BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY

Noise Robust Formant Frequency Estimation Method for Speech Recognition Based on Spectral Model of Repeated Autocorrelation of Speech.

by

Abu Shafin Mohammad Mahdee Jameel

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 **"Noise Robust Formant Frequency Estimation Method for Speech Recognition Based on Spectral Model of Repeated Autocorrelation Of Speech"** submitted by Abu Shafin Mohammad Mahdee Jameel, Student No.: 1009062035, 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

Abu Shafin Mohammad Mahdee Jameel

To all my teachers

Whose blessings carry me forward.

Contents

Acknowledgements

In the name of Allah, the most Gracious, the most Compassionate.

I would like to express my sincere gratitude to my Creator for giving me the chance to work with my supervisor Dr. Shaikh Anowarul Fattah, whose support and guidance the span of this research was extraordinary. I also want to thank Dr. Fattah for his dedication and love for his craft, which is contagious and inspires his students for greater achievements. I pray and hope I would be able to emulate him one day.

I would also like to thank the head of the department of Electrical and Electronic Engineering for allowing me to use the lab facilities, which contributed greatly in completing the work in time. I wish to express note of thanks to Dr. Celia Shahnaz, for being a continuous source of inspiration. I express appreciation to my friends Rabu and Rajib, for their suggestions, support and friendship.

Special note of thanks goes to my brothers, sisters, my grandfather, other relatives and well wishers for their continuous support and encouragement. Finally, I wish to thank the Creator again for blessing me with such wonderful parents, whose love and support are the driving forces of my life.

Abstract

Formants frequencies of the voiced utterance represent the free resonances of the human vocal tract system. They are one of the fundamental properties of human voiced speech, and for the purpose of speech analysis or speech recognition, formant frequencies play a dominant role. In this thesis, effective methods for formant estimation are developed, which work well even in the presence of significant background noise. In real life applications, very often human speech is affected by environmental noises from different sources. Hence noise robustness of formant estimation methods is a key factor. Accurate estimation of formants from given noise corrupted speech is a very difficult task. The major objective of this research is to develop an algorithm that can successfully estimate the formants in the presence of noise, overcoming the limitations of conventional methods. The autocorrelation operation on the speech signal can be viewed as a mean to overcome the adverse effects of noise, since it offers advantageous property of strengthening the dominant formant peaks, leading to better formant estimation accuracy in noise. One major idea in this research, unlike the conventional spectral domain peak picking is to develop a spectral model of autocorrelated speech signal and thereby introduce a model fitting scheme to find out the model parameters which are directly related to formants. Based on the spectral peak strengthening property of the autocorrelation operation by introducing new poles on the formant location, the idea of repeated autocorrelation is presented. The effects of repeated autocorrelation in time and frequency domains are investigated in detail, especially in noisy environments. It is observed that that in comparison to single autocorrelation, double autocorrelation function of a signal exhibits more noise immunity. A spectral model is further developed to incorporate the effects of

double autocorrelation. Finally the effect of spectral band limiting of the speech signal before performing the autocorrelation operation is investigated. It is shown that formant estimation from each band further improves the estimation performance. In order to utilize this property, a band limiting approach is developed that can adaptively filter the frequency zones where a formant frequency is most likely to be present. Spectral model for the double autocorrelation function of the band limited signal is proposed and employed in a model matching approach for estimating the formants. Several vowel sounds taken from the naturally spoken continuous speech signal are tested in the presence of noise. Vowel sounds from synthetic speech as well as naturally spoken isolated words are also considered. The experimental results demonstrate superior performance obtained by the proposed scheme in comparison to some of the existing methods at low levels of signal-to-noise ratio. The estimated formants are used in a basic vowel recognition scheme utilizing a linear discriminant analysis based classifier along with Mel frequency cepstral coefficients (MFCC), and the results demonstrate a good degree of noise robustness compared to the methods using formant values estimated using traditional formant estimation schemes.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Speech is the primary method for intelligent communication among humans. Thousands of languages all over the world continue to convey the intention of people, with words uttered by humans leaving a powerful impact on the world around us. Human communication revolves around the ability to manipulate sounds to form expressions. This type of intelligent communication using complex forms of sounds is one of the features that distinguish humans from other species, leading to the superior position.

With the advent of technology, the communication method for driving powerful machines has shifted towards written commands. For example, computers are controlled via keyboards, a device for writing commands, and mouse, which is used as a pointing device. Written communication also has the added advantage of being easy to preserve. However, people are more comfortable using voice communication, as evident by the popularity of telephones over letters. This has led to the search for developing methods for recognition of speech by machines. Speech recognition can also be easily configured to provide written outputs from oral speech, powering speech-to-text systems. Thus a good speech recognition scheme can be used to preserve dictations of speeches, control machines with voice and give feedback to people with hearing disabilities.

While designing speech recognition methods, one of the primary goals is to find out

facets of speech that are unique to each sound. In this regard, the structure of human speech has been analyzed in detail by researchers, and some of the features that can be used to identify speech have been explored. A recognition system that works with speech normally makes a comparison using these features instead of directly comparing speech utterances.

