2009

Spectral analysis of pathological acoustic speech waveforms

Priyanka Medida
University of Nevada Las Vegas

SPECTRAL ANALYSIS OF PATHOLOGICAL ACOUSTIC SPEECH WAVEFORMS

by

Priyanka Medida

Bachelor of Science
Anna University
Chennai T. N., India
2006

A thesis submitted in partial fulfillment
of the requirements for the

Master of Science Degree in Electrical Engineering
Department of Electrical and Computer Engineering
Howard R. Hughes College of Engineering

Graduate College
University of Nevada, Las Vegas
December 2009

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

THE GRADUATE COLLEGE

We recommend that the thesis prepared under our supervision by

Priyanka Sabarimala Medida

entitled

SPECTRAL ANALYSIS OF PATHOLOGICAL ACOUSTIC SPEECH WAVEFORMS

be accepted in partial fulfillment of the requirements for the degree of

Master of Science
Electrical Engineering

Eugene McGaugh, Committee Chair

Venkatesan Muthukumar, Committee Member

Emma Regentova, Committee Member

Satish Bathnagar, Graduate Faculty Representative

Ronald Smith, Ph. D., Vice President for Research and Graduate Studies
and Dean of the Graduate College

December 2009

ABSTRACT

Spectral Analysis of Pathological Acoustic Speech Waveforms

by

Priyanka Medida

Dr. Eugene McGaugh, Examination Committee Chair
Associate Professor of Electrical Engineering
University of Nevada, Las Vegas

Biomedical engineering is the application of engineering principles and techniques to the medical field. The design and problem solving skills of engineering are combined with medical and biological science, which improves medical disorder diagnosis and treatment. The purpose of this study is to develop an automated procedure for detecting excessive jitter in speech signals, which is useful for differentiating normal from pathologic speech. The fundamental motivation for this research is that tools are needed by speech pathologists and laryngologists for use in the early detection and treatment of laryngeal disorders. Acoustical analysis of speech was performed to analyze various features of a speech signal. Earlier research established a relation between pitch period jitter and harmonic bandwidth. This concept was used for detecting laryngeal disorders in speech since pathologic speech has been found to have larger amounts of jitter than normal speech.

Our study was performed using vowel samples from the voice disorder database recorded at the Massachusetts Eye and Ear Infirmary (MEEI) in1994. The KAYPENTAX company markets this database. Software development was conducted using MATLAB, a user-friendly programming language which has been applied widely for signal processing. An algorithm was developed to compute harmonic bandwidths for various speech samples of sustained vowel sounds. Open and closed tests were conducted on 23 samples of pathologic and normal speech samples each.

Classification results showed 69.56% probability of correct detection of pathologic speech samples during an open test.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to Dr. Eugene McGaugh who served as my advisor, guided me through the entire course of study. This thesis would have been beyond my capability without his help and cooperation. It has been a great experience and fun working with him in the research.

I am also grateful to Dr.Venkatesan Muthukumar, Dr. Emma Regentova and Dr. Satish Bathnagar for kindly accepting to serve on my graduate committee.

I also take this opportunity to thank everyone who directly and indirectly helped me complete this thesis in time. Finally, I would like to give special thanks to my family and friends for their continuing love and affection throughout my life and career.

CHAPTER 1

INTRODUCTION

1.1Motivation for Study

Biomedical engineering is the application of engineering principles and techniques to the medical field. The design and problem solving skills of engineering are combined with medical and biological science which improves medical disorder diagnosis and treatment. The fundamental motivation for this research is that tools are needed by speech pathologists and laryngologists for use in the early detection and treatment of laryngeal disorders. Physicians usually detect laryngeal pathologies by means of a laryngoscope or endoscope which involves inserting a device down the throat of a patient. These procedures represent effective yet intrusive methods of detecting laryngeal disorders. In the past, researchers have been able to distinguish people who have some vocal fold problems from those who do not by analyzing their acoustic speech waveforms. It is intended that the research described in this paper will be of value to the medical industry for the detection of laryngeal pathologies through the use of a non intrusive method.

It has been determined that a key factor in the speech of many patients with laryngeal pathologies is excessive amounts of pitch period jitter [8]. Jitter is the time variation between pulses in a periodic signal. Many studies have shown that the voiced speech of patients with laryngeal disorders was found to have more jitter when compared to people without laryngeal disorders. Therefore, jitter can by used as a factor for detecting abnormal speech.

1.2 Goals of Study

The main goal of this study is to develop an automated procedure for detecting excessive jitter in speech signals which is useful for differentiating normal from pathologic speech. This procedure could be used in the early detection of laryngeal pathologies and in monitoring their treatment. Software development will be conducted using MATLAB, a user-friendly programming language which has been applied widely for signal processing.

1.3 Literature Survey

Much research has been done in the past to detect laryngeal disorders by analyzing acoustic speech waveforms, which is the visual representations of speech vibrations. Philip Lieberman conducted early research in 1963 to measure jitter in continuous speech [8]. By measuring the differences between the durations of adjacent fundamental periods, pitch perturbations were computed from recorded acoustic waveforms. Laryngeal mirror was used to take high speed motion pictures of the vocal cords. It was observed that pitch perturbations reflect variations in the shape of the glottal area wave, and also variations in glottal periodicity. The pitch perturbations of 23 speakers with pathological larynges were measured. It was found that the speakers who had pathologic growths on their vocal cords had larger pitch perturbations than did normal speakers with the same median fundamental periods. It was concluded that certain types of laryngeal conditions could be detected by measuring the perturbation factor [1].

The variations in pitch period length in the human voice, has attracted most of the researchers attention. Koike's [2] research's main purpose was to improve the procedure developed by Lieberman in measuring pitch perturbations, which would help in evaluating laryngeal dysfunction. Sound was extracted through the skin and

tissues by using a contact microphone placed on the throat. Relative average perturbations (RAP) were determined from the distance of pitch periods from a smoothed trend line of fundamental pitch periods. RAP is given by :

$$RAP = \frac{\frac{1}{N-2}\sum_{i=2}^{N-1}\left|\frac{To^{(i-1)} + To^{(i)} + To^{(i+1)}}{3} - To^{(i)}\right|}{\frac{1}{N}\sum_{i=1}^{N} To^{(i)}}$$

….....1.1

where N is the number of pitch periods To [2].

It was observed that pathological voices showed significantly higher values of RAP, which also depended upon the nature and degree of the disorder.

Childers and Bae [3], discuss two procedures for the detection of laryngeal pathology: 1) a spectral distortion measure using pitch synchronous and asynchronous methods with linear predictive coding (LPC) vectors and vector quantization (VQ) and 2) analysis of the electroglottographic signal using time interval and amplitude difference measures. The procedures were conducted on 29 pathological and 52 normal voices. These procedures yielded 75.9% and 69.0% accuracies respectively with a 9.6% false alarm probability for normal subjects.

Cesar and Hugo [4] address issues like the clinical procedures for laryngeal examination being invasive in nature. They also emphasize the increased interest in the acoustic analysis of normal and pathological voices for research because it is nonintrusive in nature and provides quantitative data within a reasonable analysis time. In the same article they have described the implementation of a system for automatic detection of laryngeal pathologies using acoustic analysis of speech in the frequency domain by using different techniques like cepstrum, mel cepstrum, delta cepstrum and FFT. Using neural networks they could distinguish between normal and

pathological voices. Their research indicated that this kind of analysis provides a noninvasive way of characterizing pathological condition and the results provide an alternative support tool for the diagnosis of pathologies. A 93.5% of accuracy was obtained using their method.

Mitev and Hadjitodorov's [5] research is aimed at the development of new methods of fundamental frequency determination of voiced signal uttered by patients with severe laryngeal disorders. They mention the unsatisfactory results in cases of severely distorted periodicity of the signal in the acoustic voice analysis by classic methods. Autocorrelation and cepstral methods are proposed in this paper. Since these methods gave higher accuracy of fundamental frequency determination compared to the most commonly used methods, they were combined in a system for acoustic analysis and screening of pathological voices and thus this system is used in the everyday clinical practice.

Stefen, Boyan and Bernard [6] address issues such as the classification of normal and pathological patients. An approach based on modeling of the probability density functions of the input vectors of the normal and pathological speakers by means of two prototype distribution maps (PDM), respectively, is proposed and later applied in a consulting system for laryngeal pathology detection. The database consisted of 100 normal voices and 300 pathological voices recorded in the Phoniatric Department of the University Hospital in Sofia, Bulgaria. 95.1% of classification accuracy was achieved.

Campbell and Reynolds [7] address to the issue of using a standard speech corpora for the development and evaluation in automatic speech processing research. It allows researchers to compare performance of different techniques on common data. Speech data produced at Massachusetts Eye and Ear Infirmary is the only commercially

available database and is distributed by the KayPENTAX company. But even though this database is used, there may be differences in the way its files are chosen. To get a better comparison of two methods, one must use the same data that others have used.

Alireza, Shikanth and Narayanana [7] focused on a robust, rapid and accurate system for automatic detection of normal voice and speech pathologies. Mel-frequency filter bank cepstral coefficients and measures of pitch dynamics were modeled by Gaussian mixtures in a Hidden Markov Model classifier. A total of 700 subjects of normal and pathological voices were used to evaluate this method. The Massachusetts Eye and Ear Infirmary (MEEI) database was used for the research. The authors claimed a method 99.44% correct classification rate.

CHAPTER 2

SPEECH PRODUCTION MODEL FOR SUSTAINED VOWELS

2.1 Introduction

We speak everyday without concentrating on the process of speech production. The movement of the lips, tongue, and other organs is among the subtlest and most adept of any actions performed by human beings. Here, I discuss the mechanism of speech production which includes the human vocal organs and the discrete-time speech production model.

2.2 The Physiology of Speech Production

Figure 2.1 is a diagram of the human vocal organs.

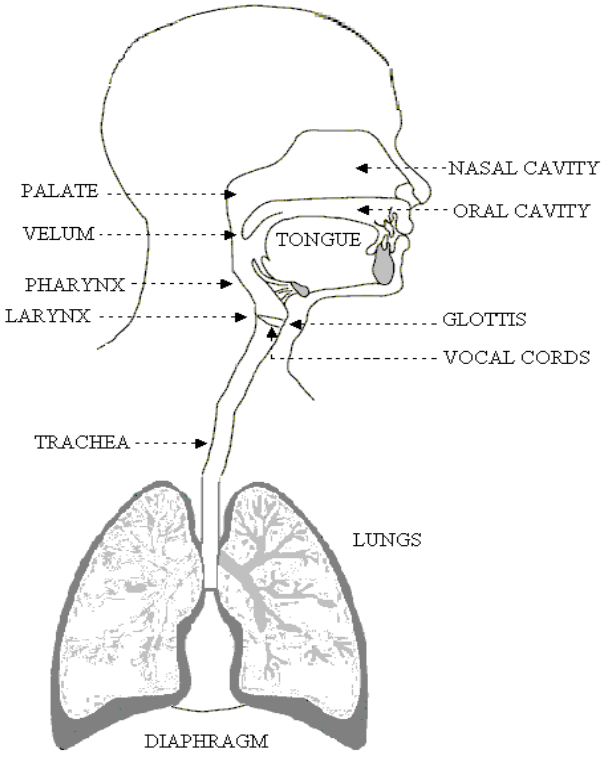

Figure 2.1 Human Vocal Organs

2.2.1 Lungs

As shown in the Figure 2.1 human speech is produced by vocal organs. The lungs and diaphragm are the main source of energy. Air enters the trachea from the lungs. Air flow is forced through the glottis between vocal cords in the larynx to the pharynx and oral and nasal cavities which are three main cavities of the vocal tract. From the oral and nasal cavities the airflow exits through the nose and mouth, respectively.

2.2.2 Larynx

The larynx is the most important organ for generating speech. Pitch and volume are manipulated there. The glottis which is a V-shaped opening between the vocal cords is the most important sound source in the vocal system. Vocal cords modulate air flow by rapid opening and closing which causes a buzzing sound. From this buzzing sound vowels and voiced consonants are produced. The fundamental frequency of vibration depends on the mass and tension and is about 110 Hz, 200 Hz, and 300 Hz with men, women, and children, respectively [10]. Consider the case for stop consonants: the vocal cords act suddenly from a completely closed position, in which they cut the air flow completely, to totally open position producing a light cough or a glottal stop. For unvoiced consonants like /s/ or /f/, they may be completely open. An intermediate position may also occur with for example phonemes like /h/ [9].

2.2.3 Vocal Tract-Pharynx, Nose, Mouth

From Figure 2.1, it is seen that the pharynx connects the larynx to the oral cavity. The pharynx has nearly fixed dimensions, but its length may be changed slightly by raising or lowering the larynx at one end and the soft palate at the other end. The route from the nasal cavity to the pharynx is either isolated or connected by the soft palate.

The epiglottis at the bottom of pharynx prevents food from reaching the larynx and isolates the esophagus acoustically from the vocal tract. The epiglottis, the false vocal cords and the vocal cords are closed during swallowing and open during normal breathing [9].

Now, let us consider the oral cavity which consists of the lips, velum, palate, tongue and teeth. Its size, shape and acoustics can be varied by its component parts. Especially the tongue is very flexible, the tip and the edges can be moved independently and the entire tongue can move forward, backward, up and down. The lips control the size and shape of the mouth opening through which speech sounds are radiated. Unlike the oral cavity, the nasal cavity has fixed dimensions and shape. Its length is about 12 cm and a volume of 60 cm^3. The soft palate controls the air stream to the nasal cavity. The pharynx and oral cavity are referred to as the vocal tract.

Figure 2.2 shows the external structure of the larynx and Figure 2.3 shows the internal structure of the larynx.

