

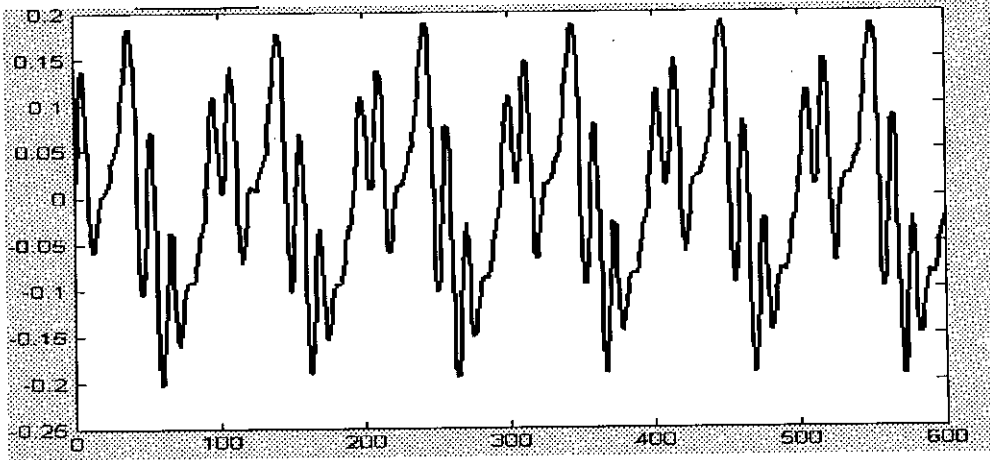


Figure47: voice : "ii" pitch : ni

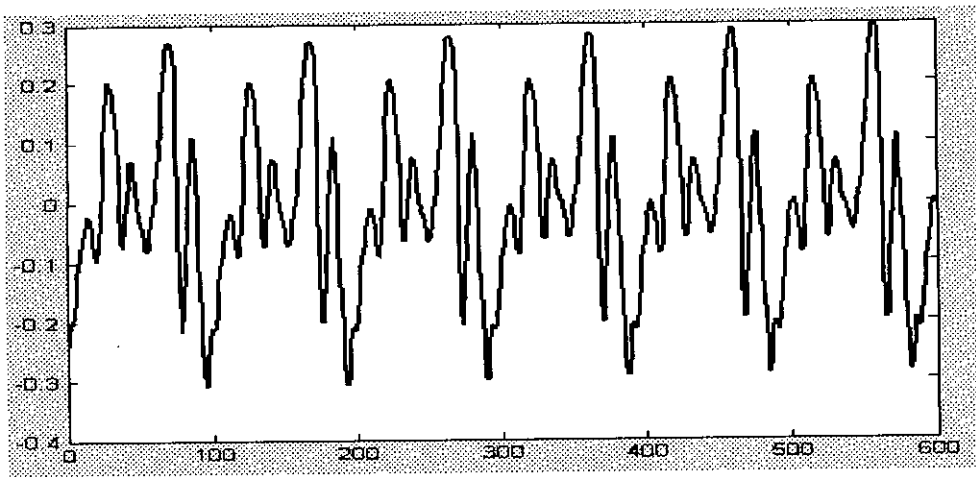


Figure48: voice : "ii" pitch : sa(hi)