The human speech can be primarily divided into two categories, voiced speech and unvoiced speech. Voiced speech is normally represented by vowels in the alphabet, and they contain the major portions of energy in speech. Voiced speech has both longer duration and higher amplitude compared to unvoiced speech. This difference arises because of the way speech is produced in the human vocal system. While producing sounds, air travels through the human vocal tract system, and the constrictions applied there are responsible for the sounds that omit from the lip. In case of voiced sounds, there is very little resistance to the flow of air, and thus the sound is louder compared to the unvoiced sounds. As air passes through the vocal tract system for a voiced sound, the structure of the path creates resonances. In the spectrum of the subsequently produced sound, these resonant frequencies are clearly distinguishable as the areas where peaks with high energy are present. These frequencies, caused by the free resonances of the vocal tract, are called formants that are important features of the voiced sounds. Particularly for different vowels, these formant frequencies can be used to distinguish one from another. Thus in cases of vowel recognition, formants are a promising feature.

However, everyday speech is not isolated from the environment in which it is produced. Conversations seldom take place in sound resistant places, and as such, various other sounds that are produces at the same time at neighboring places are mixed with voiced sounds. While performing analysis of speech, these unwanted sounds, classified as noise, play a big role. As noise free environments cannot be guaranteed for all speech utterances, speech recognition systems require features that present significant amount of intra-class compactness and inter-class separability even in the presence of noise. As discussed previously, formants are an important tool in distinguishing vowels from one another. Thus in automated vowel recognition systems formant estimation schemes that are robust

in the presence of noise are highly in demand.

1.2 Vocal Tract System

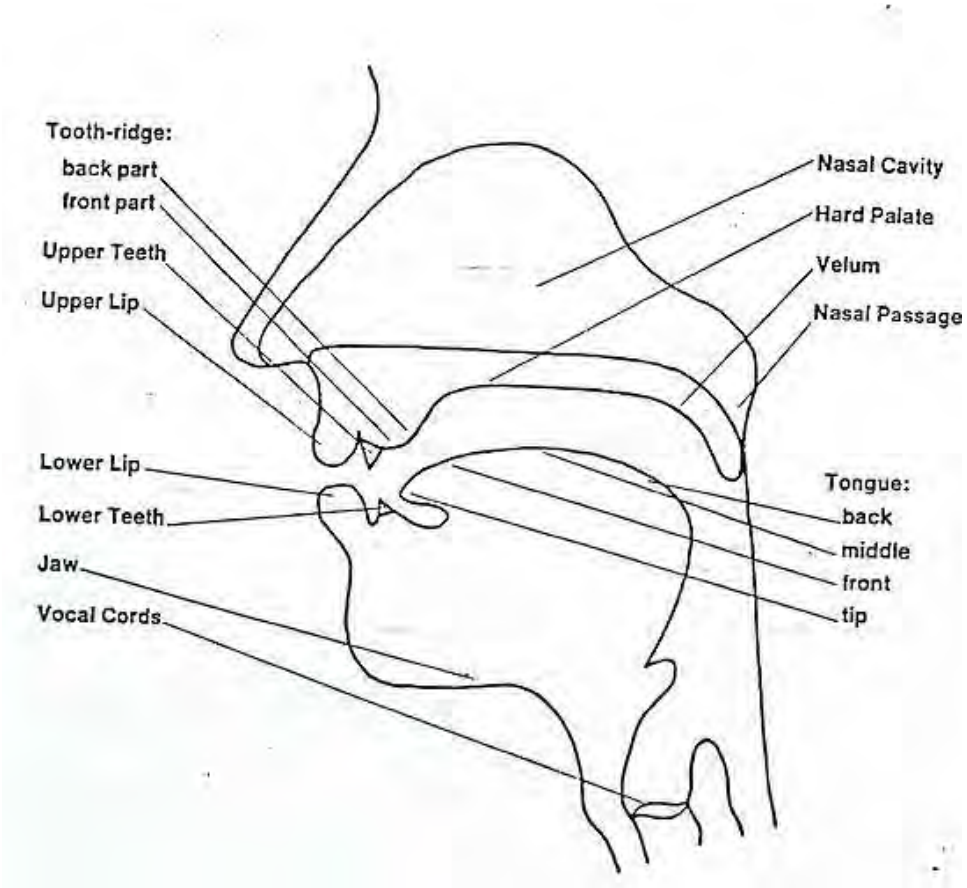


Figure 1.1: Components of the human vocal tract

The speech production in humans are carried out through a set of organs grouped as the human vocal tract . The vocal tract is made up of three cavities: the pharyngeal, oral and nasal cavities and their contribution to the acoustic properties of the vocal tract depend on their configuration and connection to the whole articulatory system. Vocal cords also play an important role in the composition of the vocal tract, as they control the inflow of air from the lungs into the cavities.