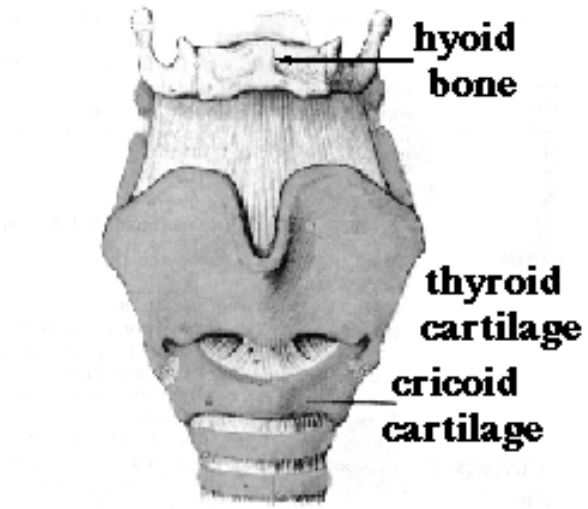

Figure 2.2 Front View of the Larynx

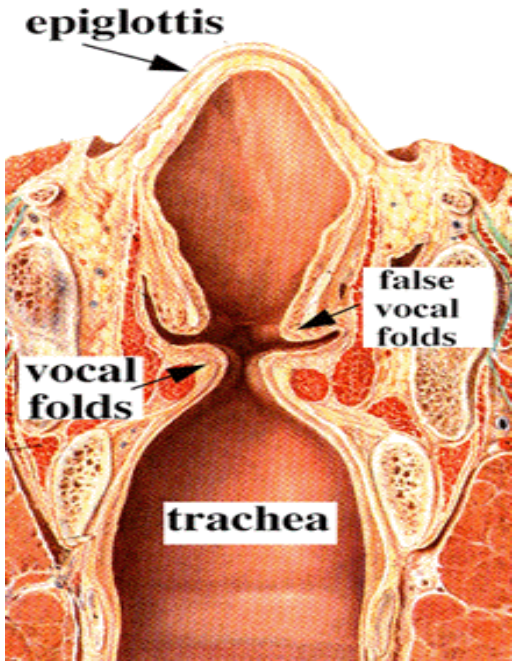

Figure 2.3 Internal Structure of the Larynx

As seen from the Figure 2.3, the space between the vocal cords is called the glottis. The vocal cords are wide open during quiet respiration preceding speech.

2.3 Continuous- Time Speech Production Model for Vowel Sounds

2.3.1 Introduction

A general continuous-time speech production model for voiced and unvoiced speech is shown in Figure 2.4. For most of the speech sounds, we can assume that the general properties of excitation and vocal tract are fixed over a period of 10 – 20 msec. Vowel sounds are usually used for laryngeal function assessment because the vocal folds are vibrating at a sustained frequency.

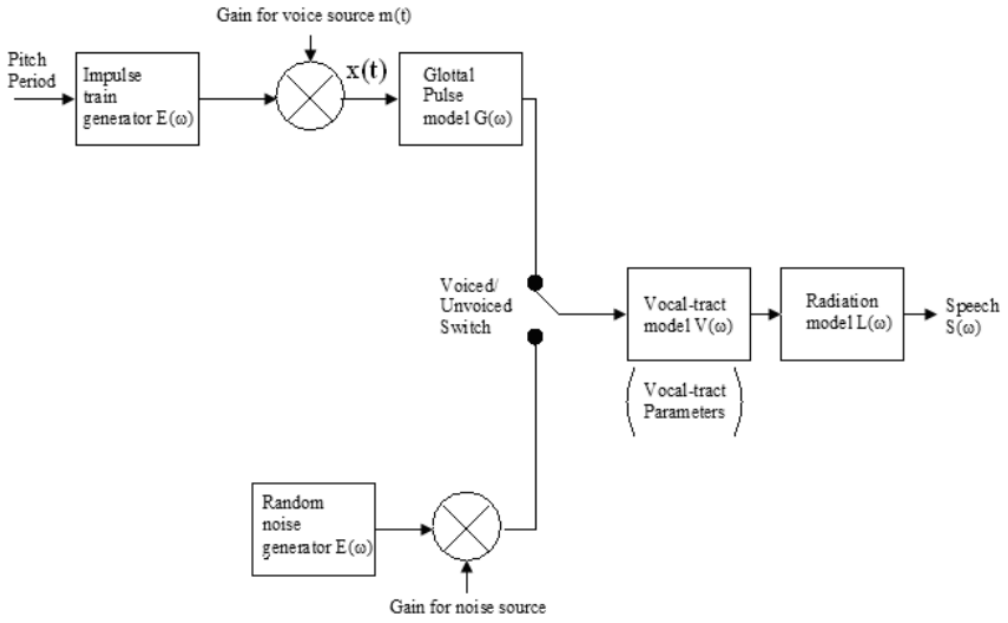

Figure 2.4 General Continuous Time Speech Production Model

Two reasons for using sustained vowels are:

1. They reflect the physical condition of vocal cords.

2. They can generally be treated as realizations from almost stationary stochastic processes.

The speech production model consists of the excitation function which is represented by a periodic impulse train E(ω). The glottal pulse, vocal tract and lip radiation are represented by G(ω),V(ω) and R(ω) respectively. The glottal excitation x(t), which is the input to the glottal pulse model, is produced from a finite sequence of impulses, e(t), having unit strength, which is modulated by function m(t) representing the strength of each pulse, as shown in Figure 2.5.

The expression for x(t) is given by

$$x(t) = \sum_{n=0}^{N_0-1} m_n \delta(t - t_n) \qquad (2.1)$$

where n = 0,1,2,3……..N_0-1 are the times at which impulses occur and m_n represents impulse strength [29].

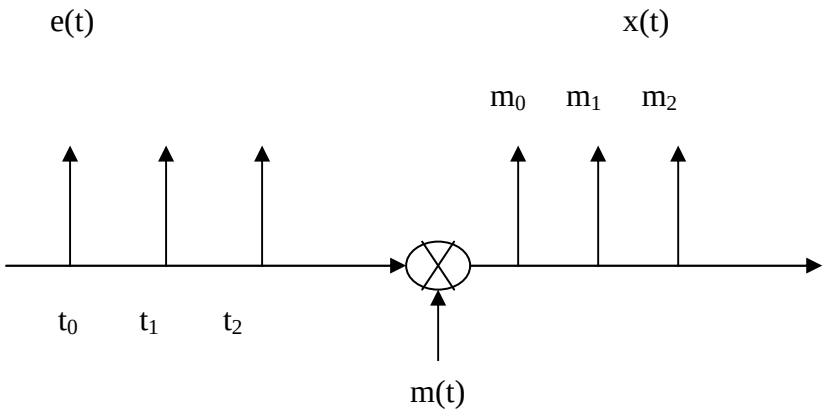

Figure 2.5 Glottal Excitation Modulation

2.3.2 Glottal Pulse Model

Glottal pulses have a short time duration with very short rise and fall times. A simple model of the glottal pulse shape filter impulse response, is given by

$$g(t) = G_0(t+1)e^{-ct}, \qquad 0 \le t < \infty \qquad (2.2)$$

where G_0 and c are constants. This model was derived by J.L.Flanagan [17]. The Fourier transform of a sampled band limited representation of g(t) is given by

$$G(\omega) = \frac{G_0}{(1-e^{-cT}e^{-j\omega})^2} \quad (2.3)$$

where T is the sampling period in seconds.

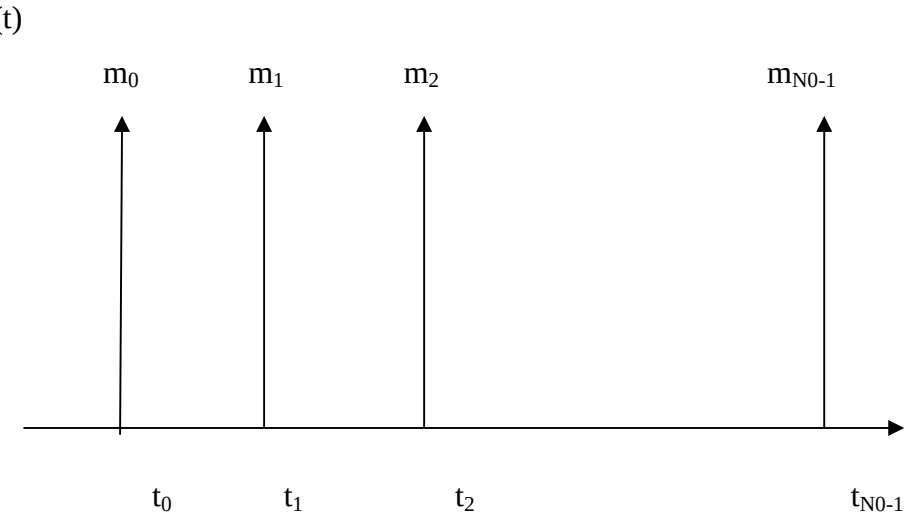

Figure 2.6 Glottal Excitation Pulse Train

2.3.3 Vocal Tract Modeling

The vocal tract impulse response v(t) is a function of the actual shape of the vocal tract and can be considered to remain stationary over 10 millisecond intervals during utterances of sentences [17]. However, for the case of sustained vowels v(t) can be assumed to remain stationary for the total time duration of the vowel. Also, for sustained vowel sounds the vocal tract is modeled as a linear time-invariant system with resonant frequencies called formants. The frequency location of formants is determined by the shape or configuration of the vocal tract and consistently occur within certain ranges with respect to specific vowels. Generally, the first three or four formants are sufficient for speech recognition. The vocal tract is relatively independent of other speech production components (i.e, glottal pulse excitation and lip radiation). The Fourier transform of a band limited sampled version of v(t) is given by:

$$V(\omega) = \prod_{i=1}^{K} \frac{1}{(1 - e^{j(\omega-\omega_i)})(1 - e^{j(\omega+\omega_i)})} \qquad (2.4)$$

where K corresponds to the number of formant frequenices ω_i [23].

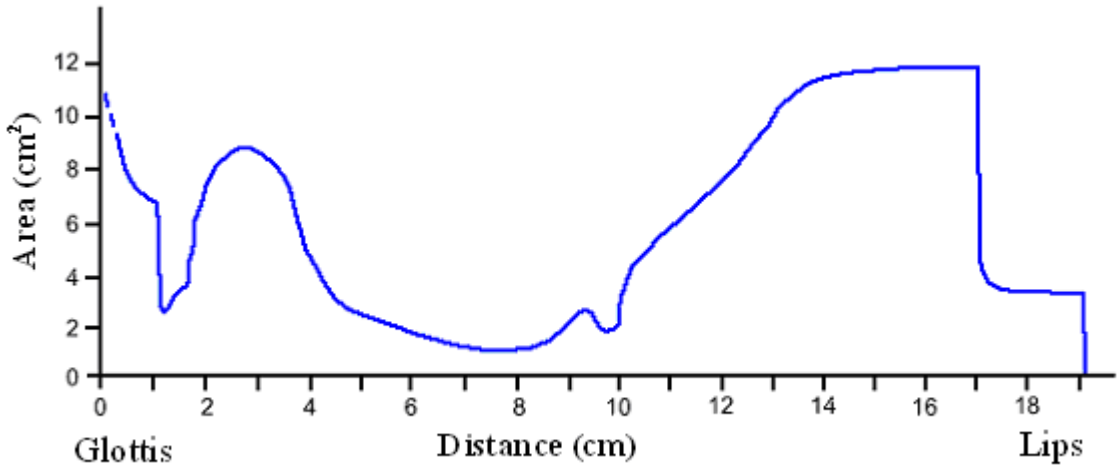

Figure 2.7 Cross-Section Area vs Vocal Tract from Glottis to Lips

The vocal tract transfer function shows resonance patterns across the spectrum for a particular articulation. A typical vocal tract spectrum is shown in the Figure 2.8

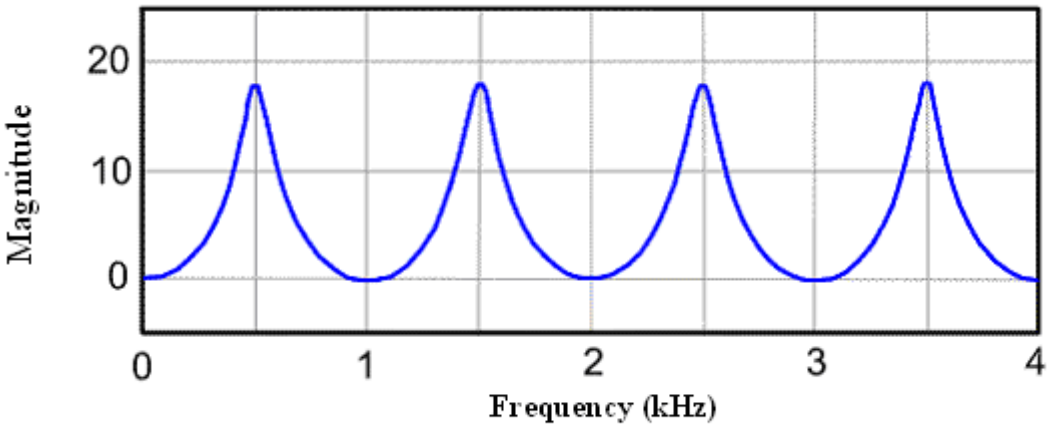

Figure 2.8 Vocal Tract Resonance Pattern

2.3.4 Glottal Excitation Modeling

Figure 2.9 represents a typical glottal excitation sequence associated with a sustained vowel where N_0 is the number of pulses,t_1 through t_{No} represents the times at which the pulses occur. Glottal pulses occur when vocal cords quasiperiodically

open to release short puffs of air which cause the vocal tract to resonate. The duration of each cycle in the speech waveform is called the glottal pulse or pitch period length. We represent the length in time of the glottal pulse or pitch period as shown in Figure 2.9

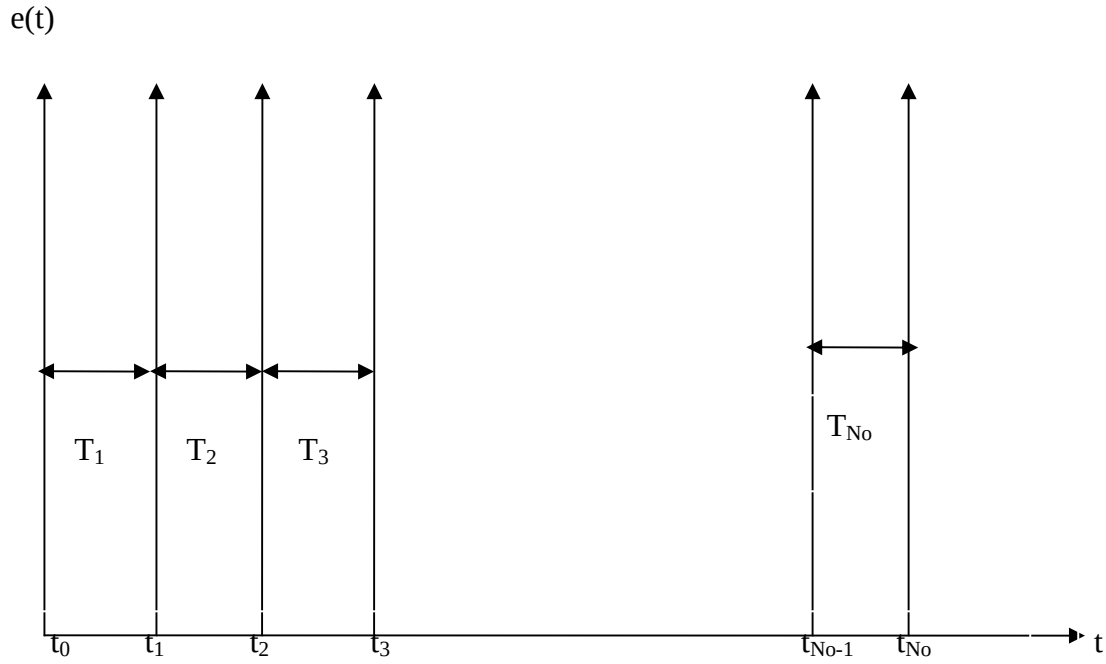

Figure 2.9 Glottal Excitation Timing

Since the vocal cord openings are independent we assume that the time period between the glottal pulses T_i to be independent. If we assume these time periods to be random variables from the same probability distribution with a mean T_o and variance σ^2_T, these periods can be expressed as [22]

$$T_i = T_o + \xi_i \qquad (2.5)$$

where T_o is constant and ξ_i is a normal random variable with mean zero and variance σ^2_T.

