BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY

Formant Estimation for Noise Robust Vowel Recognition Based on Spectral Domain Ramp Cepstrum Model

by

Rajib Goswami

MASTER OF SCIENCE IN ELECTRICAL AND ELECTRONIC ENGINEERING

Department of Electrical and Electronic Engineering
BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY

July 2012

The thesis entitled **Formant Estimation for Noise Robust Vowel Recognition Based on Spectral Domain Ramp Cepstrum Model"** submitted by Rajib Goswami, Student No.: 1009062055, Session: October, 2009 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of MASTER OF SCIENCE IN ELECTRICAL AND ELECTRONIC ENGINEERING on July 11, 2012.

BOARD OF EXAMINERS

1. _____

(Dr. Shaikh Anowarul Fattah) — **Chairman** (Supervisor)
Associate Professor
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka - 1000, Bangladesh.

2. _____

(Dr. Pran Kanai Saha) — **Member** (Ex-officio)
Professor and Head
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka - 1000, Bangladesh.

3. _____

(Dr. Mohammed Imamul Hassan Bhuiyan) — **Member**
Associate Professor
Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology
Dhaka - 1000, Bangladesh.

4. _____

(Dr. Farruk Ahmed) — **Member** (External)
Professor
School of Engineering and Computer Science (SECS)
Independent University, Bangladesh (IUB)
Block-B, Bashundhara R/A, Dhaka-1229.

CANDIDATE'S DECLARATION

I, do, hereby declare that neither this thesis nor any part of it has been submitted elsewhere for the award of any degree or diploma.

Signature of the Candidate

Rajib Goswami

To my ever caring parents

whose inspirations are behind my every success.

Contents

Acknowledgement

The opportunity to work with my supervisor Dr. Shaikh Anowarul Fattah, whose dynamic ideas, support and guidance helped me althrough this research, was immense and extraordinary. I want to express my heartiest gratitude to him.

Completing the work in time required using the lab facilities of EEE department at different times. I would also like to thank the Head of the Department of Electrical and Electronic Engineering for allowing me to use the lab facilities. I wish to express my gratitude to Dr. Celia Shahnaz, who was a continuous source of inspiration and support. I express appreciation to my friends Shafin, Raju, Partha and Rabu, for their suggestions, support and friendship.

Abstract

Formants are the distinguishing frequency components of human speech, which play an important role in characterizing different voiced sounds. Formant based speech synthesis and coding are widely used in several real life applications. such as voice operated controls and telecommunication.In almost all practical applications speech signals are affected by different kinds of background noise and estimation of formants under severe background noise is a difficult task. In this thesis efficient formant estimation is investigated and methods for formant estimation are devised with a view to improve the estimation performance under severe noisy conditions. In order to extract the formant frequencies, first a strongly voiced portion of the given speech utterance is extracted based on the energy measure. Instead of considering the whole duration of a voiced sound at a time, frame by frame analysis is performed. Within a frame of voiced speech signal, formants can be estimated by using different time or frequency domain approaches. Correlation based methods are the most common time domain approaches to estimate formants from speech signals . In linear predictive coding (LPC) based methods, from the autocorrelation function (ACF) of the given speech utterance, Yule-Walker equations are constructed and from their solutions formants can be obtained. Spectral peak picking is another extremely popular method of formant estimation, where both parametric and non parametric spectral estimation techniques are used. Recently cepstrum domain methods has been used in formant estimation . In the presence of heavy background noise, spurious peaks appear in the speech spectrum making the task of accurate formant estimation very difficult. The estimation performance of both time and frequency domain methods deteriorates drastically under heavy noisy conditions.The main goal here is to

develop a formant estimation scheme which provides satisfactory performance even at low levels of signal to noise ratio (SNR). In order to reduce the effect of noise the strength of dominant pole pairs on the spectrum of noisy speech needs to be enhanced. With a view to achieve this objective a spectral domain ramp cepstrum model of autocorrelation function of speech signal is developed. The model utilizes the advantageous property of the ACF that provides better noise immunity in comparison to the noisy signal directly. Transforming to cepstral domain from time domain offers the advantage of homomorphic deconvolution which can reduce the effect of pitch in speech analysis. In order to avoid the rapid cepstral decay, instead of cepstrum, ramp cepstrum is used. Since, the pole preserving property of the ramp cepstrum (RC) is better exploited via spectral peaks, the spectrum of RC of the ACF of speech is proposed as the desired model. In order to extract the formants from the observed noisy speech signal utilizing the derived model, model matching scheme is introduced. In the model matching technique, instead of relying on the peak picking, fitting error is minimized over a wider peak zone resulting more accurate formant frequency estimation. Finally, the estimated formants are used in vowel recognition scheme as potential features. The linear discriminant based algorithm is used for the purpose of recognition. Extensive experimentation is carried out considering different male and female vowel utterances from standard speech database under different noisy conditions. It is found that the proposed methods provide a high degree of formant estimation accuracy in comparison to that obtained by some state of the art methods, especially at very low levels of SNR.

List of Figures

List of Tables

Chapter 1

Introduction

The ability of communicating intelligently via speech is a major quality that distinguishes homo sapiens from other species. Speech is one of the earliest modes of communication in human beings and the possibility of using speech in newer and newer applications has fascinated humans for centuries. Speaking a language is an amazing skill that serves not only in communication but also in sharing experiences, feelings, thoughts and ideas among people. Because of various acoustic and articulatory features different sounds are distinguishable and are used to form different meaningful words. One major areas of research is to analyze speech characteristics which significantly helps to deal with different real life speech applications, such as voice synthesis, speech coding, speech enhancement etc.

1.1 Speech Characteristics

1.1.1 Voiced and Unvoiced Sounds

Speech is air pressure waves radiating from the mouth and nostrils of the speaker. The main components of the human speech production apparatus are the lungs, the glottis and the vocal tract. The lungs are the source of an airflow that passes through the larynx and vocal tract, before leaving the mouth as pressure variations constituting the speech signal. At the glottis, the vocal cords constrict the path from the lungs to the vocal tract.

Vocal tract is the most important component in speech production. The vocal tract has two speech functions: (1) it can modify the spectral distribution of energy in glottal sound waves, and (2) it can contribute to the generation of sound for obstruent (stop and fricative) sounds. Different sounds are primarily distinguished by their periodicity (voiced or unvoiced sounds), spectral shape (which frequencies have the most energy), and duration [1]. The vocal folds specify the voicing feature, and a sound's duration is the result of synchronized vocal tract actions, but the major partitioning of speech into sounds is accomplished by the vocal tract via spectral filtering. As volumes of air and corresponding sound pressure waves passes through the vocal tract, certain frequencies are attenuated and others are ampified, depending on the filters frequency response.

As we speak, we change the shape of the vocal tract, and thus the frequency response of the filter, in order to produce different sounds. The excitation is filtered by the vocal tract to produce the sounds. The shape of the vocal tract, i.e. the position of the jaws, the opening of the lips, the shape of the tongue and the opening or closing of the velum will determine the frequency response of the vocal tract. In voiced sounds, air pressure from the lungs build up behind the closed vocal cords, until they abruptly open to release a burst of air before closing again. The cycle repeats and produces a quasi periodic sequence of excitation pulses.The intonation of speech is determined by the variations of the fundamental frequency. Due to the closure of vocal tract voiced sounds are produced with huge energy and they are of high magnitude. The inverse of the pulse period is called the fundamental frequency, which determines the perceived pitch of the speech signal. In unvoiced sounds, the vocal cords stay open and thus the signal amplitude and energy in unvoiced sounds are much lower than that of voiced sounds.

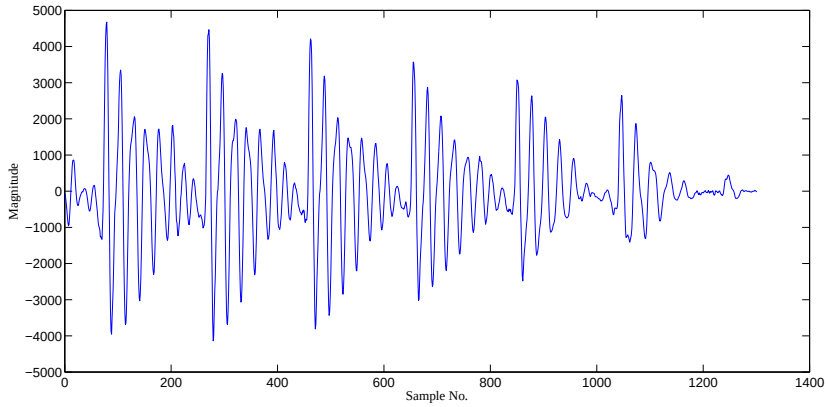

Figure 1.1: Voiced speech sound /aa/ taken form the TIMIT sentence "His head flopped back"

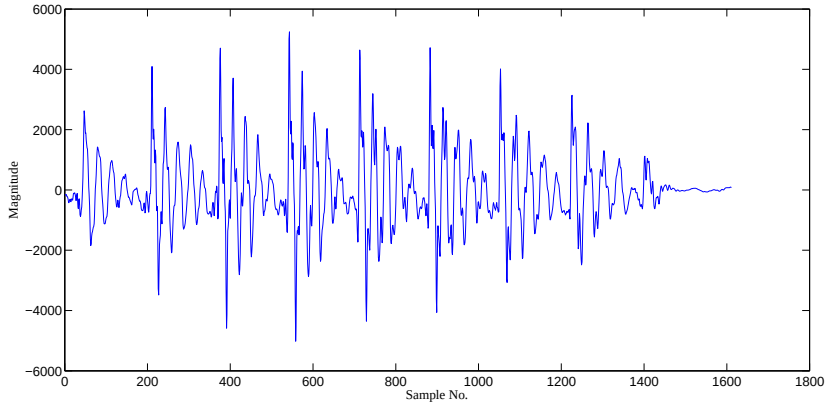

Figure 1.2: Voiced speech sound /eh/ taken form the TIMIT sentence "His head flopped back"

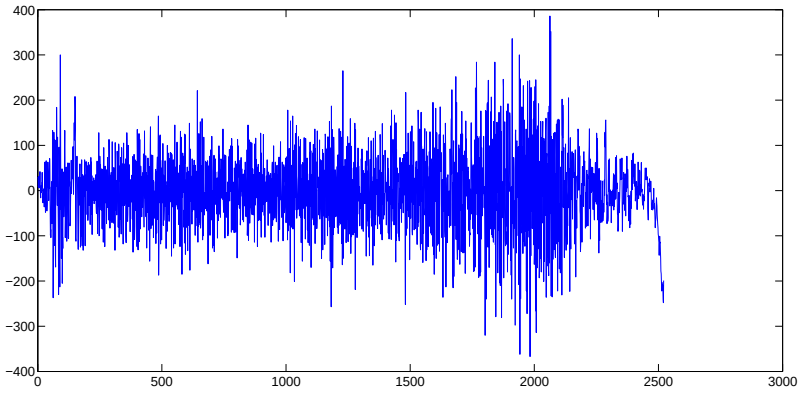

Figure 1.3: Unvoiced speech sound /f/ taken form the TIMIT sentence "His head flopped back"

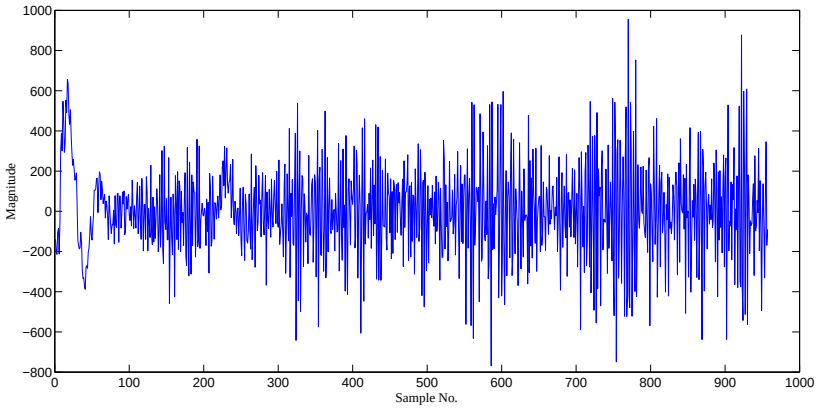

Figure 1.4: Unvoiced speech sound /z/ taken form the TIMIT sentence "His head flopped back"

In Figs. 1.1 and 1.2 voiced speech sounds /aa/ and /eh/ are shown. In Figs. 1.3 and 1.4 unvoiced speech sounds /f/ and /z/ are shown. All the sounds are taken from the sentence "His head flopped back" of the TIMIT speech corpus uttered by a male speaker. From these figures it is quite clear that the voiced sounds have much higher amplitude than the unvoiced sounds.

The smallest distinguishing unit of a language is the phoneme. The phoneme is an abstraction, covering a multitude of possible actual realizations, phones, of the sound. Different sounds are created by moving the articulators (e.g. tongue, lips, jaws). The articulators are physical entities with a mass, which means that their movement cannot be instantaneous. Thus, the realization of a phoneme will depend on the articulator positions of the preceding and the succeeding sounds. This phenomenon is called co-articulation. An accurate modeling of the co-articulation phenomena is vital for the performance of both speech recognizers and speech synthesizers.

1.1.2 Formants

The vocal tract can be modeled as an acoustic tube with resonances and antiresonances. Analysis of the vocal tract shows that the frequency response will typically be dominated by a small number of resonances, called formants. Formants are the distinguishing or meaningful frequency components of human speech. They are often mathematically de-

fined as poles of a system transfer function expressing the input-output relation of a vocal tract.

Formants are defined as the spectral peaks of the sound spectrum $|P(f)|$ of the voice. They are the characteristics of voiced sounds. They are often measured as amplitude peaks in the frequency spectrum of the sound, using a spectrogram or a spectrum analyzer, though in vowels spoken with a high fundamental frequency, as in a female or child voice, the frequency of the resonance may lie between the widely-spread harmonics and hence no peak is visible. In unvoiced sound no formants are evident.

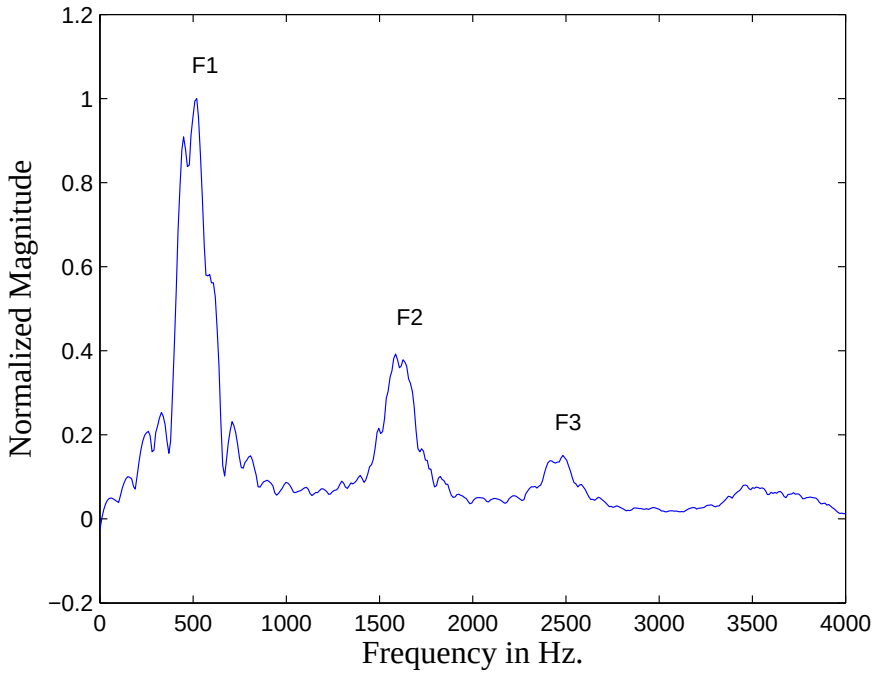

Figure 1.5: Magnitude spectrum of the voiced speech sound /eh/ showing distinct formant peaks

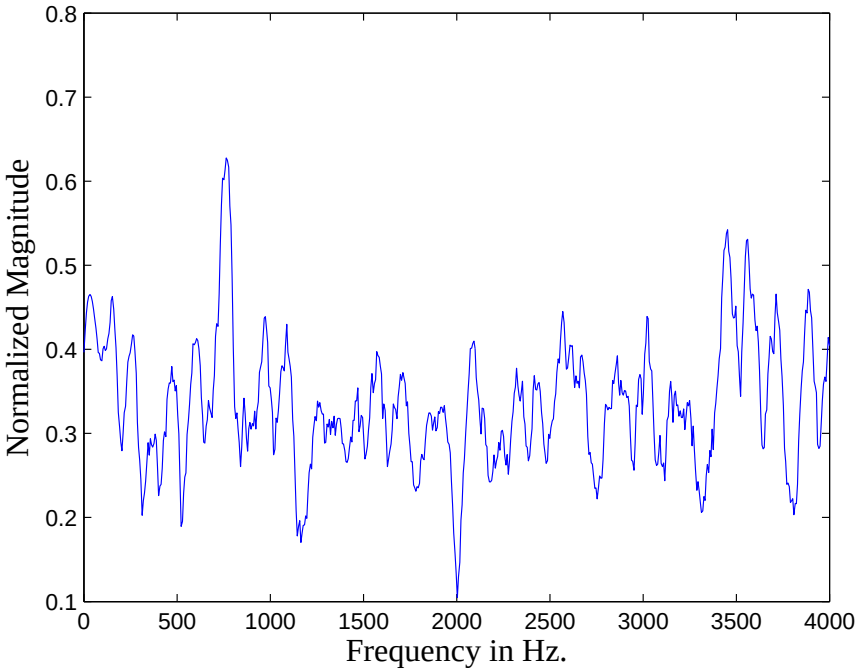

Figure 1.6: Magnitude spectrum of the voiced speech sound /f/

In Fig. 1.5 spectrum of the voiced natural sound /eh/ as presented in Fig. 1.2 is shown and in Fig. 1.6 spectrum of unvoiced natural sound /f/ as presented in Fig. 1.3 is shown. In the spectrum of voiced sound formant peaks are quite clearly identifiable, but in case of unvoiced spectrum there are no such peaks.

The formant with the lowest frequency is called F1, the second and third formants are named F2 and F3 respectively. Vowels (including diphthongs) are voiced and are the phonemes with the greatest intensity. They range in duration from 50 to 400ms in normal speech. Like all sounds excited solely by periodic glottal source, vowel energy is primarily concentrated below 1kHz and falls off at about -6dB/octave with frequency. The vowel signals are quasi-periodic due to repeated excitations of the vocal tract by vocal fold closures. Thus, vowels have line spectra with frequency spacing of $F0$ Hz (energy concentrated at multiples of $F0$). The largest harmonic amplitudes are near the low formants frequencies. Vowels are distinguished primarily by the location of their three formant frequencies.

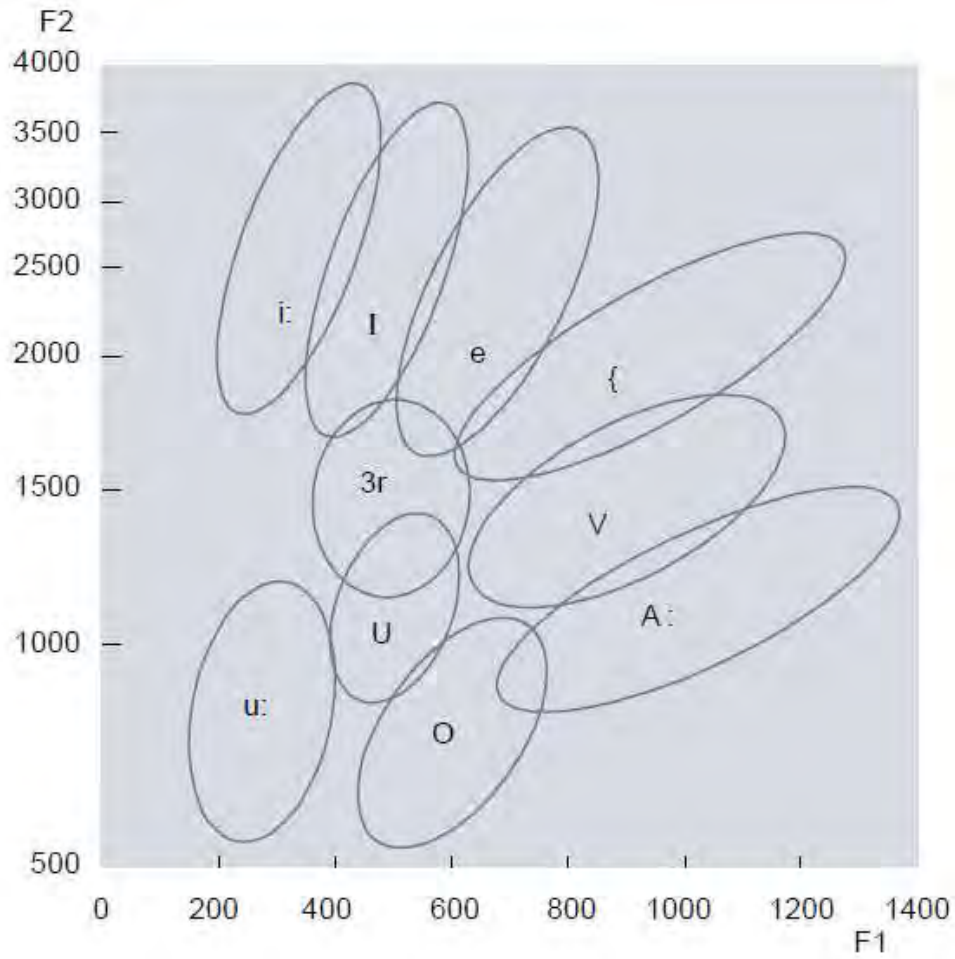

Figure 1.7: The Vowel Triangle Showing First Formant F1 on x axis and F2 on y axis

The information that humans require to distinguish between vowels can be represented purely quantitatively by the frequency content of the vowel sounds. In speech, these are the characteristic partials that identify vowels to the listener. The positions of the formant are the most significant factor in terms of human identification of speech sounds. Vowels, to a large extent can be identified on the basis of the position of the two lowest formant, F1 and F2 (Fig. 1.7). However, the distribution of F1/F2 values for the different vowels has overlapping regions where the formant information is insufficient for making unambiguous decisions on vowel identity.

1.1.3 Problems in Formant Estimation

Formants can be estimated both in time and frequency domain. Among different formant estimation techniques, correlation based methods, such as different variants of lin-

ear predictive coding (LPC) methods, are most commonly used. However, under a noisy condition , the estimation performance of the LPC based formant estimation methods deteriorate significantly. The effect of noise and variation of speech are two major problems in formant estimation. Due to noise speech is affected both in time and frequency domain. Thus in time domain parameters necessary for formant estimation becomes erroneous. In spectral domain, due to noise many spurious peaks other than the formant peaks are present which makes the task of estimating the formant peaks difficult.

The effect of fundamental frequency or pitch can introduce significant errors in formant estimation. These effects can be generally of three types. They are rapid variations of pitch, vary high pitch and very low pitch. Due to the effect of pitch formant estimation performance can be greatly reduced.

1.2 Major Areas of Applications

Speech processing is a broad and established field with a long history. All through these years several branches of speech processing has been emerged. Speech and Speaker recognition systems are one of the earliest of these branches. An important application of speaker recognition technology is in forensics. Speech synthesis is another interesting field which has many different applications. Speech to Text and Text to Speech applications are also very common.

Speech recognition, which is amongst the most popular applications of speech, is implemented in front-end or back-end of the medical documentation process in the health care domain.

The application of speech recognition systems in training air traffic controllers (ATC) is another excellent domain. Automatic Speech Recognition (ASR) in the field of telephony is now common and in the field of computer gaming and simulation it is becoming widespread. Speech is used mostly as a part of User Interface (UI), for creating predefined or custom speech commands. Speech is also used in person authentication, voice operated control etc. Speech is extensively used in speech to text and text to speech

applications.

In spite of extensive research efforts over nearly half a century and numerous applications that are seen everyday, the task of making truly speech enabled machines remains an elusive goal. Although the technology has come a long way since the first efforts at producing electronic machinery for this purpose, the state-of-the-art systems are still at an early evolutionary stage. For human beings learning to speak and comprehend a language though generally takes quite some time to master, doesn't require too much hardship while communicating once it is learnt. For example, once learned, the human capacity of recognizing speech even in noisy environments is quite amazing. Near perfect speech recognition in environments where the noise level exceeds the speech level is something most people do without too much effort. On the other hand an automatic speech recognizer will be rendered nearly useless if the SNR drops below 10–15 dB. Changes in speech patterns, topic shifts etc. are also handled without problems by humans while the machines have huge difficulties with situations that deviate from a well-defined and well known setting. Thus speech research which had a vivid past until now, has a bright albeit challenging future ahead with 'miles to go before' it can reach the destination.

1.3 Automatic Vowel Recognition

In statistical automatic vowel recognition, the human speech is represented as a stochastic process, for which an acoustic model is used to approximate the acoustic aspects (such temporal and spectral patterns) and a language model is used to deal with the linguistic aspects (such as syntax and semantics) of speech. Acoustic models ate often established in feature space, where features are meant to be salient representations of speech signals for the purpose of recognizing the embedded linguistic targets.

An vowel recognition system includes a module of feature extraction in the front end and a module of speech models in the back end. The parameters in speech models are first trained with train data and then used for test data. After an ASR system has been trained and tested, its performance can be evaluated by different performance based

on the objective of the underlying application. In summary, the first step of speech recognition is reading voiced speech and extract features, next templates are made with this features for training and testing and finally an algorithm is used for recognition.

There are mainly two types of voice recognition methods 1. Direct Matching 2. Feature based matching. Direct Matching is a simple method but high computational cost and much memory space is required. It is time consuming as well and not suitable in some practical cases like real time applications. Feature based method is more suitable for practical uses. It does not use all the values of a voice signal. Rather, it extracts some features from the speech signal. This approach requires less computational complexity, less memory and less time. Features can be found by time domain analysis, frequency domain analysis or time-frequency domain analysis. LPC based feature extraction methods require time domain analysis while spectral peak peaking and power spectral density based methods use frequency domain analysis. Cepstrum [2] and wavelets [3] are examples of time-frequency domain analysis methods.

Vowel recognition in particular means classification of different vowel groups based on some features. Male Frequency Cepstral Coefficients(MFCC) are most commonly used as features in vowel recognition. But in noisy conditions MFCC s can not be estimated properly and thus it affects the vowel recognition accuracy. Formants can be used as features in conjunction with MFCCs to increase the vowel recognition accuracy. For these combined features if noise free MFCCs are taken for both training and testing then vowel recognition accuracy mostly depends on the formant estimation performance.

1.4 Literature Review

In recent years, there has been an increasing demand for the development of accurate, efficient and compact representations of speech production systems. Such representations require the extraction of the characteristics of a vocal-tract system from speech signals. Thus, vocal-tract system identification(VTSI) received potential applications in many areas of speech processing, such as speech analysis/synthesis [4],[5],[6] speech coding [7],

speech recognition [8],[9], [10] , acoustic phonetics [11],[12], modeling of speech production process[13]. Formants are one of the most important features in speech signals, and is used in almost all the speech applications. Previous formant extraction methods can largely be classified into spectral peak picking, root extraction, and analysis by synthesis [14], [15]. The spectral peak picking methods and their variants have been widely used for a long time because of low computational complexity, but they often seriously suffer from the peak merger problems [14], [15], where two adjoining formants are identified into a single one. The spectral peak picking method and its variants have been widely used for formant extraction . In most cases, instead of the short-term spectrum itself, smoothed spectra, such as linear prediction (LP) spectrum or cepstrally smoothed spectrum are often employed. However, LP spectra are more often used for this purpose, since they show conspicuous peaks. In The root extraction methods try to find out all the locations of roots by solving a prediction error polynomial obtained from linear prediction coefficients (LPC), which obviously requires much computation [16]. An efficient method for evaluating the pole locations by iteratively computing the number of poles in a sector in the z plane has been reported in [14]. However, the accuracy of the root extraction methods can hardly be high because it is not always clear to determine whether a root obtained forms a formant or just shapes the spectrum [16].

Estimating the formants accurately becomes a much difficult task, especially in the presence of severe background noise. Among different formant estimation techniques, linear predictive coding (LPC) based methods are most commonly used, which offer a little noise immunity [17]. In LPC based methods, from the autocorelation function (ACF) of the given speech utterance, Yule-Walker equations are constructed and from their solutions formants can be obtained. Spectral peak picking is another extremely popular method of formant estimation, where both parametric and non parametric spectral estimation techniques are used[18]. As far as real-life applications are concerned, development of a formant estimation method that performs well in noise is essential but very challenging. Most of the formant frequency estimation methods so far reported are capable of handling only nise free environments [19],[20],[21]. Even some of the recent

methods on formant estimation reported results only in case of noise free environments [22]. In order to overcome the noise effect, an adaptive filter-bank method is proposed in [23], where the formants are extracted from different spectral bands of pre-processed noisy speech. In [24], modified Yule walker equations are applied on once repeated autocorrelation function for the purpose of system identification in the noisy environments. Utilizing the advantageous properties of conventional autocorrelation and cepstrum [25] in handling noisy environment, ramp-cepstrum model based autoregressive system identification methods are proposed in [26]. However, the formant estimation accuracy of these methods has not been investigated on state-of the art databases of continuous speech signals, especially under severe noisy conditions.

It is to be mentioned that in speech recognition applications, formants and most commonly the mel-frequency cepstral coefficients (MFCC) can be employed as features to deal with noise-free conditions [15]. MFCC feature derived from a perceptually warped spectrum are the most widely used feature for speech recognition. In [27], formant frequencies in combination with MFCC features are employed in classification with the condition that classification incorporates a confidence measure in each formant frequency estimation. In [28] the techniques described in [27] was extended allowing the system to choose the most appropriate parametrization depending on speech sound type which was being hypothesized. Other approaches attempt to model spectral peaks rather that looking the resonances in the signal.

However, the performance of the MFCC based recognition system drastically degrades in the presence of noise [29]. It would be interesting to analyze the performance of a vowel recognition system in noise utilizing noise robust formant estimates. In [9] it was shown that combining formants with MFCCs can give can offer good vowel recognition accuracy. Thus, development of a noise robust formant estimation algorithm for obtaining better vowel recognition accuracy in noisy environment would be a challenging task.

1.5 Objective of the Thesis

The objectives of the thesis are:

- To derive a spectral domain ramp cepstrum model of autocorrelation function of band limited speech.
- To derive a spectral domain ramp cepstrum model of once repeated autocorrelation function of band limited speech.
- To design an effective spectral domain reisdue based model fitting algorithm to obtain formant estimates in noise.
- To develop a noise robust automatic vowel recognition scheme incorporating the estimated formants.
- To investigate the performance of the proposed formant estimation as well as vowel recognition schemes in different standard real life speech databases under various noisy conditions.

1.6 Organization of the thesis

The major objective of this thesis is to develop noise robust formant estimation methods that are able to provide better formant estimation performance than the available methods. Another objective is to apply these estimated formants in vowel recognition to improve the recognition accuracy in highly noisy conditions. Spectral domain models based on autocorrelation and cepstrum have been proposed for achieving robustness to noise and pitch variation. Using these models formants are computed from a spectral domain model fitting approach. In this thesis the estimated formant frequencies are chosen as features along with conventional features like MFCC for achieving better recognition accuracy even at a very low SNR. In conjunction to commonly used features for recognition such as MFCC, which are badly corrupted by noise, relying on the performance of the proposed estimation methods, estimated formants, which are less effected by the

noise, are selected as features for vowel recognition. An LDA based classifier is used for classification.

In Chapter 2 of this thesis, firstly, a brief description of vocal tract modeling and the use of autocorrelation and cepstrum for finding better formant estimates is described. Then a spectral domain model is proposed and later it is employed in finding formants. The performance of the proposed formant estimation method in fomant based vowel recognition is also shown in comparison to some other existing methods. In Chapter 3, the effect of repeated autocorrelation both in time and frequency domain are explored and using repeated autocorrelation and cepstral domain analysis a spectral domian formant estimation scheme is formulated which gives better formant estimation accuracy in comparison to some other existing methods. In Chapter 4, banding of the speech signal is explored for the purpose of deriving a better model for formant estimation in comparison to the previous full band models. The effect of using repeated autocorrelation and cepstrum on the banded signal is explored and eventually a noise robust formant estimation is devised. The formant estimation accuracy of the proposed band limited method along with the vowel recognition performance of the estimated formants is compared with some available methods.

Finally, in Chapter 5, some concluding remarks regarding the contributions of the thesis and some future works are presented.