The vocal tract can perform two major operations in processing different types of signals: (i) to generate different types of obstacles and (ii) to modify spectral energy distribution of generated sound. In Fig. 1.1 the typical composition of the human vocal tract is shown[1]. Air enters from the lungs through the vocal cords and then passes through

the cavities and finally exits through the lip, or the nose in case of nasal phonemes. For and adult male, the vocal tract is about 17 cm long from the glottis to the lips. When the tract is considered to be a cavity resonator, then the position of the tongue, the area of opening of the mouth, and any changes which affect the volume of the cavity will re tune the resonance.

In general, the VT system exhibits the acoustical characteristics of an acoustic tube whose cross-sectional dimensions are small relative to the wavelengths of the frequencies generated. In its most simplified form, the vocal tract is modeled in the form of a chain of homogeneous tubes. In that case, the vocal tract can be modeled as a closed tube resonator, with the prominent frequency peaks in the vowel sounds corresponding to the resonances of the model. Spectral behavior of such a model can be estimated by comparing the acoustic tube model to a transmission line. However, the actual shape of the vocal tract is more complex than this simple tube model, because its walls vary in shape, and components like the tongue, upper and lower lip are subject to much movement.

From a systems perspective, the linked cylindrical tube model of the vocal tract can be represented by an all pole transfer function. An all pole or autoregressive (AR) modeling of the vocal tract considers the resonances to be caused by complex conjugate pole pairs in the transfer function.

1.3 Voiced and Unvoiced Speech

In regard to phonology, voicing refers to the articulatory process in which the vocal cords vibrate. Vowels are normally voiced sounds, while consonants are generally not voiced. In case of voiced sounds, air passes mostly unobstructed from the lungs to the lips, while in case of unvoiced sounds, the air faces obstruction in various parts of the vocal tract and as a result the resulting speech signal is weak. Compared to voiced speech signals, which show pretty strong periodicity following the pitch or fundamental frequency of speech, the unvoiced signals are quite random.