If $t_o = 0$, then,

$$t_i = \sum_{n=1}^{i} T_n \qquad i = 1 \ldots\ldots\ldots N_o \tag{2.6}$$

$$= \sum_{n=1}^{i} (T_o + \xi_n) \tag{2.7}$$

$$= T_o + \sum_{n=1}^{i} \xi_n \tag{2.8}$$

$$= (i + \sum_{n=1}^{i} \overline{\xi_n}) T_o \tag{2.9}$$

where $\overline{\xi_n} = \xi_n / T_o$

Note that while $\overline{\xi_n}$ are independent, t_i's are dependent.

Let the estimated mean $\hat{T}$ and variance ξ^2_T be computed in the following manner.

$$\hat{T} = 1/N_o \sum_{i=1}^{N_0} T_i \tag{2.10}$$

$$S^2_T = 1/(N_o - 1) \sum_{i=1}^{N_0} (T_i - \hat{T})^2 \tag{2.11}$$

$$S_T = \sqrt{S^2_T} \tag{2.12}$$

where S_T is the sample standard deviation.

The estimate of jitter $\hat{J}$ for the speech signal is the ratio is the of sample standard deviation to the sample mean.

$$\hat{J} = S_T / \hat{T} \tag{2.13}$$

$\hat{J}$ is the consistent estimate for the actual jitter J so that,

$$J = \lim_{N_o \to \infty} S_T / \hat{T}$$

$$= \sigma_T / T_o \tag{2.14}$$

Jitter amounts in sustained vowels produced by people with no laryngeal disorders have been found to be less than 0.01 while amounts greater than 0.02 have been measured in vowel sounds produced by people who have abnormal growths or

masses on their vocal cords [18]. So the technique for jitter detection must allow the user to consistently discriminate between vowel signals having these quantities of jitter present.

Assumptions about ξ_i and t_i have been made for modeling speech signal production [22]. These assumptions are:

(1)Glottal pulse periods can be treated as statistically independent random samples.

(2)Glottal pulse period samples can be treated as having a normal distribution.

These assumptions were validated by comparing the power spectrums of synthesized speech, having the above properties with the power spectrums of real speech and finding that their characteristics are consistent with the real speech signal. The results of these experiments will be discussed later.

2.3.5 Lip Radiation Modeling

The lip radiation filter represents the conversion of the volume velocity waveform at the lips to the sound waveform s(t). Davis [21] derived a simplified frequency response of this function given by

$$L(\omega) = L_0(1 - e^{-j\omega}) \qquad (2.15)$$

where L_0 is a gain factor. The continuous time interpretation of (2.15) is that the sound waveform is a scaled derivative of the volume velocity waveform at the lips with respect to time.

2.3.6 Combined Filter Response

Graphic representations of the individual filter responses are shown in Figures 2.11 (a),(b) and (c) and and the combined filter response, H(ω), is shown in Figure 2-11(d).

H(ω), is given by

$$H(\omega) = G(\omega) \cdot V(\omega) \cdot L(\omega) \qquad (2.16)$$

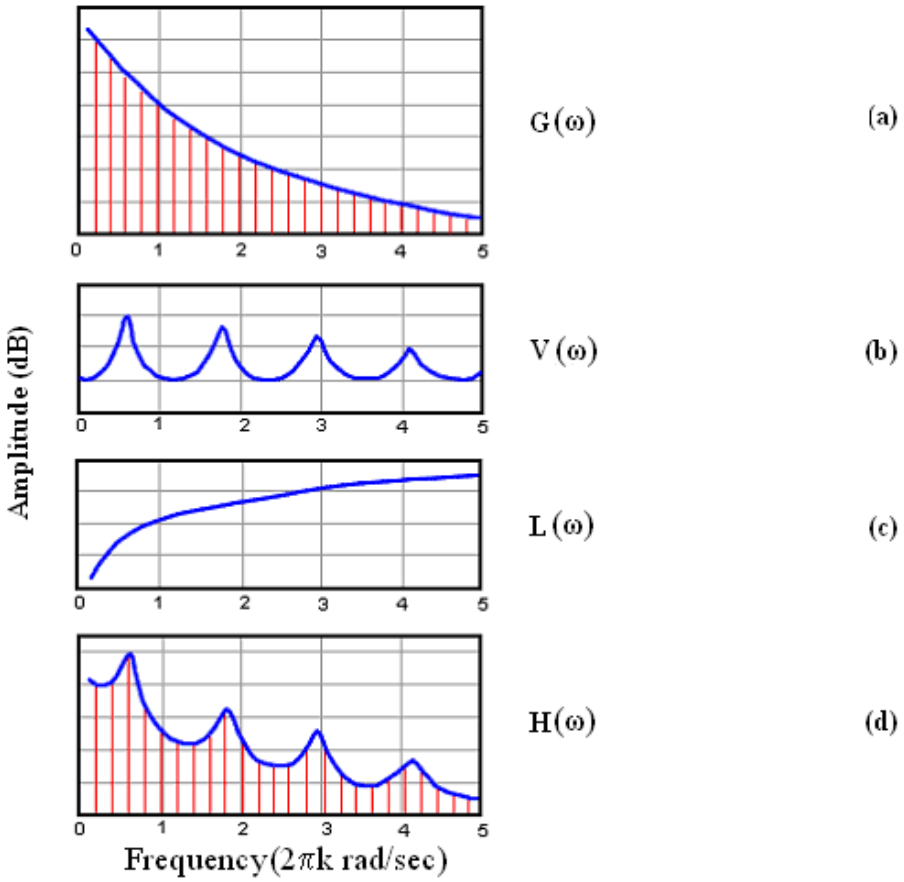

Figure 2.10 Vocal System Filter Frequency Responses

where G (ω) is the glottal pulse response, V (ω) is the vocal tract response, L (ω) is the lip radiation response and H (ω) is the combined response.

$$H(\omega) = \frac{G_O L_O \left(1 - e^{-(cT + j\omega))}\right)^{-2} \left(1 - e^{-j\omega}\right)}{\prod_{i=1}^{K} \left(1 - e^{j(\omega - \omega_i)}\right)\left(1 - e^{j(\omega + \omega_i)}\right)} \qquad (2.17)$$

Since cT is much less than unity (c = 200π/sec), two of the numerator terms cancel

allowing H(ω) to be expressed as an all-pole filter :

$$H(\omega) = \frac{G_0 L_0}{\left(1 - e^{-(cT + j\omega)}\right)^{-2} \prod_{i=1}^{K} \left(1 - e^{j(\omega - \omega_i)}\right)\left(1 - e^{j(\omega + \omega_i)}\right)} \qquad (2.18)$$

An alternative version of the speech production model is shown in Figure 2.11. The speech signal output response, S (ω), is related to the other model components through the expression

$$S(\omega) = X(\omega) \cdot G(\omega) \cdot V(\omega) \cdot L(\omega) \quad (2.19)$$

$$= X(\omega) \cdot H(\omega) \quad (2.20)$$

where $X(\omega) = E(\omega) * M(\omega)$.

The purpose of this speech model is to facilitate an understanding of the speech power spectrum so that it may be used for pathological assessment. A power spectrum expression of the glottal excitation function will be derived in the next chapter based on its assumed mathematical affect on glottal excitation spectrum analysis model representation. H (ω) will be considered for its affect on glottal excitation spectrum analysis.

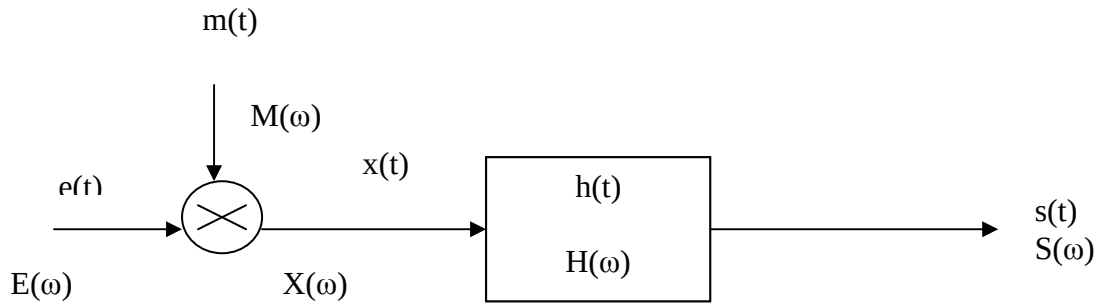

Figure 2.11 Simplified Speech Production Model

CHAPTER 3

SPECTRAL ANALYSIS OF SPEECH SIGNALS

3.0 Introduction

Techniques for spectrum estimation can generally be divided into parametric and non-parametric methods. The parametric approaches assume that the underlying stationary stochastic process has a certain structure which can be described using a small number of parameters (for example, using an auto-regressive or moving average model). In these approaches, the task is to estimate the parameters of the model that describes the stochastic process. By contrast, non-parametric approaches explicitly estimate the covariance or the spectrum of the process without assuming that the process has any particular structure. The periodogram is a classic non-parametric technique

3.1 Fourier Analysis

The periodogram is an estimate of the power spectral density (PSD) of a signal. Usually, the periodogram is computed from a finite-length digital sequence using the Fast Fourier transform (FFT). The Fourier transform is used to transform a continuous time signal into the frequency domain. It provides the continuous spectrum of a time signal. Let x(t) , $0 \leq t \leq L$, be a finite-length continuous-time signal of length L in seconds. The continous-time Fourier transform of x(t) is given by:

$$X(\omega) = \int_0^L x(t)e^{-j\omega t}\,dt\,, \quad -\infty < \omega < \infty$$

where ω is the analog frequency in radians per second. The inverse Fourier transform of $X(\omega)$ is given by:

$$x(t) = 1/2\pi \int_{-\infty}^{\infty} X(\omega)e^{j\omega t}\, d\omega$$

The Discrete Fourier Transform

A time sampled version of x(t) is given by $x[nT_s]$ where T_s is the sampling period and $0 \leq nT_s \leq (N-1)T_s$. $L = NT_s$ where N is the total number of time samples of $x[nT_s]$. The discrete-time version of $X(\omega)$ is given by:

$$X(k2\pi/NT_s) = \sum_{n=0}^{N-1} x[nT_s]e^{-jk(2\pi/NTs)nTs}$$

Using just the time and frequency indices alone, the discrete-time version of $X(k2\pi/NT_s)$ can be expressed as

$$X(k) = \sum_{n=0}^{N-1} x[n]e^{-j2\pi nk/N} \text{ where } 0 \leq k \leq N-1$$

This is the Discrete Fourier Transform (DFT) of x[n].

The inverse DFT is given by $x[n] = 1/N \sum_{n=0}^{N-1} X[k]e^{-j2\pi nk/N}$ where $0 \leq n \leq N-1$.

The Fast Fourier Transform

The *Fast Fourier transform* (FFT) is simply a class of special algorithms which implements the discrete Fourier transform with considerable savings in computational time. It must be pointed out that the FFT is not a different transform from the DFT, but rather just a means of computing the DFT with a considerable reduction in the number of calculations required.

3.1.1 Power Spectrum Estimate for Finite Length Signals

In the speech production, parameters like average pitch frequency and vocal tract shape vary with respect to time, because speech signals generally fall in to the category of non stationary random processes. In the case of long sustained vowel

sounds at a constant average pitch frequency and fixed amount of jitter, the signals can be treated as almost stationary and ergodic. This assumption allows one speech signal to be used for determining the amount of jitter which is always present in the speech produced by a particular person.

We shall now initiate the spectral analysis of sustained vowel speech signals by deriving power spectrum expressions for finite length signals. H(ω) represents the combined response of the speech production model filters of Figure2.11. The Fourier transform of a vowel of length L can be expressed as

$S_L(\omega) = X_L(\omega) * H_L(\omega)$ (3.1)

where $X_L(\omega)$ is the Fourier transform of the excitation function x(t).

It follows that the power spectrum estimate [21] for the finite length s(t) is

$$\hat{P}_S(\omega) = 1/L \; |S_L(\omega)|^2$$

$$= \frac{1}{L}\left|X_L(\omega).H_L(\omega)\right|^2$$

$$= \frac{1}{L}\left|X_L(\omega)\right|^2.\left|H_L(\omega)\right|^2$$

$$= \hat{P}_X(\omega) \cdot |H_L(\omega)|^2 \quad (3.2)$$

where $\hat{P}_X(\omega) = \frac{1}{L}\left|X_L(\omega)\right|^2$ is the power spectrum estimate for x(t).The symbol "^" is indicates that the power spectrum results are just estimates of the true power spectrum of infinite length versions of the signals. Equation (3.2) shows how the glottal excitation and the speech filter, H(ω), power spectrum combine to form the speech signal power spectrum.

The expected value of the speech signal power spectrum may be expressed as

$$E[\bar{P}_S(\omega)] = E[\hat{P}_X(\omega)] \cdot |H_L(\omega)|^2 \quad (3.3)$$

The expected value of $P_S(\omega)$ as expressed in Equation(3.3) is equivalent to computing the average over an infinite number of power spectrums of infinite length vowel sound signals with the same statistical parameters, i.e.,

$$E[P_S(\omega)] \triangleq \lim_{K\to\infty} \frac{1}{K} \sum_{i=1}^{K} \hat{P}_S^{(i)}(\omega) \quad (3.4)$$

where $\hat{P}_S^{(i)}(\omega)$ is the power spectrum estimate of the i^{th} record.