Chapter 2

Spectral Domain Ramp Cepstrum Model of Autocorrelation Function

The objective of this chapter is to develop a formant estimation scheme which can efficiently tackle the adverse effect of observation noise and provide an accurate estimate of formant frequencies of speech signals. Considering the overall human vocal-tract as an all-pole system, we propose a band limited spectral domain ramp cepstrum (SDRC) model for a single sided autocorrelation function (SSACF) of speech signals. The parameters of the proposed SDRC model provide a direct relationship with formant frequencies. A band limited model fitting based approach is introduced for formant estimation which gives better results even in the presence of severe noise. The estimated formants are used in vowel recognition scheme as potential features. The linear discriminant based algorithm is used for the purpose of recognition. Extensive experimentation is carried out considering different male and female vowel utterances from standard speech database under different noisy conditions. It is found that the proposed methods provide a high degree of formant estimation accuracy in comparison to that obtained by some state of the art methods, especially at very low levels of SNR.

2.1 Methodology

In this section, firstly vocal tract modeling for formant estimation purpose is described. Next the problem of formant estimation in noise is presented. To overcome this problem a spectral domain model based on a cepstral domain representation of single sided ACF (SSACF) of speech is introduced. Finally a model matching scheme is proposed to estimate formants in noise.

2.1.1 Background

Human vocal tract can be assumed to be a causal, stable, linear time-invariant and stationary autoregressive (AR) system, and thus a voiced speech signal constructed from it can be characterized as

$$x(n) = -\sum_{k=1}^{P} a_k x(n-k) + Gu(n), \tag{2.1}$$

where $\{a_k\}$ are the system parameters, G denotes the gain factor, P is the known system order and $u(n)$ represents the excitation to the system. The AR system transfer function $H(z)$, which in this case is the vocal tract transfer function can be expressed as

$$H(z) = \frac{G}{1 + \sum_{k=1}^{P} a_k z^{-k}} = \frac{G}{\prod_{k=1}^{P}(1 - p_k z^{-1})}, \tag{2.2}$$

where $p_k = r_k e^{j\omega_k}$ denotes the k-th pole of the AR system with magnitude r_k and angle ω_k. Formants are associated with the free resonances of the vocal tract system. In order to model each formant, a pair of complex conjugate poles is required. In (2.2), each formant corresponds to p_k and it's conjugate. Thus, for a vocal tract system modelled with P-th order AR system, there exists $P/2$ formants. Formant frequency (F_k) and bandwidth (B_k) can be expressed in terms of pole parameters as [30]

$$F_k = \frac{F_s}{2\pi}\omega_k; B_k = -\frac{F_s}{\pi} ln(r_k), \tag{2.3}$$

where F_s is the sampling frequency.

In the LPC based methods, the ACF of the given speech signal $x(n)$ is used in the Yule-Walker equations to obtain the AR parameters and thereby the poles of the vocal tract system and from the estimated poles, formants are calculated [1],[15],[29]. But in the presence of observation noise, LPC based methods fail to provide an accurate estimate of the AR parameters and thus exhibit poor formant estimation accuracy. Moreover, the effect of pitch variation may cause significant errors in the LPC based formant estimation [31]. Hence, development of a formant estimation scheme, which can estimate the formants with a higher accuracy even in the presence of severe background noise as well as handle the effect of pitch variation is in great demand.

2.1.2 Cepstral Domain Analysis

In speech analysis, in order to reduce the effect of pitch from the speech signal, cepstrum that offers the advantage of homomorphic de-convolution has been most commonly used. The principle of homomorphic de-convolution helps in separating signals that have been combined via convolution and thus it becomes a very important tool in different speech processing applications, such as speech recognition. The complex cepstrum of a signal $h(n)$ is defined as [24],[32]

$$c_{hc}(n) = F^{-1}\left\{ln(H(e^{j\omega}))\right\}, \tag{2.4}$$

where $F^{-1}\{.\}$ denotes the inverse Fourier transform and the spectrum of $h(n)$ is given by $H(e^{jw})$. Here an additional c along with h in the subscritpt is introduced just to indicate the type of the cepstrum. Considering the vocal tract (VT) system as minimum phase, $c_{hc}(n)$ is a sequence that is real and causal. Now, according to (2.2), $ln[H(e^{jw})]$ of (2.4) can be expanded as

$$\begin{aligned} ln[H(e^{j\omega})] &= -\sum_{i=1}^{P} ln(1 - p_i e^{-j\omega}) + lnG \\ &= \sum_{i=1}^{P}\sum_{n=1}^{\infty} \frac{p_i^n}{n} e^{-j\omega n} + lnG. \end{aligned} \tag{2.5}$$

In (2.5), constant term lnG exhibits only at the origin and has no effect for $n > 0$. Thus, from (2.4) and (2.5), $c_{hc}(n)$ can be expressed in terms of poles as

$$c_{hc}(n) = \sum_{i=1}^{p} \frac{p_i^n}{n},\ n > 0. \tag{2.6}$$

On the other hand, the real cepstrum of $h(n)$ is defined as

$$c_h(n) = F^{-1}\left\{ln(|H(e^{j\omega})|)\right\}. \tag{2.7}$$

In order to avoid notational complexity, instead of denoting real cepstrum as $c_{hr}(n)$, simply $c_h(n)$ is used, i.e. an additional subscript 'r' is not used hereafter. For $n > 0$, the relation between complex and real cepstra is given by $c_h(n) = 0.5c_{hc}(n)$. Hereafter, for simplicity, only real cepstrum shall be considered. In order to avoid logarithm of negative values, in practical applications real cepstra is most commonly used. Here, for real cepstrum (2.6) can be written as

$$c_h(n) = 0.5\sum_{i=1}^{P} \frac{p_i^n}{n}, n > 0. \tag{2.8}$$

The speech signal $x(n)$ given by (2.1) can be considered as a convolution sum between $h(n)$, the impulse response of the V.T. system and $u(n)$, the excitation to the V.T. system as follows

$$x(n) = h(n) * u(n). \tag{2.9}$$

Cepstral representation of $x(n)$ can be written as

$$\begin{aligned} c_x(n) &= F^{-1}\left\{ln(|X(e^{j\omega})|)\right\} \\ &= F^{-1}\left\{ln(|H(e^{j\omega})|) + ln(|U(e^{j\omega})|)\right\} \\ &= c_h(n) + c_u(n). \end{aligned} \tag{2.10}$$

Here $c_u(n)$ is the cepstrum of the excitation $u(n)$ and $U(e^{j\omega})$ is the frequency domain representations of $u(n)$. The periodic impulse-train excitation is commonly considered to model the voiced sounds. A periodic impulse-train excitation $\{u(n)\}_{n=0}^{N-1}$ with period T can be expressed as

$$u(n) = \sum_{k=0}^{\lambda-1} \delta(n - kT), \lambda = \lceil N/T \rceil, \tag{2.11}$$

where λ is the total number of impulses within the excitation. Based on (2.10), utilizing the advantage of homomorphic de-convolution, cepstral domain system identification methods have been proposed, which deal with the noise free environment [33].

In Fig. 2.1, poles of an AR(6) system excited by a periodic impulse train excitation with a period of $T = 200$ samples is presented in z domain. Here the gain G is chosen as unity. The poles are located at $0.9883e^{j0.161}$, $0.9806e^{j0.536}$ and $0.9767e^{j1.068}$. The angles corresponding to these poles are $\pm 9.2250°$, $\pm 30.7125°$ and $\pm 60.9525°$. These pole positions are chosen in such a manner that they follow the approximate pole locations as found from the spectrum of an utterance /eh/ taken from the TIMIT natural speech corpus. Spectrum of the chosen natural TIMIT speech sequence is shown in Fig. 2.2. The spectrum of a signal constructed from the system introduced in Fig.2.1 using the unit impulse train excitation given by (2.11) as input with T=200 sec is shown in Fig. 2.3. The periodic spike in the frequency spectrum is observed because of the effect of pitch. In this figure the magnitude response of the AR(6) system is also shown, which appears as the envelope of the spiky spectrum. It is observed grom the figure that both spectral peaks match very closely as expected.

It is apparent from (2.8) and (2.10) that in order to extract the system poles p_i from

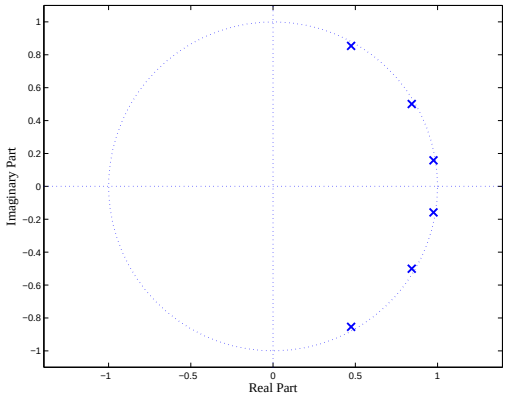

Figure 2.1: Pole plot of an AR(6) system having three pairs of complex conjugate poles

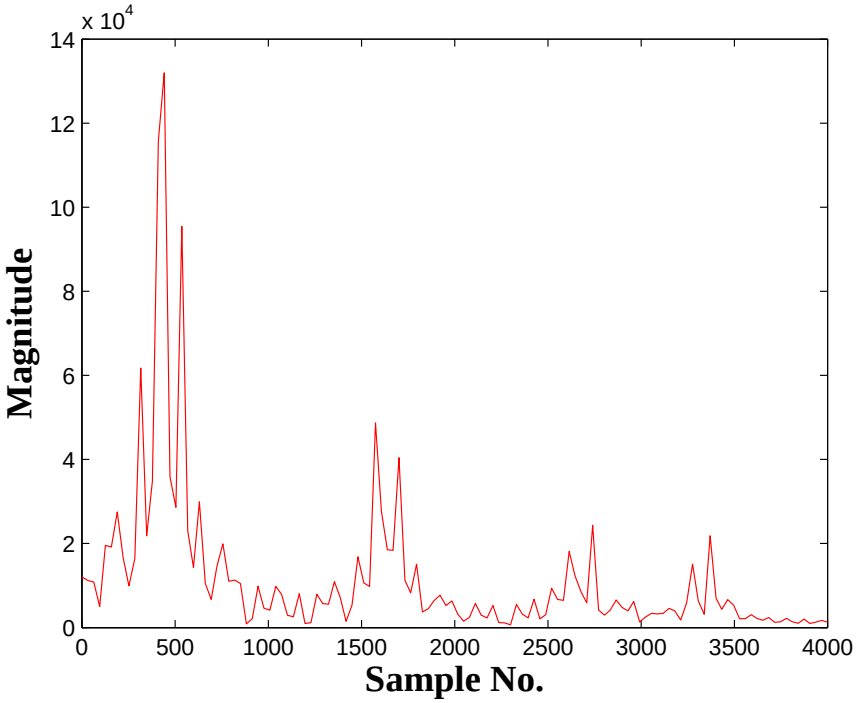

Figure 2.2: Smoothed normalized magnitude spectrum of a frame of natural voiced speech /eh/ taken from the TIMIT database

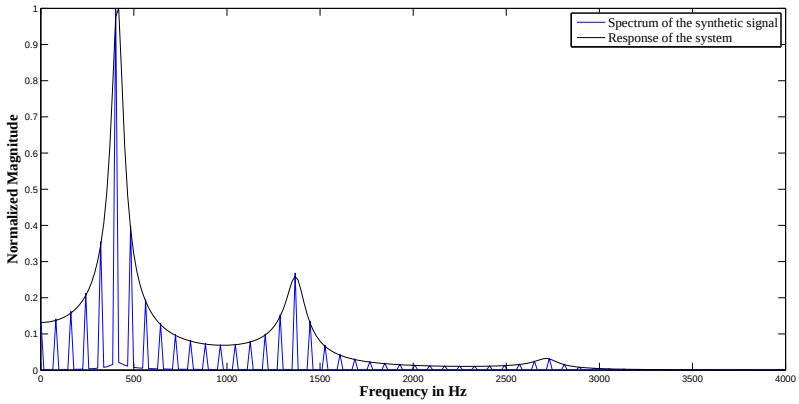

Figure 2.3: Magnitude spectrum of the signal constructed from the AR(6) system shown in Fig. 2.1 and magnitude response of the AR(6) system.

a given $c_x(n)$, one needs to extract $c_h(n)$. In case of periodic impulse train excitation, it can be shown that $c_u(n)$ contributes to $c_x(n)$ at the origin and periodically after each T interval. Thus, even in case of finite data analysis, within the range $0 < n < T$, the effect of $c_u(n)$ can be neglected, resulting

$$c_x(n) = c_h(n) = \sum_{i=1}^{P} \frac{p_i^n}{n}, 0 < n < T. \tag{2.12}$$

It indicates that the cepstral coefficients corresponding to $c_x(n)$ can be considered similar to that of $c_h(n)$ within the range $0 < n < T$.

In the presence of additive white Gaussian noise (AWGN) $v(n)$, the observed signal $y(n)$ can be written as

$$y(n) = x(n) + v(n), \tag{2.13}$$

where $v(n)$ is assumed to be zero mean stationary and independent of $u(n)$. The real cepstrum of $y(n)$ can then be expressed as

$$\begin{aligned} c_y(n) &= F^{-1}\{ln(|X(e^{j\omega})|)\} + F^{-1}\{ln(1 + \frac{|V(e^{j\omega})|}{|X(e^{j\omega})|})\} \\ &= c_x(n) + c_w(n). \end{aligned} \tag{2.14}$$

Here $c_w(n)$ appears in the presence of noise and vanishes in its absence.

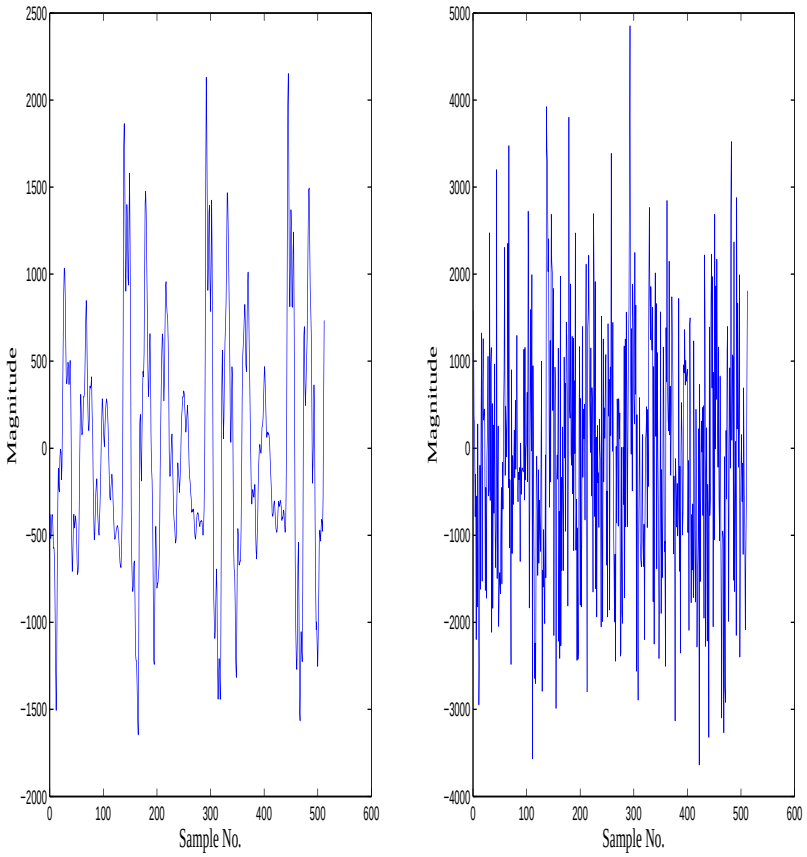

Figure 2.4: Time domain waveform of an utterance of /eh/ (a) without the background noise and (b) with -5dB background noise

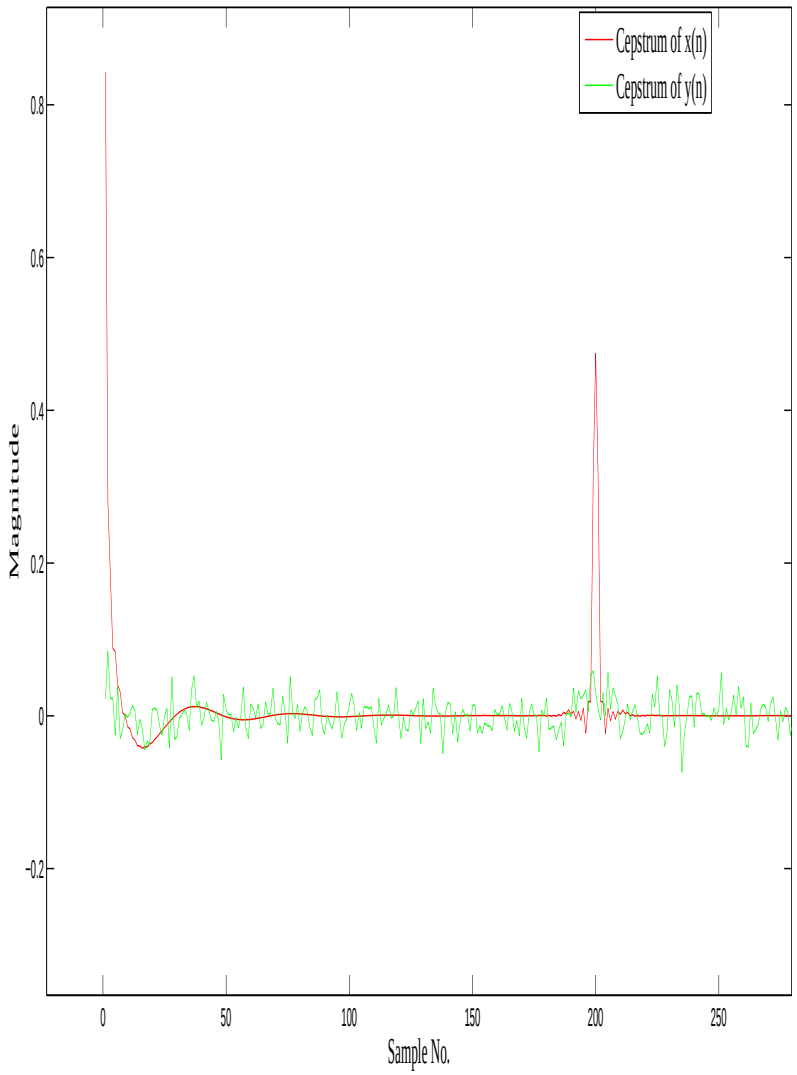

Figure 2.5: Cepstrum of $x(n)$ and $y(n)$

In Fig. 2.4, time domain waveform of an utterance of /*eh*/ is shown both in noise free and noisy conditions. The background noise added here is AWGN and the signal to noise ratio (SNR) is -5 dB. It is clearly observed that the effect of noise completely destroys the noise free signal pattern. In Fig. 2.5, comparison of $c_x(n)$ and $c_y(n)$ at SNR = -5 dB is shown. From this figure it is evident that at all samples $c_y(n)$ is significantly different from $c_x(n)$ because of $c_w(n)$. At severe noise it is very difficult to get an accurate estimate of $c_x(n)$ from $c_y(n)$, since the cepstrum decomposition techniques are very sensitive to the noise level. As a result, it is desirable to develop an algorithm that can reduce the effect of noise on the signal, thereby reducing the effect of $c_w(n)$ on $c_y(n)$ and producing more noise robust cepstral coefficients. In this regard, we propose to investigate the effect of increasing the number of poles on the formant location to enhance the strength of the formant peaks, which will be presented in the next subsection.

2.1.3 Peak Enhancement and Spectral Domain Transformation

In view of enhancing the spectral peaks corresponding to a particular frequency, one possible approach would be to introduce new poles having that frequency. In particular, if the new poles can be generated exactly at the same location of those original poles, the spectral peak corresponding to that pole location will be significantly enhanced. As only the speech signal is available at hand and one cannot change the vocal tract transfer function, it is not possible to place poles at designated places to enhance spectral peaks. As an alternate, if a signal is convolved with its folded version new poles would be introduced, which are related to the original system poles. An equivalent approach of achieving this effect is carrying out the autocorrelation operation on the signal. The ACF of $x(n)$ is defined as

$$\begin{aligned} r_{xx}(m) &= x(n) * x(-n) \\ &= E[x(n)x(n+m)], \end{aligned} \tag{2.15}$$

here $E[.]$ denotes the expectation operator. In practical applications, the ACF of $x(n)$ is computed by using the working formula given below

$$r_{xx}(n) = \frac{1}{N} \sum_{k=0}^{N-1-|n|} x(k)x(k+|n|), n = 0, 1, 2, ..., M-1, \tag{2.16}$$

where M denotes the number of lags to be considered.According to (2.15), in frequency domain $r_{xx}(n)$ can be expressed as

$$\begin{aligned} R_{xx}(e^{j\omega}) &= X(e^{j\omega}) \times X(e^{-j\omega}) \\ &= H(e^{j\omega}) \times U(e^{j\omega}) \times H(e^{-j\omega}) \times U(e^{-j\omega}) \\ &= R_{hh}(e^{j\omega}) \times R_{uu}(e^{j\omega}), \end{aligned} \tag{2.17}$$

where $R_{hh}(e^{j\omega})$ and $R_{uu}(e^{j\omega})$ are the frequency domain representations of $r_{hh}(n)$ and $r_{uu}(n)$, the ACFs corresponding to $h(n)$ and $u(n)$, respectively. According to the definition (2.15), $R_{hh}(e^{j\omega})$ can be written as

$$R_{hh}(e^{j\omega}) = H(e^{j\omega}) \times H(e^{-j\omega}). \tag{2.18}$$

Using (2.2), interms of poles $R_{hh}(e^{j\omega})$ can be expressed as

$$R_{hh}(e^{j\omega}) = \frac{C_1}{\prod_{i=1}^{P}(1 - p_i e^{-j\omega})(1 - p_i^* e^{j\omega})}. \tag{2.19}$$

Here for each pole $p_i = r_i e^{j\theta}$, there exists a pole $1/p_i^*$ which is placed at conjugate reciprocal locations. From (2.19) it is clearly seen that total number of poles in $R_{hh}(e^{j\omega})$ is $2P$, which is twice as the number of poles in $H(e^{j\omega})$. Due to the autocorrelation operation new P poles are introduced in $R_{hh}(e^{j\omega})$ which are conjugate reciprocal to the original P poles of $H(e^{j\omega})$, i.e. the new poles are located at the original pole angles as expected.

Pole plots of an all pole system having three pole pairs is shown in Fig. 2.6. In order to demonstrate the effect of autocorrelation operation on system poles, in Fig. 2.7 another all pole system is shown having all three pole pairs of the system considered in

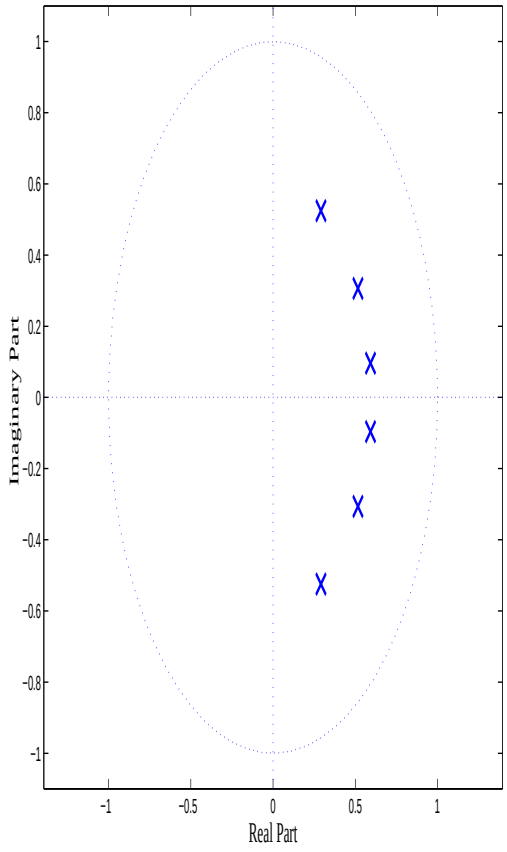

Figure 2.6: An all pole system consisting of three pole pairs

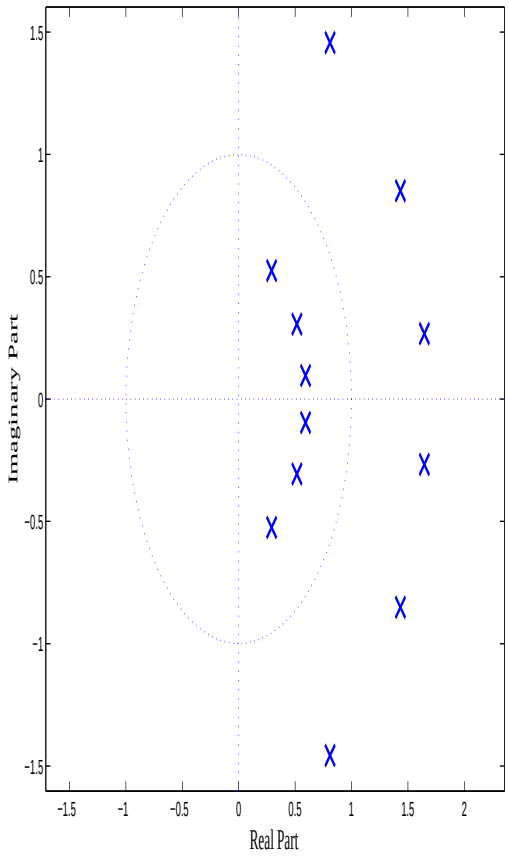

Figure 2.7: An all pole system having original poles of the system shown in Fig. 2.6 along with their conjugate reciprocal poles.

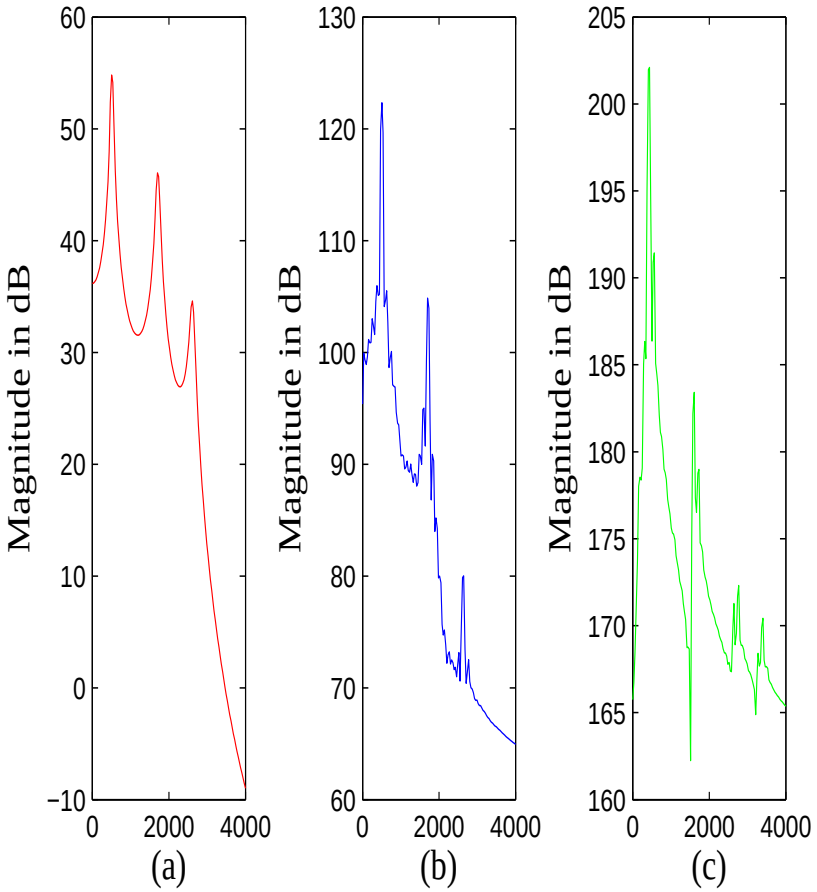

Figure 2.8: Magnitude Response of (a) h(n), (b) ACF of the synthetic speech signal presented in 2.3 and (c) ACF of the TIMIT signal presented in 2.2

Fig. 2.6 along with their complex conjugate poles. From the figure it is seen that at each angular position of the original poles, one new pole is generated outside the unit circle. Obviously, with the increase in number of poles at a particular angular position, the spectral energy corresponding to that particular frequency will be significantly increased. Especially in the presence of noise this can help in finding out the formant peaks in spite of the presence of several unwanted noise peaks. In order to present the effect of spectral peak strengthening, in Fig. 2.8 spectra corresponding to $h(n)$,$r_{synx}(n)$, $r_{xx}(n)$ are shown. It is to be mentioned that the synthetic speech considered here to calculate the ACF $r_{synx}(n)$ is the one that is used in Fig. 2.3. The same natural sound /eh/ as shown in Fig. 2.2 is used here to obtain the spectrum of $r_{xx}(n)$ in Fig. 2.8(c). From the figure it can be easily seen that due to autocorrleation the magnitude of the formant peaks are enhanced.