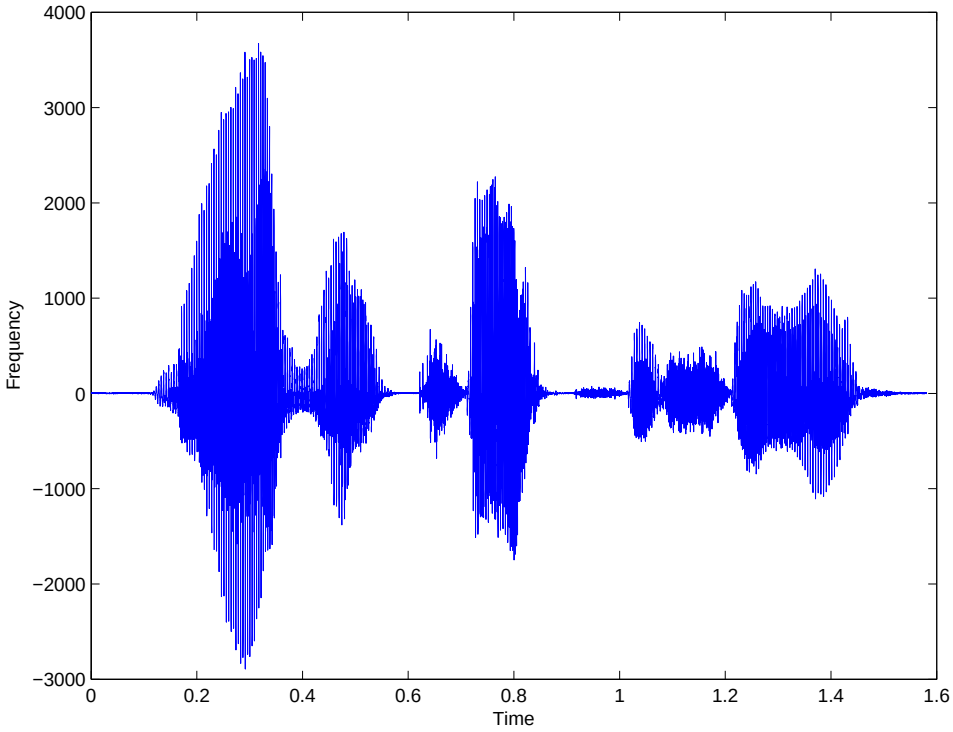


Figure 1.2: A sentence with voiced and unvoiced speech frames

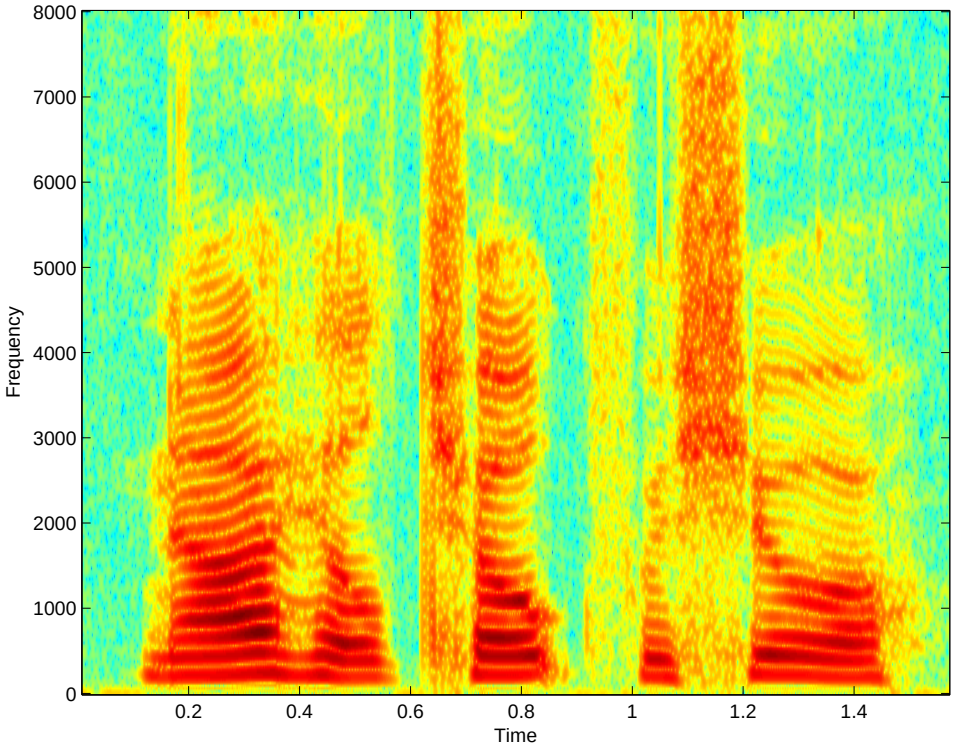


Figure 1.3: Spectrogram corresponding to the speech shown in Fig.1.2

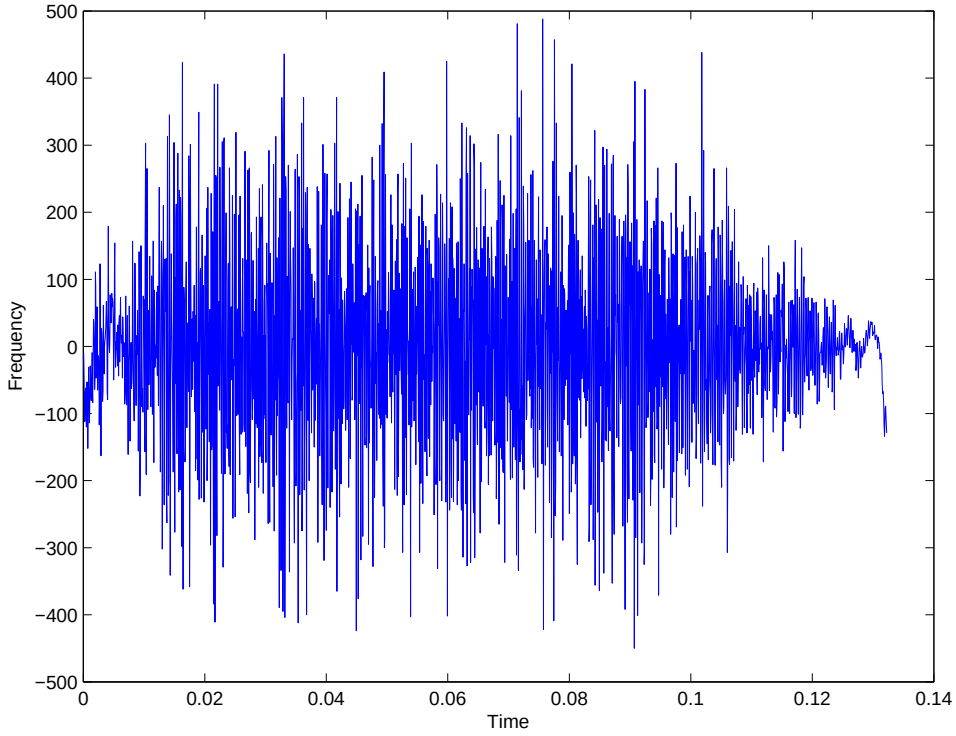


Figure 1.4: A frame showing unvoiced speech frame /s/

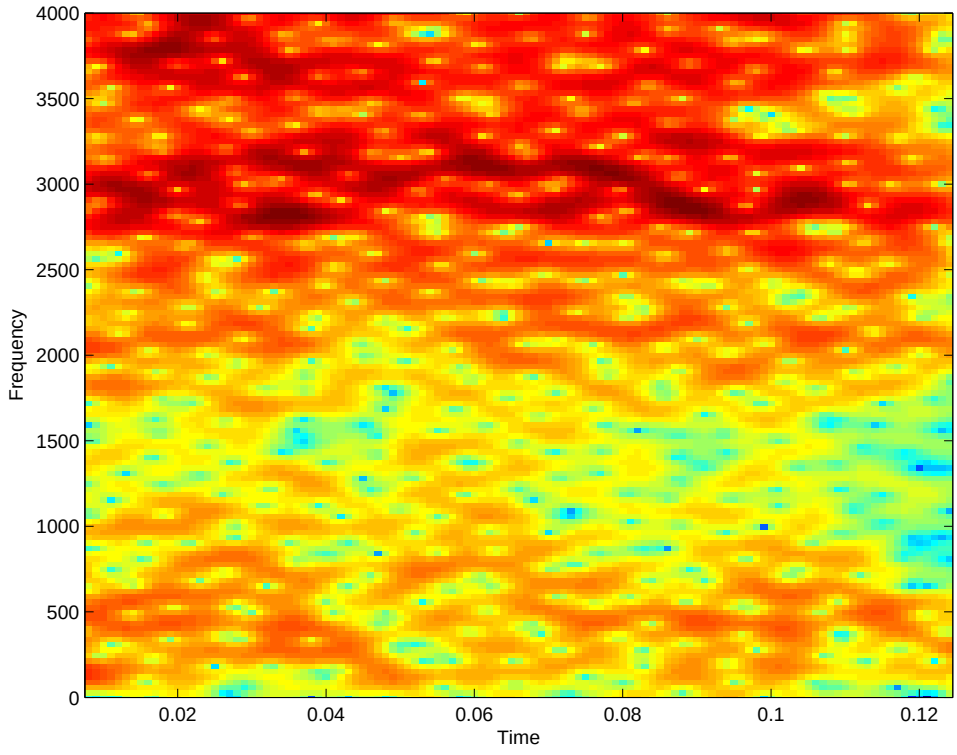


Figure 1.5: Spectrogram for unvoiced speech frames /s/ presented in Fig. 1.4

In Fig. 1.2 the time domain waveform of the sentence 'Now he'll choke for sure' is shown. The high amplitude waveforms in the sentence are caused by the vowels present in the sentence. For instance, the area with the highest amplitude, starting at around 200 ms and continuing up to 400 ms in all the sentence correspond to the vowel /aa/ in 'now'.

The high energy and duration of voiced speech frames can be more clearly understood by looking at the spectrogram of the same sentence, as presented in Fig. 1.3. Spectrogram is a time-frequency representation, where the variation in spectral energy is depicted (in Y-axis) wit respect to time (X-axis). A color map is used to represent the energy variation in frequency axis. Here the dark areas correspond to high energy voiced parts of the utterance. It can be observed from the spectrogram that the voiced phonemes are also longer in duration compared to the low energy unvoiced portions. For further clarification of the random nature of unvoiced sounds, the phoneme /s/ of the word 'sure' form the sentence presented in Fig. 1.2 is taken and its separate time domain waveform is presented in Fig. 1.4. For a better understanding of the spectral energy of this unvoiced frame, its spectrogram is presented in Fig. 1.5.

In the AR modeling of the vocal tract, the excitation that causes voiced speech is assumed to be an impulse train, while the excitation for unvoiced speech is assumed to be random noise.

1.3.1 Formants

Formants are the free resonances of the vocal tract. The resonances occur when the air passes through the vocal tract with no or little resistance. Thus these formant frequencies are clearly distinguishable in the spectra of voiced speech. As these resonant frequencies contain a high amount of energy, in the spectrum of the output speech, prominent peaks are present at formant frequency locations. Formant frequencies act as the primary distinguishing feature for vowels. More specifically, the first three formant frequencies tend to follow a distinct path for different vowels. As the fundamental frequency or pitch has totally different ranges for male and female speakers, the ranges of formants for male

and female speakers is also different.