3.1.1.1 Derivation of Expected Power Spectrum for Glottal Excitation

To derive expression for $\hat{P}_X(\omega)$ and its expected value, let us assume that all glottal impluses have unit strength(no shimmer),then Equation (2.1) becomes

$$x(t) = \sum_{n=0}^{N_O-1} \delta(t - t_n) \quad (3.5)$$

and the Fourier transform of x(t) is

$$X(\omega) = \sum_{n=0}^{N-1} e^{-j\omega t_n} \quad (3.6)$$

It follows that

$$\hat{P}_X(\omega) = \frac{1}{L} |X_L(\omega)|^2 \quad (3.7)$$

$$= \frac{1}{L} \left(\sum_{n=0}^{N_o-1} e^{-j\omega t_n} \right) \left(\sum_{k=0}^{N_o-1} e^{j\omega t_k} \right) \quad (3.8)$$

where t_n is the same as t_i in Equation(2.6),i.e.,

$$t_n = (n + \sum_{i=1}^{n} \bar{\xi}_i)\, T_o \quad (3.9)$$

$$\hat{P}_X(\omega) = \left\{ N_0 + \sum \sum_{n\neq k}^{N_0-1} \cos[\omega(k - n + \sum_{i=1}^{k} \bar{\xi}_i - \sum_{i=1}^{n} \bar{\xi}_i) T_0] \right\} \quad (3.10)$$

If we assume that $\overline{\xi_i}$ has a normal distribution with zero mean and variance σ^2 then it can be shown that the expected value of $\hat{P}_X(\omega)$ may be expressed as [22]

$$E[\hat{P}_X(\omega)] = \frac{1}{L}\left\{ N_0 + \sum_{n\neq k}^{N_0-1}\sum \cos[\omega T_0(k-n)] e^{-\omega^2(\sigma T_0)^2|k-n|/2} \right\} \quad (3.11)$$

A discrete representation of (3.11) is obtained by sampling it at intervals of $\Delta\omega = \frac{2\pi}{L}$. This allows samples of $E[\hat{P}_X(\omega)]$ to be taken at $\omega = \omega_m = m\Delta\omega$.

m = 0,1,2,3,……, as follows.

$$E[\hat{P}_x(\omega=\omega_m)] = \frac{1}{L}\left\{ N_0 + \sum_{n\neq k}\sum \cos[m\Delta\omega T_0(k-n)] e^{-(m\Delta\omega)^2(\sigma T_0)^2|k-n|/2} \right\} \quad (3.12)$$

L is chosen such that $L = N_0T_0$ where T_0 is the mean pitch period length.

$$E[\hat{P}_x(m)] = \frac{1}{L}\left\{ N_0 + \sum_{n\neq k}\sum \cos[(\frac{2\pi\, m}{N_0T_0})T_0(k-n)] e^{-(\frac{2\pi m}{N_0T_0})^2(\sigma T_0)^2|k-n|/2} \right\} \quad (3.13)$$

$$E[\hat{P}_x(m)] = \frac{1}{L}\left\{ N_0 + \sum_{n\neq k}\sum \cos[(\frac{2\pi\, m}{N_0})(k-n)] e^{-(\frac{2\pi m}{N_0})^2(\sigma)^2|k-n|/2} \right\} \quad (3.14)$$

where m is a frequency indexing number which refers to the frequency $\frac{2\pi m}{L}$. Since (3.14) is an even function of k and n, it can be rewritten in the following form:

$$E[\hat{P}_x(m)] = \frac{1}{L}\left\{ N_0 + \sum_{l=-(N_0-1)}^{N_0-1} [N_0 - |l|] \cos(\frac{2\pi\, m}{N_0} l) e^{-(\frac{2\pi m}{N_0})^2(\sigma)^2|l|/2} \right\} \quad (3.15)$$

where l = k-n. Obviously, a considerable amount of computation time is saved by using (3.15) in place of (3.14)

3.1.2 How Jitter Affects the Power Spectrum of the Glottal Excitation

Figure 3-1 shows a sketch of Equation 3.15, which consists of a periodic sequence of "bell" shaped pulses(harmonics) centered at integer multiples of $\frac{2\pi}{T_0}$ (mean fundamental frequency). The bandwidth of these pulses is proportional to the variance of the random variable $\overline{\xi}_i$ and increases at higher frequencies. Also, the pulse amplitudes decay with frequency at a rate proportional to the variance of $\overline{\xi}_i$. To show that the width of the "bell" shaped pulses increases as the variance of $\overline{\xi}_i$ increases, consider what happens around the first harmonic of E[$\hat{P}_X(m)$] or E[$\hat{P}_X(m = N_0)$], as shown in Figure 3.2.

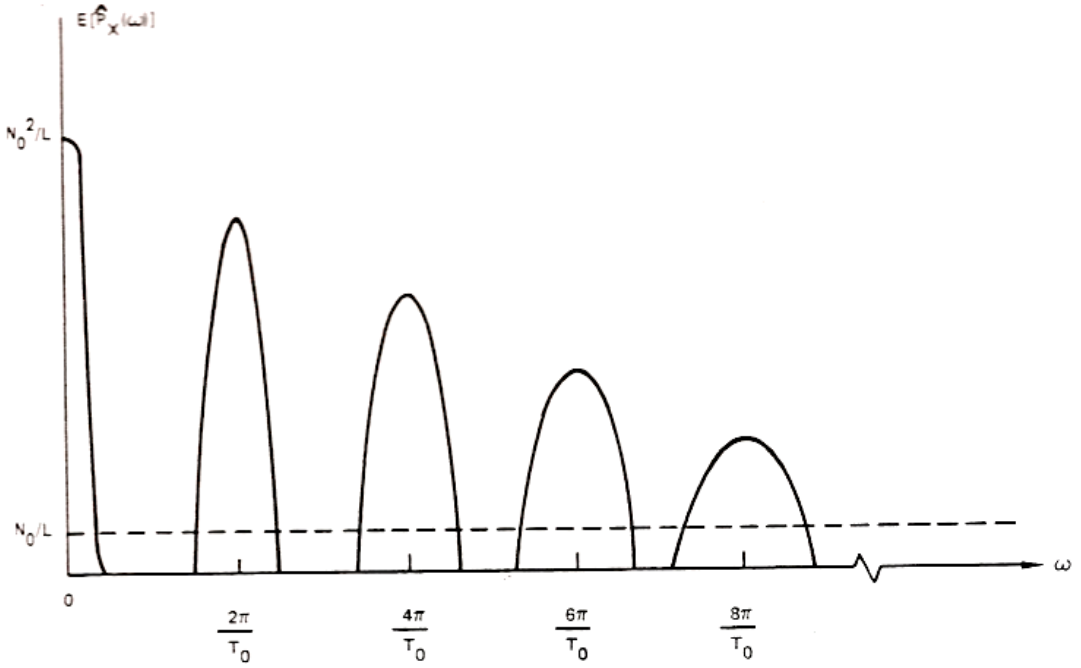

Figure 3.1 Expected Value of the Excitation Power Spectrum Estimate

$$E[\hat{P}_x(m = N_0)] = \frac{1}{L}\sum_{l}[N_0 - |\, l\,|]\cos[\,2\pi\, l\,]e^{-(2\sigma\pi)^2|l|/\,2}$$

(3.16)

At one frequency increment away from m= N_0 or m= $N_0 \pm 1$

$$E[\hat{P}_X(m = N_0 \pm 1)] = \frac{1}{L}\sum_{l}[N_0 - |\, l\,|\cos[\frac{2\pi\, l}{N_0}]e^{-(1\pm\frac{1}{N_0})^2(2\sigma\pi)^2|l|/\,2} \qquad (3.17)$$

It can be shown that as σ^2 increases, the difference between (3.16) and (3.17) diminishes which implies that the magnitude of the slope of the harmonic pulses diminishes as σ^2 increases and bandwidth of these pulses increases with increase in jitter which is represented in Figure 3.2

A MATLAB program was written to plot Equation 3.17 as a function of jitter. Three first harmonic pulses for different jitter values were generated. Figure 3.2 shows that as jitter increases, the harmonic bandwidths also increase.

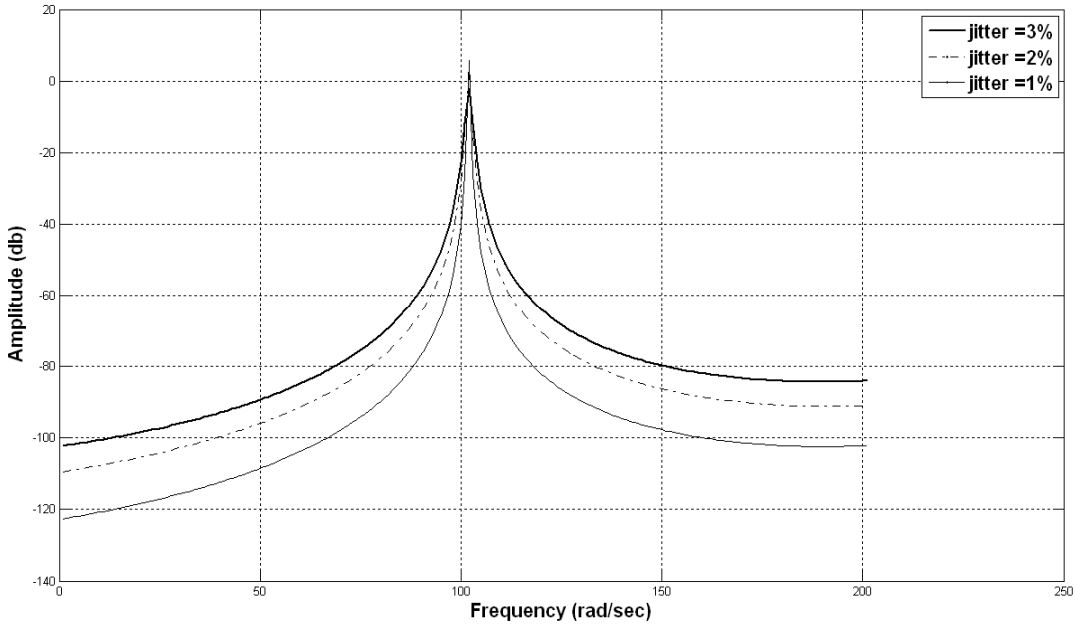

Figure 3.2 Relation Between Harmonic Bandwidth and Jitter

3.2 Maximum Entropy Spectral Analysis

3.2.1 The Concept

In 1967, Burg developed a nonlinear procedure for spectral estimation with increased resolution called the maximum entropy method (MEM). The major attraction of this procedure is that it provides high resolution spectral estimates for relatively short record lengths without using window functions. Methods used prior to this method, basically calculated the auto correlation estimate and then windowed the autocorrelation function estimate ,appended zeros, and performed the Fourier transform. The window is optimized to give as much resolution as possible with little leakage.

In MEM method suggests instead of appending zeros to increase the length of the estimated autocorrelation function, that the estimated autocorrelation function should be extrapolated beyond the data limited range. The principle used for extrapolation is that either the spectral estimated must be the most random or have the maximum entropy of any power spectrum which is consistent with the sample values of the calculated autocorrelation function [19].

The Burg algorithm is probably the most widely known AR spectral estimation procedure. Because of its derivation in the context of maximum entropy methods, the algorithm is sometimes designated "MEM". The procedure computes the AR coefficients directly from the data by estimating the reflection coefficients (partial autocorrelations) at successive orders. Since the computed coefficients are the harmonic mean between the forward and backward partial autocorrelation estimates, the Burg procedure is also known as the "Harmonic" algorithm.

3.2.2 Predictor Filter Coefficient Calculations

The linear prediction method predicts the n^{th} value of the sequence by

$$\hat{x} = \sum_{k=1}^{p} a_k x_{n-k}$$

where P represents the number of past samples in the data and they are presumed known.

Error between predicted value and true value is

$$\hat{u} = x_n - \hat{x}_n$$

or

$$x_n = \hat{x}_n + u_n$$

$$= \sum_{k=1}^{p} a_k x_{n-k} + u_n \qquad [19]$$

The predicted value is calculated by convolving the P "prediction filter" coefficients a_k with the past P values of the data x_{n-k}. This shows that the MEM spectrum is modeled as an all pole spectrum with P poles. Let X(z) be the z-transform of x_n and assume that u_n is unit white noise.

$$|X(z)|^2 = 1/|1 - \sum_{1}^{P} a_k z^{-k}|^2 \qquad [19]$$

The fundamental equation to be solved for the estimated $\hat{a}_k$ are

$$\sum_{k=1}^{p} \hat{a}_k R_{|i-k|} = R_i \qquad 1 \leq i \leq P$$

where R_i are autocorrelation coefficients estimated from the data record. These equations will be recognized as the discrete counter part of the Wiener prediction filter equation [19].

The MEM finally leads to the auto correlation prediction equations:

$$\hat{R}_1 = \sum_{k=1}^{p} a_k R_{|i-k|}, l \geq P+1$$

where the $\hat{R}_1$ (l ≥ P+1) are the predicted autocorrelation values.

CHAPTER 4

CLASSIFICATION

4.1 Introduction

Laryngeal pathology detection requires classification between normal and pathologic speech. Classification is based on the feature extracted from measurements of the data. It mainly depends on selecting a good feature that can significantly contribute to classification performance. Classifier selection is also important. As mentioned in Chapter 3, the speech spectral harmonic bandwidth (HBW) is our selected feature.

4.2 Discriminant Function

Discriminant function analysis is used to assign the feature measurements into categories. Only if the discriminant function analysis is effective for a set of data the classification estimates will yield a high percentage of correctness.

The main purpose of the discriminant analysis is

1. To classify samples into groups.
2. To test the classifier by observing whether samples are correctly assigned to groups.

The Discriminant score is the value resulting from applying a discriminant function formula to the data for a given case. The samples are classed based on the discriminant score.

4.3 Classifier Performance Evaluation.

Performance of a classifier is decided based on the amount of false alarms or the misclassification it is producing. After the classifier is designed for the samples selected, tests are performed on the classifier. Purpose of the test is to observe how correctly the classifier can distinguish between the two categories (classes). With the

prior knowledge of the class, a sample is chosen and passed through the classifier for identification.