At this point, it can be shown that to deal only with the causal part of the signal if single sided ACF (SSACF) is considered, the dominant peaks would become more distinct [34]. Since our objective is to handle the severe noisy condition, the use of SSACF would be a better choice. The SSACF of $x(n)$ namely $r^{+}_{xx}(m)$, can be obtained from the double sided ACF (DSACF) as

$$r^{+}_{xx}(m) = \begin{cases} r_{xx}(m)\,, & m > 0 \\ 0.5 r_{xx}(m)\,, & m = 0 \\ 0\,, & m < 0 \end{cases} \tag{2.20}$$

Since the DSACF is symmetric about the zero lag ($m = 0$), it can be computed using (2.15). The fourier transform of $r^{+}_{xx}(m)$ is a complex spectrum $R^{+}_{xx}(e^{j\omega})$ and its spectral envelope is defined as

$$E(e^{j\omega}) = |R^{+}_{xx}(e^{j\omega})| \tag{2.21}$$

Due to the large dynamic range of speech spectra, the envelope of $R^{+}_{xx}(e^{j\omega})$ strongly enhances the highest power frequency bands with respect to the spectrum of $r_{xx}(m)$,

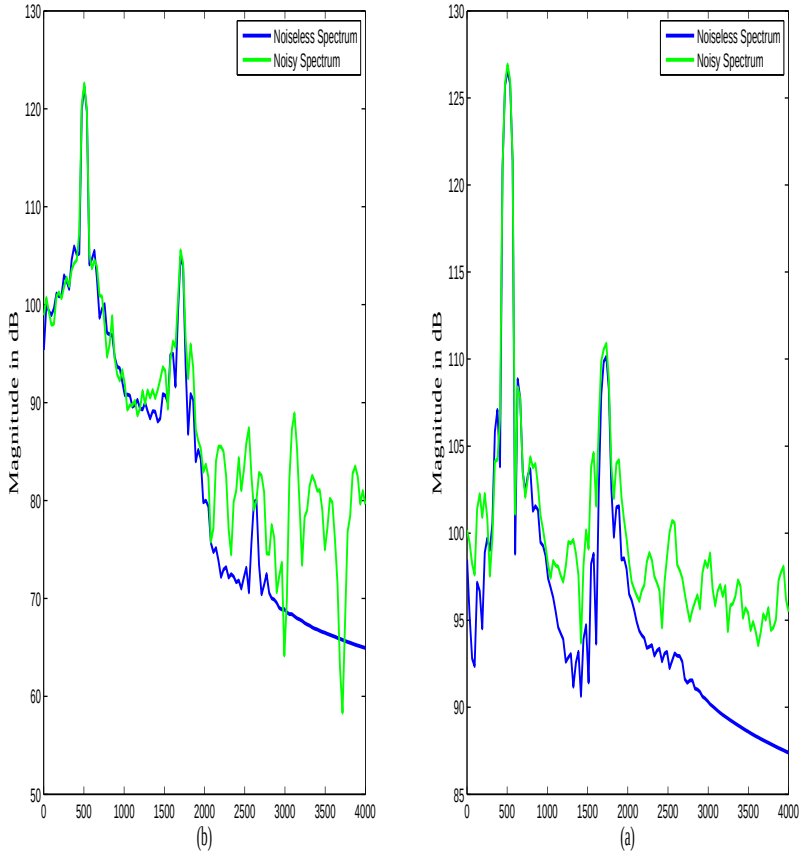

Figure 2.9: Magnitude Response of (a)ACF and (b)SSACF of the synthetic speech signal presented in 2.3 at noiseless condition and noisy condition with SNR=0 dB.

namely $R_{xx}(e^{j\omega})$. Consequently, the noise components lying outside the enhanced frequency bands are largely attenuated in $E(e^{j\omega})$ with respect to $R_{xx}(e^{j\omega})$, and thus use of the envelope of $R^{+}_{xx}(e^{j\omega})$ is more robust to broadband noise than using $R_{xx}(e^{j\omega})$. In addition to the noise robustness, there is another well known advantage, that the SSACF and the original signal $x(n)$ have the same poles[34], [35]. These two properties, i.e. robustness to noise and pole preservation, suggest that AR parameters of the speech signal can be more reliably estimated from the SSACF. The robustness of the SSACF to additive white noise is illustrated in Fig. 2.9. As can be seen from this figure that the envelope of the squared magnitude spectrum of the SSACF shows a prominent first formant, and the whole curve is more robust to additive white noise in comparison to that obtained by using the DSACF. It can be shown that similar to (2.15), the SSACF of $x(n)$, can be expressed as the convolution between $r^{+}_{hh}(m)$ and $r^{+}_{uu}(m)$, which are single sided autocorrelation sequences generated from $h(n)$ and $u(n)$, respectively within the limit $0 \leq m < T$, where T is the time period of the impulse train $u(n)$. This relation is expressed in the following manner

$$r^{+}_{xx}(m) = r^{+}_{hh}(m) * r^{+}_{uu}(m), \;\; 0 \leq m < T \tag{2.22}$$

Here, $r^{+}_{uu}(m)$ is a periodic sequence which has the same priodicity as $u(n)$. From (2.22) it is now obvious that transferring to the cepstral domain can provide the opportunity of source signal separation using the propetry of homomorphic deconvolution. In cepstral domain, (2.22) can be written as

$$c_{r^{+}_{xx}}(m) = c_{r^{+}_{hh}}(m) + c_{r^{+}_{uu}}(m), \tag{2.23}$$

where $c_{r^{+}_{xx}}(m)$, $c_{r^{+}_{hh}}(m)$ and $c_{r^{+}_{uu}}(m)$ are the real cepstra corresponding to $r^{+}_{xx}(m)$, $r^{+}_{hh}(m)$ and $r^{+}_{uu}(m)$, respectively, computed in the same manner as (2.7). In Fig. 2.10, $c_{r^{+}_{xx}}(m)$ and $c_{r^{+}_{hh}}(m)$, computed from the signal as considered in Fig. 2.5 are shown. From this figure it is observed that $c_{r^{+}_{xx}}(m)$ is approximately equal to $c_{r^{+}_{hh}}(m)$ within $0 < m < T$ and (2.23) can be written as

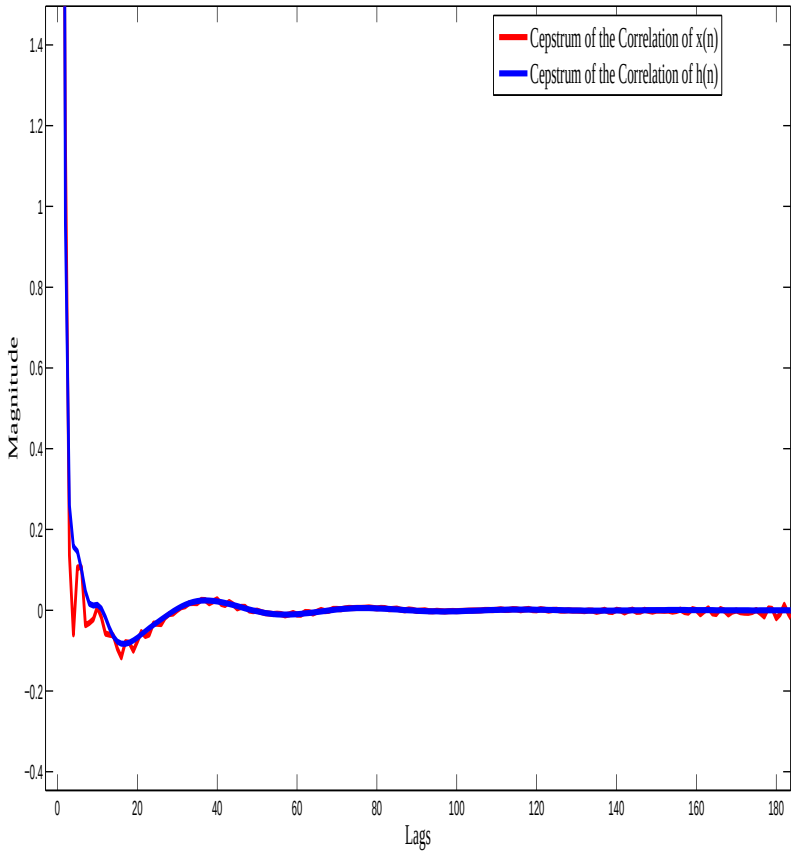

Figure 2.10: Cpestrum of $C_{r_{xx}}(m)$ and $C_{r_{hh}}(m)$

$$c_{r^+_{xx}}(m) \approx c_{r^+_{hh}}(m), 0 < m < T \tag{2.24}$$

Since, $r^+_{xx}(m)$ and the original signal $x(n)$ have the same poles, the fourier transform of $r^+_{hh}(m)$ will have poles inside the unit circle similar to $H(e^{j\omega})$. Hence, (2.10) the complex cepstra corresponding to $r^+_{hh}(m)$ can be repesented as

$$\begin{aligned} c_{r^+_{hh}}(m) &= F^{-1}[ln(R^+_{hh}(e^{j\omega}))] \\ &= F^{-1}[ln(H(e^{j\omega}))] \\ &= 0.5c_h(m), m > 0. \end{aligned} \tag{2.25}$$

Using 2.8 and (2.25), one can write

$$c_{r_{hh}^{+}}(m) = 0.5 \sum_{k=1}^{P} \frac{p_k^m}{m}, m > 0. \tag{2.26}$$

It is now evident that $c_{r_{hh}^{+}}(m)$ is directly related to the system poles and according to (2.24) within the range $0 < m < T$ the same relationship to the system poles holds true for $c_{r_{xx}^{+}}(m)$. Hence, with a view to developing a spectral domain scheme for formant frequency estimation it would be sufficient to consider the detailed analysis of $c_{r_{hh}^{+}}(m)$ instead of $c_{r_{xx}^{+}}(m)$. As is seen from (2.26), cepstrum decays rapidly with m, which makes it difficult to use in estimating the system poles. In order to overcome this problem, an easy-to-handle ramp cepstrum is proposed as,

$$\phi_h(m) = m c_{r_{hh}^{+}}(m) = 0.5 \sum_{k=1}^{P} p_k^m, m > 0. \tag{2.27}$$

According to (2.24) the ramp cepstrum corresponding to $r_{xx}^{+}(m)$, namely $\phi_x(m) = m c_{r_{xx}^{+}}(m)$ can be expressed as

$$\phi_x(m) \approx \phi_h(m) = 0.5 \sum_{k=1}^{P} p_k^m, 0 < m < T. \tag{2.28}$$

Using (2.13) and (2.15), the ACF of noisy speech $y(n)$ can be expressed as

$$r_{yy}(n) = r_{xx}(n) + r_{ww}(n), \tag{2.29}$$

where

$$r_{ww}(n) = r_{vv}(n) + r_{vx}(n) + r_{xv}(n). \tag{2.30}$$

Here, $r_{vv}(n)$ is the ACF of noise $v(n)$ and $r_{vx}(n)$ and $r_{xv}(n)$ are the cross correlation terms. Since $v(n)$ is uncorrelated with $x(n)$, it is expected that the values of the cross-correlation terms, in comparison to that of $r_{xx}(n)$, will be negligible. On the other hand, the ACF of the AWGN $v(n)$ generally exhibits a peak at the zero lag and the values at all other lags should be very small and ideally should be zero.

In Figs. 2.11(a)-2.11(f), different ACFs, namely $r_{xx}(n)$, $r_{yy}(n)$, $r_{ww}(n)$, $r_{vv}(n)$, $r_{xv}(n)$ and $r_{vx}(n)$ are plotted at SNR$= -5$ dB. From Figs. 2.11(e) and 2.11(f), it can be observed

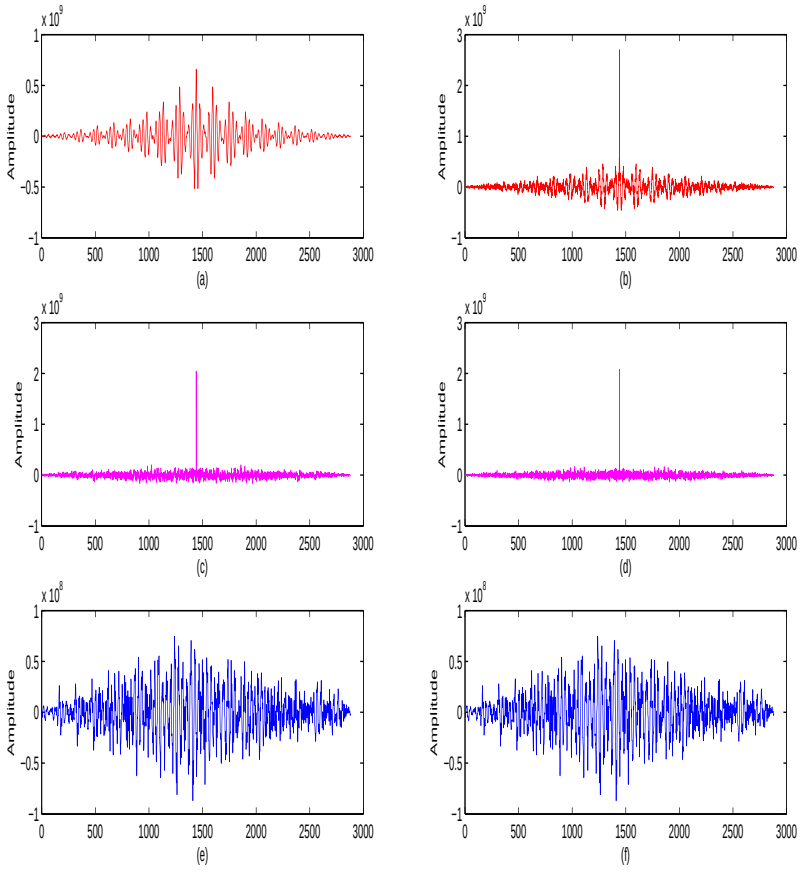

Figure 2.11: Effect of noise in the autocorrelation domain: plot of different autocorrelation functions (a)$r_{xx}(n)$, (b) $r_{yy}(n)$, (c) $r_{ww}(n)$, (d) $r_{vv}(n)$, (e)$r_{xv}(n)$ and (f) $r_{vx}(n)$

that the values of the cross correlation terms are very small as expected. As seen in Fig. 2.11(d), $r_{vv}(n)$ although exhibits very large peak at zero lag, nonzero small values exist at all other lags because of the finite data length. It is also observed in Fig. 2.11(c) that $r_{ww}(n)$ exhibits the maximum value at the zero lag and the values at other lags are comparatively very small. From these figures, it can be concluded that in comparison to the effect of $v(n)$ on $x(n)$ as shown in Fig. 2.4, the effect of $r_{ww}(n)$ on $r_{xx}(n)$ is drastically reduced because of the autocorrelation operation. Now, the cepstral coefficients of $r_{yy}(n)$ can be expressed as

$$\begin{aligned} c_{r_{yy}}(n) &= F^{-1}\{ln|R_{xx}(e^{j\omega})|\} + F^{-1}\{ln(1+\frac{|R_{ww}(e^{j\omega})|}{|R_{xx}(e^{j\omega})|})\} \\ &= c_{r_{xx}}(n) + c_{r_{w1}}(n). \end{aligned} \tag{2.31}$$

Here $c_{r_{xx}}(n)$ is the cepstrum of $r_{xx}(n)$, $c_{r_{w1}}(n)$ is the cepstrum corresponding to the additional term in (2.29) and $R_{xx}(e^{j\omega})$ represents the Fourier transform of $r_{xx}(n)$. In time domain, as described before that the effect of $r_{ww}(n)$ on $r_{xx}(n)$ is smaller in comparison to the effect of $v(n)$ on $x(n)$. For all lags other than the zero lag, the energy ratio of $r_{ww}(n)$ to $r_{xx}(n)$ is much smaller than that of $v(n)$ to $x(n)$. Based on Parseval's theorem, it can be inferred that over the entire range of frequency overall spectral energy of $|R_{ww}(e^{j\omega})|$ is lower in comparison to that of $|R_{xx}(e^{j\omega})|$. Hence, it is expected that the effect of $c_{r_{w1}}(n)$ on $c_{r_{yy}}(n)$ in (2.31) is smaller than the effect of $c_w(n)$ on $c_y(n)$ in (2.14). In this case, (2.31) can be rewritten as

$$c_{r_{yy}}(n) \approx c_{r_{xx}}(n). \tag{2.32}$$

This relation holds true also for the cepstrum computed using the SSACF, which as stated earlier provides more noise robustness. Hence, the cepstrum of the SSACF of the noisy signal can be rewritten as

$$c_{r^+_{yy}}(n) \approx c_{r^+_{xx}}(n). \tag{2.33}$$

Corresponding relationship in ramp cepstral domain as per (2.28) can be written as

$$\phi_y(m) \approx \phi_x(m) = 0.5 \sum_{k=1}^{P} p_k^m, m > 0. \tag{2.34}$$

Here, $\phi_y(m) = mc_{r_{yy}^+}(m)$ is the ramp cepstrum of $r_{yy}^+(m)$. Hence, it is expected that given noisy speech, if ramp cepstrum of its single sided correlation sequence is computed, depending on the level of noise, it may exhibit more noise immunity in comparison to time domain analysis.

As in (2.28) it is shown that $\phi_h(m)$ is directly related to system poles, the corresponding frequency domain representation is given by

$$\Phi_h(e^{j\omega}) = \sum_{i=1}^{P} \frac{C_i}{(1 - p_i e^{-j\omega})} = \sum_{i=1}^{P/2} \{\frac{C_i}{(1 - p_i e^{-j\omega})} + \frac{C_i'}{(1 - p_i^* e^{-j\omega})}\}, \tag{2.35}$$

where $\Phi_h(e^{j\omega})$ is the Fourier transform of $\phi_h(m)$ and C_i and C_i' are gain factors and p_i and $p_i^*$are a complex conjugate pole pair. As seen from (2.35) the system corresponding to the SSACF of $h(n)$, namely $R_{hh}^+(e^{j\omega})$, has $P/2$ pairs of complex conjugate poles.

Based on (2.28), (2.34) and (2.35) it is expected that in noisy environment it is advantageous to use the spectrum of $\phi_y(m)$, within $0 < m < T$, which exhibits more noise robustness in comparison to $r_{yy}(m)$ and can be approximated as

$$\Phi_y(e^{j\omega}) \approx \Phi_h(e^{j\omega}), \tag{2.36}$$

where $\Phi_y(e^{j\omega})$ is the Fourier transform of $\phi_y(m)$, which can be computed from $r_{yy}^+(m)$ in the following manner

$$\Phi_y(e^{j\omega}) = F\left[m \times F^{-1}\{ln|F\left[r_{yy}^+(m)\right]|\}\right] \tag{2.37}$$

Here, $F[.]$ denotes Fourier transform.

In Fig. 2.12, for the natural voiced speech /eh/ as shown in Fig. 2.2, a comparison between the noiseless and noisy spectra at SNR = 0 dB is shown. In Fig. 2.13, spectrum of the ACF of the noisy signal presented in Fig. 2.12, is shown and in Fig. 2.14, the

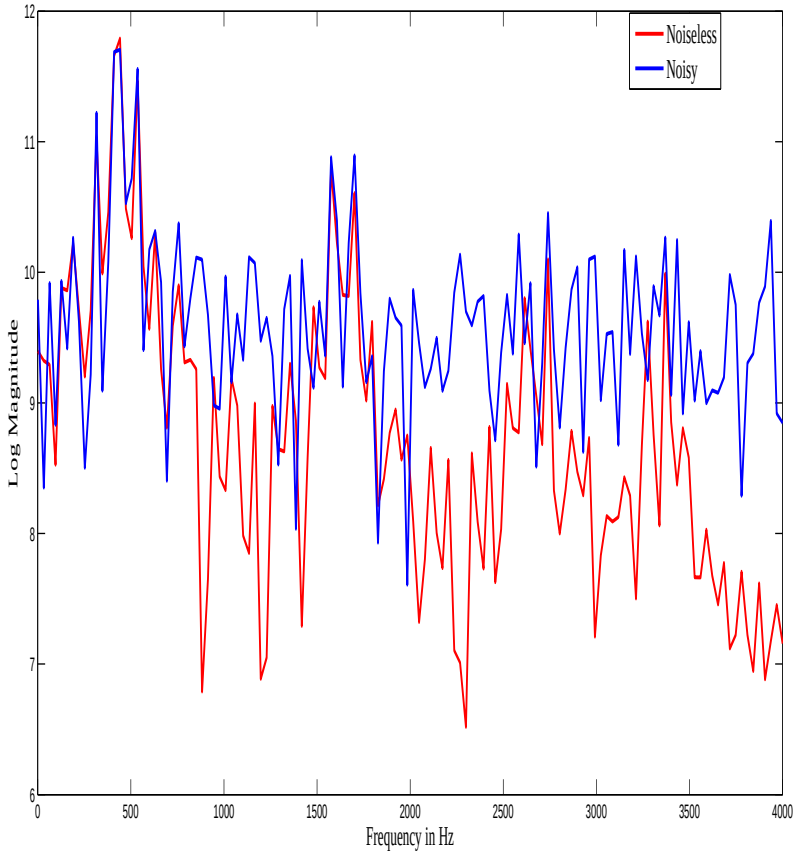

Figure 2.12: Spectrum of the noisy and noiseless signal

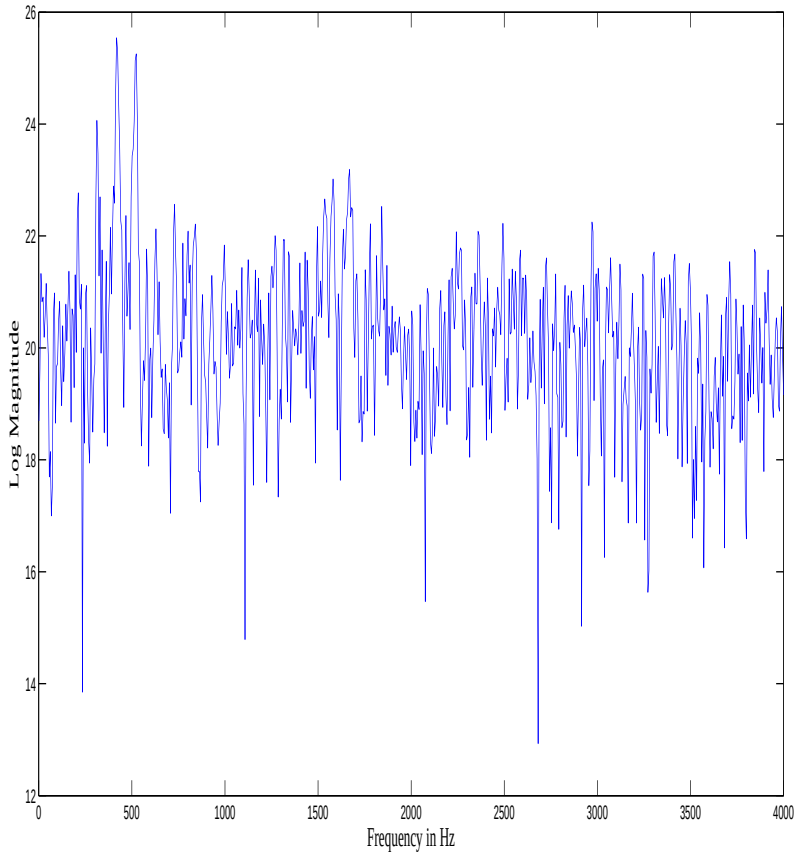

Figure 2.13: Spectrum of the ACF of the noisy signal presented in 2.12

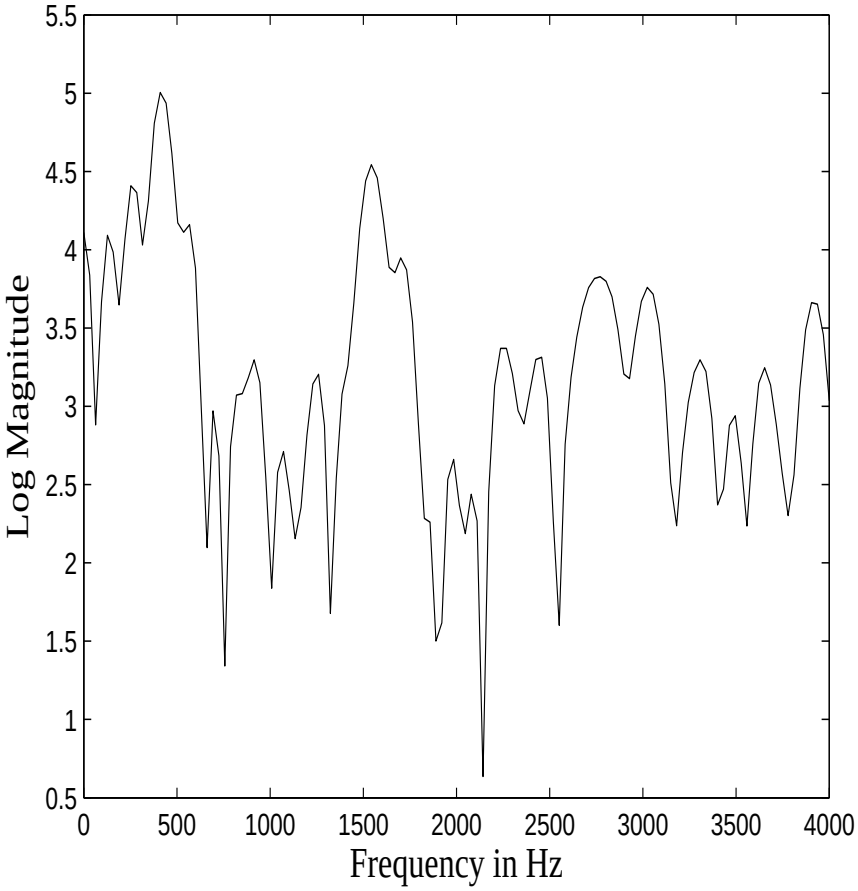

Figure 2.14: Spectrum of the ramp cepstrum of SSACF of the noisy signal

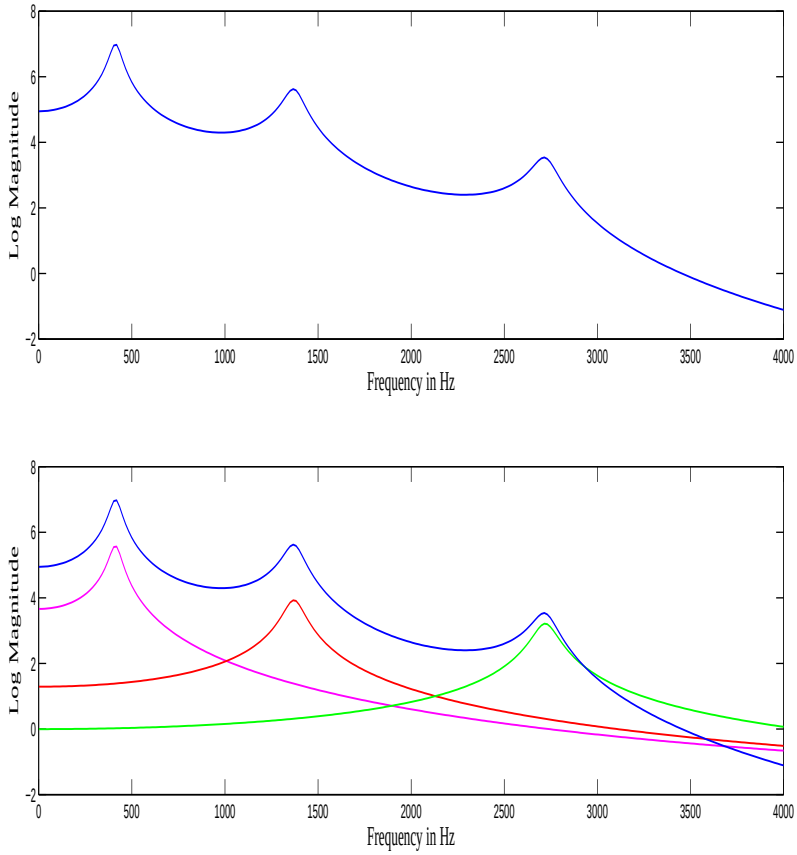

Figure 2.15: Response of the system in Fig. 2.1 and multiplied response of three subsystems each consisting of a pair of complex conjugate poles.

spectrum of the ramp cepstrum of the SSACF of the noisy signal is shown. From these figures it is evident that the spectrum of the ramp cepstrum of the SSACF retains the dominat peaks of the origianl signal and exhibits less spurious peaks than the spectrum of the ACF.

2.1.4 Model Generation and Model matching

As given by (2.2), the transfer function of the VT system if modeled as an AR(P) system , one can consider it as a cascade of $P/2$ blocks where each block consists of a pair of complex conjugate poles. In Fig. 2.15, the magnitude response of an AR(6) system with three complex conjugate poles is shown along with the magnitude responses of the three individual pole pairs. From this figure it is clearly observed that the response of the AR

system exhibits three prominent peaks corresponding to the three formants each of which is related to a particular pole pair. Considering the vocal tract as an AR system, a pair of complex conjugate poles is responsible for generating a dominant peak in the spectral domain. Although the effect of other pole pairs, unless otherwise located at a very close vicinity, may enhance the spectral level, dominance of a particular formant peak is mostly because of the pole pair located in that particular formant frequency. For real life speech applications the first three formants are mostly considered. Thus considering only the first three formants, the cascaded spectrum representation of (2.2) can be written as

$$\begin{aligned} H(e^{j\omega}) &= \frac{C_i}{\prod_{i=1}^{3}(1-p_i e^{-j\omega})(1-p_i^* e^{-j\omega})} \\ &= H_1(e^{j\omega})H_2(e^{j\omega})H_3(e^{j\omega}) \end{aligned} \tag{2.38}$$

Hence, the ramp cepstrum corresponding to the SSACF of (2.38)can be written as follows

$$\phi_h(m) = \phi_{h1}(m) + \phi_{h2}(m) + \phi_{h3}(m) \tag{2.39}$$

According to (2.35) it can be shown that the spectral domain representation of (2.39) is obtained as

$$\begin{aligned} \Phi_h(e^{j\omega}) &= F[\phi_h(m)],\ m > 0 \\ &= \Phi_{h1}(e^{j\omega}) + \Phi_{h2}(e^{j\omega}) + \Phi_{h3}(e^{j\omega}) \end{aligned} \tag{2.40}$$

The first formant peak is prominent in the spectrum of ramp cepstrum of SSACF presented in Fig. 2.14, indicating that the effect of $\Phi_{h2}(e^{j\omega})$ and $\Phi_{h3}(e^{j\omega})$ are negligible on $\Phi_{h1}(e^{j\omega})$. Using this property, it can be assumed that the output response closely match $\Phi_{h1}(e^{j\omega})$ around the first formant peak. Thus instead of conventional peak picking, in this chapter, the task of formant estimation is carried out through spectral model fitting, which ensures that both the frequency and bandwidth of formant peaks are matched.

However, in noisy environments, presence of spurious peaks may cause difficulties in

identification of formant peaks even in the case of band limited signals. As discussed in the previous section, the autocorrelation operation can reduce the effect of noise. Moreover, performing the ramp cepstrum operation on the SSAC sequence will definitely exhibit significant noise reduction. In order to identify the formant peaks, especially under noisy condition, one possibility is to consider a transfer function which can produce an impulse response that closely matches the ramp cepstrum of the SSACF of the most prominent subsystem whose frequency domain representation is $\Phi_{h1}(e^{j\omega})$. By limiting the comparison to only the zone where only the first formant frequency should be present, the spectrum corresponding to that transfer function can then be used in a spectral matching technique along with the spectrum obtained from the ramp cepstrum of SSACF of the noise corrupted signal. In this case, the transfer function of the subsystem responsible for the spectrum of the ramp cepstrum of the SSACF around the first formant peak as per (2.35) can be represented as

$$\Phi_{h1}(e^{j\omega}) = \frac{C_1}{(1-p_1e^{-j\omega})(1-p_1^*e^{-j\omega})}, \tag{2.41}$$

where $C_1 = 1 - Re[p_1]e^{-j\omega}$.

As, in the previous section a direct relationship between $\Phi_y(e^{j\omega})$ and $\Phi_h(e^{j\omega})$ is developed, (2.41) can be used to derive a model for the ramp cepstrum of the SSACF of a noisy sequence for the first formant as follows

$$\Phi^1_{model}(e^{j\omega}) = \frac{C_1}{(1-p_1e^{-j\omega})(1-p_1^*e^{-j\omega})}, \tag{2.42}$$

where $\Phi^1_{model}(e^{j\omega})$ is the representation for the first band.

In the proposed formant estimation method, a spectral model corresponding to the first formant zone of the spectrum of the ramp cepstrum of SSACF of the speech signal is introduced, which is utilized in a model matching technique to find out the model parameters that in turn will provide the first formant frequency. In what follows the proposed approach of model matching will be elaborated in detail where each formant will be esti-

mated once at a time. In the estimation of each formant, one such model corresponding to that specific formant is required. Similar to (2.42) for the first formant, for estimating each formant one such model is required and the i-th model can be represented as

$$\begin{aligned} \Phi^i_{model}(e^{j\omega}) &= \frac{C_i}{(1-p_i e^{-j\omega})(1-p_i^* e^{-j\omega})}, \\ p_i &= r_i e^{j\theta_i},\ p_i^* = r_i e^{-j\theta_i} \end{aligned} \tag{2.43}$$

The spectrum $\Phi^i_y(e^{j\omega})$ of the ramp cepstrum of the SSACF of the observed noisy signal $y(n)$ is used in conjunction with the proposed model $\Phi^i_{model}(e^{j\omega})$ to form an objective function and for the first formant with $i = 1$ based on the absolute difference of these spectra, namely

$$e^i_{min}(r_j, \theta_j) = \underset{\substack{r_l < r_i < r_h \\ \theta_l < \theta_i < \theta_h}}{min} \sum_{\omega=\omega_{lc}}^{\omega_{hc}} \left(|\Phi^i_{model}(e^{j\omega})| - |\Phi^i_y(e^{j\omega})| \right) \tag{2.44}$$

Note that here the superscript i is introduced to control the step by step algorithm. The algorithm for the first formant where $i = 1$, is given below in brief.

1. From given noisy speech $y(n)$ computing $\Phi^i_y(e^{j\omega})$ using (2.37)
2. Generating $\Phi^i_{model}(e^{j\omega})$ using the model of (2.43)
3. Minimizing the objective function in 2.44 within a restricted frequency range ω_{lc} to ω_{hc} which depends on the range of each formant zone.

One may utilize the $-3dB$ points on the lower and higher sides of the peak in the spectrum of the model to extract ω_{lc} and ω_{hc}. Within that specified range $\omega_{lc} \leq \omega \leq \omega_{hc}$, the optimum values of the two variables r_i and θ_i are obtained at the minimum of absolute differences. Based on the fundamental knowledge of traditional range of formants, one may restrict the search range for the two variables i.e., $r_l \leq r \leq r_h$ and $\theta_l \leq \theta \leq \theta_h$ or adopt a coarse and fine search approach [36]. Formant frequencies are estimated from the pole angle θ_j that produces the best match between the spectra using (2.44) .

Once the first formant frequency $F1$ is obtained, (2.43) is utilized to estimate the

second formant frequency $F2$. $\Phi_y^i(e^{j\omega})$ can be written as the sum of $\Phi_y^1(e^{j\omega})$, $\Phi_y^2(e^{j\omega})$ and $\Phi_y^3(e^{j\omega})$ alike (2.40). From the magnitude spectrum of $\Phi_y^i(e^{j\omega})$ the estimated model spectrum $\Phi_{model}^1(e^{j\omega})$ is subtracted such that the resulting spectrum closely resembles the sum of $\Phi_y^2(e^{j\omega})$ and $\Phi_y^3(e^{j\omega})$. Hence $\Phi_y^i(e^{j\omega})$ in general for estimating second and third formant can be expressed as

$$\Phi_y^i(e^{j\omega}) = \Phi_y^{i-1}(e^{j\omega}) - \Phi_{model}^{i-1}(e^{j\omega}),\ i > 1 \tag{2.45}$$

Then similar to the matching in the first formant zone, matching is performed in the second formant zone and $F2$ is estimated. Then from the magnitude spectrum of $\Phi_y^2(e^{j\omega})$ the estimated model spectrum $\Phi_{model}^2(e^{j\omega})$ is subtracted to obtain $\Phi_y^3(e^{j\omega})$. According to the simplified modeling of the vocal tract presented above, $\Phi_y^3(e^{j\omega})$ should closely match with $\Phi_{model}^3(e^{j\omega})$, leading to a similar approach as described in (2.43) and (2.44) to obtain $F3$.

2.1.5 Formant Based Vowel Recognition

After estimating formants in the described manner, in the proposed scheme they are employed in a vowel recognition system as potential features along with the commonly used Mel frequency cepstral coefficients (MFCC).

A typical ASR system includes a module of feature extraction in the front end and a module of speech models in the back end. The parameters in speech models are first trained with trained data and then used for test data. After an ASR system has been trained and tested, its performance can be evaluated by different performance measures based on the objective of the underlying application. In summary, the first step of speech recognition is reading voiced speech and extract features, next templates are made with this features for training and testing and finally an algorithm is used for recognition.