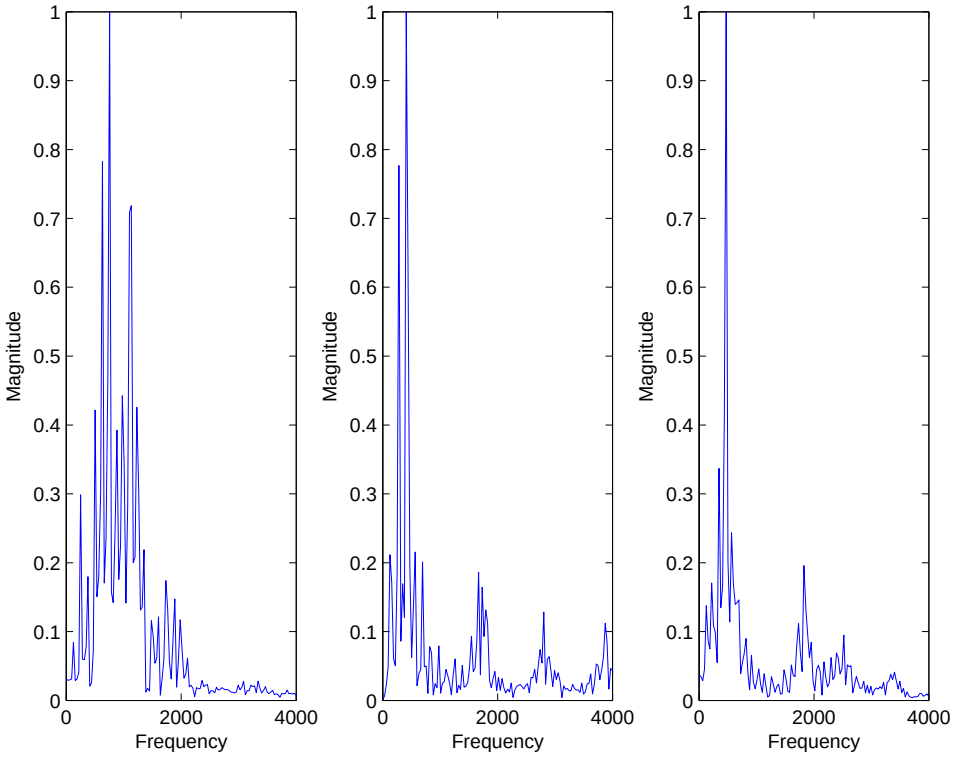


Figure 1.6: Spectra of three different vowels /a/, /i/ and /e/

For example, normalized spectrum of three natural vowels /a/, /i/ and /e/ are presented in Figs. 1.6(a)-1.6(c). It can be clearly observed that all the spectrum show peaks at their formant frequencies. For a clear representation of the formant locations, the It is also evident that the locations of the formant frequencies for these three vowels are different. In practice, a plot of the mean values of the first two formants $F1$ and $F2$ exhibit a condition known as the vowel triangle, where the point vowels /i/, /a/ and /u/ have extreme values and most other vowels have formant values lying close to the sides of the triangle [1]. For example, the /i/-/a/ axis lies close to the front vowels and the /u/-/a/ axis lies close to the back vowels. The formant values are closely related to the articulation of vowels. The vowels for which the tongue has a higher position normally have low $F1$ values, while the second formant $F2$ has a close relation with the forward and backward positions of the tongue. Thus the high front vowel /i/ has the lowest $F1$ and highest $F2$ among all vowels. Again vowel intensity decreases with the increase in tongue height, resulting in very little energy outside the first formant range, as observer

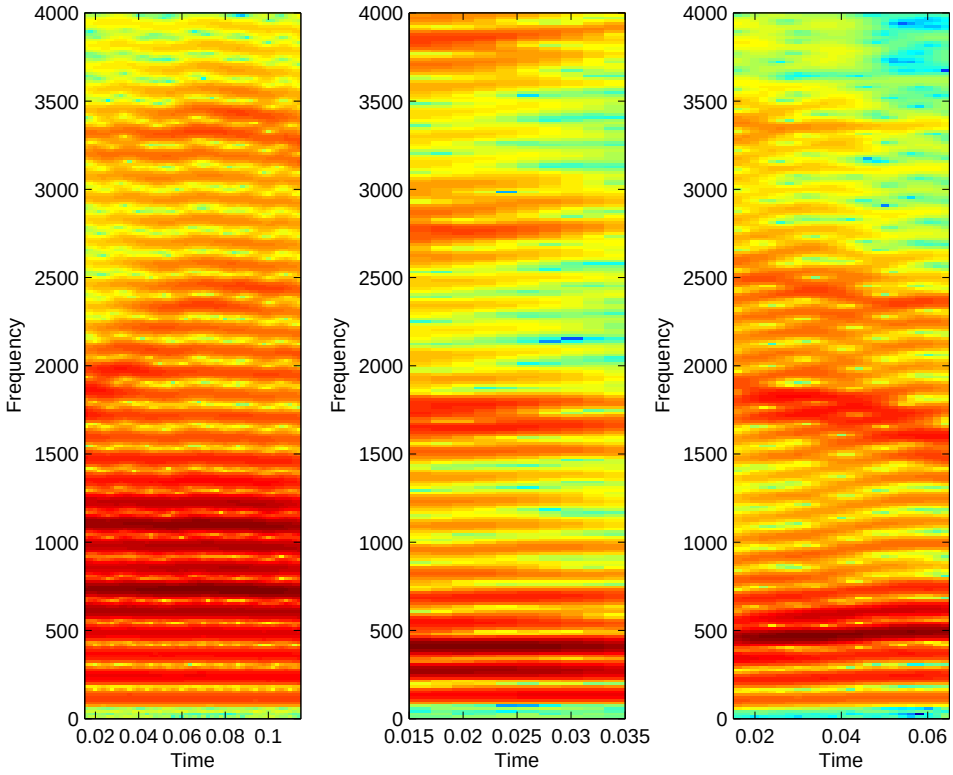


Figure 1.7: Spectrograms for the three vowels whose spectra are presented in Fig. 1.6