For n samples from each class for test, if k samples are correctly classified then the percentage of correct detection of the classifier is given by (k/n)*100 %.

The classifier performance is evaluated based on the value of the above percentage. Higher percentage shows that the classifier is good.

4.4 Bayes Decision Criterion

4.4.1 Maximum Likelihood Classification

The Bayes decision theory has three distinct spaces.

1. Observation (measurements)
2. Parameters (unknown)
3. Decisions

The main criteria used for the selection is maximum likelihood criterion. Without prior information we use the maximum likelihood approach. It is a model that maximizes the probability of correct detection.

The likelihood function is calculated for the feature x extracted from a k dimensional class as

$g(x) = p(h_i \mid x)$

where $p(h_i \mid x)$ is the posterior conditional density of the class parameter vector h_i for class i given feature vector x.

This is calculated using Bayes rule

$$p(h_i \mid x) = \frac{p(h_i) * p(x \mid h_i)}{p(x)}$$

where $p(h_i)$ is the a priori density of class i and $p(x \mid h_i)$ is the a priori conditional density of x given the parameter vector h_i for class i and p(x) is the probability density of the features.

If the parameter vector h_i of the a priori conditional density is unknown, it is estimated from the feature vectors belonging to the class using the maximum likelihood technique. The estimate maximizes the conditional density $p(x \mid h_i)$. If the a priori conditional densities are assumed to have normal distributions, the likelihood function is expressed as,

$$g(x) = 1/\sqrt{2\pi\sigma^2}\exp\left[-1/2\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

where $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

4.4.1.1 Likelihood Ratio

In statistics likelihood ratio is the ratio of the maximum probability of a result under two different hypotheses. It is used for a statistical test to make a decision between two classes. For a two class problem the likelihood criterion is expressed as a likelihood ratio by,

$g(x) = g_1(x)/g_2(x)=[p_1(h_1 \mid x)]/[p_2(h_2 \mid x)]$

where $p_1=1-p_2$

4.4.1.2 Threshold

Class one is chosen if the ratio is greater than one. The decision rule can be alternately stated by as, choose class one if $[p_1(x \mid h_1)]/[p_2(x \mid h_2)] > T$

where T is a threshold value chosen to maximize the probability of correct detection. If a threshold value is varied over a range and results are tabulated, any false alarm probability may be realized.

4.4.1.3 Logarithm of Likelihood Ratios

If the logarithm of likelihood ratio is taken then,

$$\log \{[p_1(x \mid h_1)]/[p_2(x \mid h_2)]\} > \log_e y$$

$$d_1(x) - d_2(x) > \log_e y$$

where

$$d_1(x) = \{-\log(w_1) + ([x-m_1]^2/w_1)\}/2$$

$$d_2(x) = \{-\log(w_2) + ([x-m_2]^2/w_2)\}/2$$

When x is the feature from one class, w_1 and w_2 are variances of classes one and two class respectively. m_1 and m_2 are the mean values computed for the features from class one and two respectively.

When $d_1(x)$ and $d_2(x)$ are computed the decision can be made based on the threshold value chosen.

CHAPTER 5

PROCEDURE

5.1 Introduction

This chapter describes how optimum speech power spectrum estimates were produced and the classification of the spectral results. An algorithm was developed to compute harmonic bandwidth, which is the pre-selected feature. A relation between this feature and jitter was established in Section 3.1.2. The algorithm processes speech to compute the HBW which can differentiate normal speech from abnormal speech. The Maximum Entropy power spectrum requires optimization of the filter order and signal length parameters for the best spectral resolution. Optimum filter order and signal length were determined using synthesized speech. Classification of spectral measurements is shown in Chapter 6. All the experiments were initially conducted on a synthesized speech because speech parameters like fundamental frequency, signal length and jitter could be controlled for the signal. Harmonic bandwidth was computed for different amounts of jitter. Once optimum spectral parameters were determined, they were applied to real speech samples. Classification was performed as discussed in Chapter 4 and the results are shown in Chapter 6.

5.2 Data Description

5.2.1 Development and Use of the Exponential Pulse Sequence

As previously mentioned, synthesized speech was used for experimental purposes. To produce a synthesized speech signal, an exponential pulse train was developed using MATLAB. The i^{th} pulse of the exponential pulse train may be expressed as:

$$q(n) = A\, e^{-(n\Delta t - t_{i-1})/a}, \quad t_{i-1} < n\Delta t < t_i \qquad (5.1)$$

where A- amplitude

n- time index

Δt – sampling interval in time

a – time constant

$$t_i = \sum_{k=1}^{i} T_k \qquad (5.2)$$

where T_k is the time duration of the k^{th} pitch period.

A synthesized speech record can be generated for any number of pitch periods N_o. Mean pitch period length $T_o = 1/F_o$ where F_o is the fundamental frequency. The sampling frequency is given by $F_s = 1/\Delta t$. A Gaussian random number generator was used to add desired levels of jitter to the pitch periods T_k previously described by Equation 5.1. An exponential pulse train sample is shown in Figure 5.1. Once the exponential pulse train x(n) was generated, spectrum analysis was performed. The spectrum analysis results will be discussed in Chapter 6.

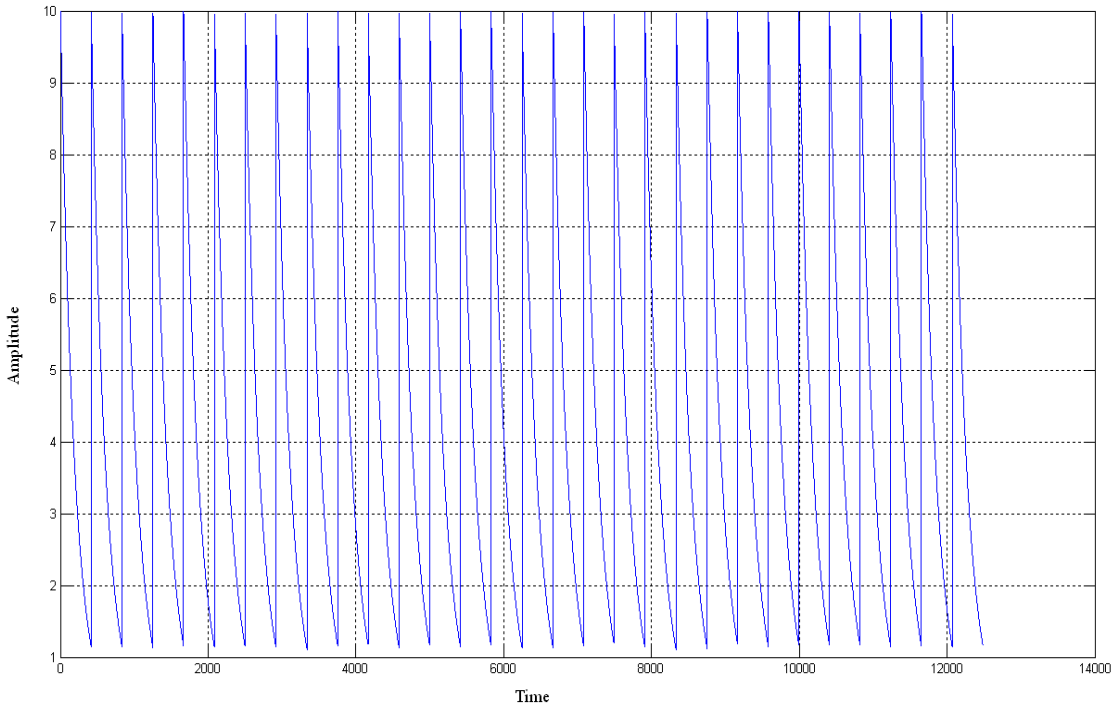

Figure 5.1 Exponential Pulse Train

5.3 Real Speech Data

5.3.1 Kaypentax Database Description

Our study was performed using vowel samples from a database consisting of real speech samples recorded at the Massachusetts Eye and Ear Infirmary (MEEI). Sustained samples of the vowel /a/ were recorded from both normal and pathological speakers who had a variety of pathologies including vocal nodules, paralysis etc. The database was created by Dr. Robert E. Hillman [20].

Normal speech samples were sampled at a rate of 50 kHz and abnormal speech samples were sampled at 25 kHz. The duration of these vowel samples was 3s for normal speakers and 1s for abnormal speakers. Vowel samples in the database appear to include only the stable part of the phonation.

The speech database was acquired from KayPENTAX company. The file format was .NSP, which is a Kay Elemetrics format. The database files had to be converted into a format compatible with MATLAB. Hence, the database files were converted to wave format. For our experimental purposes we needed two sets of data, one for the closed test and the other for open test. Information of these selected data groups is shown in Tables 5.1 and 5.2 where F_o is the fundamental frequency and RAP is the relative average perturbation. RAP is a measurement of pitch period jitter [2].

Identification	Fo(Hertz)	RAP
BJV1NAL	247.134	0.098
CAD1NAL	302.78	0.156
DAJ1NAL	210.022	0.285
DFP1NAL	216.849	0.4888
DMA1NAL	239.3	0.238
DWS1NAL	184.855	0.266
EDC1NAL	217.661	0.421
EJC1NAL	143.738	0.484
FMB1NAL	168.449	0.173
GPC1NAL	132.492	0.37
HBL1NAL	236.561	0.54
JAF1NAL	211.764	0.24
JAN1NAL	260.528	0.279
JAP1NAL	240.484	0.45
JEG1NAL	241.538	0.3
JMC1NAL	173.188	0.166
JTH1NAL	298.351	0.131
JXC1NAL	238.614	0.275
KAN1NAL	122.232	0.111
LAD1NAL	240.883	0.4
LDP1NAL	316.504	0.2
LLA1NAL	258.633	0.235
LMV1NAL	303.744	0.38

Table 5.1 Data Group of Normal Speech Samples for Closed Test.

Identification	Fo(Hertz)	RAP
AAT30AN	104.403	3.049
AAT31AN	103.797	3.287
ASR20AN	106.145	3.965
BRT18AN	303.04	3.078
BSA08AN	85.254	3.088
BXD17AN	122.161	3.74
CAR10AN	198.78	3.472
CXP02AN	199.331	3.909
DJM28AN	188.485	4.946
EED07AN	507.207	3.709
FLW13AN	231.849	4.134
FMC08AN	195.574	3.211
FRH18AN	148.563	3.595
IGD16AN	178.716	3.217
JCL50AN	170.424	4.344
JJD29AN	132.554	3.167
LBA24AN	220.949	3.303
MMD01AN	225.826	3.714
AMC23AN	196.57	2.277
AXT11AN	184.529	2.305
BMM09AN	233.269	2.284
CMS25AN	184.001	2.806
CXL08AN	170.731	0.17783

Table 5.2 Data Group of Abnormal Speech Samples for Closed Test.

5.4 Power Spectrum Estimation

5.4.1 Fourier Spectrum

The MATLAB FFT was used to compute the speech power spectrum estimate. The results will be discussed in Chapter 6.

5.4.2 Maximum Entropy Spectrum

The predictor error filter coefficients for Maximum Entropy Power Spectrum estimation were computed using an algorithm developed by Burg which is based on a

least squares solution for the coefficients. Once the coefficients α_m are computed, they are plugged into the expression:

$$P(k) = \frac{1}{|1+\sum_{m=1}^{M}\alpha_m e^{-\frac{j2\Pi km}{Nc}}|^2} \qquad (5.1)$$

where m is the number of coefficients, k is the frequency index and c is a constant which determines the frequency spacing between samples of the spectrum, c = 1 provides the typical radian frequency spacing of $\Delta\omega = \frac{2\pi}{N}$

where N is the number of samples in the time record.

The predictor error filter coefficients for Maximum Entropy Power Spectrum estimation were computed as discussed in Section 3.2. MATLAB has an inbuilt function PBURG, which can perform Burg spectrum analysis on speech signals. This function was used for obtaining the power spectrum of our speech signals. The ME method requires parameter optimization.

5.4.2.1 Maximum Entropy Spectrum Optimization

The ME spectrum optimization procedure involved determining the best analysis parameter values of signal length and filter order for spectrum estimation. This method consisted of classification on data samples obtained with known characteristics. In optimizing the spectrum parameters for harmonic bandwidth measurements, exponential pulse train samples with 1% and 2% jitter were used in closed tests. Sequence length N was always selected to be in integral multiples of the mean pitch period To (i.e . , N = No*To). This enables the observed harmonic bandwidths to be strictly a function of jitter and the number No of pitch periods,

alone. This was performed by varying the filter order from 0.9*To to 1.3*To, where To is the number of points in the pitch period.

5.5 Classification

A MATLAB program was written for speech classification based on the Bayes decision criterion as discussed in Chapter 4. This MATLAB program accepts two groups x_1 and x_2 which are the harmonic bandwidths of normal (1% jitter) and abnormal (2% jitter) speech, respectively. Feature mean and variance are computed for each group. As discussed in Chapter 4, this program uses Bayes decision criteria. Based on the threshold given it classifies a given, input harmonic bandwidth sample as either normal or abnormal.

The probability of correct detection (PCD) is calculated based on the number of correctly assigned samples. The results are shown in Chapter 6.

5.6 Measurement of Harmonic Bandwidth

5.6.1 Fast Fourier Transform

The MATLAB FFT was used to compute the power spectrum estimate for the synthesized speech signal. The FFT power spectrum estimate was performed on exponential pulse train samples containing 1 % and 2 % jitter values. Once the spectrum was obtained, first harmonic bandwidth at10 db below the peak was calculated as shown in the Figure 5.2.