There are mainly two types of voice recognition methods, direct and feature based. Direct Matching is a simple method but high computational cost and much memory space is required. It is time consuming as well and not suitable in some practical cases like real time applications. Feature based method is more suitable for practical uses. It

does not use all the values of a voice signal. Rather, it extracts some features from the speech signal. This approach requires less computational complexity, less memory and less time. Features can be found by time domain analysis, frequency domain analysis or time-frequency domain analysis. LPC based feature extraction methods require time domain analysis while spectral peak peaking and power spectral density based methods use frequency domain analysis. Cepstrum and wavelets are examples of time-frequency domain analysis methods.

The task of vowel recognition in particular means classification of different vowel groups based on some features. Male Scale Cepstral Coefficients(MFCC) are most commonly used as features in vowel recognition. But in noisy conditions MFCC s can not be estimated properly and thus it affects the vowel recognition accuracy. Formants can be used as features in conjunction with MFCCs to increase the vowel recognition accuracy. For these combined features if noise free MFCCs are taken for both training and testing then vowel recognition accuracy mostly depends on the formant estimation performance.

In the proposed scheme for the purpose of recognition two major steps are followed. First given the train data set for differents vowels, formants and MFCC features are extracted. For each vowel a number of samples (tokens) are considered in the training stage. during the testing phase, the similar features are extracted from the test vowels. Utilizing the Linear discriminant analysis (LDA) based classifier, the label of the unknown test vowel is identified. It is to be noted that the use of formants increases the dimension by 3. However, as can be seen from the experimental result, it will offer a huge increase in estimation accuracy.

LDA based discriminants take into account the intra-cluster scatter matrix computed from the training vectors pertaining to each of the classes. For our proposed scheme, a frame by frame classification method is used, which offers vowel recognition results for each voiced frame independently. The classifier classifies the data into different groups generally, depending on the significant characteristics of the group members. The quality of a classifier depends on its ability to provide the compactness among the member within a cluster and the separation between the members of different clusters in terms of

feature characteristics. The task of recognizer is to identify the class label of a test sample utilizing the classified data. In a feature based scheme, classification is performed utilizing the extracted features of the data, instead of directly employing the data themselves. In the proposed method, the LDA is used to classify the vowel among the different classes (in our case, vowel) available. In LDA, the total scatter matrix is a scaled covarience matrix, defined as

$$S = \sum_{i=1}^{N} [x_i - \mu][x_i - \mu]^T \tag{2.46}$$

where μ denotes the global mean of the entire set of the training vector. The between-class scatter matrix is denoted as

$$S_b = N_+[\mu_+ - \mu][\mu_+ - \mu]^T + N_-[\mu_- - \mu][\mu_- - \mu]^T \tag{2.47}$$

Here the three points (μ, μ_+ and μ_-) are collinear, meaning that

$$[\mu_+ - \mu] = \frac{N_-}{N}[\mu_+ - \mu_-] \tag{2.48}$$

and

$$[\mu_- - \mu] = -\frac{N_+}{N}[\mu_+ - \mu_-] \tag{2.49}$$

using the values obtain from (2.48,2.49) in (2.47), the between class scatter matrix is obtained as

$$S_b = \frac{N_- N_+}{N}[\mu_+ - \mu_-][\mu_+ - \mu_-]^T \tag{2.50}$$

in addition, the within class scatter matrix is defined as

$$S_w = \sum [x_i - \mu_+][x_i - \mu_+]^T + \sum [x_i - \mu_-][x_i - \mu_-]^T \tag{2.51}$$

The goal of LDA is to find out the linear projection w_{opt} using these relationships that maximized a special kind of signal to noise ratio. Here the signal is represented by the

projected inter-cluster distance and the noise by the projected intra-cluster variance.The objective function is based on determining a projection direction w to maximize the Fisher's discriminant defined as [37]

$$J(w) = \frac{w^T S_b w}{w^T S_w w} \tag{2.52}$$

2.2 Results and Simulation

For evaluating the formant estimation performance of the proposed method, numerous experiments have been conducted using the voiced speech signals taken from the TIMIT acoustic-phonetic continuous speech corpus, which has jointly been developed by Massachusetts Institute of Technology (MIT), Stanford Research Institute (SRI) and Texas Instruments (TI) [38]. The TIMIT database contains a large collection of sentences uttered by both male and female English speakers using various dialects. A total of 6300 sentences, with 10 sentences spoken by each of the speakers are present in the database. Phonetic description of the database is also available, which helps in identifying voiced and unvoiced frames. However, as the TIMIT database does not contain reference values of formants, in order to compare estimated results, the most commonly used reference formant database for the TIMIT speech corpus is chosen, where formant frequencies are estimated based on vocal tract resonances (VTR) with manual adjustment [39]. The formant estimates reported in [39] are taken as ground truth and the estimation performance of different methods is evaluated at different levels of signal to noise ratios (SNR). The VTR reference database of TIMIT speech corpus contains 376 sentences across the training set, representing 173 speakers. These sentences contain 18 voiced phonemes, out of which, the diphthongs have been ignored, and 11 phonemes are considered. A total of 2726 utterances of phonemes are used from the VTR subset, out of which 1583 are from male and 1143 are from female speakers, have been analysed. In VTR database, formant estimates are reported for every 10 ms interval. However, vowel duration is generally much larger than 10 ms. In the frame by frame formant analysis, when the size of

analysis frame is larger than 10 ms, the estimated formants are then compared with the average VTR formant values obtained over the different 10 ms frames within the duration of that formant under investigation. For the purpose of performance comparison, first the most widely used LPC based formant estimation method [40] is chosen, where the order of the LPC is chosen as 12. It is to be noted that the performance of the most widely used LPC-10 based formant estimation is also investigated and most of the cases it is found that the performance of the LPC-12 method is comparatively better.

Apart from the LPC method, a state of the art adaptive filter bank (AFB) method is also chosen. In the AFB method, formant estimation is carried out in sample by sample basis, and for the purpose of comparison, average estimated formant values over a period is considered [23].

Table 2.1: Performance comparison in terms of mean error(%) for synthetic speech

Vowels		$5dB$			$-5dB$		
		Proposed	LPC	AFB	Proposed	LPC	AFB
/a/	F1	4.45	20.24	46.90	4.97	20.46	49.77
	F2	9.31	65.23	32.58	13.79	113.79	30.99
	F3	2.92	17.80	8.45	3.13	34.02	9.84
/o/	F1	10.37	49.53	128.07	12.27	78.29	18.29
	F2	9.62	138.88	20.42	17.08	133.29	46.61
	F3	1.24	39.93	9.56	2.26	36.28	12.53
/u/	F1	9.60	72.96	109.00	10.96	98.29	12.98
	F2	7.98	116.33	14.62	14.82	121.92	33.72
	F3	1.41	52.31	11.40	1.72	40.60	13.74

In the proposed model fitting scheme, the range of the model parameters are set according to the general behavior of the vocal tract. The possible range of the parameter r is changed within the limit 0.8 to 0.99, which covers even a very rapidly decaying impulse for the purpose of our simulation. The search range for θ is set according to the determined formant band. Search resolutions for r and θ are chosen as$\triangle r = 0.01$ and $\triangle\theta = 0.001\pi$, respectively. In our experiments in order to obtain a noisy signal, noise sequence of a particular SNR is added with the clean (noise-free) signal. Noisy signals are generated according to (2.13) where the noise variance σ_v is appropriately determined according to a specified level of SNR defined as

$$SNR = 10log_{10}\frac{\sum_{n=0}^{N-1} x(n)^2}{\sum_{n=0}^{N-1} v(n)^2} \tag{2.53}$$

At first results for three synthetic vowels /a/, /o/ and /u/ are presented in Table 2.1. Vowels with duration of 80 ms are synthesized using the Klatt synthesizer considering the pitch values of 220 Hz . Estimation eroor for the first three formants are taken into consideration after performing estimation for 10 independent trials. Here the estimation error, the mean average deviation between the estimated formant frequency f_E and the reference formant frequency f_R is defined as

$$E = |\frac{f_E - f_R}{f_R}| \times 100\% \tag{2.54}$$

In Table 2.1, the estimatin error is shown for the three synthesized vowels at the presence of white Gaussian noise with a SNR of $5dB$ and$-5dB$ for both male and female sounds, respectively. It is clearly observed that the proposed method is able to reduce estimation error significantly in the case of noisy environments.

Table 2.2: Performance comparison in terms of mean error(%) for male speakers

Vowel		-5 dB			5 dB		
		Proposed	AFB	LPC	Proposed	AFB	LPC
	F1	16.11	30.88	30.53	17.01	17.74	26.48
/aa/	F2	16.65	36.42	82.19	10.63	21.87	45.44
	F3	21.15	15.47	43.35	16.83	17.07	39.80
	F1	18.02	24.12	31.64	17.37	16.31	24.65
/ah/	F2	12.90	28.88	57.43	9.72	24.41	35.57
	F3	19.00	13.09	39.21	14.16	11.61	37.72
	F1	18.36	35.49	22.63	18.06	37.77	19.72
/ow/	F2	16.03	26.03	47.20	12.08	24.65	41.67
	F3	19.02	14.20	36.68	17.22	14.00	37.74
	F1	18.79	36.49	20.14	19.24	36.55	19.49
/uh/	F2	12.34	23.49	38.02	12.21	23.23	37.48
	F3	14.60	13.89	37.24	14.09	13.86	37.33
	F1	18.64	36.72	29.66	18.28	39.58	20.09
/uw/	F2	13.40	23.14	40.36	12.23	22.49	36.45
	F3	16.51	14.53	39.48	15.29	14.50	38.25

Table 2.3: Performance comparison in terms of mean error(%) for female speakers

Vowel		-5 dB			5 dB		
		Proposed	AFB	LPC	Proposed	AFB	LPC
	F1	14.23	19.40	45.04	13.74	12.33	25.84
/aa/	F2	15.34	69.07	27.67	9.24	21.02	20.53
	F3	11.51	30.96	12.77	11.56	20.79	11.95
	F1	15.29	14.68	36.14	13.51	10.79	17.91
/ah/	F2	16.52	37.14	21.46	11.90	13.37	20.44
	F3	10.76	23.14	15.75	9.32	19.79	12.44
	F1	12.34	11.75	26.75	12.20	10.82	15.70
/ow/	F2	15.94	43.25	26.84	10.56	27.30	20.96
	F3	10.35	18.63	15.23	10.21	17.84	10.00
	F1	13.02	12.91	18.07	13.14	10.54	16.46
/uh/	F2	13.52	23.46	21.40	11.49	18.72	20.04
	F3	8.97	18.93	10.40	8.83	18.40	9.54
	F1	13.37	9.47	16.67	13.26	9.12	16.37
/uw/	F2	13.18	17.18	20.33	10.57	16.46	18.49
	F3	9.55	18.12	9.82	9.61	18.19	9.42

The simulation results for TIMIT database are presented next. The estimation errors obtained by the proposed method and that by the other two methods are presented under the influence of white gaussian noise conditions for male and female speakers in Tables 2.2 and 2.3.

For Table2.2 and Table 2.3 SNR levels $5dB$ and $0dB$ are considered. For each vowel, the estimation errors for three different formants, namely $F1$,$F2$ and $F3$ are listed. As can be seen from the tables, the proposed method offers better performance than both the12 order LPC and the AFB methods under presence of background noise. It can be observed that the estimation error obtained by the proposed method in comparison to that of the other methods is extremely lower in such severe noisy conditions.

In some cases it is found that the estimation accuracy decreases for the cases when the two formants are very closely spaced, for example in case of vowel /ih/, though, considering the level of noise, the estimation accuracy obtained by the proposed method is quite acceptable. It is also observed that the estimation error relatively increases in

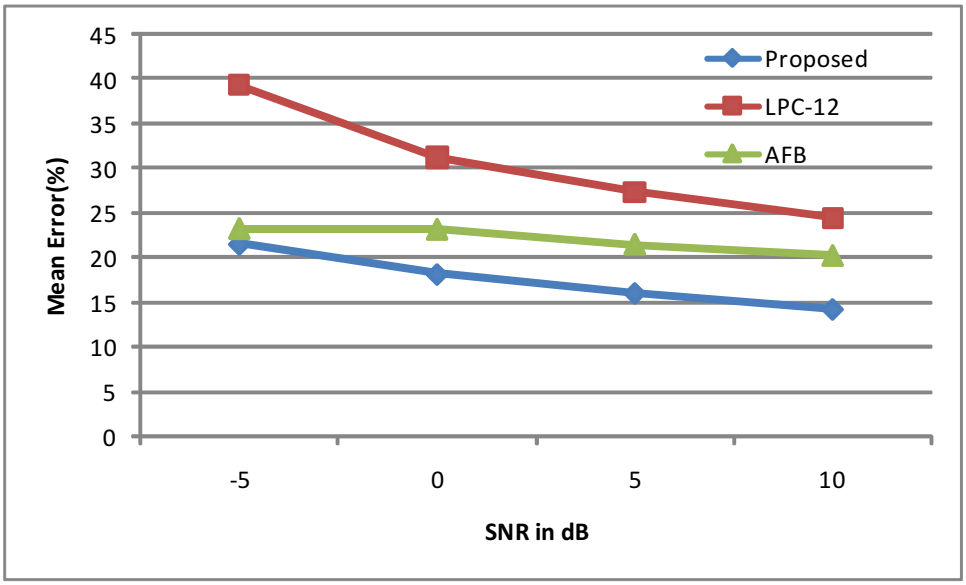

Figure 2.16: Error comparison of all formants for different methods

case of high pitch female speakers. It is clearly observed that the estimation performance for the third formant, which is by nature very difficult to estimate because of low spectral magnitude, is significantly enhanced by the proposed method. Hence, overall it can be said that, the proposed method increases formant estimation performance.

In order to present the overall formant estimation errors over a large range of SNRs considered in the experimental setup, in Fig. 2.16 the overall estimation error for all vowels are shown. In a simiar way, in order to present the overall formant estimation errors over a large range of SNRs considered in the experimental setup, in Fig. 2.17 the overall estimation error for all vowels are shown. It is observed that the formant estimation performance obtained by the three methods remains similar in case of high level of SNR. However, with the decrease in SNR level, the estimation performance of the other two methods deteriorates in comparison to that of the proposed method. The performance of the proposed method remains quite consistent even in the low levels of SNRs and level of performance degradation is not very significant till -15 dB. However, beyond that the performance of the proposed method is not satisfactory because of the severe noise corruption, leading to complete failure for the conventional methods.

In the proposed method formant estimation is carried out frame by frame with a frame length of 512 samples and 10 ms overlap between the successive frames. As a

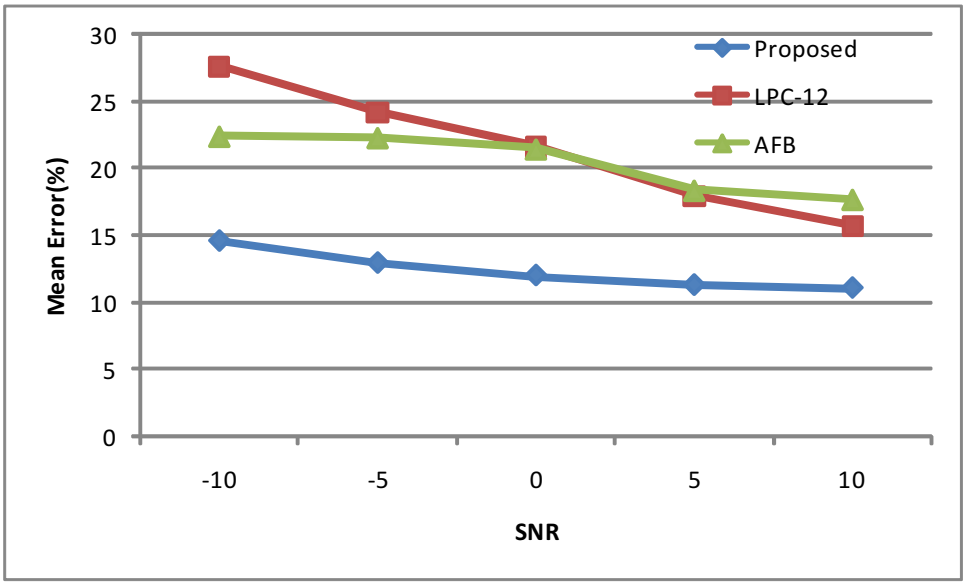

Figure 2.17: Error comparison of all formants for different methods for female

result for a vowel sound of duration of about 80 ms, 5 frames are analyzed. It is to be noted that, because of the inherent characteristics of the fast Fourier transform (FFT) operation, there exists an inherent error caused by the minimum width of the FFT bin. For instance, when a 512 point FFT is performed on a speech frame with sampling frequency of 16 kHz, the resulting FFT has a resolution of 15.6 Hz.

Table 2.4: Recognition Accuracy

	Feature Vector	
SNR	MFCC + Proposed Method	MFCC + LPC-12
-10 dB	68.00	60.00
-5 dB	82.67	80.00
5 dB	89.33	85.33

By incorporating the estimated formants in a feature vector along with traditional MFCC, significantly better vowel recognition accuracies are achieved compared to a feature vector consisting of MFCC and formants estimated by LPC, especially under the influence of noise. By using these formants along with the traditional 12 MFCC coefficients as a feature vector, vowel recognition was performed for the vowels /aa/, /ux/

and /ix/ from the TIMIT database. As formant ranges for male and female vowels vary significantly, they are considered as separate classes for this LDA based classification operation. There are 20 utterances for male and 20 utterances for female considered for each vowel. Accuracies are calculated by leaving one sample out while training the classifier and then testing the left out sample. This check is performed for all the samples in the database, and it is found that the proposed feature vector offers better performance in noisy conditions. The recognition accuracies for different vowels is presented in Table 2.4. It can be concluded from the table that the proposed noise robust formant estimation method, when used for vowel recognition, increases the recognition accuracy for vowel recognition systems under the influence of noise.

As seen from these analyses, the proposed method offers a better performance over the LPC and AFB methods in noise free as well as in noisy conditions. In order to demonstrate the effectiveness of our proposed method, a spectrogram of the sentence "His head flopped back", uttered by a male speaker taken from the TIMIT database is shown in Fig. 2.18. The formant frequencies estimated at different frames using the proposed method uder SNR= 0 dB are shown over the spectrogram of clean speech. In the tracking, only the estimated formants of the vowels are shown. It can be observed from the figure that the proposed method tracks the formant frequencies quite accurately even in noisy speech.

2.3 Conclusion

In this chapter it is shown that the spectrum of the ramp cepstrum of the SSACF of the system impulse response exhibits direct relationship with the system poles. Hence, a spectral model of the RC using the SSACF is proposed to use as a target function to extract formants from given noisy speech observations. A residue based spectral domain model matching scheme is introduced where the spectral error between the proposed model and the spectrum of the RC of the SSACF of noisy speech signal is minimized. In order to reduce the computational burden, in the proposed residue based spectral

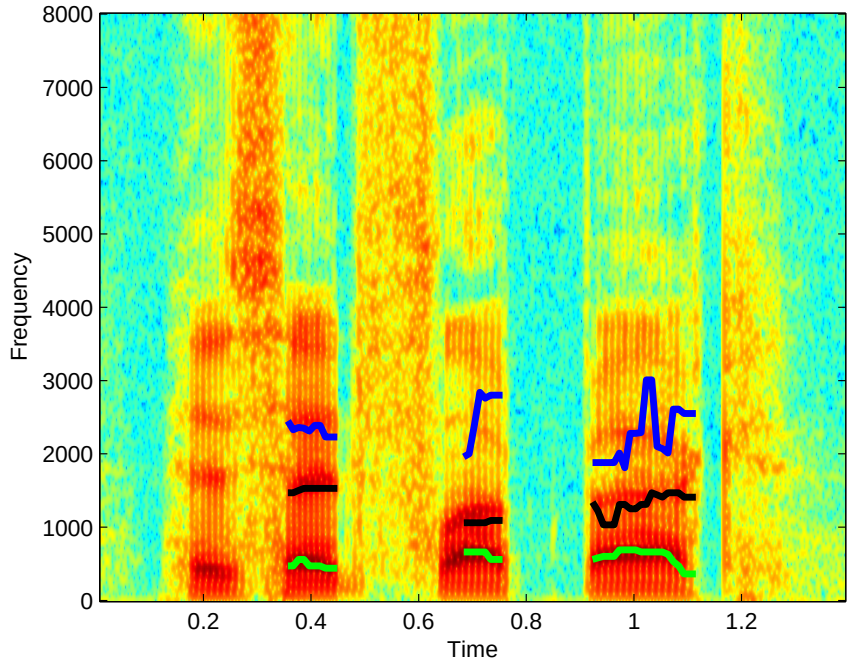

Figure 2.18: Spectrogram of the sentence "His head flopped back" at 0 dB noise with tracked formant by the proposed method

matching scheme fitting operation is carried out for each formant separately restricting the search within a formant zone. Formant estimation is performed considering different vowels uttered. Estimation performance of the proposed method is compared with the widely used LPC and AFB formant estimation methods.

Chapter 3

Spectral Domain Ramp Cepstrum Model of Repeated ACF

The objective of this chapter is to develop a formant estimation scheme which can provide an accurate estimate of formant frequencies of speech signals even under severe noisy condition. Motivated by the advantageous properties of the ACF such as the peak strengthening and noise reduction, the use of repeated ACF in ramp-cepstral domain is introduced. Considering the overall human vocal-tract as an all-pole system, we propose a spectral domain ramp cepstrum (SDRC) model for a once repeated single sided autocorrelation function (ORSSACF) of speech signals. Since, parameters of the proposed RC model provide a direct relationship with formant frequencies; the residue based model fitting approach is employed for formant estimation.

3.1 Methodology

3.1.1 Background

As shown in the previous chapter, vocal tract can be assumed to be a causal, stable, linear time-invariant and stationary autoregressive (AR) system, and thus a voiced speech signal constructed from it can be characterized as

$$x(n) = -\sum_{k=1}^{P} a_k x(n-k) + Gu(n), \tag{3.1}$$

where $\{a_k\}$ are the system parameters, G denotes the gain factor, P is the known system order and $u(n)$ represents the excitation to the system. The AR system transfer function $H(z)$, which in this case is the vocal tract transfer function can be expressed as

$$H(z) = \frac{G}{1 + \sum_{k=1}^{P} a_k z^{-k}} = \frac{G}{\prod_{k=1}^{P}(1 - p_k z^{-1})}, \tag{3.2}$$

where $p_k = r_k e^{j\omega_k}$ denotes the k-th pole of the AR system with magnitude r_k and angle ω_k. Formants are associated with the free resonances of the vocal tract system. In order to model each formant, a pair of complex conjugate poles is required. In (3.2), each formant corresponds to p_k and it's conjugate. Thus, for a vocal tract system modeled with P-th order AR system, there exists $P/2$ formants. Formant frequency (F_k) and bandwidth (B_k) can be expressed in terms of pole parameters as [30]

$$F_k = \frac{F_s}{2\pi}\omega_k; B_k = -\frac{F_s}{\pi} ln(r_k), \tag{3.3}$$

where F_s is the sampling frequency.

In the LPC based methods, the ACF of the given speech signal $x(n)$ is used in the Yule-Walker equations to obtain the AR parameters and thereby the poles of the vocal tract system and from the estimated poles, formants are calculated. But in the presence of observation noise, LPC based methods fail to provide an accurate estimate of the AR parameters and thus exhibits poor formant estimation accuracy. Moreover, the effect of pitch variation may cause significant errors in the LPC based formant estimation. Hence, development of a formant estimation scheme, which can estimate the formants with a higher accuracy even in the presence of severe background noise as well as handle the effect of pitch variation is in great demand.

3.1.2 Cepstral Domain Analysis

As introduced in the previous chapter, in order to reduce the effect of pitch from the speech signal, cepstrum that offers the advantage of homo-morphic de-convolution has been most commonly used. The principle of homomorphic deconvolution helps in separating signals that have been combined via convolution and thus it become a very important tool in different speech processing applications, such as speech recognition. The complex cepstrum of a signal $h(n)$ is defined as [24]

$$c_{hc}(n) = F^{-1}\left\{ln(H(e^{j\omega}))\right\}, \tag{3.4}$$

where $F^{-1}\{.\}$ denotes the inverse Fourier transform and the spectrum of $h(n)$ is given by $H(e^{jw})$. Considering the VT system as minimum phase, using (3.2) $c_{hc}(n)$ is a sequence that is real and causal. Recalling from the previous chapter,$c_{hc}(n)$ can be expressed in terms of poles as

$$c_{hc}(n) = \sum_{i=1}^{P} \frac{p_i^n}{n}, n > 0. \tag{3.5}$$

On the other hand, the real cepstrum of $h(n)$ is defined as

$$c_h(n) = F^{-1}\left\{ln(|H(e^{j\omega})|)\right\}. \tag{3.6}$$

In order to avoid notational complexity, instead of denoting real cepstrum as $c_{hr}(n)$, simply $c_h(n)$ is used, i.e. an additional subscript 'r' is not used hereafter. For $n > 0$, the relation between complex and real cepstra is given by $c_h(n) = 0.5c_{hc}(n)$. Hereafter, alike previous chapter, only real cepstrum shall be considered. In order to avoid logarithm of negative values, in practical applications real cepstra is most commonly used. Here, for real cepstrum (3.5) can be written as

$$c_h(n) = 0.5\sum_{i=1}^{P} \frac{p_i^n}{n}, n > 0. \tag{3.7}$$

The speech signal $x(n)$ given by (3.1) can be considered as a convolution sum between $h(n)$, the impulse response of the V.T. system and $u(n)$, the excitation to the V.T. system

as follows

$$x(n) = h(n) * u(n). \tag{3.8}$$

Thus, one can write the corresponding cepstral representation of $x(n)$ as

$$c_x(n) = c_h(n) + c_u(n). \tag{3.9}$$

Here, $c_h(n)$ is the cepstrum of $h(n)$, $c_u(n)$ is the cepstrum of the excitation $u(n)$ and $H(e^{j\omega})$ and $U(e^{j\omega})$ are frequency domain representations of $h(n)$ and $u(n)$. As periodic impulse-train excitation is commonly considered to model the voiced sounds, here a periodic impulse-train excitation $\{u(n)\}_{n=0}^{N-1}$ with period T is considered, which can be expressed as

$$u(n) = \sum_{k=0}^{\lambda-1} \delta(n - kT), \lambda = \lceil N/T \rceil. \tag{3.10}$$

Here, λ is the total number of impulses within the excitation. As introduced in the previous chapter, based on (3.6), utilizing this advantage of homomorphic deconvolution, cepstral domain system identification methods have been proposed which deal with the noise free environment . It is apparent from(3.5) and (3.9) that in order to extract the system poles p_i, from a given $c_x(n)$ one needs to extract $c_h(n)$. However, as stated in (3.9) , $c_h(n)$ is mixed with $c_u(n)$ resulting $c_x(n)$. In case of periodic impulse train excitation, it can be shown that $c_u(n)$ contributes to $c_x(n)$ at the origin and periodically after each T interval. Thus, even in case of finite data analysis, within the range $0 < n < T$, the effect of $c_u(n)$ can be neglected, resulting

$$c_x(n) = c_h(n) = 0.5 \sum_{i=1}^{P} \frac{p_i^n}{n}, 0 < n < T \tag{3.11}$$

It indicates that the cepstral coefficients corresponding to $c_x(n)$ can be considered similar to that of $c_h(n)$ within the range $0 < n < T$.

3.1.3 Effect of Noise

In the presence of additive white Gaussian noise (AWGN) $v(n)$, the observed signal $y(n)$ can be expressed as

$$y(n) = x(n) + v(n), \tag{3.12}$$

where $v(n)$ is assumed to be zero mean stationary and independent of $u(n)$. The cepstral coefficients of $y(n)$ can then be expressed as

$$\begin{aligned} c_y(n) &= F^{-1}\{ln(|X(e^{j\omega})|)\} + F^{-1}\{ln(1 + \frac{|V(e^{j\omega})|}{|X(e^{j\omega})|})\} \\ &= c_x(n) + c_w(n). \end{aligned} \tag{3.13}$$

$c_w(n)$ appears in the presence of noise and vanishes in its absence. As shown in the previous chapter, at severe noise it is very difficult to get an accurate estimate of $c_x(n)$ from $c_y(n)$, since the cepstrum decomposition techniques are very sensitive to the noise level. As a result, it is desirable to develop an algorithm that can reduce the effect of noise on the signal, thereby reducing the effect of $c_w(n)$ on $c_y(n)$ and producing more noise robust cepstral coefficients. In this regard, we propose to investigate the effect of increasing the number of poles on the formant location to enhance the strength of the formant peaks.

In view of enhancing the spectral peaks corresponding to a particular frequency, one possible approach would be to introduce new poles having that frequency. In particular, if the new poles can be generated exactly at the same location of those original poles, the spectral peak corresponding to that pole location will be significantly enhanced. As only the speech signal is available at hand and the VT transfer function can not be changed, it is not possible to place poles at designated places to enhance spectral peaks. As an alternate, if a signal is convolved with its folded version new poles would be introduced, which should be related to the original system poles. An equivalent approach is to achieve this effect by simply doing the autocorrelation operation on the signal. The autocorrelation function (ACF) of $x(n)$ is defined as

$$\begin{aligned} r_{xx}(m) &= x(n) * x(-n) \\ &= E[x(n)x(n+m)]. \end{aligned} \tag{3.14}$$

Here $E[.]$ denotes the expectation operator.

As, shown in the previous chapter, according to (3.14) and (3.8), the z transform of $r_{xx}(n)$ can be expressed as

$$R_{xx}(e^{j\omega}) = R_{hh}(e^{j\omega}) \times R_{uu}(e^{j\omega}), \tag{3.15}$$

where $R_{hh}(e^{j\omega})$ and $R_{uu}(e^{j\omega})$ are the frequency domain representations of $r_{hh}(n)$ and $r_{uu}(n)$, the ACFs corresponding to $h(n)$ and $u(n)$, respectively. According to the definition (3.14), $R_{hh}(e^{j\omega})$ can be written as

$$R_{hh}(e^{j\omega}) = H(e^{j\omega}) \times H(e^{-j\omega}). \tag{3.16}$$

Using (3.2), interms of poles $R_{hh}(e^{j\omega})$ can be expressed as

$$R_{hh}(e^{j\omega}) = \frac{C_1}{\prod_{i=1}^{P}(1 - p_i e^{-j\omega})(1 - p_i^* e^{j\omega})}. \tag{3.17}$$

Here for each pole $p_i = r_i e^{j\theta}$, there exists a pole $1/p_i^*$ which is placed at conjugate reciprocal locations. From (3.17) it is clearly seen that total number of poles in $R_{hh}(e^{j\omega})$ is $2P$, which is twice as the number of poles in $H(e^{j\omega})$. Due to the autocorrelation operation new P poles are introduced in $R_{hh}(e^{j\omega})$ which are conjugate reciprocal to the original P poles of $H(e^{j\omega})$, i.e. the new poles are located at the original pole angles as expected.

Using (3.12) and (3.14), the ACF of noisy speech $y(n)$ can be expressed as

$$r_{yy}(n) = r_{xx}(n) + r_{ww}(n). \tag{3.18}$$

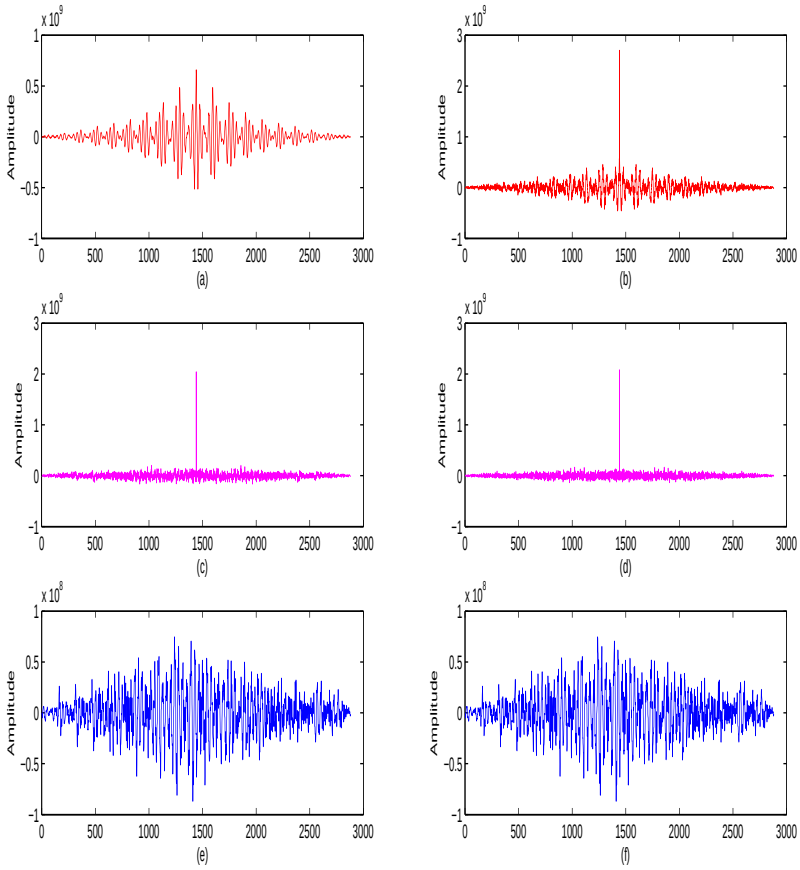

Figure 3.1: Effect of noise in the autocorrelation domain: plot of different autocorrelation functions (a)$r_{xx}(n)$, (b) $r_{yy}(n)$, (c) $r_{ww}(n)$, (d) $r_{vv}(n)$, (e)$r_{xv}(n)$ and (f) $r_{vx}(n)$

where,

$$r_{ww}(n) = r_{vv}(n) + r_{vx}(n) + r_{xv}(n). \tag{3.19}$$

Here $r_{vv}(n)$ is the ACF of noise $v(n)$and $r_{vx}(n)$ and $r_{xv}(n)$are the cross correlation terms. Since $v(n)$ is uncorrelated with $x(n)$, it is expected that the values of the cross-correlation terms, in comparison to that of $r_{xx}(n)$,will be negligible. On the other hand, the ACF of the AWGN $v(n)$ generally exhibits a peak at the zero lag and the values at all other lags should be very small and ideally should be zero.