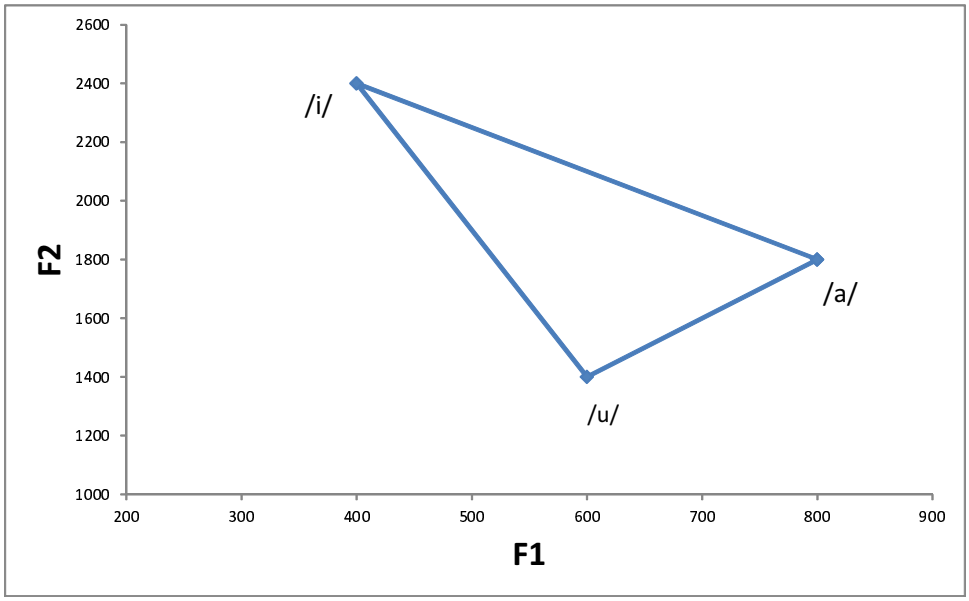


Figure 1.8: The vowel triangle

in Fig. 1.6(b). In order to further clarify the representations of formant locations, in Fig. 1.7, the spectrograms for the same natural vowel utterances are presented. Here it is clearly observed that the formant bands contain very high energy and the dark lines track the variations in formant frequencies with time.

Normally, vowel formants occur on average every 1 kHz for adult males. Thus, formant frequencies are one of the most distinguishing features for vowels.

1.3.2 Problems in Formant Estimation

However, there are some problems regarding formant estimation that arise due to the nature of human speech. As vowel spectrum have a decrease in intensity with the increase in energy, the higher formants have comparatively low energy values, especially for high vowels like /i/. Due to this, identifying the second and third peak from the spectrum becomes difficult. In case of the formant frequencies for female speakers, the formant frequency zones are quite diverse. For instance, the female vowel /i/ has its first formant frequency in the range of 300 Hz while its second formant frequency peak is found at around 2800 Hz, meaning that its second formant frequency is higher than the third formant frequency for the vowel /u/. On the other extreme, male vowel /a/ has a first formant value around 700 Hz and a second formant range at around 1100 Hz, resulting in closely spaced formant peaks.

For instance, the spectrum of a vowel /a/ from a male speaker is presented in Fig. 1.9. It can be seen here that the first and second formant frequencies are very closely positioned and the energy from the first formant affects the second formant, too. So while estimating the formants, it is quite difficult to separate the two formants.

For the task of formant estimation, the conventional methods can be mainly divided into two types, ones that depend on finding the roots for the system and ones that depend on finding the peaks of the spectrum. One of the most basic formant estimation methods was the analysis by synthesis method, where synthetic speech is produced with a varied formant value and is matched with the voiced speech. In linear predictive coding based methods, an estimation is made for relevant speech parameters based on the output