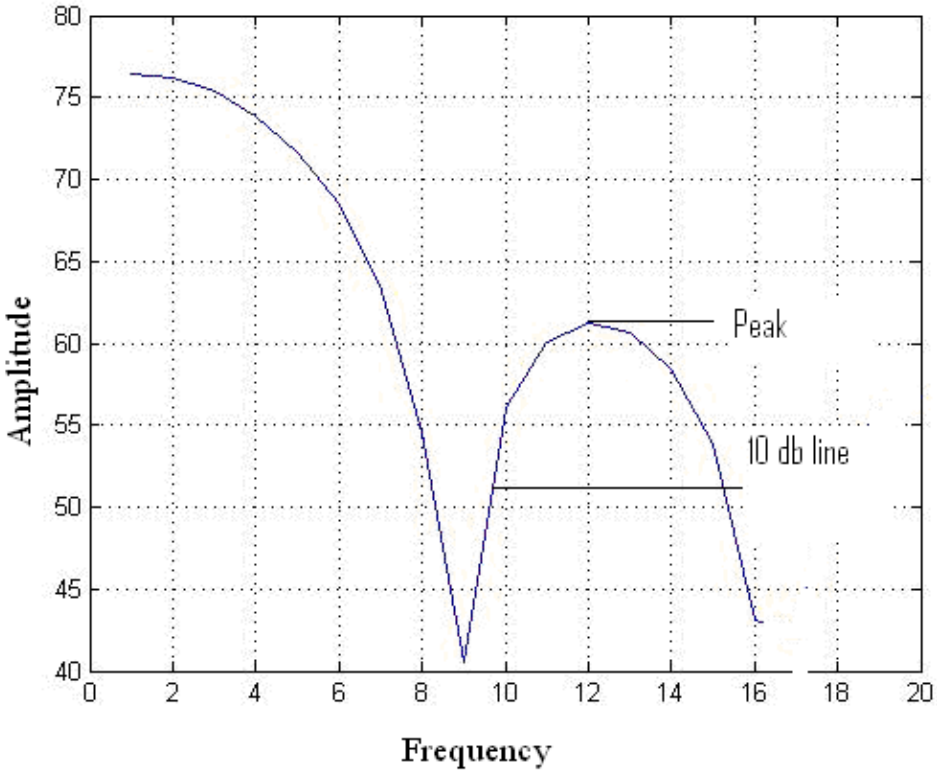

Figure 5.2 First Harmonic Bandwidth Measurement

5.6.2 Maximum Entropy Harmonic Bandwidth

Maximum Entropy spectrum estimates were computed using the method described in Section 5.4.2 The best parameter values were selected by using the method discussed in Section 5.4.2.1 The input data consisted of 60 exponential pulse trains with 1% jitter and 2% jitter levels. Again, first harmonic bandwidths at 10 db below the peak were measured. Once the harmonic bandwidth measurements for synthesized speech were completed, real speech data was processed. Classification was performed as discussed in Chapter 4 for all the data acquired.

CHAPTER 6

RESULTS AND DISCUSSION

6.1 Introduction

This chapter presents the results obtained using procedures described in Chapter 5. ME-spectrum parameter optimization and classification testing results for synthesized speech are given.

6.2 FFT Results.

As discussed in Chapter 5, the FFT power spectrum was computed for synthesized speech for 4000 FFT points and some results are shown here. Figure 6.1 shows the first harmonic for 1% jitter and Figure 6.2 shows the first harmonic for 2% jitter.

FFT failed to give good results even after increasing the number of FFT points and using optimum windowing techniques like Gaussian and Hamming. It failed to allow clear differentiation between signals with 1% jitter and 2% levels of jitter.

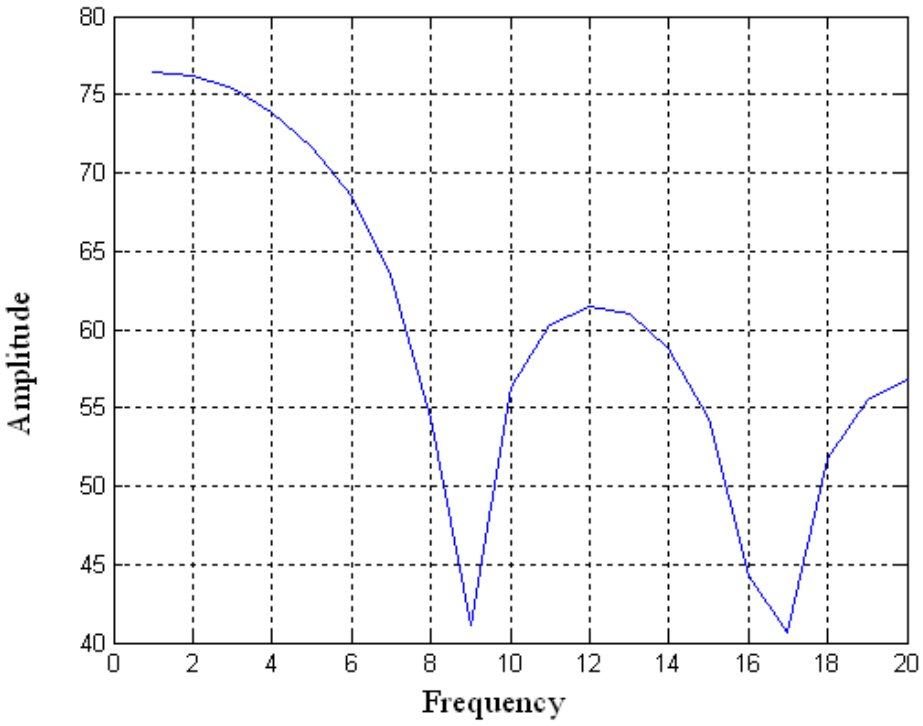

Figure 6.1 First Harmonic Obtained Using FFT for 1% Jitter

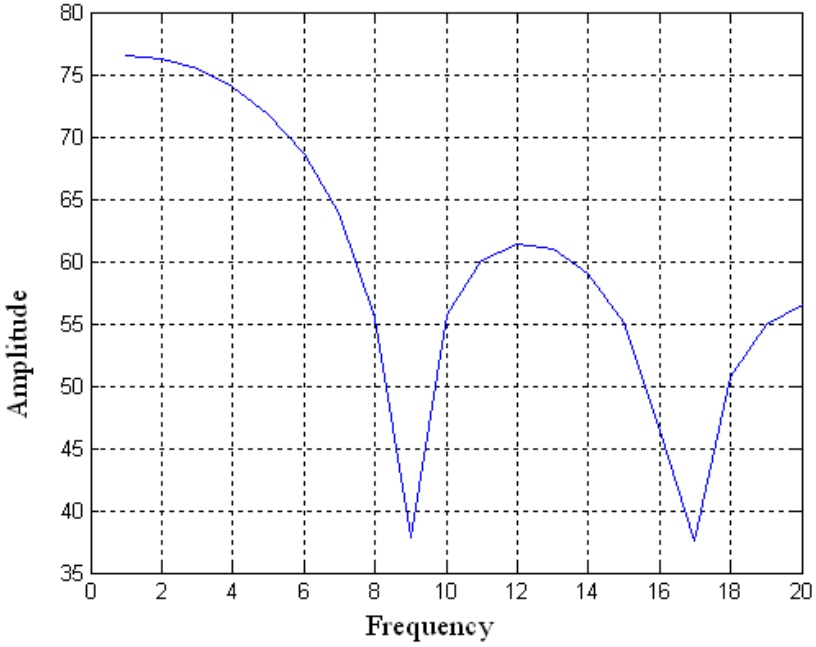

Figure 6.2 First Harmonic Obtained Using FFT for 2% jitter

The above two Figures 6.1 and 6.2 are very similar and there is no difference in 10 dB harmonic bandwidth measurements. It shows that sufficient resolution is not obtained using the FFT.

Non-parametric methods like FFT require long signals for good resolution and more over there is spectral leakage when using a rectangular window. When we use other windows we may reduce the leakage but in this process we will degrade the resolution.

6.3 ME Spectrum Optimization Results

As discussed in Chapter 5, ME harmonic bandwidth measurements were taken on synthesized speech. ME filter orders ranged from 0.9*To to 1.3*To for a constant record length of 30 pitch periods. After determining the optimum filter order, that value was used to find the optimum signal length. The signal length was varied from 20 to 40 pitch periods. Closed test classification was performed using zero dB

threshold value. Figure 6.3 shows graph of normalized filter orders versus PCD for a fixed signal length of 30*To for filter orders 0.9*To, 1.0*To, 1.1*To, 1.2*To and 1.3*To.

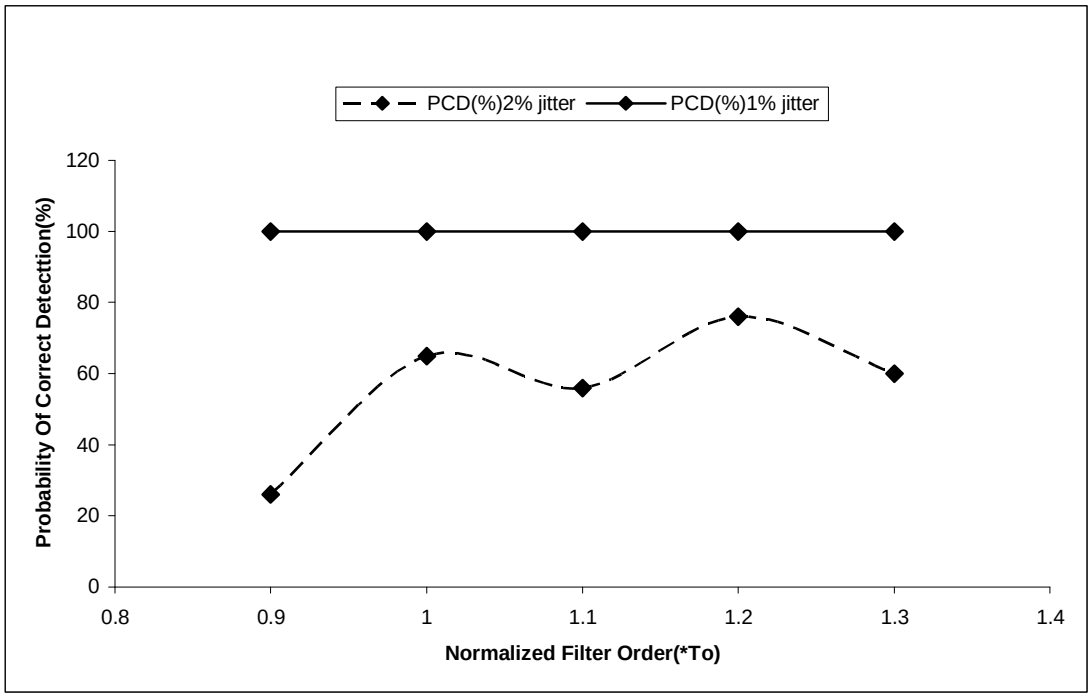

Figure 6.3 PCD vs Normalized Filter Order (length 30*To).

Table 6.1 shows probability of correct detection values for different normalized filter orders Fo for a fixed length of 30 pitch periods.

Fo(*To)	PCD(%) 2% jitter	PCD(%) 1% jitter
0.9	26	100
1	65	100
1.1	56	100
1.2	76	100
1.3	60	100

Table 6.1 PCD Values for Different Normalized Filter Orders Fo (length 30*To).

From the Table 6.1, we see that 1.2 is the best normalized filter order. In order to optimize the length, the same experiment is repeated using different signal length for the fixed filter order of 1.2*To. A plot of PCD vs normalized signal length is shown in Figure 6.4. Best results are achieved for a normalized signal length of 40 pitch periods.

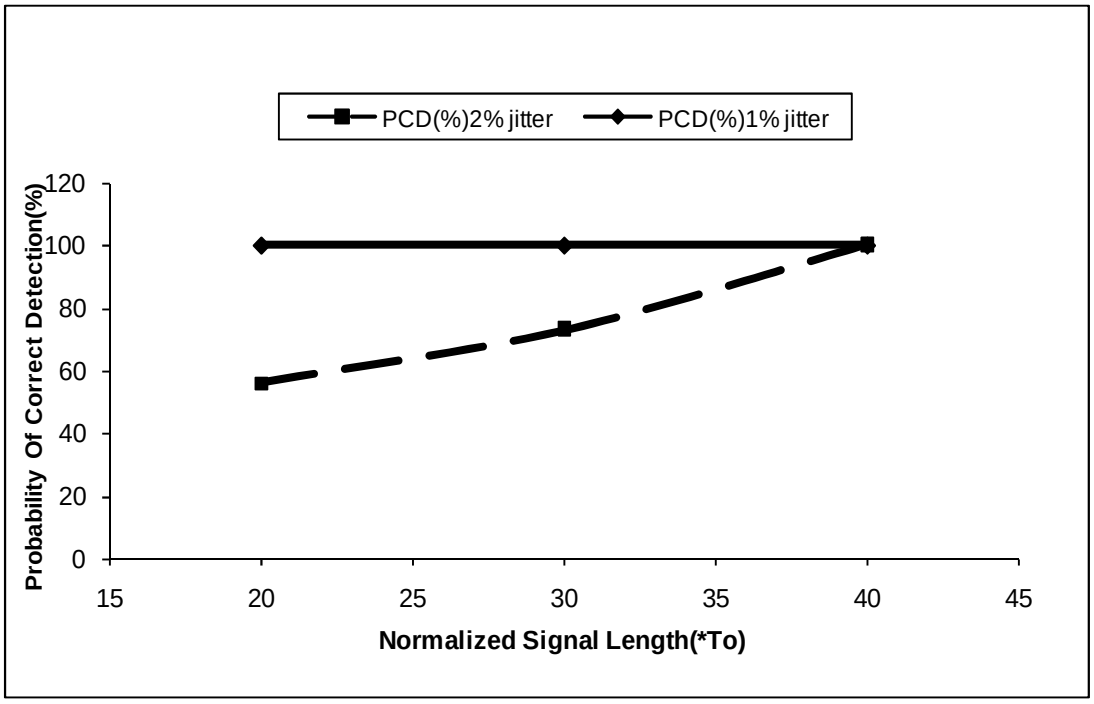

Figure 6.4 Normalized Signal Length vs. PCD

6.3.1 Relation Between Filter Order and Power Spectrum Resolution

The graphs in Figure 6.5 and Figure 6.6 show the relation between resolution and filter order in a power spectrum for synthesized speech and real speech respectively. It shows that the resolution increases as the filter order increases.