In Figs. 3.1(a)-3.1(f), different ACFs, namely $r_{xx}(n)$, $r_{yy}(n)$, $r_{ww}(n)$, $r_{vv}(n)$, $r_{xv}(n)$ and $r_{vx}(n)$ are plotted at SNR$= -5$ dB. From Figs. 3.1(e) and 3.1(f), it can be observed that the values of the cross correlation terms are very small as expected. As seen in Fig. 3.1(d), $r_{vv}(n)$ although exhibits very large peak at zero lag, nonzero small values

exist at all other lags because of the finite data length. It is also observed in Fig. 3.1(c) that $r_{ww}(n)$ exhibits the maximum value at the zero lag and the values at other lags are comparatively very small. From these figures, it can be concluded that in comparison to the effect of $v(n)$ on $x(n)$ as shown in Fig. 2.4, the effect of $r_{ww}(n)$ on $r_{xx}(n)$ is significantly reduced because of the autocorrelation operation.

3.1.4 Effect of Double Autocorrelation

Realizing the effect of spectral peak strengthening and reduction of noise due to autocorrelation as described in the previous section, we propose to generate more poles at the location of the original poles to further strengthen the spectral peaks. In view of achieving this objective, the ACF operation can be repeated, which not only strengthens the dominant peaks but also preserves pole locations. Performing further autocorrelation operation on an ACF of a noise corrupted speech signal will imitate duplication of poles at the original locations of the system. Hence, the resulting double correlated signal is expected to exhibit more noise immunity and in its spectrum, even under heavy noisy condition, the formant peaks will be significantly enhanced. Hence, use of the spectrum corresponding to the double correlated signal, instead of that corresponding to the noisy signal, would be much convenient for formant estimation. According to the definition of the ACF given in (3.14), the ACF of $r_{xx}(m)$, namely the repeated ACF of $x(n)$ can be expressed as

$$\rho_{xx}(m) \quad = \quad r_{xx}(m) * r_{xx}(-m). \tag{3.20}$$

Transferring (3.20) into z domain yields the following,

$$\begin{aligned} P_{xx}(e^{j\omega}) &= R_{xx}(e^{j\omega}) \times R_{xx}(e^{-j\omega}) \\ &= R_{hh}(e^{j\omega}) \times R_{hh}(e^{-j\omega}) \times R_{uu}(e^{j\omega}) \times R_{uu}(e^{-j\omega}) \\ &= P_{hh}(e^{j\omega}) \times P_{uu}(e^{j\omega}), \end{aligned} \tag{3.21}$$

where $P_{hh}(e^{j\omega})$ and $P_{uu}(e^{j\omega})$ are the frequency domain representations of $\rho_{hh}(n)$ and $\rho_{uu}(n)$, the ACFs corresponding to $r_{hh}(n)$ and $r_{uu}(n)$, respectively.

Using (3.2) and (3.17) interms of poles $P_{hh}(e^{j\omega})$ can be expressed as

$$P_{hh}(e^{j\omega}) = \frac{C_1}{\prod_{i=1}^{2P}\{(1-p_i e^{-j\omega})(1-p_i^* e^{j\omega})\}} = \frac{C_1}{\prod_{i=1}^{P}\{(1-p_i e^{-j\omega})(1-p_i^* e^{j\omega})\}^2}. \tag{3.22}$$

Here for each pole $p_i = r_i e^{j\theta}$, there exists a pole $1/p_i^*$ which is placed at conjugate reciprocal locations. From (3.22) it is clearly seen that total number of poles in $P_{hh}(e^{j\omega})$ is $4P$, which is twice as the number of poles in $R_{hh}(e^{j\omega})$. Due to the autocorrelation operation new $2P$ poles are introduced in $P_{hh}(e^{j\omega})$ which are conjugate reciprocal to the original $2P$ poles of $R_{hh}(e^{j\omega})$. Hence, new poles are located at the original pole angles as expected and for each original pole location of $R_{hh}(e^{j\omega})$, both inside and outside unit circle, in $P_{hh}(e^{j\omega})$ there exists two poles.

In order to demonstrate the effect of double autocorrelation operation on system poles, in Fig. 3.2 another all pole system is shown having all six pole pairs of the system considered in Fig. 2.7 along with their complex conjugate poles. From the figure it is seen that at each angular position of the original poles, one new pole is generated both inside and outside the unit circle. Obviously, with the increase in number of poles at a particular angular position, the spectral energy corresponding to that particular frequency will be significantly increased. Especially in the presence of noise this can help in finding out the formant peaks in spite of the presence of several unwanted noise peaks. In order to present the effect of spectral peak strengthening, in Fig. 3.3 spectra corresponding to the synthetic speech $x(n)$, its ACF $r_{synx}(m)$, and ORACF $\rho_{synx}(m)$ are

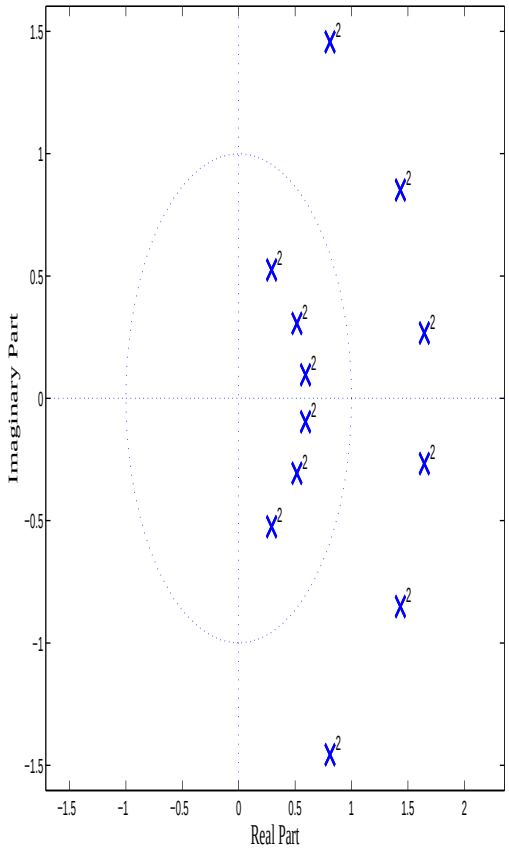

Figure 3.2: Pole locations of ORACF

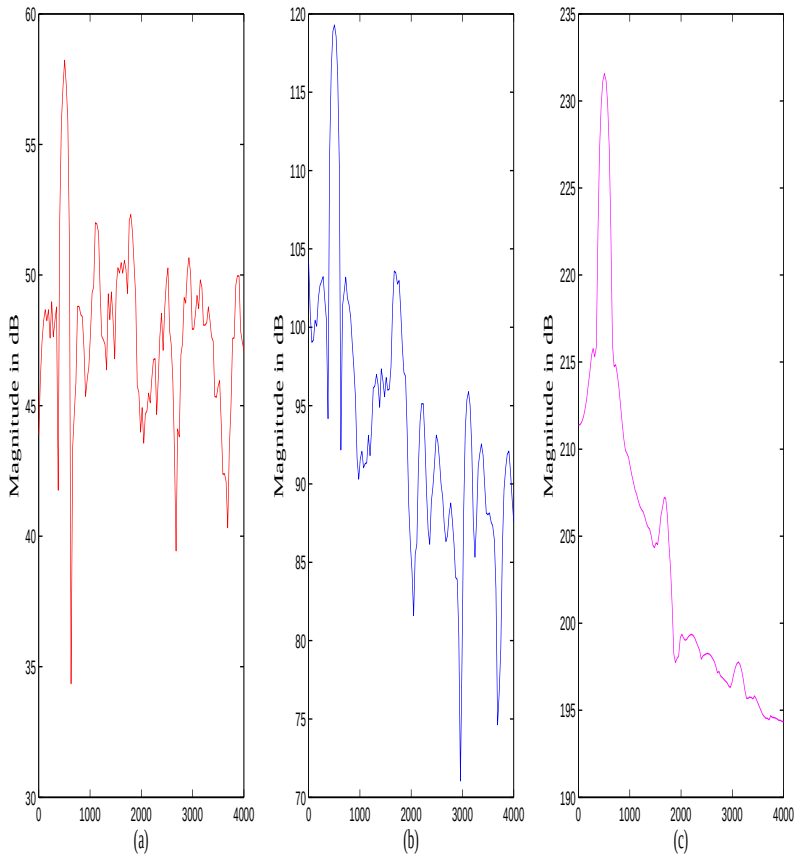

Figure 3.3: Magnitude Spectra of (a) $x(n)$, (b) ACF of $x(n)$ and (c) ORACF $x(n)$ for the synthetic signal used in Fig. 2.3 all at $SNR = -5dB$.

shown at $SNR = -5dB$. It is to be mentioned that the synthetic speech considered here is the one that is used in Fig. 2.3. From the figure it is quite clear that ORACF much more noise robust not only than the noisy signal $x(n)$ but also from its ACF $r_{synx}(m)$. Moreover, ORACF shows lesss spurious peaks than ACF

At this point if only the causal part of the ACF signal i.e. the single sided ACF (SSACF) is considered, the dominant peaks would become more distinct [34]. Since our objective is to handle the severe noisy condition, the use of SSACF would be a better choice. Following (2.20) for the SSACF of $x(n)$, namely $r^+_{xx}(m)$, the ORSSACF of $x(n)$, namely $\rho^+_{xx}(m)$ can be obtained from the double sided ACF (DSACF) as

$$\rho^+_{xx}(m) = \begin{cases} \rho_{xx}(m)\,, & m > 0 \\ 0.5\rho_{xx}(m)\,, & m = 0 \\ 0\,, & m < 0 \end{cases} \tag{3.23}$$

Since the DSACF is symmetric about the zero lag ($m = 0$), it can be computed using (3.20). The fourier transform of $\rho^+_{xx}(m)$ is a complex spectrum $P^+_{xx}(e^{j\omega})$ and its spectral envelope is defined as

$$E^2(e^{j\omega}) = |P^+_{xx}(e^{j\omega})|$$

It can be shown that due to the large dynamic range of speech spectra, the envelope of the ORSSACF spectrum, $P^+_{xx}(e^{j\omega})$ enhances the highest power frequency bands with respect to the spectrum of $\rho_{xx}(m)$, namely $P_{xx}(e^{j\omega})$, just like the SSACF spectrum $R^+_{xx}(e^{j\omega})$ enhances the highest power frequency bands with respect to $P_{xx}(e^{j\omega})$ [41]. Consequently, the noise components lying outside the enhanced frequency bands are largely attenuated in $E^2(e^{j\omega})$ with respect to $P_{xx}(e^{j\omega})$, and thus use of the envelope of $P^+_{xx}(e^{j\omega})$ is more robust to broadband noise than using $P_{xx}(e^{j\omega})$. As shown in the previous chapter, the SSACF and the original signal $x(n)$ have the same poles[34], [35]. SSAC sequence only has the causal part of the double sided sequence and doesnot include the poles outside the unit circle. Similarly, ORSSAC sequence has only the causal part of the repeated ACF

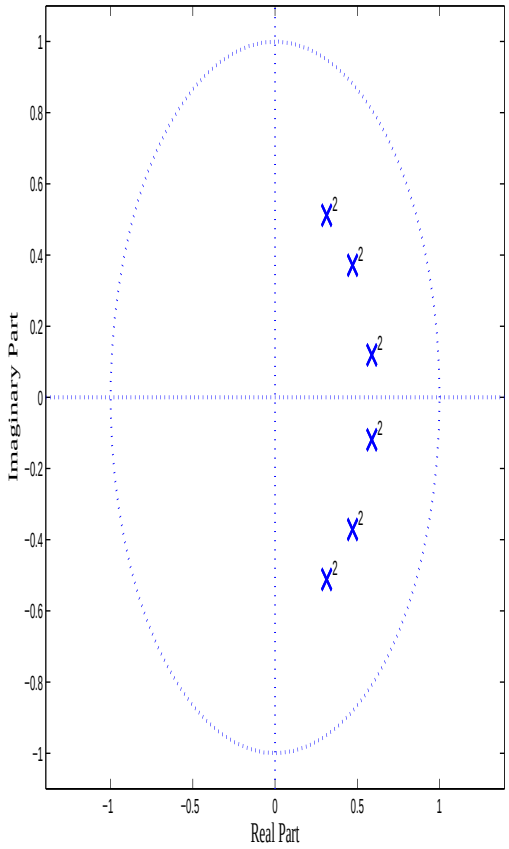

Figure 3.4: Pole locations of ORSSACF

sequence and thus it should include only the poles inside the unit circle. Hence, following (3.22) in terms of poles the frequency domain representation of the transfer function in relation to ORSSAC sequence is given below,

$$P_{hh}^{+}(e^{j\omega}) = \frac{G_{hh}}{\prod_{i=1}^{2P}(1-p_i e^{-j\omega})} = \frac{G_{hh}}{\prod_{i=1}^{P}\{(1-p_i e^{-j\omega})(1-p_i^{*} e^{-j\omega})\}^2}, \tag{3.24}$$

It is to be noted that as $P_{hh}^{+}(z)$ has $2P$ poles all inside the unit circle, it has$P$ complex conjugate pole pairs in $P/2$ locations. Thus in each pole locations it has two poles.

In Fig. 3.4 pole locations of a ORSSAC sequence is demonstrated where in each of the original pole locations two poles are found and there are no poles outside the unit circle which indicates the system is causal and stable. Obviously, with the increase in number of poles at a particular angular position, the spectral energy corresponding to that particular frequency will be significantly increased. Now, properties of ORSSACF like

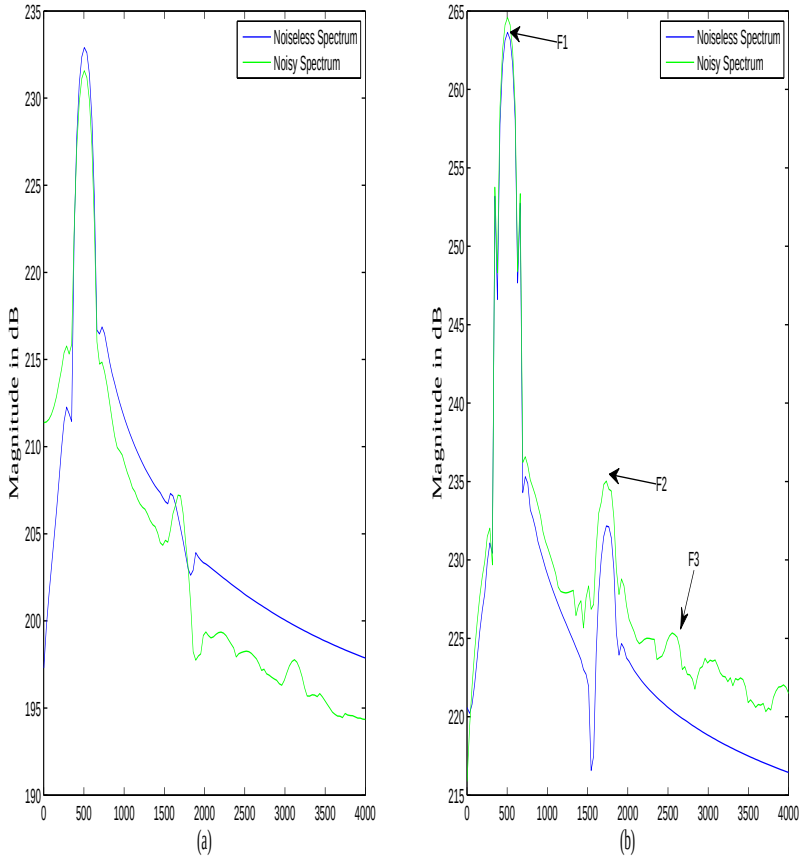

Figure 3.5: Magnitude Spectra of (a) ORACF, (b) ORSSACF of $x(n)$ for the synthetic signal used in Fig. 2.3 at noiseless conditions and SNR= -5 dB.

noise robustness and pole preservation, suggest that AR parameters of the speech signal can be more reliably estimated from the ORSSACF. The robustness of the ORSSACF to additive white noise at SNR=0 dB is illustrated in Fig. 3.5. As can be seen from this figure that the envelope of the squared magnitude spectrum of the ORSSACF shows a prominent first formant, and the whole curve is more robust to additive white noise in comparison to that obtained by using the once repeated ACF (ORACF).

Using (3.21) it can be shown that in time domain similar to (3.20), the ORSSACF of $x(n)$, can be expressed as the convolution between $\rho^{+}_{hh}(m)$ and $\rho^{+}_{uu}(m)$, which are single sided autocorrelation sequences generated from $r_{hh}(m)$ and $r_{uu}(m)$, respectively within the limit $0 \leq m < T$, where T is the time period of the impulse train $u(n)$. This relation is expressed in the following manner

$$\rho^{+}_{xx}(m) = \rho^{+}_{hh}(m) * \rho^{+}_{uu}(m).\ 0 \leq m < T \tag{3.25}$$

Here, $\rho^{+}_{uu}(m)$ is a periodic sequence which has the same priodicity as $u(n)$.From (3.25) it is obvious that transferring to the cepstral domain can provide the opportunity of source signal separation using the property of homomorphic deconvolution. In cepstral domain, 3.25 can be written as

$$c_{\rho^{+}_{xx}}(m) = c_{\rho^{+}_{hh}}(m) + c_{\rho^{+}_{uu}}(m), \tag{3.26}$$

where, $c_{\rho^{+}_{xx}}(m)$, $c_{\rho^{+}_{hh}}(m)$ and$c_{\rho^{+}_{uu}}(m)$ are the real cepstra of $\rho^{+}_{xx}(m)$,$\rho^{+}_{hh}(m)$ and $\rho^{+}_{uu}(m)$.

In Fig. 3.6 comparison of $c_h(n)$ and $c_x(n)$ and in Fig. 3.7 $c_{\rho^{+}_{xx}(m)}$ and $c_{\rho^{+}_{hh}(m)}$ of a signal $x(n)$ constructed from the system in Fig. 2.1is shown. From these two figures it is observed that $c_{\rho^{+}_{xx}(m)}$ follows $c_{\rho^{+}_{hh}(m)}$ within $0 < m < T$ in a much clearer manner than $c_h(n)$ follows $c_x(n)$ and thus 3.26 can be approximated as

$$c_{\rho^{+}_{xx}}(m) \approx c_{\rho^{+}_{hh}}(m), \tag{3.27}$$

Using (3.24), the complex cepstrum corresponding to $\rho^{+}_{hh}(m)$ can be repesented in the same manner of (2.25) and (2.26) as

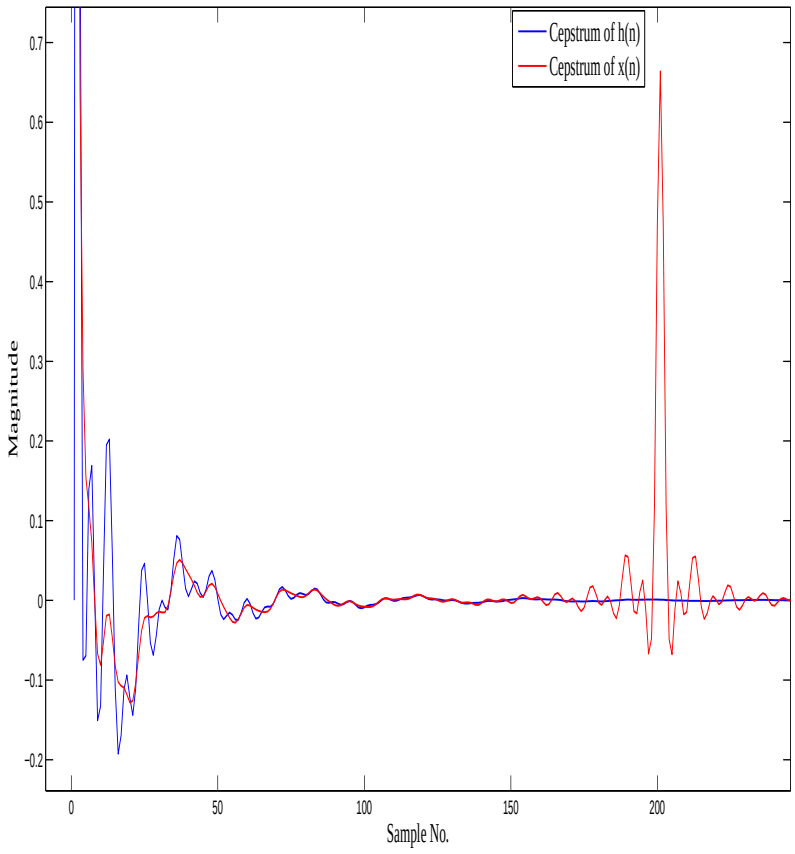

Figure 3.6: Comparison of $c_h(n)$ and $c_x(n)$

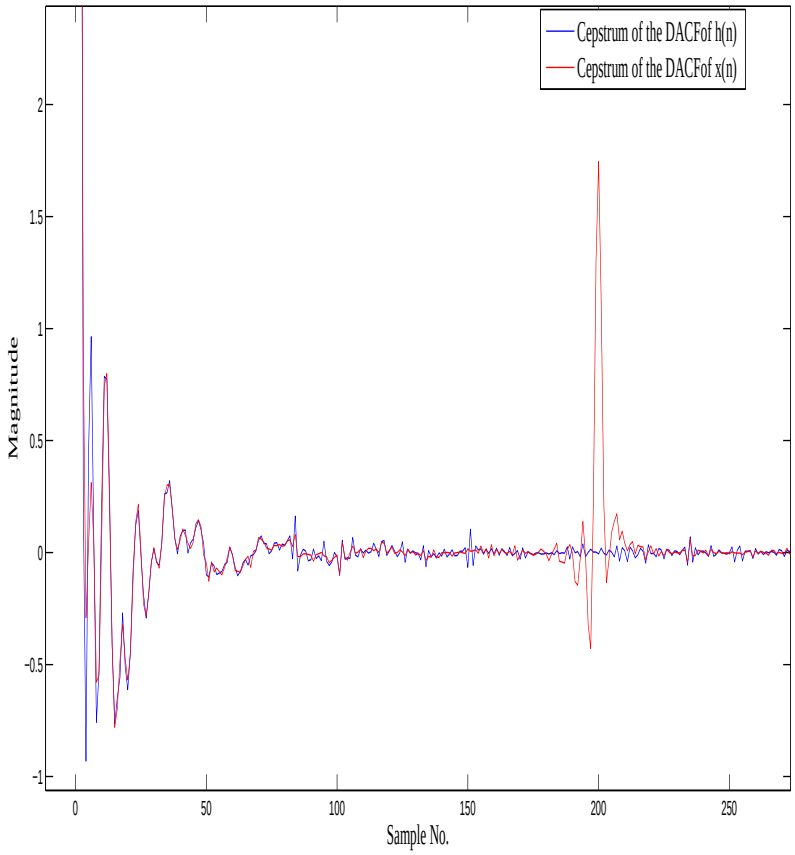

Figure 3.7: $c_{\rho_{xx}(m)}$ and $c_{\rho_{hh}(m)}$ of a signal $x(n)$

$$
\begin{aligned}
c_{\rho^+_{hh}}(m) &= F^{-1}[ln(P^+_{hh}(e^{j\omega}))] \\
&= F^{-1}[\sum_{i=1}^{2P}\sum_{n=1}^{\infty}\frac{p_i^n}{n}e^{-j\omega n} + lnG_2] \\
&= \sum_{i=1}^{2P}\frac{p_i^m}{m}, m > 0. \qquad (3.28)
\end{aligned}
$$

From this relation it is obvious that the cepstrum of the ORSSACF is directly related to the system poles. But, as is seen from (3.28), $c_{\rho^+_{hh}}(m)$ decay rapidly with m, which makes it difficult to estimate the system poles from it. In order to overcome this problem, an easy-to-handle ramp cepstrum [26] is proposed as

$$\mu_h(m) = mc_{\rho^+_{hh}}(m) = \sum_{i=1}^{2P} p_i^m, m > 0 \qquad (3.29)$$

According to (3.27) the ramp cepstrum corresponding to $\rho^+_{xx}(m)$, namely $\mu_x(m) = mc_{\rho^+_{xx}}(m)$ can be expressed as

$$\mu_x(m) \approx \mu_h(m) = \sum_{i=1}^{2P} p_i^m, 0 < m < T. \qquad (3.30)$$

In case of noisy signals further application of ACF on the noise corrupted signal $r_{yy}(n)$ produces $\rho_{yy}(n)$ which can be expressed as

$$
\begin{aligned}
\rho_{yy}(n) &= \rho_{xx}(n) + \rho_c(n) \\
\rho_c(n) &= \rho_{ww}(n) + \rho_{wx}(n) \qquad (3.31)
\end{aligned}
$$

where $\rho_{xx}(n)$ and $\rho_{ww}(n)$are the ACF of $r_{xx}(n)$ and $r_{ww}(n)$ and $\rho_{xw}(n)$ and $\rho_{wx}(n)$ are cross correlation terms. It is expected that the effect of $\rho_c(n)$ on $\rho_{xx}(n)$ is very negligible, as there exists very little correlation between $r_{xx}(n)$ and $r_{ww}(n)$, and $r_{ww}(n)$ is quite insignificant at points other than the zero lag.

In Figs. 3.8(a)-(c), the DACFs $\rho_{yy}(n)$,$\rho_{xx}(n)$,$\rho_c(n)$, are shown. It is clearly observed that the values of $\rho_c(n)$ are extremely small at all lags except the zero lag in comparison to that of $\rho_{xx}(n)$ as expected. From these figures, it can be concluded that in comparison

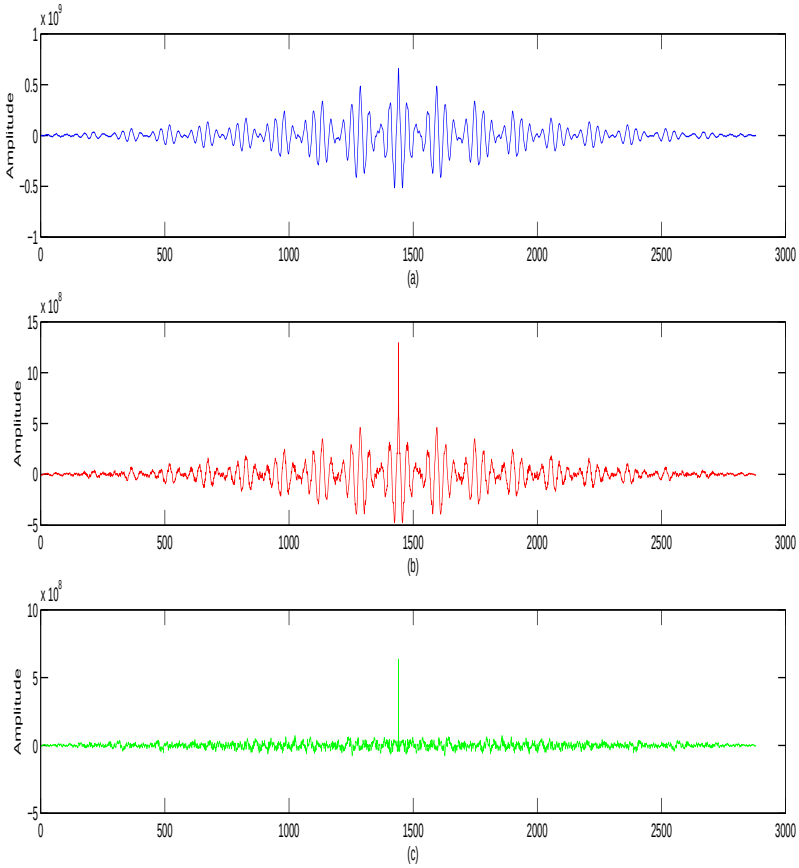

Figure 3.8: Comparison of $\rho_{yy}(n)$,$\rho_{xx}(n)$,$\rho_c(n)$

to the effect of $r_{ww}(n)$ on $r_{xx}(n)$ as shown in Fig. 2.11, the effect of $\rho_c(n)$ on $\rho_{xx}(n)$ is significantly reduced because of the repeated autocorrelation operation. Now, let us consider the cepstral coefficients of $\rho_{yy}(n)$ which can be expressed as

$$\begin{aligned} c_{\rho_{yy}}(n) &= F^{-1}\{ln|P_{xx}(e^{j\omega})|\} + F^{-1}\{ln(1+\frac{|P_C(e^{j\omega})|}{|P_{xx}(e^{j\omega})|})\} \\ &= c_{\rho_{xx}}(n) + c_{\rho_c}(n). \end{aligned} \tag{3.32}$$

Here, $c_{\rho_{yy}}(n)$ is the cepstrum of $\rho_{yy}(n)$, $c_{\rho_{xx}}(n)$ is the cepstrum of $\rho_{xx}(n)$, $c_{\rho_c}(n)$ is the cepstrum of $\rho_c(n)$. In time domain, the effect of $\rho_c(n)$ on $\rho_{xx}(n)$ is smaller than the effect of $r_{ww}(n)$ on $r_{xx}(n)$, because for lags greater than zero the energy ratio of $\rho_c(n)$ to $\rho_{xx}(n)$ is smaller than the energy ratio of $r_{ww}(n)$ to $r_{xx}(n)$. Thus alike the previous case of single autocorrelation, according to Parseval's theorem, in frequency domain the effect of $P_C(e^{j\omega})$ on $P_{xx}(e^{j\omega})$ in 3.32 is smaller than the effect of $R_{ww}(e^{j\omega})$ on $R_{xx}(e^{j\omega})$ shown in the previous chapter. In this case, 3.32 can be rewritten as

$$c_{\rho_{yy}}(n) \approx c_{\rho_{xx}}(n). \tag{3.33}$$

This relation holds true also for the cepstrum computed using the ORSSACF, which as stated earlier provides more noise robustness. Hence, the cepstrum of the ORSSACF of the noisy signal can be rewritten as

$$c_{\rho^+_{yy}}(n) \approx c_{\rho^+_{xx}}(n). \tag{3.34}$$

Corresponding relationship in ramp cepstral domain as per (3.29) can be written as

$$\mu_y(m) \approx \mu_x(m) = 0.5\sum_{k=1}^{2P} p_k^m, m > 0. \tag{3.35}$$

Here, $\mu_y(m) = mc_{\rho^+_{yy}}(m)$ is the ramp cepstrum of $\rho^+_{yy}(m)$. Hence, it is expected that given noisy speech, if ramp cepstrum of its ORSSAC sequence is computed, depending on the level of noise, it may exhibit more noise immunity in comparison to time domain

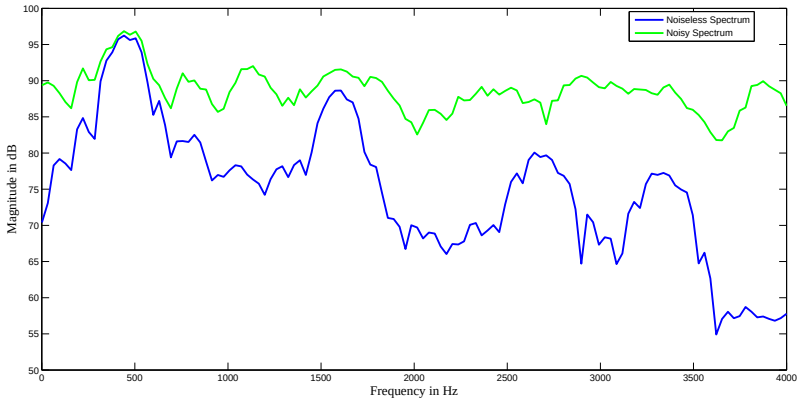

Figure 3.9: Spectrum of the noisy and noiseless signal as used in 2.2 at $SNR = -5dB$

analysis.

As from (3.30) it is shown that $\mu_h(m)$ is directly related to system poles, the corresponding frequency domain representation is given by

$$\mu_H(e^{j\omega}) = \sum_{i=1}^{2P} \frac{G_{hh}}{(1 - p_i e^{-j\omega})}, \tag{3.36}$$

where $\mu_H(e^{j\omega})$ is the Fourier transform of $\mu_h(m)$, G_{hh} is a gain factor and p_i is accompanied by its complex conjugate pole p_i^*. As seen from (3.36) the system corresponding to the ORSSACF of $h(n)$, namely $P_{hh}^{+}(e^{j\omega})$, has P pairs of complex conjugate poles.