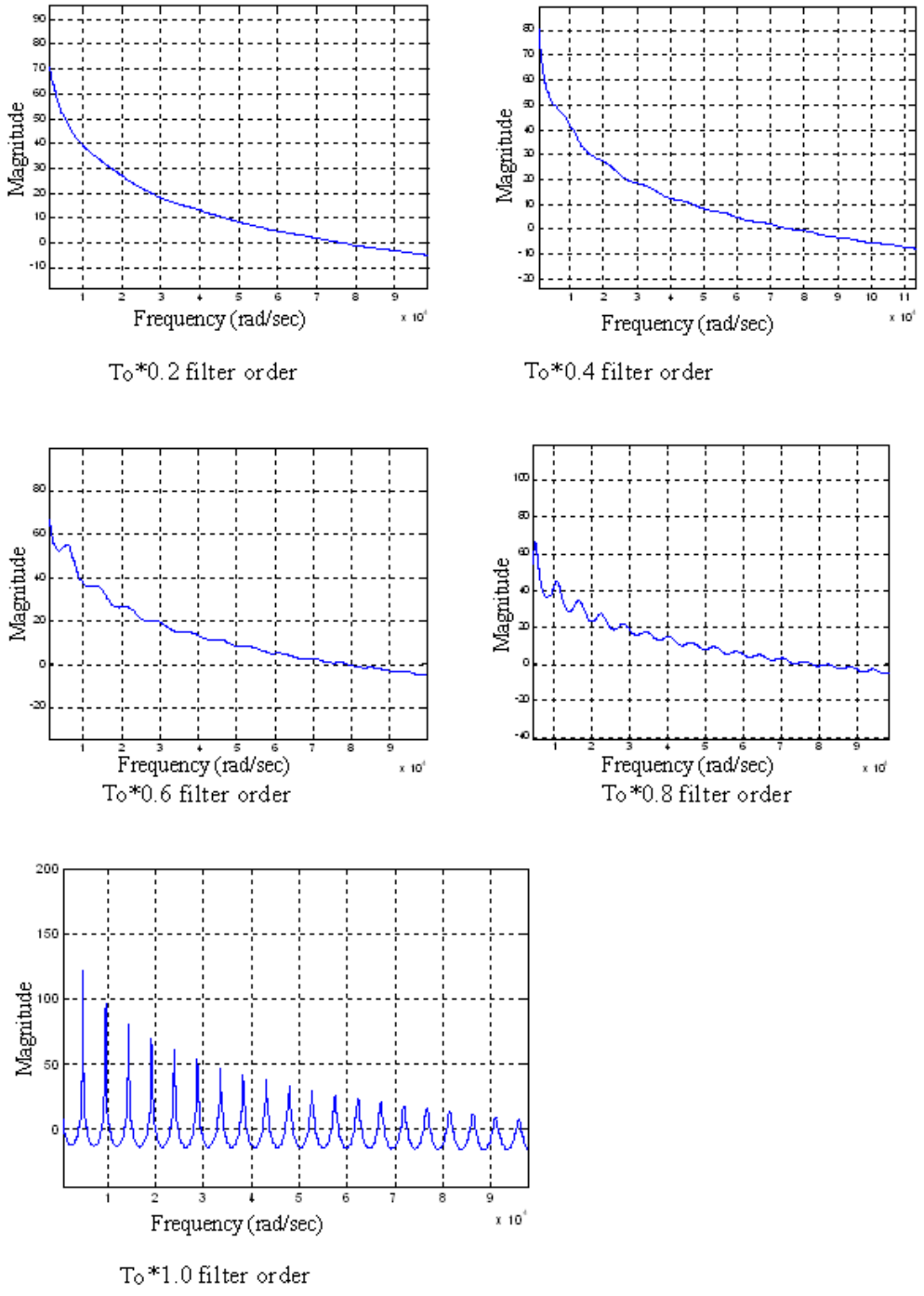

Figure 6.5 Inverse Filter Spectra as a Function of Filter Order: Synthesized Speech

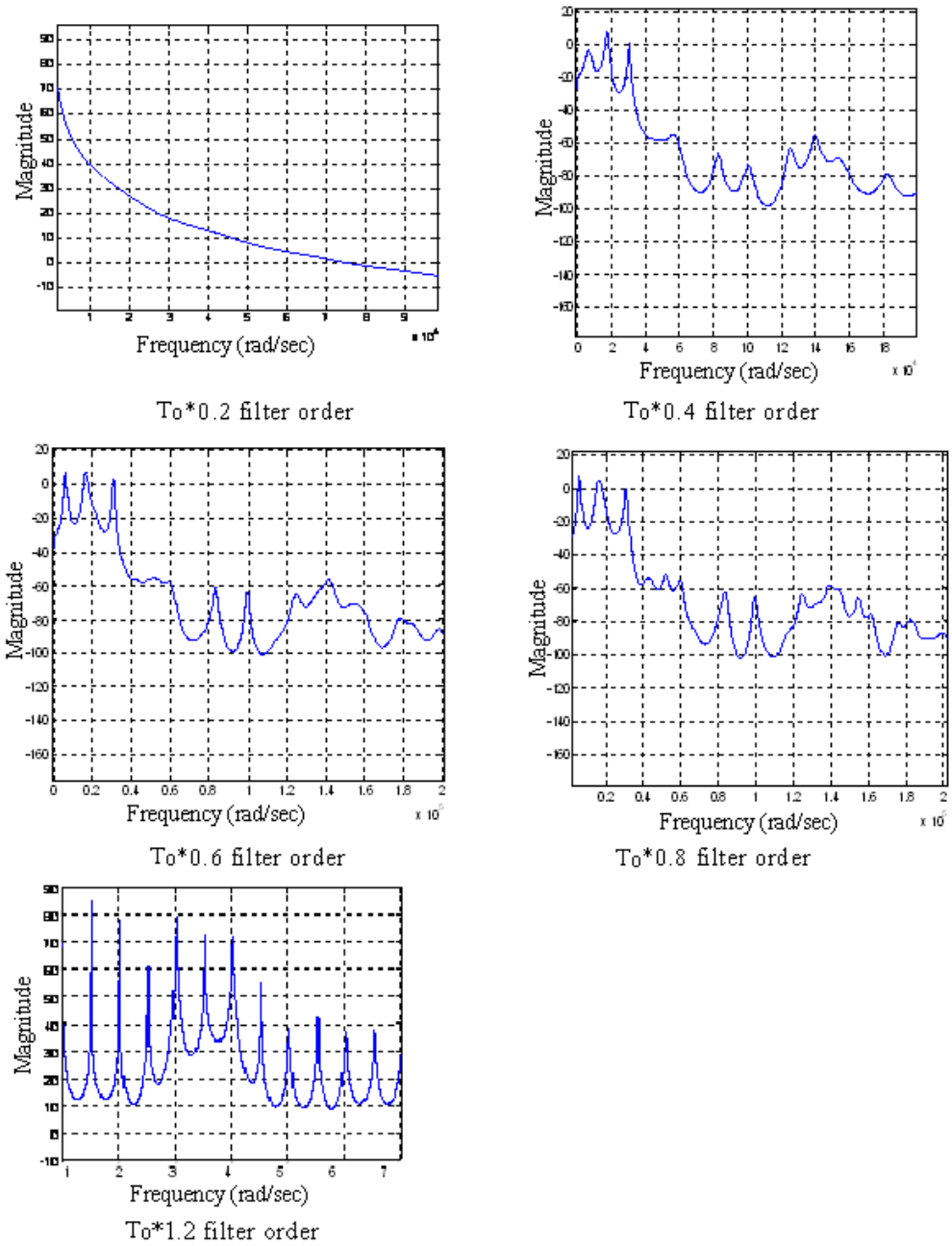

Figure 6.6 Inverse Filter Spectra as a Function of Filter Order using Real Speech

6.4 Burg Spectrum Estimate Results

The FFT failed to produce a useful power spectrum estimate for our analysis. One second of speech is not long enough to provide sufficient FFT resolution. A minimum of 10 seconds of speech is required to provide 0.1 Hz of FFT spectral

resolution. It is difficult to sustain a vowel sound for 10 seconds to produce a stationary signal. Optimized filter order and signal length which were chosen in order to get the best spectral performance from the PBURG power spectrum. From the results shown in Section 6.3, it is clear that, a filter order of 1.2 and a length of 40 is an optimum selection for the Burg analysis. Hence, these parameters were included in PBURG for real speech. Once the spectrum was obtained, harmonic bandwidth calculations were made on the first harmonic at 10 db below the peak value. Harmonic bandwidth in terms of digital frequency Dhb, is obtained from the plot. To convert Dhb into analog frequency, it was multiplied by $F_s/2$, where F_s is the sampling frequency. This algorithm was used to compute the harmonic bandwidth for synthesized speech samples with jitter levels of 1% and 2%, using PBURG. Figure 6.7 is a plot of probability of correct detection versus threshold values obtained from classifying synthesized speech samples with 1% and 2% levels of jitter.

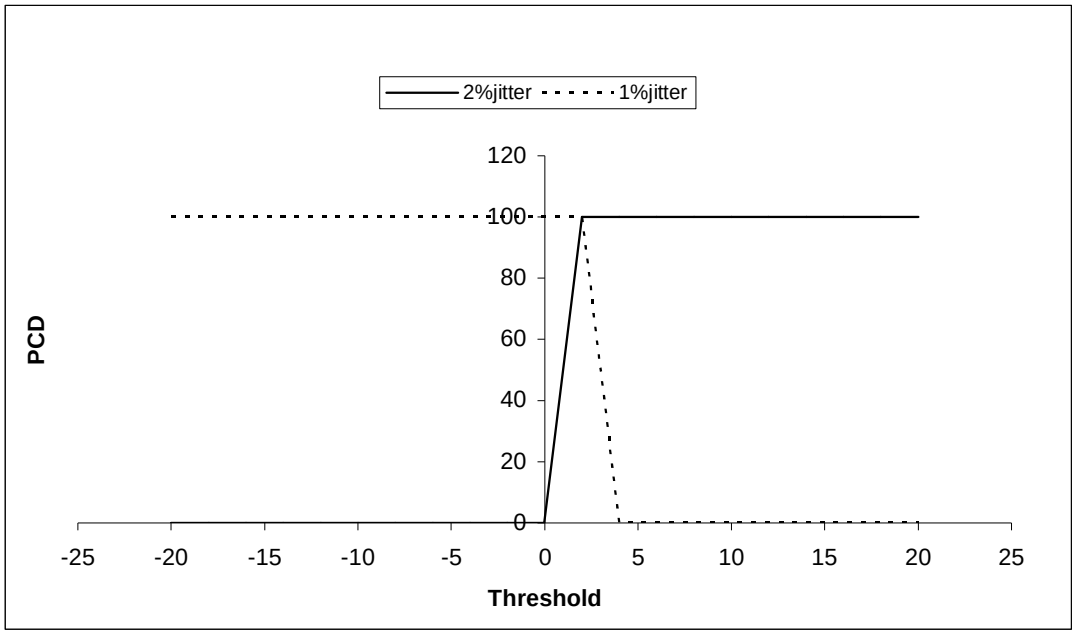

Figure 6.7 PCD vs. Threshold Values for Synthesized Speech - Closed Test

From Figure 6.7, it is clear that using the optimum parameter values, 100% PCD results were obtained. Hence, the same test was performed using real speech data which is discussed in section 6.5

6.5 Real Speech Results

Using MATLAB's inbuilt function, the wave files were read and speech signals in the time domain were plotted. In order to compare normal and abnormal signals, normal signals which were originally sampled at 50 kHz, were down sampled so that both groups would have the sampling rate of 25 kHz. Bandwidth was computed for real speech signals containing 40 periods. As discussed in Chapter 5 closed test and open test were performed and Table 6.2 and 6.3 show the normal and abnormal harmonic bandwidths for 23 samples of real speech for each, where Fo is the fundamental frequency and RAP is the relative average perturbation. RAP is a measurement of pitch period jitter [2].

Identification	Fo (Hertz)	RAP	BW(Hertz)
BJV1NAL	247.134	0.098	0.09007
CAD1NAL	302.78	0.156	0.2523
DAJ1NAL	210.022	0.285	0.18593
DFP1NAL	216.849	0.4888	0.192
DMA1NAL	239.3	0.238	0.2991
DWS1NAL	184.855	0.266	0.1059
EDC1NAL	217.661	0.421	0.136
EJC1NAL	143.738	0.484	0.0672
FMB1NAL	168.449	0.173	0.11407
GPC1NAL	132.492	0.37	0.9385
HBL1NAL	236.561	0.54	0.24643
JAF1NAL	211.764	0.24	0.35297
JAN1NAL	260.528	0.279	0.1628
JAP1NAL	240.484	0.45	0.17537
JEG1NAL	241.538	0.3	0.33967
JMC1NAL	173.188	0.166	0.06313
JTH1NAL	298.351	0.131	1.5384
JXC1NAL	238.614	0.275	0.12427
KAN1NAL	122.232	0.111	0.3756
LAD1NAL	240.883	0.4	1.70627
LDP1NAL	316.504	0.2	0.41213
LLA1NAL	258.633	0.235	0.1213
LMV1NAL	303.744	0.38	0.17403

Table 6.2 Harmonic Bandwidths for Normal Speech

Identification	Fo(Hertz)	RAP	BW (Hertz)
AAT30AN	104.403	3.049	16.0176
AAT31AN	103.797	3.287	14.99107
ASR20AN	106.145	3.965	1.6751
BRT18AN	303.04	3.078	3.56707
BSA08AN	85.254	3.088	1.93153
BXD17AN	122.161	3.74	1.27887
CAR10AN	198.78	3.472	1.37697
CXP02AN	199.331	3.909	0.46713
DJM28AN	188.485	4.946	0.31413
EED07AN	507.207	3.709	4.75507
FLW13AN	231.849	4.134	1.41283
FMC08AN	195.574	3.211	0.76397
FRH18AN	148.563	3.595	1.60763
IGD16AN	178.716	3.217	0.1955
JCL50AN	170.424	4.344	1.5793
JJD29AN	132.554	3.167	5.4057
LBA24AN	220.949	3.303	0.33373
MMD01AN	225.826	3.714	0.63517
AMC23AN	196.57	2.277	0.36857
AXT11AN	184.529	2.305	0.12493
BMM09AN	233.269	2.284	0.63177
CMS25AN	184.001	2.806	0.0767
CXL08AN	170.731	0.17783	0.17783

Table 6.3 Harmonic Bandwidths for Abnormal Speech

Once the harmonic bandwidths were computed, classification was performed at different threshold values. Table 6.4 shows the threshold values and the probability of correct detection values for close test. Figure 6.5 shows the graph plotted for the values in Table 6.5.

Threshold	PCD(%) Abnormal	Threshold	PCD(%) Normal
-20	30.434	-20	100
-18	30.434	-18	100
-16	30.434	-16	100
-14	34.78	-14	100
-12	34.78	-12	100
-10	43.47	-10	100
-8	47.82	-8	100
-6	52.17	-6	100
-4	52.17	-4	100
-2	52.17	-2	100
0	56.52	0	100
2	82.6	2	100
4	100	4	60.86
6	100	6	0
8	100	8	0
10	100	10	0
12	100	12	0
14	100	14	0
16	100	16	0
18	100	18	0
20	100	20	0

Table 6.4 PCD vs Threshold for Closed Test.