Based on (3.35) and (3.36) it is expected that in noisy environment it is advantageous to use the spectrum of $\mu_y(m)$, within $0 < m < T$, which exhibits more noise robustness in comparison to $\rho_{yy}(m)$ and can be approximated as

$$\mu_Y(e^{j\omega}) \approx \mu_H(e^{j\omega}), \tag{3.37}$$

where $\mu_Y(e^{j\omega})$ is the Fourier transform of $\mu_y(m)$, which can be computed from $\rho_{yy}^{+}(m)$ in the following manner

$$\mu_Y(e^{j\omega}) = F\left[m \times F^{-1}\{ln|F\left[\rho_{yy}^{+}(m)\right]|\}\right] \tag{3.38}$$

Here, $F[.]$ denotes Fourier transform.

In Fig. 3.9, for the natural voiced speech /eh/ as shown in Fig. 2.2, a comparison

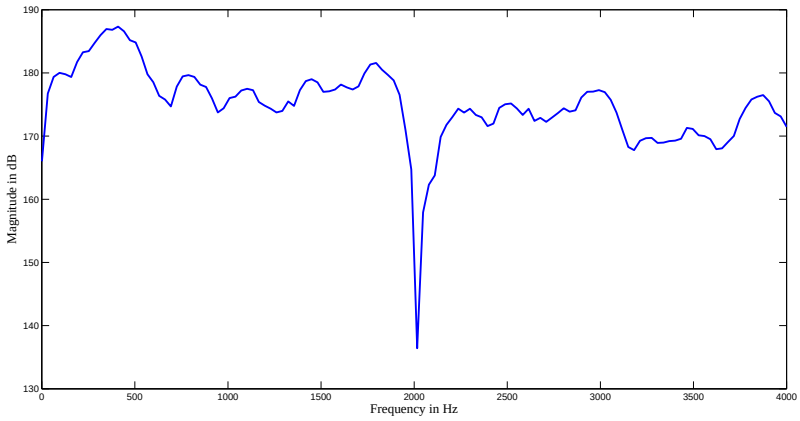

Figure 3.10: Spectrum of the ORACF of the noisy signal presented in 3.9

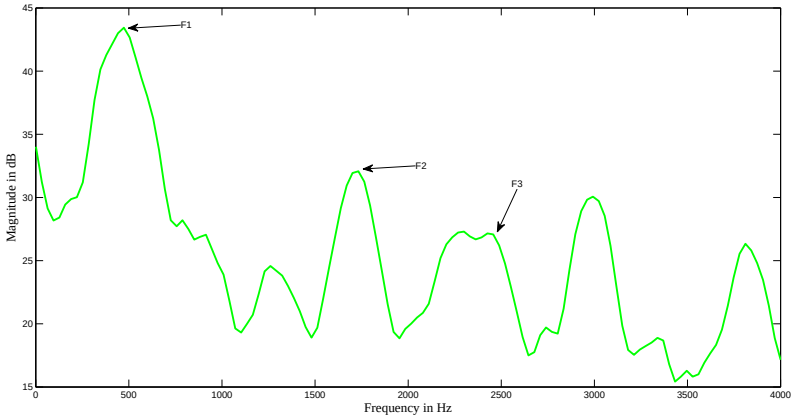

Figure 3.11: Spectrum of the ramp cepstrum of ORSSACF of the signal at SNR=-5 dB

between the noiseless and noisy spectra at $SNR = -5dB$ is shown. In Fig. 3.10, spectrum of the ORACF of the noisy signal presented in Fig. 3.9, is shown and in Fig. 3.11, the spectrum of the ramp cepstrum of the ORSSACF of the noisy signal is shown. From these figures it is evident that the spectrum of the ramp cepstrum of the ORSSACF retains the dominat peaks of the origianl signal and exhibits less spurious peaks than the spectrum of the ORACF. Thus spectrum of the ramp cepstrum of the ORSSACF can be used as a model in a spectral domain model matching scheme for noise robust formant estimation.

3.2 Model matching

As given by (3.2), the transfer function of the VT system if modeled as an $AR(P)$ system, one can consider it as a cascade of $P/2$ blocks where each block consists of a pair of complex conjugate poles. From Fig. 2.15 of previous chapter where the magnitude response of an $AR(6)$ system with three complex conjugate poles is shown along with the magnitude responses of the three individual pole pairs, it is clearly observed that the response of the AR system exhibits three prominent peaks corresponding to the three formants each of which is related to a particular pole pair. Considering the vocal tract as an AR system, a pair of complex conjugate poles is responsible for generating a dominant peak in the spectral domain. Although the effect of other pole pairs, unless otherwise located at a very close vicinity, may enhance the spectral level, dominance of a particular formant peak is mostly because of the pole pair located in that particular formant frequency. As, for real life speech applications the first three formants are mostly considered, taking only the first three formants into consideration alike previous chapter, the cascaded spectrum representation of (3.2) can be written as

$$\begin{aligned} H(e^{j\omega}) &= \frac{C_i}{\prod_{i=1}^{3}(1-p_i e^{-j\omega})(1-p_i^* e^{-j\omega})} \\ &= H_1(e^{j\omega})H_2(e^{j\omega})H_3(e^{j\omega}) \end{aligned} \tag{3.39}$$

Hence, the ramp cepstrum corresponding to the ORSSACF of (3.39) can be written as follows

$$\mu_h(m) = \mu_{h1}(m) + \mu_{h2}(m) + \mu_{h3}(m) \tag{3.40}$$

According to (3.36) it can be shown that the spectral domain representation of 3.40 is obtained as

$$\begin{aligned} \mu_H(e^{j\omega}) &= F[\mu_h(m)],\ m > 0 \\ &= \mu_{H1}(e^{j\omega}) + \mu_{H2}(e^{j\omega}) + \mu_{H3}(e^{j\omega}) \end{aligned} \tag{3.41}$$

The first formant peak is prominent in the spectrum of ramp cepstrum of ORSSACF presented in Fig. 2.14, indicating that the effect of $\mu_{H2}(e^{j\omega})$ and $\mu_{H3}(e^{j\omega})$ are negligible on $\mu_{H1}(e^{j\omega})$. Using this property, it can be assumed that the output response closely match $\mu_{H1}(e^{j\omega})$ around the first formant peak. Thus instead of conventional peak picking, in this chapter, the task of formant estimation is carried out through spectral model fitting, which ensures that both the frequency and bandwidth of formant peaks are matched.

However, in noisy environments, presence of spurious peaks may cause difficulties in identification of formant peaks even in the case of band limited signals. As discussed in the previous section, the autocorrelation operation can reduce the effect of noise. Moreover, performing the ramp cepstrum operation on the ORSSAC sequence will definitely exhibit significant noise reduction. In order to identify the formant peaks, especially under noisy condition, one possibility is to consider a transfer function which can produce an impulse response that closely matches the ramp cepstrum of the ORSSACF of the most prominent subsystem whose frequency domain representation is $\mu_{H1}(e^{j\omega})$. By limiting the comparison to only the zone where only the first formant frequency should be present, the spectrum corresponding to that transfer function can then be used in a spectral matching technique along with the spectrum obtained from the ramp cepstrum of ORSSACF of the noise corrupted signal. In this case, the transfer function of the subsystem responsible for the spectrum of the ramp cepstrum of the ORSSACF around the first formant peak as per (2.35) can be represented as

$$\mu_{H1}(e^{j\omega}) = \frac{C_1}{\{(1-p_1e^{-j\omega})(1-p_1^*e^{-j\omega})\}}, \tag{3.42}$$

where $C_1 = 1 - Re[p_1]e^{-j\omega}$.

As, in the previous section a direct relationship between $\mu_Y(e^{j\omega})$ and $\mu_H(e^{j\omega})$ is developed, (3.42) can be used to derive a model for the ramp cepstrum of the ORSSACF of a noisy sequence for the first formant as follows

$$\mu^1_{model}(e^{j\omega}) = \frac{C_1}{\{(1-p_1e^{-j\omega})(1-p_1^*e^{-j\omega})\}}, \tag{3.43}$$

where $\mu^1_{model}(e^{j\omega})$ is the representation for the first band.

In the proposed formant estimation method, a spectral model corresponding to the first formant zone of the spectrum of the ramp cepstrum of ORSSACF of the speech signal is introduced, which is utilized in a model matching technique to find out the model parameters that in turn will provide the first formant frequency. In what follows the proposed approach of model matching will be elaborated in detail where each formant will be estimated once at a time. In the estimation of each formant, one such model corresponding to that specific formant is required. Similar to (3.43) for the first formant, for estimating each formant one such model is required and the i-th model can be represented as

$$\begin{aligned} \mu^i_{model}(e^{j\omega}) &= \frac{C_i}{\{(1-p_ie^{-j\omega})(1-p_i^*e^{-j\omega})\}}, \\ p_i &= r_ie^{j\theta_i},\ p_i^* = r_ie^{-j\theta_i} \end{aligned} \tag{3.44}$$

The spectrum $\mu^i_Y(e^{j\omega})$ of the ramp cepstrum of the ORSSACF of the observed noisy signal $y(n)$ is used in conjunction with the proposed model $\mu^i_{model}(e^{j\omega})$ to form an objective function and for the first formant with $i = 1$ based on the absolute difference of these spectra, namely

$$e^i_{min}(r_j,\theta_j) = \underset{\substack{r_l<r_i<r_h \\ \theta_l<\theta_i<\theta_h}}{min} \sum_{\omega=\omega_{lc}}^{\omega_{hc}} \left(|\mu^i_{model}(e^{j\omega})| - |\mu^i_y(e^{j\omega})|\right) \tag{3.45}$$

Note that here the superscript i is introduced to control the step by step algorithm. The algorithm for the first formant where $i = 1$, is given below in brief.

1. From given noisy speech $y(n)$ computing $\mu^i_Y(e^{j\omega})$ using (2.37)
2. Generating $\mu^i_{model}(e^{j\omega})$ using the model of (3.44)
3. Minimizing the objective function in (3.45) within a restricted frequency range ω_{lc} to ω_{hc} which depends on the range of each formant zone.

One may utilize the $-3dB$ points on the lower and higher sides of the peak in the spectrum of the model to extract ω_{lc} and ω_{hc}. Within that specified range $\omega_{lc} \leq \omega \leq \omega_{hc}$, the optimum values of the two variables r_i and θ_i are obtained at the minimum of absolute differences. Based on the fundamental knowledge of traditional range of formants, one may restrict the search range for the two variables i.e., $r_l \leq r \leq r_h$ and $\theta_l \leq \theta \leq \theta_h$ or adopt a coarse and fine search approach [36]. Formant frequencies are estimated from the pole angle θ_j that produces the best match between the spectra using (3.45) .

Once the first formant frequency $F1$ is obtained, (3.44) is utilized to estimate the second formant frequency $F2$. $\mu^i_Y(e^{j\omega})$ can be written as the sum of $\mu^1_Y(e^{j\omega})$, $\mu^2_Y(e^{j\omega})$ and $\mu^3_Y(e^{j\omega})$ alike (3.41). From the magnitude spectrum of $\mu^i_Y(e^{j\omega})$ the estimated model spectrum $\mu^1_Y(e^{j\omega})$ is subtracted such that the resulting spectrum closely resembles the sum of $\mu^2_Y(e^{j\omega})$ and $\mu^3_Y(e^{j\omega})$. Hence $\mu^i_Y(e^{j\omega})$ in general for estimating second and third formant can be expressed as

$$\mu^i_Y(e^{j\omega}) = \mu^{i-1}_Y(e^{j\omega}) - \mu^{i-1}_{model}(e^{j\omega}),\ i > 1 \tag{3.46}$$

Then similar to the matching in the first formant zone, matching is performed in the second formant zone and $F2$ is estimated. Then from the magnitude spectrum of $\mu^2_y(e^{j\omega})$

the estimated model spectrum $\mu^2_{model}(e^{j\omega})$ is subtracted to obtain $\mu^3_y(e^{j\omega})$. According to the simplified modeling of the vocal tract presented above, $\mu^3_y(e^{j\omega})$ should closely match with $\mu^3_{model}(e^{j\omega})$, leading to a similar approach as described in (3.44) and (3.45) to obtain $F3$.

3.2.1 Formant Based Vowel Recognition

After estimating formants in this manner, in the proposed scheme they are employed in vowel recognition as features along with the commonly used mel frequency cepstral coefficients (MFCC) coefficients. Linear discriminant analysis (LDA) based classifier is used to accomplish this task. For our proposed scheme, a frame by frame classification method is used, which offers vowel recognition results for each voiced frame independently.

The classifier classifies the data into different groups generally, depending on the significant characteristics of the group members. The quality of a classifier depends on its ability to provide the compactness among the member within a cluster and the separation between the members of different clusters in terms of feature characteristics. The task of recognizer is to identify the class label of a test sample utilizing the classified data. In a feature based scheme, classification is performed utilizing the extracted features of the data, instead of directly employing the data themselves. In the proposed method, the LDA is used to classify the vowel among the different classes (in our case, vowel) available. In LDA, a linear projection is determined that maximizes a ratio between the signal, represented by the projected inter-cluster distance and the noise, represented by the projected intra-cluster variance. Here the objective function is based on determining a projection direction w to maximize the Fisher's discriminant defined as

$$J(w) = \frac{w^T S_b w}{w^T S_w w} \tag{3.47}$$

where S_w and S_b are within- and between-class scatter matrices, respectively [37].

3.3 Results and Simulation

In order to evaluate the recognition performance of the proposed methods, numerous experiments have been conducted on the TIMIT acoustic-phonetic continuous speech corpus, which has jointly been developed by Massachusetts Institute of Technology (MIT), Stanford Research Institute (SRI) and Texas Instruments (TI) [38]. The TIMIT database contains a large collection of sentences uttered by both male and female English speakers using various dialects. A total of 6300 sentences, with 10 sentences spoken by each of the speakers are present on the database. Voiced and unvoiced portions of speech are clearly marked on accompanying phone files. However, as TIMIT does not contain reference values of formants, to compare estimated results, the most commonly used formant database is chosen, where formant frequencies are estimated based on vocal tract resonances (VTR) with manual correction [39]. The formant estimates reported in [39] are taken as ground truth and the estimation performance of different methods is evaluated at different levels of signal to noise ratios (SNR). This VTR subset of TIMIT database contains 376 sentences across the training set, representing 173 speakers. These sentences contain 18 voiced phonemes, out of which, the diphthongs have been ignored, and 11 phonemes are considered. A total of 2726 utterances of phonemes are used from the VTR subset, out of which 1583 are from male and 1143 are from female speakers, have been analysed. In VTR database, formant estimates are reported for every 10 ms interval. However, vowel duration is generally much larger than 10 ms. In the frame by frame formant analysis, when the size of analysis frame is larger than 10 ms, the estimated formants are then compared with the average VTR formant values obtained over the different 10 ms frames within the duration of that formant under investigation. For the purpose of performance comparison, first the most widely used LPC based formant estimation method [40] is chosen, where the order of the LPC is chosen as 12. Apart from the LPC method, a state of the art adaptive filter bank (AFB) method is also chosen. In the AFB method, formant estimation is carried out in sample by sample basis, and for the purpose of comparison, average estimated formant values over a period is considered [23].

Table 3.1: Performance comparison in terms of mean error(%) for synthetic speech

Vowels		$5dB$			$-5dB$		
		Proposed	LPC	AFB	Proposed	LPC	AFB
/a/	F1	3.58	20.24	46.90	3.14	20.46	49.77
	F2	9.45	65.23	32.58	7.75	113.79	30.99
	F3	16.82	17.80	8.45	4.58	34.02	9.84
/o/	F1	12.44	49.53	128.07	15.10	78.29	18.29
	F2	3.21	138.88	20.42	26.04	133.29	46.61
	F3	14.85	39.93	9.56	5.38	36.28	12.53
/u/	F1	12.31	72.96	109.00	13.78	98.29	12.98
	F2	4.32	116.33	14.62	20.30	121.92	33.72
	F3	13.06	52.31	11.40	2.91	40.60	13.74

In the proposed model fitting scheme, the range of the model parameters are set according to the general behavior of the vocal tract. The possible range of the parameter r is changed within the limit 0.8 to 0.99, which covers even a very rapidly decaying impulse for the purpose of our simulation. The search range for θ is set according to the determined formant band. Search resolutions for r and θ are chosen as$\triangle r = 0.01$ and $\triangle\theta = 0.001\pi$, respectively. In our experiments in order to obtain a noisy signal, noise sequence of a particular SNR is added with the clean (noise-free) signal. Noisy signals are generated according to 3.12, where the noise variance σ_v is appropriately determined according to a specified level of SNR defined as

$$SNR = 10log_{10}\frac{\sum_{n=0}^{N-1} x(n)^2}{\sum_{n=0}^{N-1} v(n)^2} \tag{3.48}$$

At first results for three synthetic vowels /a/, /o/ and /u/ are presented in Table 3.1. Vowels with duration of 80 ms are synthesized using the Klatt synthesizer considering the pitch values of 220 Hz . Estimation eroor for the first three formants are taken into consideration after performing estimation for 10 independent trials. The estimatin error is shown for the three synthesized vowels a SNRs of $5dB$ and$-5dB$ for both male and female sounds, respectively. It is clearly observed that the proposed method is able to reduce estimation error significantly in comparison to the other methods, even with an

increase in the level of background noise.

Table 3.2: Performance comparison in terms of mean error(%) for male speakers

Vowel		-5 dB			5 dB		
		Proposed	AFB	LPC	Proposed	AFB	LPC
	F1	16.33	30.88	30.53	16.37	17.74	26.48
/aa/	F2	13.10	36.42	82.19	9.38	21.87	45.44
	F3	20.06	15.47	43.35	11.15	17.07	39.80
	F1	16.07	24.12	31.64	15.43	16.31	24.65
/ah/	F2	9.62	28.88	57.43	9.16	24.41	35.57
	F3	16.90	13.09	39.21	11.11	11.61	37.72
	F1	20.20	35.49	22.63	17.14	37.77	19.72
/ow/	F2	15.66	26.03	47.20	11.65	24.65	41.67
	F3	18.02	14.20	36.68	11.86	14.00	37.74
	F1	17.44	36.49	20.14	16.84	36.55	19.49
/uh/	F2	12.20	23.49	38.02	11.92	23.23	37.48
	F3	11.90	13.89	37.24	11.50	13.86	37.33
	F1	16.76	36.72	29.66	16.55	39.58	20.09
/uw/	F2	13.12	23.14	40.36	12.63	22.49	36.45
	F3	14.09	14.53	39.48	11.37	14.50	38.25

The estimation errors obtained by the proposed method and that by the other two methods are presented, under the influence of white gaussian noise conditions for male and female speakers of TIMIT database, in Tables 3.2 and 3.3 . Here the estimation error, the mean average deviation between the estimated formant frequency f_E and the reference formant frequency f_R is defined as

$$E = |\frac{f_E - f_R}{f_R}| \times 100\% \tag{3.49}$$

For Table 3.2and Table 3.3SNR levels $5dB$ and $0dB$ are considered. For each vowel, the estimation errors for three different formants, namely $F1$,$F2$ and $F3$ are listed. As can be seen from the tables, the proposed method offers better performance than both the12 order LPC and the AFB methods under presence of background noise. It can be observed that the estimation error obtained by the proposed method in comparison to

Table 3.3: Performance comparison in terms of mean error(%) for female speakers

Vowel		-5 dB			5 dB		
		Proposed	AFB	LPC	Proposed	AFB	LPC
	F1	12.79	19.40	45.04	13.19	12.33	25.84
/aa/	F2	13.60	69.07	27.67	10.79	21.02	20.53
	F3	13.62	30.96	12.77	12.53	20.79	11.95
	F1	14.17	14.68	36.14	13.47	10.79	17.91
/ah/	F2	17.30	37.14	21.46	17.86	13.37	20.44
	F3	12.01	23.14	15.75	10.90	19.79	12.44
	F1	12.08	11.75	26.75	12.61	10.82	15.70
/ow/	F2	17.49	43.25	26.84	17.52	27.30	20.96
	F3	12.69	18.63	15.23	11.02	17.84	10.00
	F1	13.07	12.91	18.07	13.13	10.54	16.46
/uh/	F2	17.90	23.46	21.40	19.11	18.72	20.04
	F3	11.63	18.93	10.40	10.27	18.40	9.54
	F1	13.25	9.47	16.67	13.13	9.12	16.37
/uw/	F2	19.02	17.18	20.33	19.07	16.46	18.49
	F3	11.96	18.12	9.82	10.76	18.19	9.42

that of the other methods is extremely lower in such severe noisy conditions.

In some cases it is found that the estimation accuracy decreases for the cases when the two formants are very closely spaced, for example in case of vowel /ih/, though, considering the level of noise, the estimation accuracy obtained by the proposed method is quite acceptable. It is also observed that the estimation error relatively increases in case of high pitch female speakers. It is clearly observed that the estimation performance for the third formant, which is by nature very difficult to estimate because of low spectral magnitude, is significantly enhanced by the proposed method. Hence, overall it can be said that, the proposed method increases formant estimation performance.

In order to present the overall formant estimation errors over a large range of SNRs considered in the experimental setup, in Fig. 3.12 the overall estimation error for all vowels are shown. In a simiar way, in order to present the overall formant estimation errors over a large range of SNRs considered in the experimental setup, in Fig. 3.13 the overall estimation error for all vowels are shown. It is observed that the formant

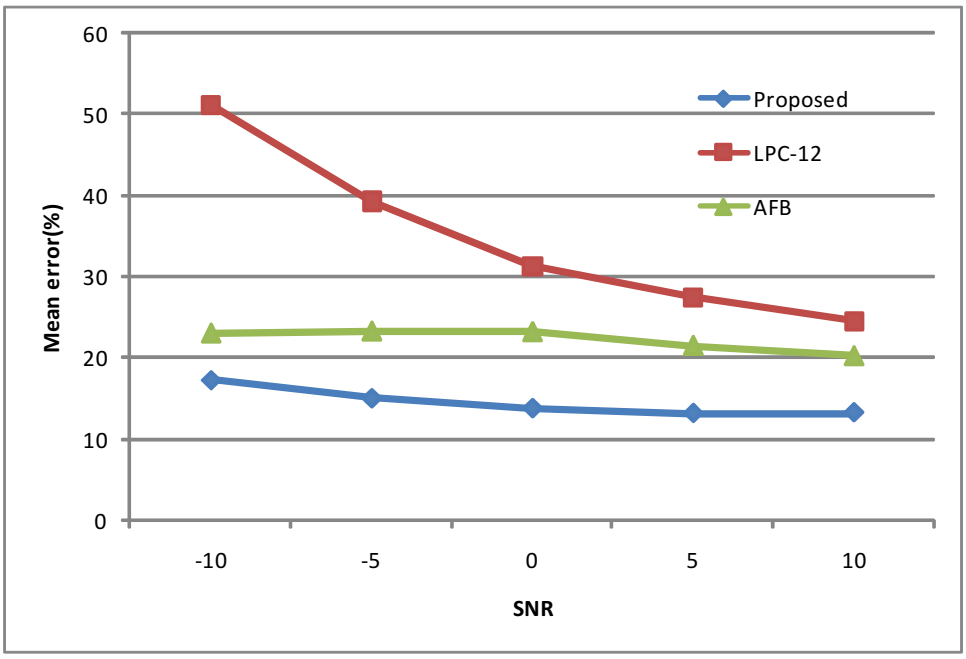

Figure 3.12: Error comparison of all formants for different methods

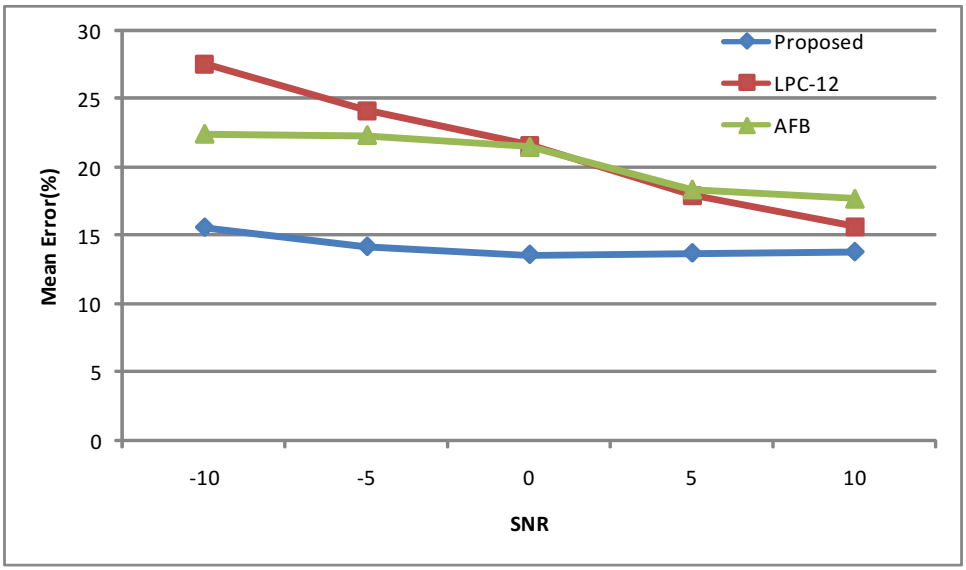

Figure 3.13: Error comparison of all formants for different methods for female

estimation performance obtained by the three methods remains similar in case of high level of SNR. However, with the decrease in SNR level, the estimation performance of the other two methods deteriorates in comparison to that of the proposed method. The performance of the proposed method remains quite consistent even in the low levels of SNRs and level of performance degradation is not very significant till -15 dB. However, beyond that the performance of the proposed method is not satisfactory because of the severe noise corruption, leading to complete failure for the conventional methods.

In the proposed method formant estimation is carried out frame by frame with a frame length of 512 samples and 10 ms overlap between the successive frames. As a result for a vowel sound of duration of about 80 ms, 5 frames are analyzed. It is to be noted that, because of the inherent characteristics of the fast Fourier transform (FFT) operation, there exists an inherent error caused by the minimum width of the FFT bin. For instance, when a 512 point FFT is performed on a speech frame with sampling frequency of 16 kHz, the resulting FFT has a resolution of 15.6 Hz.

Table 3.4: Recognition Accuracy

	Feature Vector	
SNR	MFCC + Proposed Method	MFCC + LPC-12
-10 dB	66.67	60.00
-05 dB	85.00	80.00
05 dB	90.00	85.33

By incorporating the estimated formants in a feature vector along with traditional MFCC, significantly better vowel recognition accuracies are achieved compared to a feature vector consisting of MFCC and formants estimated by LPC, especially under the influence of noise. By using these formants along with the traditional 12 MFCC coefficients as a feature vector, vowel recognition was performed for the vowels /aa/, /ux/ and /ix/ from the TIMIT database. As formant ranges for male and female vowels vary significantly, they are considered as separate classes for this LDA based classification operation. There are 20 utterances for male and 20 utterances for female considered for each vowel. Accuracies are calculated by leaving one sample out while training the classifier

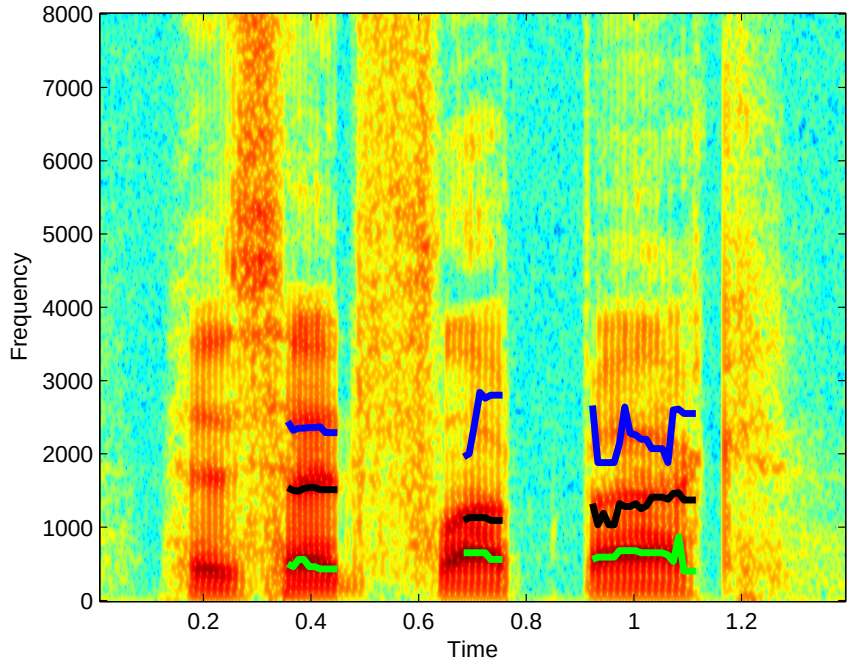

Figure 3.14: A spectrogram of the sentence "His head flopped back" with tracked formant by the proposed method at 0 dB SNR

and then testing the left out sample. This check is performed for all the samples in the database, and it is found that the proposed feature vector offers better performance in noisy conditions. The recognition accuracies for different vowels is presented in Table 2.4. It can be concluded from the table that the proposed noise robust formant estimation method, when used for vowel recognition, increases the recognition accuracy for vowel recognition systems under the influence of noise.

As seen from these analyses, the proposed method offers a better performance over the LPC and AFB methods in noise free as well as in noisy conditions. In order to demonstrate the effectiveness of our proposed method, a spectrogram of the sentence "His head flopped back", uttered by a male speaker taken from the TIMIT database is shown in Fig. 3.14. The formant frequencies estimated at different frames using the proposed method uder SNR= 0 dB are shown over the spectrogram of clean speech. In the tracking, only the estimated formants of the vowels are shown. It can be observed from the figure that the proposed method tracks the formant frequencies quite accurately even in noisy speech.

3.4 Conclusion

In this chapter, the advantageous effect of double autocorrelatioon in comparison to the single autocorrelation is explored. It is found that the once repeated ACF is capable of providing better noise reduction and spectral peak strengthening. It is shown that similar to the previous chapter, in order to estimate the formants a residue based spectral domain model matching scheme is employed where the proposed spectral model is fitted with the spectrum of the spectrum of the ORSSACF of noisy speech. In order to reduce the computational burden, in the proposed residue based spectral matching scheme fitting operation is carried out for each formant separately restricting the search within a formant zone. Formant estimation is performed considering different vowels uttered. Estimation performance of the proposed method is compared with the widely used LPC and AFB formant estimation methods.

Chapter 4

Spectral Domain Ramp Cepstrum Model Using Band Limted Speech Signals

The objective of this chapter is to develop a noise robust formant estimation scheme considering the band limited signals. As shown in the previous chapter that the spectral peak strengthening effect of the correlation or repeated correlation operation is more prominent in the spectral region containing dominant poles. Hence, in this chapter, the spectral models are developed considering band-limited signals where the bands are chosen in such a manner that in each band mainly the formant peak dominates. Spectral models of RC of ORSSACF are derived for the band limited signal. However, instead of the residue based iterative spectral model fitting appraoch used in the previous chapter a direct spectral error minimization scheme is used to determine the model parameters. It is shown that the proposed method can efficiently tackle the adverse effect of observation noise and provide an accurate estimate of formant frequencies of speech signals.

4.1 Methodology

4.1.1 Background

As shown in previously, vocal tract can be assumed to be a causal, stable, linear time-invariant and stationary autoregressive (AR) system, and thus a voiced speech signal constructed from it can be characterized as

$$x(n) = -\sum_{k=1}^{P} a_k x(n-k) + Gu(n), \tag{4.1}$$

where $\{a_k\}$ are the system parameters, G denotes the gain factor, P is the known system order and $u(n)$ represents the excitation to the system. The AR system transfer function $H(z)$, which in this case is the vocal tract transfer function can be expressed as

$$H(z) = \frac{G}{1 + \sum_{k=1}^{P} a_k z^{-k}} = \frac{G}{\prod_{k=1}^{P}(1 - p_k z^{-1})}, \tag{4.2}$$

where $p_k = r_k e^{j\omega_k}$ denotes the k-th pole of the AR system with magnitude r_k and angle ω_k. Formants are associated with the free resonances of the vocal tract system. In order to model each formant, a pair of complex conjugate poles is required. In (4.2), each formant corresponds to p_k and it's conjugate. Thus, for a vocal tract system modeled with P-th order AR system, there exists $P/2$ formants. Formant frequency (F_k) and bandwidth (B_k) can be expressed in terms of pole parameters as [30]

$$F_k = \frac{F_s}{2\pi}\omega_k; B_k = -\frac{F_s}{\pi} ln(r_k), \tag{4.3}$$

where F_s is the sampling frequency.

In the LPC based methods, the ACF of the given speech signal $x(n)$ is used in the Yule-Walker equations to obtain the AR parameters and thereby the poles of the vocal tract system and from the estimated poles, formants are calculated. But in the presence of observation noise, LPC based methods fail to provide an accurate estimate of the AR parameters and thus exhibits poor formant estimation accuracy. Moreover, the effect of pitch variation may cause significant errors in the LPC based formant estimation. Hence,

development of a formant estimation scheme, which can estimate the formants with a higher accuracy even in the presence of severe background noise as well as handle the effect of pitch variation is in great demand.

As introduced in the previous chapter, in order to reduce the effect of pitch from the speech signal, cepstrum that offers the advantage of homo-morphic de-convolution has been most commonly used. The principle of homomorphic deconvolution helps in separating signals that have been combined via convolution and thus it become a very important tool in different speech processing applications, such as speech recognition. The complex cepstrum of a signal $h(n)$ is defined as [24]

$$c_{hc}(n) = F^{-1}\left\{ln(H(e^{j\omega}))\right\}, \tag{4.4}$$

where $F^{-1}\{.\}$ denotes the inverse Fourier transform and the spectrum of $h(n)$ is given by $H(e^{jw})$. Considering the VT system as minimum phase, using (3.2) $c_{hc}(n)$ is a sequence that is real and causal. Recalling from the previous chapter,$c_{hc}(n)$ can be expressed in terms of poles as

$$c_{hc}(n) = \sum_{i=1}^{P} \frac{p_i^n}{n}, n > 0. \tag{4.5}$$

On the other hand, the real cepstrum of $h(n)$ is defined as

$$c_h(n) = F^{-1}\left\{ln(|H(e^{j\omega})|)\right\}. \tag{4.6}$$

In order to avoid notational complexity, instead of denoting real cepstrum as $c_{hr}(n)$, simply $c_h(n)$ is used, i.e. an additional subscript 'r' is not used hereafter. For $n > 0$, the relation between complex and real cepstra is given by $c_h(n) = 0.5c_{hc}(n)$. Hereafter, alike previous chapter, only real cepstrum shall be considered. In order to avoid logarithm of negative values, in practical applications real cepstra is most commonly used. Here, for real cepstrum (4.5) can be written as

$$c_h(n) = 0.5 \sum_{i=1}^{P} \frac{p_i^n}{n}, n > 0. \tag{4.7}$$

The speech signal $x(n)$ given by (3.1) can be considered as a convolution sum between $h(n)$, the impulse response of the V.T. system and $u(n)$, the excitation to the V.T. system as follows

$$x(n) = h(n) * u(n). \tag{4.8}$$

Thus, one can write the corresponding cepstral representation of $x(n)$ as

$$c_x(n) = c_h(n) + c_u(n). \tag{4.9}$$

Here, $c_h(n)$ is the cepstrum of $h(n)$, $c_u(n)$ is the cepstrum of the excitation $u(n)$ and $H(e^{j\omega})$ and $U(e^{j\omega})$ are frequency domain representations of $h(n)$ and $u(n)$. As periodic impulse-train excitation is commonly considered to model the voiced sounds, here a periodic impulse-train excitation $\{u(n)\}_{n=0}^{N-1}$ with period T is considered, which can be expressed as

$$u(n) = \sum_{k=0}^{\lambda-1} \delta(n - kT), \lambda = \lceil N/T \rceil. \tag{4.10}$$

Here, λ is the total number of impulses within the excitation. As introduced in the previous chapter, based on (4.6), utilizing this advantage of homomorphic deconvolution, cepstral domain system identification methods have been proposed which deal with the noise free environment . It is apparent from(4.5) and (4.9) that in order to extract the system poles p_i, from a given $c_x(n)$ one needs to extract $c_h(n)$. However, as stated in (4.9) , $c_h(n)$ is mixed with $c_u(n)$ resulting $c_x(n)$. In case of periodic impulse train excitation, it can be shown that $c_u(n)$ contributes to $c_x(n)$ at the origin and periodically after each T interval. Thus, even in case of finite data analysis, within the range $0 < n < T$, the

effect of $c_u(n)$ can be neglected, resulting

$$c_x(n) = c_h(n) = 0.5 \sum_{i=1}^{P} \frac{p_i^n}{n}, 0 < n < T \tag{4.11}$$

It indicates that the cepstral coefficients corresponding to $c_x(n)$ can be considered similar to that of $c_h(n)$ within the range $0 < n < T$.

In the presence of additive white Gaussian noise (AWGN) $v(n)$, the observed signal $y(n)$ can be expressed as

$$y(n) = x(n) + v(n), \tag{4.12}$$

where $v(n)$ is assumed to be zero mean stationary and independent of $u(n)$. The cepstral coefficients of $y(n)$ can then be expressed as

$$\begin{aligned} c_y(n) &= F^{-1}\{ln(|X(e^{j\omega})|)\} + F^{-1}\{ln(1 + \frac{|V(e^{j\omega})|}{|X(e^{j\omega})|})\} \\ &= c_x(n) + c_w(n). \end{aligned} \tag{4.13}$$

$c_w(n)$ appears in the presence of noise and vanishes in its absence. As shown in the previous chapter, at severe noise it is very difficult to get an accurate estimate of $c_x(n)$ from $c_y(n)$, since the cepstrum decomposition techniques are very sensitive to the noise level. As a result, it is desirable to develop an algorithm that can reduce the effect of noise on the signal, thereby reducing the effect of $c_w(n)$ on $c_y(n)$ and producing more noise robust cepstral coefficients. In this regard, we propose to investigate the effect of increasing the number of poles on the formant location to enhance the strength of the formant peaks.

In view of enhancing the spectral peaks corresponding to a particular frequency, one possible approach would be to introduce new poles having that frequency. In particular, if the new poles can be generated exactly at the same location of those original poles, the spectral peak corresponding to that pole location will be significantly enhanced. As only the speech signal is available at hand and the VT transfer function can not be

changed, it is not possible to place poles at designated places to enhance spectral peaks. As an alternate, if a signal is convolved with its folded version new poles would be introduced, which should be related to the original system poles. An equivalent approach is to achieve this effect by simply doing the autocorrelation operation on the signal. The autocorrelation function (ACF) of $x(n)$ is defined as

$$\begin{aligned} r_{xx}(m) &= x(n) * x(-n) \\ &= E[x(n)x(n+m)]. \end{aligned} \tag{4.14}$$

Here $E[.]$ denotes the expectation operator.

As, shown in the previous chapter, according to (4.14) and (3.8), the z transform of $r_{xx}(n)$ can be expressed as

$$R_{xx}(e^{j\omega}) = R_{hh}(e^{j\omega}) \times R_{uu}(e^{j\omega}), \tag{4.15}$$

where $R_{hh}(e^{j\omega})$ and $R_{uu}(e^{j\omega})$ are the frequency domain representations of $r_{hh}(n)$ and $r_{uu}(n)$, the ACFs corresponding to $h(n)$ and $u(n)$, respectively. According to the definition (4.14), $R_{hh}(e^{j\omega})$ can be written as

$$R_{hh}(e^{j\omega}) = H(e^{j\omega}) \times H(e^{-j\omega}). \tag{4.16}$$

Using (3.2), interms of poles $R_{hh}(e^{j\omega})$ can be expressed as

$$R_{hh}(e^{j\omega}) = \frac{C_1}{\prod_{i=1}^{P}(1 - p_i e^{-j\omega})(1 - p_i^* e^{j\omega})}. \tag{4.17}$$

Here for each pole $p_i = r_i e^{j\theta}$, there exists a pole $1/p_i^*$ which is placed at conjugate reciprocal locations. From (4.17) it is clearly seen that total number of poles in $R_{hh}(e^{j\omega})$ is $2P$, which is twice as the number of poles in $H(e^{j\omega})$. Due to the autocorrelation operation new P poles are introduced in $R_{hh}(e^{j\omega})$ which are conjugate reciprocal to the original P poles of $H(e^{j\omega})$, i.e. the new poles are located at the original pole angles as

expected.

Using (4.12) and (4.14), the ACF of noisy speech $y(n)$ can be expressed as

$$r_{yy}(n) = r_{xx}(n) + r_{ww}(n). \tag{4.18}$$

where,

$$r_{ww}(n) = r_{vv}(n) + r_{vx}(n) + r_{xv}(n). \tag{4.19}$$

Here $r_{vv}(n)$ is the ACF of noise $v(n)$and $r_{vx}(n)$ and $r_{xv}(n)$are the cross correlation terms. Since $v(n)$ is uncorrelated with $x(n)$, it is expected that the values of the cross-correlation terms, in comparison to that of $r_{xx}(n)$,will be negligible. On the other hand, the ACF of the AWGN $v(n)$ generally exhibits a peak at the zero lag and the values at all other lags should be very small and ideally should be zero.

4.1.2 Peak Enhancement By Repeated ACF

Realizing the effect of spectral peak strengthening and reduction of noise due to auto-correlation as described in the previous section, we propose to generate more poles at the location of the original poles to further strengthen the spectral peaks. In view of achieving this objective, the ACF operation can be repeated, which not only strengthens the dominant peaks but also preserves pole locations. Performing further autocorrelation operation on an ACF of a noise corrupted speech signal will imitate duplication of poles at the original locations of the system. Hence, the resulting double correlated signal is expected to exhibit more noise immu-nity and in its spectrum, even under heavy noisy condition, the formant peaks will be significantly enhanced. Considering the same natural sound /eh/ as shown in Fig. 2.2, the spectral domain effect of repeated ACF on this speech signal is shown in Fig.4.1. It is observed from this figure that because of the repeated ACF the formant peaks become enhanced in all the formant positions. The enhancement of the first formant is quite prominent in Fig. 4.1. However, for each ACF the dominance of second and third formant peak reduces furter in comparison to the most dominant first peak. Thus simply repeating the ACF would not completely solve the

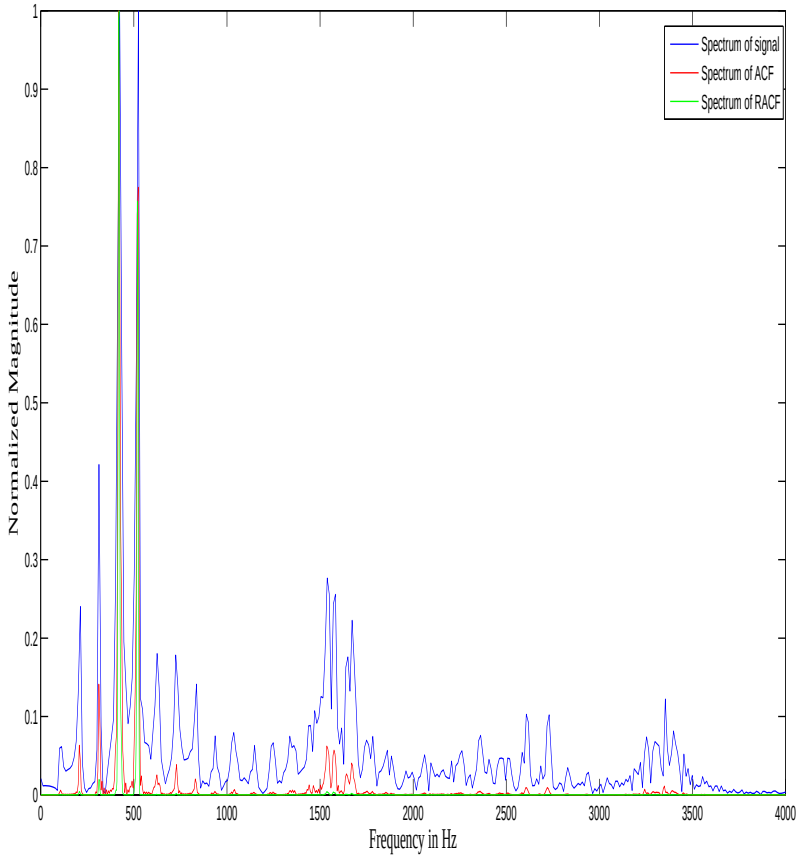

Figure 4.1: Comparison of signal spectra with ACF and RACF

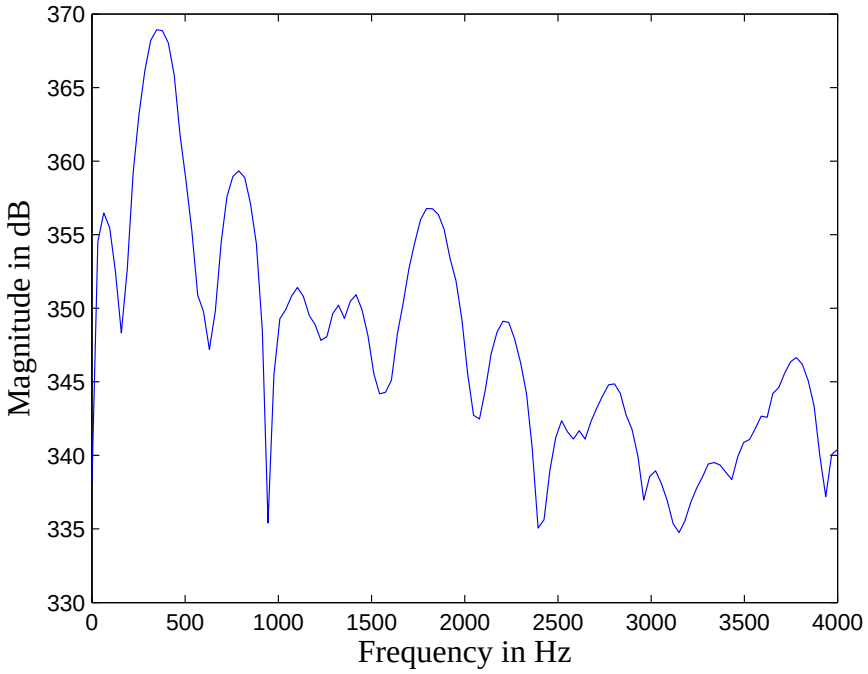

Figure 4.2: Spectrum of RACF of the signal considered in Fig. 4.1 at $SNR = -5dB$

problem of estimating second and third formant erroneously in severe noisy environment.

In Fig. 4.2, spectrum of RACF of the signal considered in Fig. 4.1 at 0dB SNR, where the magnitude is shown in log scale. From this Fig. it is clearly observed that in comparison to the spectra corresponding to $y(n)$, the first peak in the spectra corresponding to $\rho_{yy}(n)$ exhibits a extremely large peak in comparison to other peaks. One major concern in double autocorrelation operation is that it makes the effect of a strong pole more stronger shadowing the effect of relatively weak poles. This phenomenon is also observed in Fig. 4.1 and in Fig. 4.2. In comparison to the increase in the first formant peak, the spectral peaks corresponding to other formants remain very weak. This becomes a great problem in case of severe noise if spectral peak picking is used for formant estimation. In that case, several spurious peaks may appear in the spectrum with magnitudes greater than the desired peaks. In view of overcoming this problem, one practical solution is to divide the full band signal into a number of sub-bands. The sub-bands should be formulated in such a way that each sub-band corresponds to approximately one formant, in other words it should contain the effect of one dominant pole pair only. Number of sub-bands to be made depends on the number of formants to be estimated. Higher formants become increasingly weak due to their low energy concentration and the tilt caused by the lip radiation. Thus, the first three formants are mostly considered for real life applications. Unlike conventional formant analysis methods, in this paper, the task of formant estimation is carried out on the band-limited speech signal instead of the full-band signal.

4.1.3 Banding of the full-band signal

Performing autocorrelation on a speech segment significantly increases the strength of the most dominant peak with respect to other peaks, thus amplifying the effect of first formant with respect to other formants on the spectrum of a voiced speech segment. Although the autocorrelation operation can significantly reduce the effect of noise on the first formant peak, it obscures the second and third formant peaks. In order to overcome this problem, a method of localized searching for each formant based on filtered speech

signal is proposed. It offers the advantage of dealing with a band limited speech signal possessing only one dominant peak within a band. In this regard, a set of band-pass filters must be employed to extract the band-limited signal from the per-processed speech signal, where each filter corresponds to a conventional band of frequency for respective formants. It is expected that the filters utilized for the purpose of band-limiting exhibit sharp cut-offs and low pass-band ripples. The main advantage of dealing with a band-limited signal for extracting a specific formant lying in a particular band is its robustness against the interference of nearby formants and other spurious frequencies that may exhibit in the presence of noise. The band-limited signal is obtained by applying band pass filters that are tuned to the first three formant frequency bands.The z-transform of the band-limited signal $x_i(n)$ obtained by using the i-th filter transfer function $B_i(z)$ is given by

$$X_i(z) = X(z)B_i(z) \tag{4.20}$$

In the proposed method, in order to obtain the sharp cutoff and low ripple while keeping the filter order low, instead of using a bandpass filter, separate lowpass and highpass filters are employed. In view of designing the required bandpass filter, highpass and lowpass filters are used in cascade. Different types of filters with varying filter orders are tested. It is found that the elliptic filters with order 10 can provide the most satisfactory filter characteristics. In case of cascaded configuration, the filter transfer function $B_i(z)$ can be represented as

$$B_i(z) = B_{ih}(z)B_{il}(z) \tag{4.21}$$

where $B_{ih}(z)$ and $B_{il}(z)$ correspond to the transfer function of the highpass and low-pass filters, respectively.

As mentioned previously, it is more insightful to investigate the effects of filtering on the impulse response of the vocal tract system instead of the speech signal for the purpose of formant estimation. In that case, within a particular formant band, if the effect of frequency peaks outside the band is neglected, one can assume that a pair of pole of the

vocal tract system is mainly responsible for the frequency spectrum of a band-limited signal. As a result, the spectrum corresponding to the bad-limited signal, denoted by $X_i(e^{j\omega})$, will exhibit formant peaks at exactly the same location of the spectrum for $H_i(z)$ where it is assumed that the bandlimiting operation on $H(z)$ with the i-the filter produces $H_i(z)$. It is to be mentioned that the DACF operation which offers more peak-strengthening effect in comparison to the SACF, is more capable of handling the severe noisy condition. Thus before performing the autocorrelation operation on the speech frame, it would be definitely advantageous to extract the band-limited signal containing only the region that is directly associated with a single formant. However, formant frequencies and bandwidths vary widely between different phonemes, and across genders. Therefore, the upper and lower cutoffs for the filters have to be adjusted for frequency domain characteristics of individual frames. First each formant band is selected as per the conventional global formant band limits expected to be suitable for all voiced sounds [1], which are typically broad frequency bands. Within such a wide band, the region of interest for searching the formant could be a smaller zone containing higher spectral energy. In the proposed method, instead of considering the broad bands, a spectral energy based adaptive searching is carried out to determine such narrow bands, which are then used in the model matching algorithm for formant estimation.

In this approach, problems arise due to overlapping formant zones. For instance, for the phonemes uttered by female speakers, in case of /u/ the second formant is at around 950 Hz, and the third formant is at around 2600 Hz, while for /i/, the second formant is at around 2800 Hz and the third formant is at around 3300 Hz. On the other hand, for male /u/, the first three formants are located at around 400 Hz, 950 HZ and 2200 Hz. Therefore setting up a hard limit for formant boundaries is not a good approach, rather an adaptive band limiting algorithm is required. The proposed adaptive band selection algorithm consists of two major steps, namely, gender detection and correction of false band selections. One major advantage of prior gender detection is that it greatly reduces the complexity arising due to overlapping formant ranges. Even then, situations may arise when no formants are present within the broad search area. Then the selected

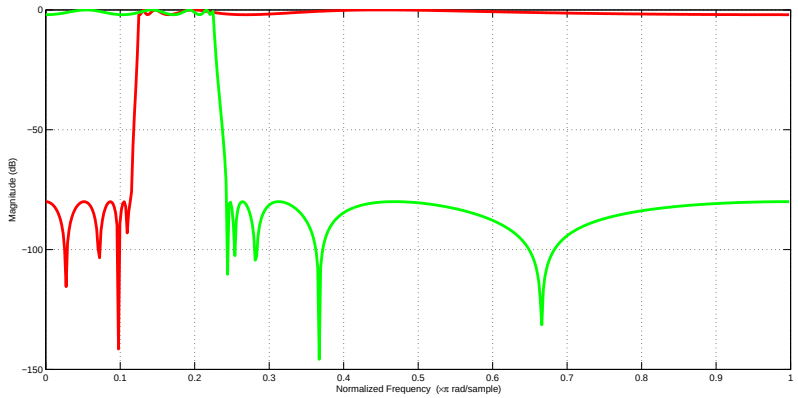

Figure 4.3: Response of a typical BPF used for filtering devised from a cascaded combination of a low pass and a high pass filter

high energy frequency zone eventually may not provide an estimate of the true formant. Once the three high energy frequency zones are selected, an adaptive control algorithm is developed to avoid false zone selection. Due to the natural spectral roll off, spectral energy around the formant decreases with the increase in frequency. In view of utilizing such spectral energy property, the pre-emphasis operation is avoided. According to this property, if the estimated third formant zone contains higher spectral energy compared to that of the estimated second formant zone, the estimated second formant zone is considered as a false estimation and therefore, the third formant zone is treated as the new estimation for the second one. Then a search for the third formant zone is performed in frequencies higher than the new second formant zone. This ensures that banding works even under extreme cases.

Normally the first formant is the most dominant one, and therefore the effect of banding is not very prominent. The banding is done with the help of one high pass and one low pass filter with sharp cut-offs. The filter response obtained by cascaded combination of these two filters is shown in Fig. 4.3. The filter responses for the three sub bands are shown in Fig 4.4. After applying this filter to the /eh/ waveform whose spectrum was presented on Fig. 4.1 the response of the three banded signals is shown in Fig. 4.5. In Fig. 4.6 ACF of the three sub-band signals is shown and in Fig. 4.7 response of these banded autocorrelation sequences are shown. Performing RACF on the subband

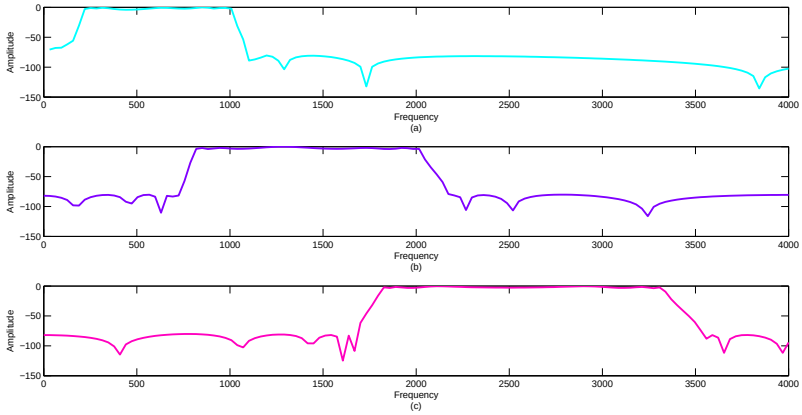

Figure 4.4: Response of three band pass filters used for the three formant bands

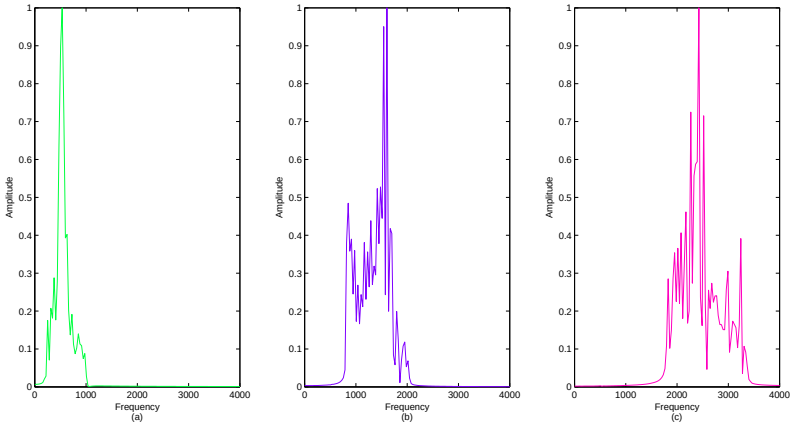

Figure 4.5: Response of the three band limited signals

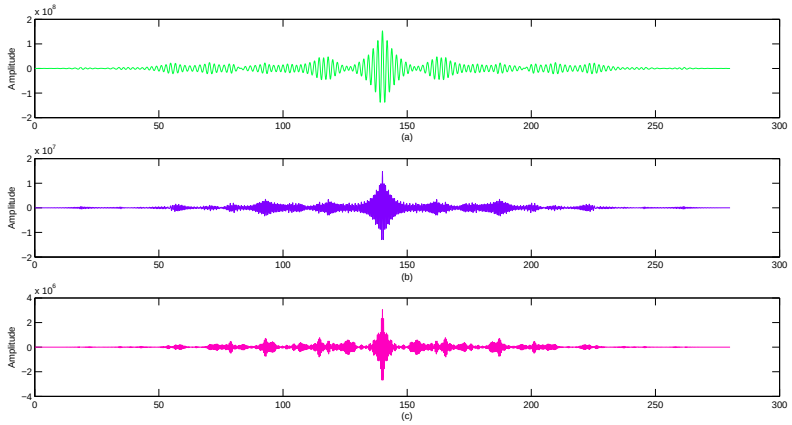

Figure 4.6: ACF of the three band limited signals

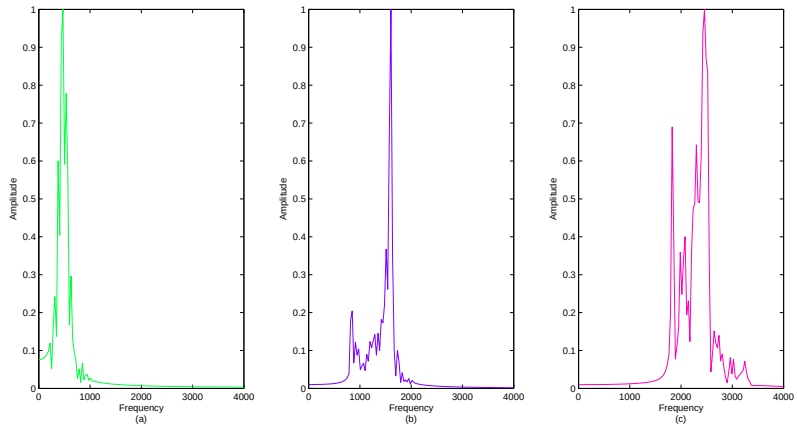

Figure 4.7: Response of the ACF of the three band limited signals

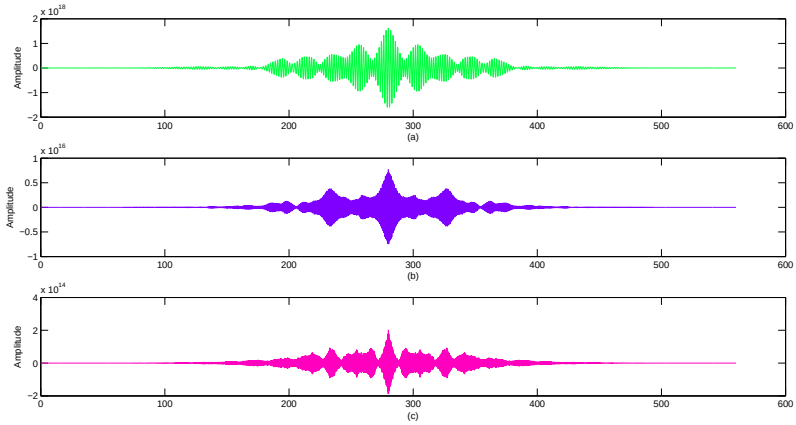

Figure 4.8: RACF of the three band limited signals

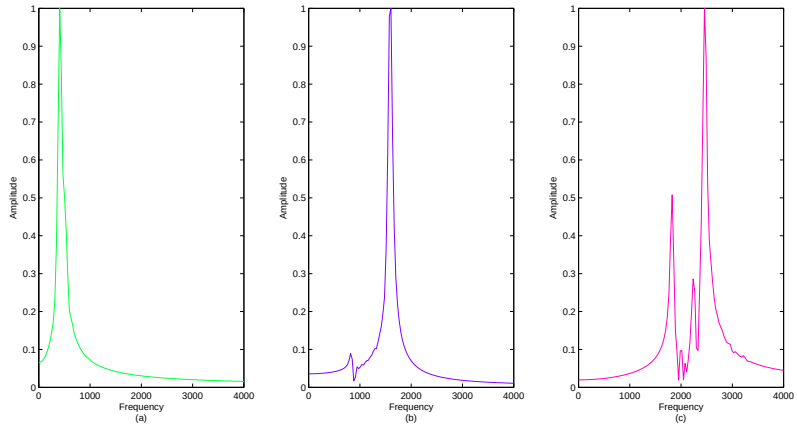

Figure 4.9: Response of the RACF of the three band limited signals

signals, the signals in Fig. 4.8 are obtained whose spectra can be seen at Fig. 4.9.

After comparing Fig. 4.5 with Fig. 4.1, it can be seen that the banding process has achieved its goal by removing the effect of dominant poles from the poles of the second formant band. While in Fig. 4.1, the second formant peak is barely visible, in Fig. 4.5(b), the second formant peak is dominant. This is also true for the third formant frequency range, whose filtered RACF spectrum is shown in Fig. 4.5(c). It is also seen that while for the full band signal due to each ACF the domiance of the second and third formant peaks were reduced, for the banded signals dominance of each banded signals are increased due to ACF. Thus the goal of peak enhancement is fully achieved for banded signals which is evident from Figs. 4.7 and 4.9. Here, due to ACF and RACF on the banded signals the resultant spectra become smooth and spurious peaks become reduced in number.

4.1.4 Formulation of Proposed Model

After band limiting and performing DACF, formant estimation is performed by performing parameter extraction through matching a frequency domain model with the filtered DACF signal.

After performing the banding as described, it is assumed that there is only one dominant complex pole pair in a sub-band. Following (4.2), the transfer function for each sub-band can be written as,

$$H_i(e^{j\omega}) = \frac{C_i''}{\prod_{k=1}^{2}(1 - p_k e^{-j\omega})}, \tag{4.22}$$

Keeping the fact in mind that autocorrelation introduces new conjugate reciprocal poles, transfer function of the two sided ACF for banded speech signal can be defined as follows,

$$R_{hh}^{i}(e^{j\omega}) = \frac{C_i'}{\prod_{k=1}^{2}(1 - p_k e^{-j\omega})(1 - p_k^{*} e^{j\omega})}, \tag{4.23}$$

Here poles, conjugate reciprocal to the original poles are introduced due to autocorrelation. The signal of (4.23) has both causal and anti-causal parts. Anti-causal parts

are introduced due to the poles outside the unit circle. Equation (4.23) represents the double sided ACF of the function in (4.22).

Following the previous chapter for a full band signal with P poles($P/2$ conjugate pole pairs), the transfer function for the once repeated ACF (ORACF) can be written as

$$P_{hh}(e^{j\omega}) = \frac{C}{\prod_{i=1}^{2P}\{(1-p_i e^{-j\omega})(1-p_i^* e^{j\omega})\}} = \frac{C}{\prod_{i=1}^{P}\{(1-p_i e^{-j\omega})(1-p_i^* e^{j\omega})\}^2} \quad (4.24)$$

Here for each pole $p_i = r_i e^{j\theta}$, there exists a pole $1/p_i^*$ which is placed at conjugate reciprocal locations. From (3.22) it is clearly seen that total number of poles in $P_{hh}(e^{j\omega})$ is $4P$, which is twice as the number of poles in $R_{hh}(e^{j\omega})$. Due to the autocorrelation operation new $2P$ poles are introduced in $P_{hh}(e^{j\omega})$ which are conjugate reciprocal to the original $2P$ poles of $R_{hh}(e^{j\omega})$. Hence, new poles are located at the original pole angles as expected and for each original pole location of $R_{hh}(e^{j\omega})$, both inside and outside unit circle, in $P_{hh}(e^{j\omega})$ there exists two poles.

Thus for a the subband signal in relation to (4.23), the transfer function of (4.24) can be expressed as,

$$P_{hh_i}(e^{j\omega}) = \frac{C_i}{\prod_{k=1}^{2}\{(1-p_k e^{-j\omega})(1-p_k^* e^{j\omega})\}^2}, \quad (4.25)$$

Here, two new pairs of poles are introduced both inside and outside of the unit circle. All of these new pole pairs are reciprocal to original four pole pairs.