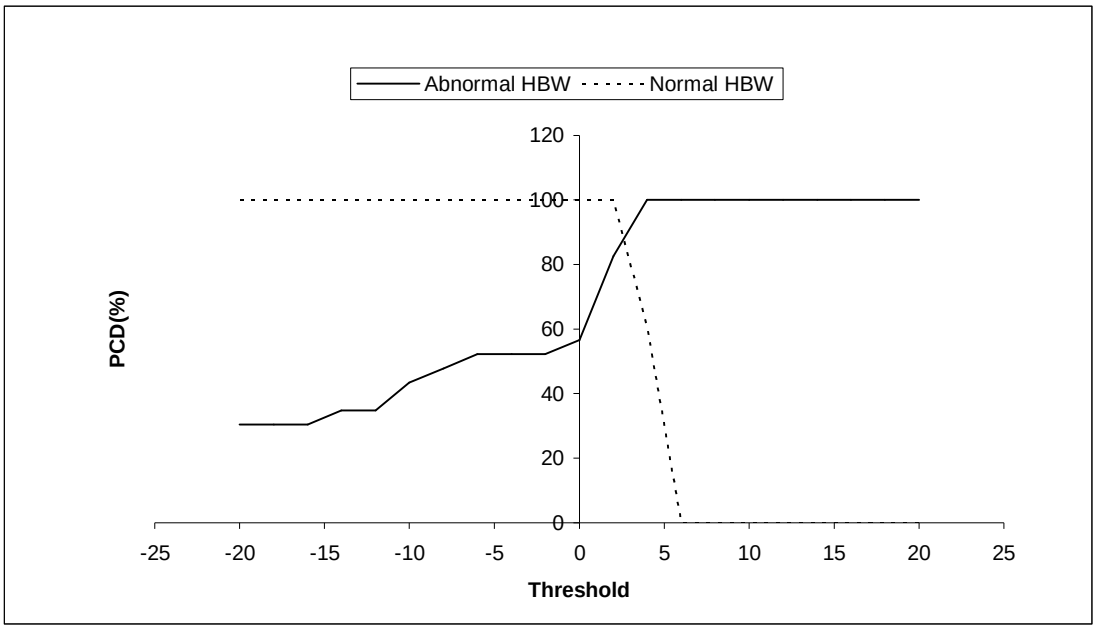

Figure 6.8 PCD vs Threshold Values for Real Speech - Closed Test.

From the Figure 6.8 it is clear that closed test results gave 82.6% results for abnormal speech and 100 % results for normal speech at a threshold value of 2.

The same tests were performed on another set of harmonic bandwidth values shown in Table 6.5 and Table 6.6 respectively for normal and abnormal speech samples. Table 6.7 shows the probability of detection values and threshold values for open test. Figure 6.9 shows the graph plotted for the values in Table 6.7

Identification	Fo(Hertz)	RAP	BW(Hertz)
LMW1NAL	224.929	0.382	0.2226
MCB1NAL	257.011	0.209	0.8701
MFM1NAL	151.24	0.324	0.1645
MJU1NAL	140.49	0.214	0.27807
MXZ1NAL	230.232	0.545	0.2758
NJS1NAL	241.156	0.418	0.1884
OVK1NAL	121.102	0.199	0.3406
OVK1NAL	121.102	0.199	0.3406
PBD1NAL	247.085	0.376	0.3346
RHM1NAL	120.394	0.087	0.22573
RJS1NAL	124.716	0.229	0.37677
SCT1NAL	225.387	0.494	0.35217
SEB1NAL	237.029	0.372	0.53087
SIS1NAL	129.366	0.086	0.18867
SLC1NAL	240.885	0.251	0.33873
SXV1NAL	188.554	0.137	0.07857
TXN1NAL	122.293	0.147	0.64967
VMC1NAL	219.61	0.17	0.2745
DJG1NAL	121.805	0.849	0.406
JKR1NAL	240.348	0.641	0.1377
MAM1NAL	250.87	0.218	0.13067
WDK1NAL	146.242	0.224	0.10667
RHG1NAL	132.452	0.443	0.4989

Table 6.5 Harmonic Bandwidth Values for Normal Speech - Open Test

Identification	Fo(Hertz)	RAP	BW(Hertz)
CXM14AN	221.94	2.203	0.52017
CXN14AN	221.94	2.203	0.52017
DGL30AN	205.131	2.29	0.22437
DRG19AN	111.804	2.179	0.5474
EDG19AN	188.345	2.513	1.16737
EEB24AN	160.206	2.807	0.9512
EGW23AN	217.944	2.984	0.61297
EXS07AN	212.004	2.623	5.76383
GEK02AN	130.997	2.016	1.15987
GLB22AN	96.46	2.402	0.55263
GSB11AN	159.759	2.022	0.17473
HMG03AN	180.268	2.055	0.2441
JAB08AN	128.523	2.452	0.34807
JAF15AN	143.896	2.645	0.29227
JCL20AN	149.002	2.58	0.22503
JLC08AN	189.058	2.466	0.5613
JXS09AN	106.105	2.339	8.74813
KCG23AN	240.922	2.093	0.22583
KJM08AN	130.756	2.017	0.37457
KMC19AN	210.304	2.802	0.43813
LBA15AN	231.476	2.583	0.09643
LCW30AN	190.973	2.872	0.30833
MAB06AN	200.338	2.726	0.14607

Table 6.6 Harmonic Bandwidth Values for Abnormal Speech -Open Test

Threshold	PCD(%) normal	Threshold	PCD(%) abnormal
-20	100	-20	43.47
-18	100	-18	47.82
-16	100	-16	47.82
-14	100	-14	52.17
-12	100	-12	52.17
-10	100	-10	52.17
-8	100	-8	52.17
-6	100	-6	52.17
-4	100	-4	52.17
-2	100	-2	52.17
0	100	0	69.56
2	0	2	100
4	0	4	100
6	0	6	100
8	0	8	100
10	0	10	100
12	0	12	100
14	0	14	100
16	0	16	100
18	0	18	100
20	0	20	100

Table 6.7 Threshold vs Probability of Correct Detection for - Open Test.

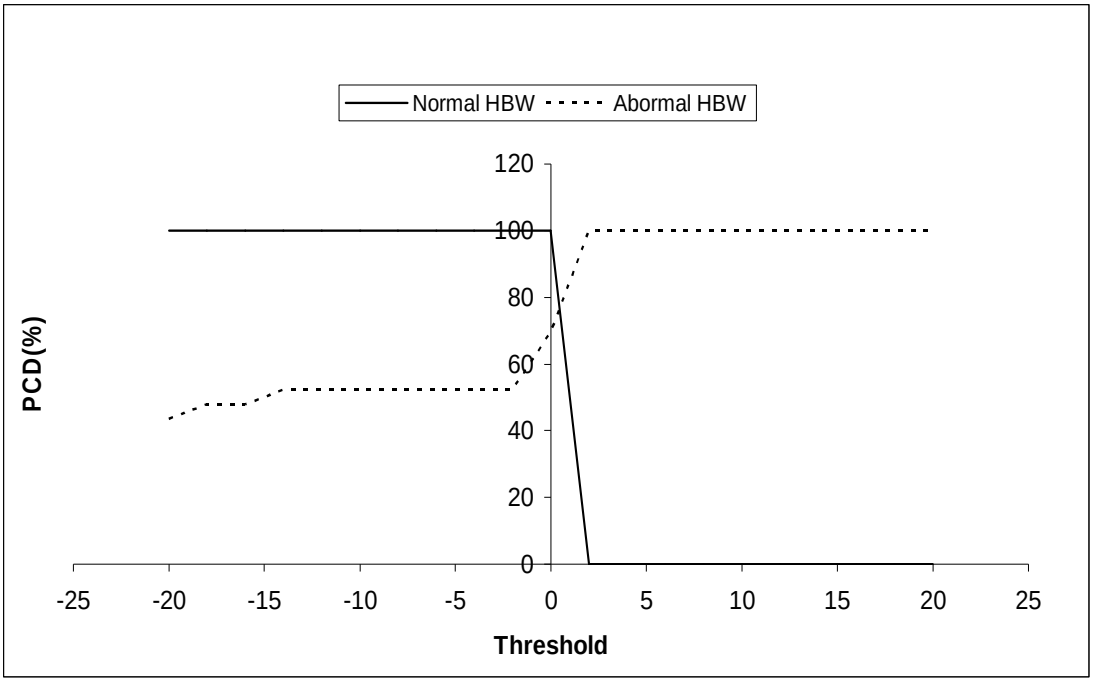

Figure 6.9 Probability of Correct Detection vs Threshold Values for Open Test.

From Figure 6.7, it is shown that for a threshold value of 0 a PCD of 69.56% resulted for abnormal speech and 100% resulted for normal speech.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

The objective of this research was to develop a method of detecting certain laryngeal pathologies through harmonic bandwidth measurements in speech signals. It has been determined that laryngeal disorders in speech can result in excessive amount of pitch period jitter in speech which causes a widening of harmonic bandwidths. Because of the narrow harmonic bandwidths of speech, high resolution power spectra are required for discriminating between speech containing normal and abnormal levels of jitter. This research focused on producing high resolution power spectrum estimates of speech signals and the classification of their harmonic bandwidth measurements. The FFT periodogram was not useful for providing sufficient spectral resolution for discriminating between signals containing 1% and 2% levels of jitter. After optimizing the filter order and record length parameters of the Burg Maximum Entropy spectrum using synthesized speech, we achieved a closed test probability of correct detection (PCD) of 82.6% and an open test PCD of 69.56 % using real speech with 0.3 % false alarm rate. All of the real speech data was acquired from the KayPENTAX Company. Jitter values in the form of Relative Average Perturbation (RAP) quotients were listed for each speech record which were /ah/ vowel sounds. We did not confirm these RAP values through our own measurements. Any future work with this database should include some confirmation of the jitter values, because some of spectral results for real speech were not consistent with results obtained for comparable synthesized speech.

BIBLIOGRAPHY

[1] Philip Lieberman, “Some Acoustic Measures of the Fundamental Periodicity of Normal and Pathologic Larynges”, The Journal of the Acoustical Society of America, vol. 35, No.3, pp.344-353, March 1953.

[2] Koike,Y., “Application of Some Acoustic Measures for thr Evaluation of Laryngeal Dysfunction,” Studia Phonologica,Vol.7.pp.17-23,1973.

[3] Donald G. Childers, J., “Electroglottography for Laryngeal Function Assessment and Speech Analysis”, IEEE Transactions on Biomedical Engineering, vol. 31, no. 12, pp. 807-817, March 2007.

[4] Martinez Cesar E., Rufiner Hugo L., “Acoustic Analysis of Speech for Detection of Laryngeal Pathologies”, Proc. of the 22nd Annual Int. Conference of the IEEE Engineering in Medicine and Biology Society, vol No.3, pp. 2369-2372, August 2008.

[5] Peter Mitev, Stefan Hadjitodorov, “Fundamental frequency estimation of voice of patients with laryngeal disorders”, Information Sciences, vol. no. 156, pp. 3-19, November 2003.

[6] Stefan Hadjitodorov, Boyan Boyanov, and Bernard Teston, “Laryngeal Pathology Detection by Means of Class-Specific Neural Maps”, IEEE Trans. On Inf. Technology Biomedical Engineering, vol.4, pp. 68-73, March 2000.

[7] Alireza A. Dibazar, S. Narayanan, T. W. Berger, “Feature Analysis for Automatic Detection of Pathological Speech”, Proc. of Second Joint EMBS/BMES Conference, vol.1, pp. 182-183, January 2003.

[8] Joseph P. Campbell, Jr. and Douglas A. Reynolds, “Corpora for the Evaluation of Speaker Recognition Systems”, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, vol. 2, pp. 829-832, August 2002.

[9] “Folds, Nodules, Polyps, Cysts and Reactive Lesions: Treatment”, http://www.voiceproblem.org/pdfs/vocal_fold_lesions.pdf.

[10] Vieira Maurilio N., Mcinnes Fergus R., Jack Mervyn A., “On The Influence Of Laryngeal Pathologies on Acoustic and Electroglottographic Jitter Measures”, The Journal of the Acoustical Society of America, vol. 111, pp. 1045-1055, 2002.

[11] David Sorrensen and Yoshiyuki Horii, “Effects of Laryngeal Topical Anesthesia on Voice Fundamental Frequency Perturbation”, Journal of Speech and Hearing Research, Vol.23, pp. 274-283, June 1980.

[12] “Phonetics and Theory of Speech Production”, http://www.acoustics.hut.fi/publications/files/theses/lemmetty_mst/chap3.html.

[13] “Voice Recognition”, http://www.barcode.ro/tutorials/biometrics/voice.html.

[14] Robert Mannell, “Speech Acoustic Vocal Tract Resonance”, http://clas.mq.edu.au/acoustics/frequency/vocal_tract_resonance.html.

[15] Alan V. Oppenheim and Ronald W. Schafer, “Digital Signal Processing”, Prentice Hall, April 1998.

[16] Wang Hongwei, “FFT Basics and Case Study using Multi-Instrument”, May 2009.

[17] Speech Synthesis *by J. L.* Flanagan , October 1982.

[18] Perceptual Differentiation of Vocal Fry and Harshness by *John F. Michel, 1964*

[19] Donal G.Childers, *Modern Spectrum Analysis-IEEE Press* 1978.

[20] Dr.Juan I. Godino-Llorente “Methodological issues in the development of automatic systems for voice pathology detection”-February 2006

[21] M.Mezzalama, P.Prinetto, B.Morra “Experiments in automatic classification of laryngeal pathology” April 1982

[22] McGaugh, Eugene, “The Detection of Laryngeal Disorders Through the Spectral Analysis of Speech Signals” Ph.D. Dissertation, University of Kansas, 1982.

VITA

Graduate College
University of Nevada, Las Vegas

Priyanka Medida

Degrees: Bachelor of Engineering, 2006
Anna University, Chennai, India

Thesis Title: Spectral Analysis of Pathological Acoustic Speech Waveforms

Thesis Examination Committee:
Chairperson, Dr. Eugene McGaugh, Ph.D.
Committee Member, Dr.Venkatesan Muthukumar, Ph.D.
Committee Member, Dr. Emma Regentova, Ph.D.
Graduate Faculty Representative, Dr. Satish Bathnagar, Ph.D.