Previously it was shown that to deal only with the causal part of the signal if single sided ACF (SSACF) from this banded once repeated correlation signal is considered the dominant peaks become more distinct as in this case poles inside the unit circle are taken into consideration only. The property of original pole retention and noise robustness of SSAC sequence was described in [34]. Since our objective is to handle the severe noisy condition, the use of SSACF would be a better choice. Thus, in order to consider only the causal parts we should discard the poles outside the unit circle and the transfer function for the ORSSACF can be modeled as,

$$P^{+}_{hh_i}(e^{j\omega}) = \frac{C_i}{\prod_{k=1}^{2}\{(1-p_k e^{-j\omega})\}^2}, \tag{4.26}$$

In (3.35) it was shown that within a specified range the ramp cepstrum of the ORSSACF of the noisy speech signal, $\mu_y(m)$ is approximately equal to $\mu_x(m)$, the ramp cepstrum of the ORSSACF of $x(n)$, corresponding to the original AR system. This relation should hold true for ORSSACF of a band limited signal also. Thus for the ORSSACF of a band limited signal following (3.36) it can be written that

$$\mu^i_y(m) \approx \mu_{x_i}(m) = \sum_{k=1}^{2} p_k^m, m > 0. \tag{4.27}$$

In (3.36) it can be found that $\mu_H(e^{j\omega})$ is directly related to system poles. Now, following (4.27) and (3.37) of the previous chapter, where a relationship between the spectrum of $\mu_y(m)$ and the spectrum of $\mu_h(m)$, the ramp cepstrum of the ORSSACF of $h(n)$ is established, a relation between the band limited spectra can be expressed as follows

$$\mu^i_Y(e^{j\omega}) = \mu_{H_i}(e^{j\omega}) = \sum_{k=1}^{2} \frac{G_{hh}}{(1-p_k e^{-j\omega})}, \tag{4.28}$$

where $\mu^i_Y(e^{j\omega})$ is the Fourier transform of the ramp cepstrum of the ORSSACF of the banded noisy speech signal, $\mu^i_y(m)$, $\mu_{H_i}(e^{j\omega})$ is the Fourier transform of $\mu_{h_i}(m)$, the transfer function for the ramp cepstrum of the ORSSACF of a banded signal, G_{hh} is the gain factor and for i-th band p_1 and p_2 are a complex conjugate pole pair where the band consists of two pair of poles such as these.

4.1.5 Proposed Model Fitting Approach

In the proposed formant estimation method, a spectral model corresponding to the spectrum of the ramp cepstrum of ORSSACF of the band limited speech signal is introduced, which is utilized in a model matching technique to find out the model parameters that in turn will provide the formant frequency corresponding to a band. The ramp cepstrum of ORSSACF of each banded noisy speech frame $y_i(n)$ is computed and used in the proposed

model matching technique.

Following (4.28) a spectral model for the ramp cepstrum of the ORSSACF of a bandlimited noisy sequence can be derived as follows

$$\mu^i_{model}(e^{j\omega}) = \sum_{k=1}^{2} \frac{G_{hh}}{(1 - p_k e^{-j\omega})}. \qquad (4.29)$$
$$p_k = r_k e^{j\theta_k}$$

The spectrum $\mu^i_Y(e^{j\omega})$ of the ramp cepstrum of the ORSSACF of the observed noisy band limited signal $y_i(n)$ is used in conjunction with the proposed model $\mu^i_{model}(e^{j\omega})$ to form an objective function for the first formant zone based on the square of absolute difference of these spectra, namely

$$e_{min}(r_j, \theta_j) = \underset{\substack{r_l < r_i < r_h \\ \theta_l < \theta_i < \theta_h}}{min} \sum_{\omega=\omega_{lc}}^{\omega_{hc}} \left(|\mu^i_{model}(e^{j\omega})| - |\mu^i_Y(e^{j\omega})|\right) \qquad (4.30)$$

Minimization of the objective function is carried out within a restricted frequency range ω_{lc} to ω_{hc} which depends on the range of the first formant zone. One may utilize the -3dB points on the lower and higher sides of the peak in the spectrum of the model to extract ω_{lc} and ω_{hc}. Within that specified range $\omega_{lc} \leq \omega \leq \omega_{hc}$, the optimum value of the two variables r_i and θ_i is obtained at the minimum square absolute difference. Based on the fundamental knowledge of traditional range of formants, one may restrict the search range for the two variables i.e., $r_l \leq r \leq r_h$ and $\theta_l \leq \theta \leq \theta_h$ or adopt a coarse and fine search approach [36]. Formant frequencies are estimated from the pole angle θ_j that produces the best match between the spectra using (4.30).

Starting from the first band to find the first formant F1, (4.30) is used in second and third band to find out the formants correspoding to those bands, namely F2 and F3.

One major advantage of the proposed model fitting approach over the conventional peak picking method lies in the fact that an entire formant band is taken into consider-

ation instead of relying only on the magnitude of the peaks, which are extremely noise sensitive. As a result the formant frequency that is chosen as the desired estimate should provide the best match between the spectra within a formant band. This spectral matching is very suitable especially when the level of noise is very severe and the formants are very closely spaced.

4.1.6 Formant Based Vowel Recognition

After estimating formants in this manner, in the proposed scheme they are employed in vowel recognition as features along with the commonly used mel frequency cepstral coefficients (MFCC) coefficients. Linear discriminant analysis (LDA) based classifier is used to accomplish this task. For our proposed scheme, a frame by frame classification method is used, which offers vowel recognition results for each voiced frame independently.

The classifier classifies the data into different groups generally, depending on the significant characteristics of the group members. The quality of a classifier depends on its ability to provide the compactness among the member within a cluster and the separation between the members of different clusters in terms of feature characteristics. The task of recognizer is to identify the class label of a test sample utilizing the classified data. In a feature based scheme, classification is performed utilizing the extracted features of the data, instead of directly employing the data themselves. In the proposed method, the LDA is used to classify the vowel among the different classes (in our case, vowel) available. In LDA, a linear projection is determined that maximizes a ratio between the signal, represented by the projected inter-cluster distance and the noise, represented by the projected intra-cluster variance. Here the objective function is based on determining a projection direction w to maximize the Fisher's discriminant defined as

$$J(w) = \frac{w^T S_b w}{w^T S_w w} \tag{4.31}$$

where S_w and S_b are within- and between-class scatter matrices, respectively [37].

4.2 Results and Simulation

In order to evaluate the recognition performance of the proposed methods, numerous experiments have been conducted on the TIMIT acoustic-phonetic continuous speech corpus, which has jointly been developed by Massachusetts Institute of Technology (MIT), Stanford Research Institute (SRI) and Texas Instruments (TI) [38]. The TIMIT database contains a large collection of sentences uttered by both male and female English speakers using various dialects. A total of 6300 sentences, with 10 sentences spoken by each of the speakers are present on the database. Voiced and unvoiced portions of speech are clearly marked on accompanying phone files. However, as TIMIT does not contain reference values of formants, to compare estimated results, the most commonly used formant database is chosen, where formant frequencies are estimated based on vocal tract resonances (VTR) with manual correction [39]. The formant estimates reported in [39] are taken as ground truth and the estimation performance of different methods is evaluated at different levels of signal to noise ratios (SNR). This VTR subset of TIMIT database contains 376 sentences across the training set, representing 173 speakers. These sentences contain 18 voiced phonemes, out of which, the diphthongs have been ignored, and 11 phonemes are considered. A total of 2726 utterances of phonemes are used from the VTR subset, out of which 1583 are from male and 1143 are from female speakers, have been analysed. In VTR database, formant estimates are reported for every 10 ms interval. However, vowel duration is generally much larger than 10 ms. In the frame by frame formant analysis, when the size of analysis frame is larger than 10 ms, the estimated formants are then compared with the average VTR formant values obtained over the different 10 ms frames within the duration of that formant under investigation. For the purpose of performance comparison, first the most widely used LPC based formant estimation method [40] is chosen, where the order of the LPC is chosen as 12. Apart from the LPC method, a state of the art adaptive filter bank (AFB) method is also chosen. In the AFB method, formant estimation is carried out in sample by sample basis, and for the purpose of comparison, average estimated formant values over a period is considered [23].

Table 4.1: Performance comparison in terms of mean error(%) for synthetic speech

Vowels		$5dB$			$-5dB$		
		Proposed	LPC	AFB	Proposed	LPC	AFB
/a/	F1	4.04	20.24	46.90	9.93	20.46	49.77
	F2	15.84	65.23	32.58	8.76	113.79	30.99
	F3	7.33	17.80	8.45	8.90	34.02	9.84
/o/	F1	11.63	49.53	128.07	14.19	78.29	18.29
	F2	4.85	138.88	20.42	19.14	133.29	46.61
	F3	6.74	39.93	9.56	6.74	36.28	12.53
/u/	F1	10.59	72.96	109.00	17.16	98.29	12.98
	F2	5.57	116.33	14.62	19.70	121.92	33.72
	F3	4.44	52.31	11.40	4.78	40.60	13.74

In the proposed model fitting scheme, the range of the model parameters are set according to the general behavior of the vocal tract. The possible range of the parameter r is changed within the limit 0.8 to 0.99, which covers even a very rapidly decaying impulse for the purpose of our simulation. The search range for θ is set according to the determined formant band. Search resolutions for r and θ are chosen as$\triangle r = 0.01$ and $\triangle\theta = 0.001\pi$, respectively. In our experiments in order to obtain a noisy signal, noise sequence of a particular SNR is added with the clean (noise-free) signal. Noisy signals are generated according to 3.12, where the noise variance σ_v is appropriately determined according to a specified level of SNR defined as

$$SNR = 10log_{10}\frac{\sum_{n=0}^{N-1} x(n)^2}{\sum_{n=0}^{N-1} v(n)^2} \tag{4.32}$$

At first results for three synthetic vowels /a/, /o/ and /u/ in the presence of white Gaussian noise with SNR $5dB$ and $-5dB$ are presented in Table 4.1 where the estimation error, the mean average deviation between the estimated formant frequency f_E and the reference formant frequency f_R is defined as

$$E = |\frac{f_E - f_R}{f_R}| \times 100\% \tag{4.33}$$

It is observed that the proposed method offers far superior performance in the presence

Table 4.2: Performance comparison in terms of mean error(%) for male speakers

Vowel		-5 dB			5 dB		
		Proposed	AFB	LPC	Proposed	AFB	LPC
	F1	15.77	30.88	30.53	14.77	17.74	26.48
/aa/	F2	13.36	36.42	82.19	12.05	21.87	45.44
	F3	17.38	15.47	43.35	13.17	17.07	39.80
	F1	16.26	24.12	31.64	15.60	16.31	24.65
/ah/	F2	13.23	28.88	57.43	12.09	24.41	35.57
	F3	14.39	13.09	39.21	10.82	11.61	37.72
	F1	19.28	35.49	22.63	19.40	37.77	19.72
/ow/	F2	14.08	26.03	47.20	11.56	24.65	41.67
	F3	15.60	14.20	36.68	12.08	14.00	37.74
	F1	18.85	36.49	20.14	19.32	36.55	19.49
/uh/	F2	11.77	23.49	38.02	11.62	23.23	37.48
	F3	11.11	13.89	37.24	10.81	13.86	37.33
	F1	18.21	36.72	29.66	18.39	39.58	20.09
/uw/	F2	13.06	23.14	40.36	12.22	22.49	36.45
	F3	12.54	14.53	39.48	11.04	14.50	38.25

of noise for the synthetic vowels.

Next the estimation errors obtained by the proposed method and that by the other two methods are presented under the influence of white gaussian noise conditions for male and female speakers in Tables 4.2, and 4.3 .

For Table 4.2and Table3.3 SNR levels $5dB$ and $0dB$ are considered. For each vowel, the estimation errors for three different formants, namely $F1$,$F2$ and $F3$ are listed. As can be seen from the tables, the proposed method offers better performance than both the12 order LPC and the AFB methods under presence of background noise. It can be observed that the estimation error obtained by the proposed method in comparison to that of the other methods is extremely lower in such severe noisy conditions.

In some cases it is found that the estimation accuracy decreases for the cases when the two formants are very closely spaced, for example in case of vowel /ih/, though, considering the level of noise, the estimation accuracy obtained by the proposed method is quite acceptable. It is also observed that the estimation error relatively increases in

Table 4.3: Performance comparison in terms of mean error(%) for female speakers

Vowel		-5 dB			5 dB		
		Proposed	AFB	LPC	Proposed	AFB	LPC
	F1	13.20	19.40	45.04	12.13	12.33	25.84
/aa/	F2	14.04	69.07	27.67	9.91	21.02	20.53
	F3	11.07	30.96	12.77	9.87	20.79	11.95
	F1	13.67	14.68	36.14	12.96	10.79	17.91
/ah/	F2	14.78	37.14	21.46	9.50	13.37	20.44
	F3	10.92	23.14	15.75	8.61	19.79	12.44
	F1	11.82	11.75	26.75	11.95	10.82	15.70
/ow/	F2	14.38	43.25	26.84	11.44	27.30	20.96
	F3	10.24	18.63	15.23	9.25	17.84	10.00
	F1	12.11	12.91	18.07	12.18	10.54	16.46
/uh/	F2	11.99	23.46	21.40	10.78	18.72	20.04
	F3	8.35	18.93	10.40	7.86	18.40	9.54
	F1	12.41	9.47	16.67	12.53	9.12	16.37
/uw/	F2	11.57	17.18	20.33	10.32	16.46	18.49
	F3	8.40	18.12	9.82	8.56	18.19	9.42

case of high pitch female speakers. It is clearly observed that the estimation performance for the third formant, which is by nature very difficult to estimate because of low spectral magnitude, is significantly enhanced by the proposed method. Hence, overall it can be said that, the proposed method increases formant estimation performance.

In order to present the overall formant estimation errors over a large range of SNRs considered in the experimental setup, in Fig. 4.10 the overall estimation error for all vowels are shown. In a simiar way, in order to present the overall formant estimation errors over a large range of SNRs considered in the experimental setup, in Fig. 4.11 the overall estimation error for all vowels are shown. It is observed that the formant estimation performance obtained by the three methods remains similar in case of high level of SNR. However, with the decrease in SNR level, the estimation performance of the other two methods deteriorates in comparison to that of the proposed method. The performance of the proposed method remains quite consistent even in the low levels of SNRs and level of performance degradation is not very significant till -15 dB. However,

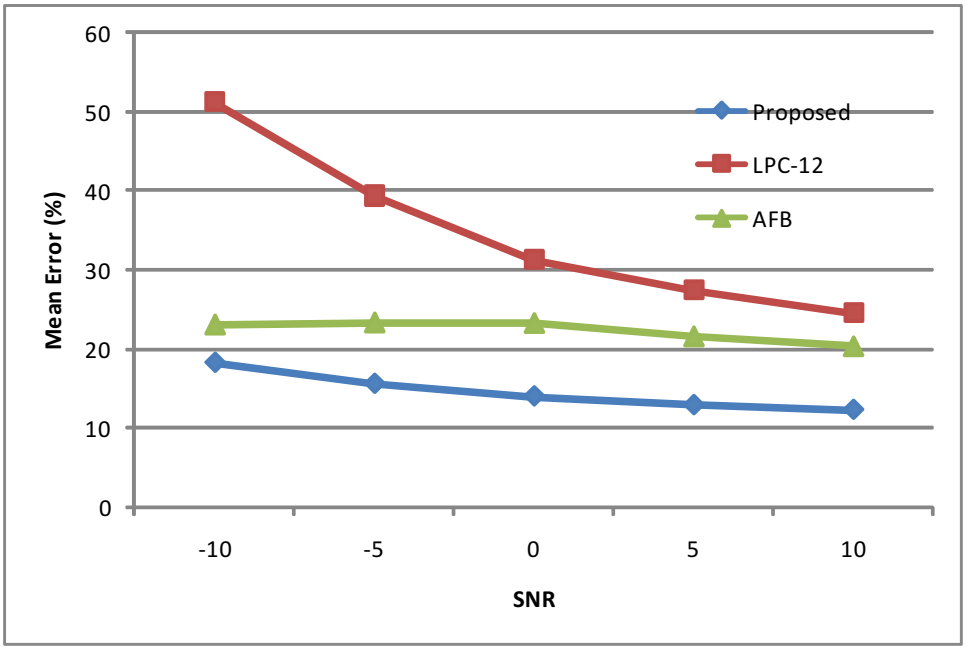

Figure 4.10: Error comparison of all three formants for different methods

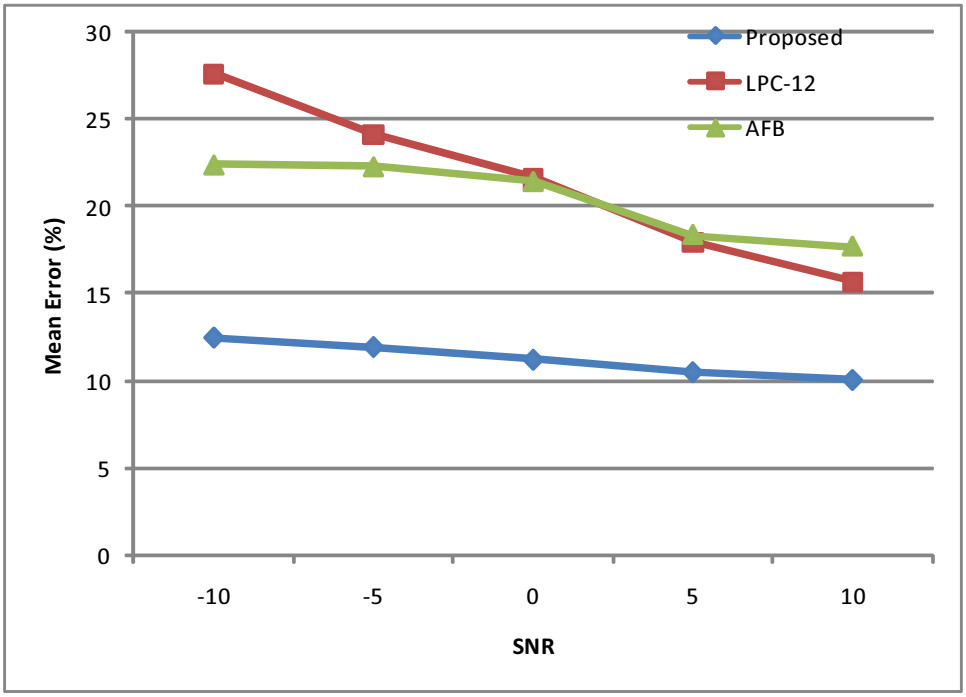

Figure 4.11: Error comparison of all three formants for different methods for female

beyond that the performance of the proposed method is not satisfactory because of the severe noise corruption, leading to complete failure for the conventional methods.

In the proposed method formant estimation is carried out frame by frame with a frame length of 512 samples and 10 ms overlap between the successive frames. As a result for a vowel sound of duration of about 80 ms, 5 frames are analyzed. It is to be noted that, because of the inherent characteristics of the fast Fourier transform (FFT) operation, there exists an inherent error caused by the minimum width of the FFT bin. For instance, when a 512 point FFT is performed on a speech frame with sampling frequency of 16 kHz, the resulting FFT has a resolution of 15.6 Hz.

Table 4.4: Recognition Accuracy

	Feature Vector	
SNR	MFCC + Proposed Method	MFCC + LPC-12
-10 dB	64	60.00
0 dB	85.33	84
5 dB	89.44	85.33

By incorporating the estimated formants in a feature vector along with traditional MFCC, significantly better vowel recognition accuracies are achieved compared to a feature vector consisting of MFCC and formants estimated by LPC, especially under the influence of noise. By using these formants along with the traditional 12 MFCC coefficients as a feature vector, vowel recognition was performed for the vowels /aa/, /ux/ and /ix/ from the TIMIT database. As formant ranges for male and female vowels vary significantly, they are considered as separate classes for this LDA based classification operation. There are 20 utterances for male and 20 utterances for female considered for each vowel. Accuracies are calculated by leaving one sample out while training the classifier and then testing the left out sample. This check is performed for all the samples in the database, and it is found that the proposed feature vector offers better performance in noisy conditions. The recognition accuracies for different vowels is presented in Table 4.4. It can be concluded from the table that the proposed noise robust formant estimation method, when used for vowel recognition, increases the recognition accuracy for vowel

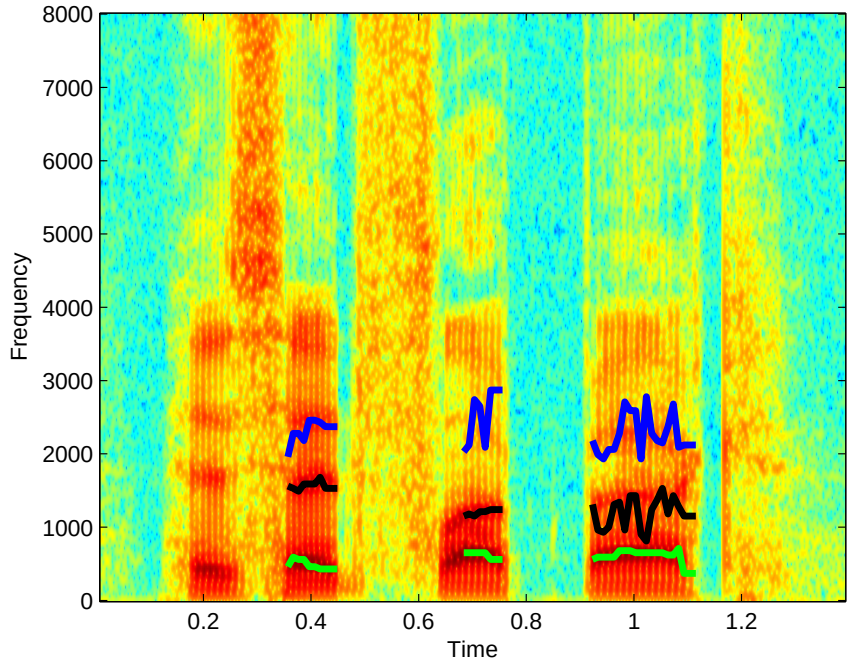

Figure 4.12: A spectrogram of the sentence "His head flopped back" with tracked formant by the proposed method at 0 dB SNR

recognition systems under the influence of noise.

As seen from these analyses, the proposed method offers a better performance over the LPC and AFB methods in noise free as well as in noisy conditions. In order to demonstrate the effectiveness of our proposed method, a spectrogram of the sentence "His head flopped back", uttered by a male speaker taken from the TIMIT database is shown in Fig. 4.12. The formant frequencies estimated at different frames using the proposed method uder SNR= 0 dB are shown over the spectrogram of clean speech. In the tracking, only the estimated formants of the vowels are shown. It can be observed from the figure that the proposed method tracks the formant frequencies quite accurately even in noisy speech.

4.3 Conclusion

In this chapter, in view of better exploiting the dominant peak strengthening effect of autocorrelation spectrum, band pass filters centered at different formant peaks are used to obtain band limited signals. Considering the human vocal tract as a cascade of these

subsystems, unlike the previous chapter, impulse response corresonding to each subsystem is considered in the model derivation. Repeated autocorrelation, which strengthens the dominant poles, and exponentially increases the peak-valley ratio at formant frequencies of the magnitude response, cancelling out the effects of noise, is then performed on the signal corresponding to each subband. It is shown that better formant estimation accuracy can be achieved with this algorithm especially under severe noisy conditions. Comparisons between standard LPC based formant estimation techniques as well as recent methods like AFB have been shown.

Chapter 5

Conclusion

5.1 Contributions of the Thesis

- The main goal of this thesis work is to develop a noise robust formant frequency estimation method that can offer robust performance in the presence of background noise. Instead of directly using the noisy observations, the autocorrelation function of the noisy signal is utilized. In time domain autocorrelation reduces the effect of noise by concentrating all the noise energy at the zeroth lag. Thus ACF on a noisy signal gives the opportunity to estimate formants robustly than estiamting formants from the nosiy speech.

- It is found that repeated autocorrelation can provide more noise immunity than single autocorrelation. It is presented in a method for estimating formants at a very low SNR. Due to its spectral strengthening effect the method utilizing repeated autocorrelation can successfully estimate formants even in a severe noisy condition.

- One major contribution is to use cepstral domain analysis in correaltion and repeated correaltion domain. The use of cepstrum not only provides logarithmic spectral smoothing , it also offers the advantage of homomorphic deconvolution to overcome the effect of pitch variation. It is shown that within the pitch period cepstral coefficients of the speech signal are approximately equal to the cepstral coefficients of its system response.

- Conventional cepstrum decays rapidly with the increase in time index, which makes it difficult to use in estimating the system poles and eventually formants. Ramp cepstrum is introduced to solve this problem and it is shown that ramp cepstrum of the speech signal can be directly related to system poles. Thus ramp cepstrum was successfully utilized in deriving a model for formant estimation.

- Use of banding on speech for formant estimation purpose was explored. It is shown that band limiting the speech before performing repeated autocorrelation and cepstral analysis can further improve the estimation performance. Thus, a band limiting approach is developed that can adaptively filter the frequency zones where a formant frequency is most likely to be present. Natural vowels as well as some naturally spoken sentences in noisy environments are tested. The experimental results show that the performance obtained by the proposed scheme is better in comparison to some of the existing methods even at a very low level of signal-to-noise ratio.

- Instead of using conventional peak picking to find formants from the spectrum of the autocorrelation function, a spectral model of the cepstrum of autocorrelated speech signal for a single formant is developed and model fitting is employed to find out model parameters which lead to the estimation of formants. The focus of the proposed work is to formulate and explicate robust formant estimation methods in order to achieve better formant estimation performance than the available methods in the presence of noise. Mitigating the effect of fundamental frequency variation on formant estimation is another objective of this work.

- The estimated formant frequencies are chosen as features along with conventional features like MFCC for achieving better recognition accuracy even at a low SNR. In conjunction to commonly used features for recognition such as MFCC, which are badly corrupted by noise, relying on the performance of the proposed estimation methods, the estimated formants, which are less effective by the noise, are selected as features for recognition. All the methods presented in this thesis provide satisfactory results in the case of noisy as well as noise free voiced speech. In comparison

between the proposed methods it is seen that due to the advantageous properties of cepstrum on repeated autocorrealtion of the banded speech signal, the method utilizing repeated autocorrelation of band limited speech signals can give a better estimate of formants.

5.2 Future Works

The objective of the scheme proposed in this thesis is to estimate formants even in the presence of severe background noise. In the process of formant estimation no conventional noise reduction method is used. Another potential approach is to use a noise reduction block first and then apply the proposed method to estimate formants which could be explored in future. Moreover, the proposed method can be investigated in the presence of colored real life practical noises.

In the matching section single variable matching is performed for finding the desired solution from the objective function . Exploring nonlinear complex multivariable solution techniques like the fuzzy logic, genetic algorithm, neural network etc. could be employed for finding the solution. This can be a potential future work of interest though these methods might be computationally expensive and could give false solution in cases of very low SNR and weak second and third formants.

Bibliography

[1] D. O'shaughnessy, *Speech communications: human and machine*. Universities press, 2000.

[2] W. Verhelst and O. Steenhaut, "A new model for the short-time complex cepstrum of voiced speech," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 34, no. 1, pp. 43–51, Feb. 1986.

[3] D. Veselinovic and D. Graupe, "A wavelet transform approach to blind adaptive filtering of speech from unknown noises," *IEEE Trans. Circuits Syst. II*, vol. 50, no. 3, pp. 150–154, Mar. 2003.

[4] G. E. Kopec, A. V. Oppenheim, and J. M. Tribolet, "Speech analysis by homomorphic prediction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 25, no. 1, pp. 40–49, Feb. 1977.

[5] K. Steiglitz, "On the simultaneous estimation of poles and zeros in speech analysis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 25, no. 3, pp. 229–234, June 1977.

[6] Y. Miyanaga, N. Miki, N. Nagai, and K. Hatori, "A speech analysis algorithm which eliminates the influence of pitch using the model reference adaptive system," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 30, no. 1, pp. 88–96, Feb. 1982.

[7] M. Yang, "Low bit rate speech coding," *IEEE Potentials*, vol. 23, no. 4, pp. 32–36, Oct. 2004.

[8] T.-H. Hwang, L.-M. Lee, and H.-C. Wang, "Cepstral behaviour due to additive noise and a compensation scheme for noisy speech recognition," *IEE Proc. Vis. Image Signal Process.*, vol. 145, no. 5, pp. 316–321, Oct. 1998.

[9] Z. Ben Messaoud and A. Ben Hamida, "Combining formant frequency based on variable order lpc coding with acoustic features for timit phone recognition," *International Journal of Speech Technology*, pp. 1–11, 2011.

[10] D. O'Shaughnessy, "Interacting with computers by voice: automatic speech recognition and synthesis," *Proc. IEEE*, vol. 91, no. 9, pp. 1272–1305, Sept. 2003.

[11] J. M. Hillenbrand, L. A. Getty, M. J. Clark, and K. Wheeler, "Acoustic characteristics of American English vowels," *J. Acoust. Soc. Am.*, vol. 97, no. 5, pp. 3099–3111, May 1995.

[12] J. Xu, J. Cheng, and Y. Wu, "A cepstral method for analysis of acoustic transmission characteristics of respiratory system," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 5, pp. 660–664, May 1998.

[13] H. Morikawa and H. Fujisaki, "System identification of the speech production process based on a state-space representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. 2, pp. 252–262, Apr. 1984.

[14] R. Snell and F. Milinazzo, "Formant location from lpc analysis data," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 2, pp. 129–134, Apr. 1993.

[15] L. Welling and H. Ney, "Formant estimation for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 6, no. 1, pp. 36–48, Jan. 1998.

[16] J. R. D. Jr., J. H. L. Hansen, and J. G. Proakis, *Discrete-time processing of speech signals*. NY: IEEE Press, 2000.

[17] B. Yegnanarayana and R. N. J. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 4, pp. 313–327, July 1998.

[18] B. Chen and P. C. Loizou, "Formant frequency estimation in noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, vol. 1, Montreal, Canada, May 2004, pp. 581–584.

[19] T. Wang and T. Quatieri, "High-pitch formant estimation by exploiting temporal change of pitch," *IEEE Trans. Audio Speech Lang. Processing*, vol. 18, no. 4, pp. 171–186, 2010.

[20] D. J. Nelson, "Cross-spectral based formant estimation and alignment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, vol. 2, Montreal, Canada, Mar. 2004, pp. 621–624.

[21] A. Watanabe, "Formant estimation method using inverse-filter control," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 4, pp. 317–326, May 2001.

[22] E. Ozkan, I. Ozbek, and M. Demirekler, "Dynamic speech spectrum representation and tracking variable number of vocal tract resonance frequencies with time-varying dirichlet process mixture models," *IEEE Trans. Audio Speech Lang. Processing*, vol. 17, no. 8, pp. 1518–1532, 2009.

[23] K. Mustafa and I. C. Bruce, "Robust formant tracking for continuous speech with speaker variability," *IEEE Trans. Audio, Speech Language Processing*, vol. 14, no. 2, pp. 435–444, Mar. 2006.

[24] S. Fattah, W. Zhu, and A. Ahmad, "Noisy autoregressive system identification by the ramp cepstrum of one-sided autocorrelation function," in *Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on.* IEEE, 2005, pp. 3147–3150.

[25] C. I. Byrnes, P. Enqvist, and A. Lindquist, "Cepstral coefficients, covariance lags, and pole-zero models for finite data strings," *IEEE Trans. Signal Processing*, vol. 49, no. 4, pp. 677–693, Apr. 2001.

[26] S. Fattah, W. Zhu, and M. Ahmad, "Identification of autoregressive systems in noise based on a ramp-cepstrum model," *Circuits and Systems II: Express Briefs, IEEE Transactions on*, vol. 55, no. 10, pp. 1051–1055, 2008.

[27] J. Holmes, W. Holmes, and P. Garner, "Using formant frequencies in speech recognition," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[28] N. Wilkinson and M. Russell, "Improved phone recognition on timit using formant frequency data and confidence measures," in *Seventh International Conference on Spoken Language Processing*, 2002.

[29] C. Chan, P. Ching, and T. Lee, "Noisy speech recognition using de-noised multiresolution analysis acoustic features," *The Journal of the Acoustical Society of America*, vol. 110, pp. 2567–2574, 2001.

[30] B. Yegnanarayana and R. Veldhuis, "Extraction of vocal-tract system characteristics from speech signals," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, no. 4, pp. 313–327, 1998.

[31] R. M Shahidur and S. Tetsuya, "Linear prediction using refined autocorrelation function," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, 2007.

[32] B. Bhanu and J. H. McClellan, "On the computation of the complex cepstrum," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 5, pp. 583–585, Oct. 1980.

[33] Z. Huang, X. Yang, X. Zhu, and A. Kuh, "Homomorphic linear predictive coding: a new estimation algorithm for all-pole speech modeling," *IEE Proc. Comm. Speech Vis.*, vol. 137, no. 2, pp. 103–108, April 1990.

[34] J. Hernando and C. Nadeu, "Linear prediction of the one-sided autocorrelation sequence for noisy speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 5, no. 1, pp. 80–84, 1997.

[35] D. McGinn and D. Johnson, "Reduction of all-pole parameter estimator bias by successive autocorrelation," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'83.*, vol. 8. IEEE, 1983, pp. 1088–1091.

[36] S. A. Fattah, W. P. Zhu, and M. O. Ahmad, "An approach to formant frequency estimation at a very low signal-to-noise ratio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 4, Honolulu, HI, Apr. 2007, pp. 469–472.

[37] R. Duda and P. Hart, *Pattern Classification.* John Wiley, 2001.

[38] J. Garofolo, L. L. W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," in *Proc. Ling. Data Consort.*, 1993.

[39] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, Toulouse, France, Apr. 2006, pp. 369–372.

[40] S. M. Kay, *Modern Spectral Estimation, Theory and Application.* Englewood Cliffs, NJ: Prentice-Hall Ltd., 1988.

[41] J. Hernando and C. Nadeu, "Speech recognition in noisy car environment based on osalpc representation and robust similarity measuring techniques," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, vol. 2. IEEE, 1994, pp. II–69.