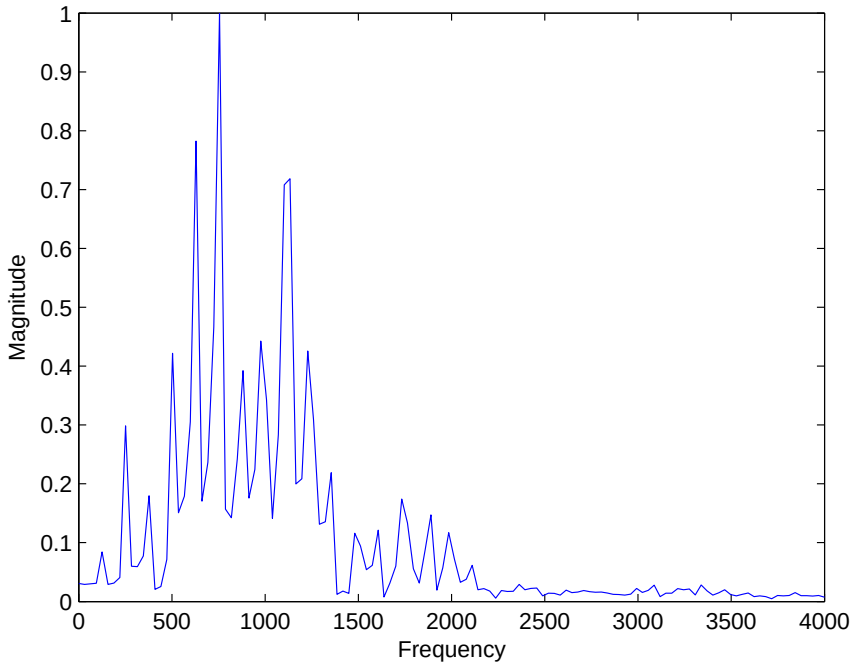


Figure 1.9: Spectrum of a natural vowel sound /a/ uttererd by a male speaker

speech. The roots of the model responsible for the output speech spectrum are predicted and formant frequencies are estimated based on that prediction. On the other hand, peak picking based methods work by finding out the peaks in the speech spectrum. These methods suffer when the peaks in speech spectrum are not distinct.

1.4 Vowel Recognition

In statistical automatic speech recognition (ASR), 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. Language models are often built in the discrete space of word sequence, with the goal of assigning most of the probability mass to the well-formed and meaningful word sequences (i.e. syntactically and semantically correct sentences) while maintaining non-zero (albeit very

small) probabilities to the ill-formed ones. An ASR system thus includes a module of features 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 based on the objective of the underlying application.

There are mainly two types of voice recognition methods , (i) direct matching and (ii) feature based matching. Direct Matching is a simple method but high computation required, more memory space and time consuming. And it is not suitable in some practical cases. Direct method needs all the data value of the speech signal. One of the examples of direct method is cross correlation. It required a machine which has high computational capability. Feature based methods are more suitable for practical uses. It does not use all the values of voice signal but extracts some feature parameters which is used for the voice recognition. This required less computational complexity, less memory and less time consuming. These features can be found in two ways , namely time domain analysis and frequency domain analysis. Again, based on the type of approach used, the ASR systems can be further classified between those with an acoustic phonetic approach, pattern recognition approach and an artificial intelligence approach.

Vowel recognition systems are a specific subset of ASR systems that are concerned with the recognition of voiced vowels only. As these systems deal with a smaller subset, they can be implemented with fast responses. These systems have specific application on digit recognition systems, where the numbers can be identified based on the dominant vowel.

Furthermore, based on the nature of the speech frame on which recognition is carried out, the recognition systems can be divided into isolated speech recognizers, continuous speech recognizers and spontaneous speech recognizers. Isolated speech recognizers usually require each utterance to have quiet (lack of an audio signal) on both sides of the sample window. It accepts single words or single utterance at a time. These systems have "Listen/Not-Listen" states, where they require the speaker to wait between utterances (usually doing processing during the pauses). Continuous speech recognizers allow

users to speak almost naturally, while the computer determines the content. Recognizers with continuous speech capabilities are some of the most difficult to create because they utilize special methods to determine utterance boundaries. Finally, an ASR system with spontaneous speech ability should be able to handle a variety of natural speech features such as words being run together, and even slight stutters.

1.5 Literature Review

1.5.1 Formant Estimation Methods

In recent years, there has been an increasing demand for the development of the accurate, efficient, and compact representations of speech dynamics. Such representations call for the extraction of characteristics of the vocal-tract system from speech signals. Thus vocal-tract system identification has been received potential applications in many research areas of speech processing, such as, speech analysis/synthesis, speech coding, speech recognition, acoustics phonetics, modeling of speech production process. Several system identification methods have been used for the estimation of the vocal tract system parameters. For example, pitch synchronous analysis [2, 3, 4], predictive deconvolution [5], homomorphic deconvolution [6], iterative inverse filtering [7], conventional SI-based methods [8, 9].

Free resonances of the vocal-tract system are called formants. They are associated with the peaks in the power spectrum of speech. Estimation of formant parameters is a difficult problem for which there have been many proposed solutions. Among them, linear predictive coding (LPC)-based methods have received considerable attention [10, 11, 12]. Cepstrum features are also used in the determination of the formants [13, 14, 15].

Most of the formant estimation methods so far reported, including some recent techniques, deal only with the noise-free environments [10, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 12] Recently, there has been a renewed interest in emphasizing formant information, particularly in the context of noise-corrupted speech in different applications. Algorithms, based on the LPC analysis, are very sensitive to the noise level. Very few

research results are so far reported that deal with formant estimation in case of noise-corrupted speech [22, 23, 24, 25]. Multicyclic covariance method, reported in [22], is based on the Prony method. It is shown that using this method the poles of the vocal tract system (equivalently the formant parameters) can be estimated only above SNR of $40dB$. In [23], new spectral analysis methods, based on the statistical properties of the zero-crossing intervals of a noisy signal, are proposed. The SVD based spectral estimation techniques, where parametric models such as AR or ARMA are used, have also been successfully applied in speech signals [26]. However, with the addition of noise, the advantage of the new methods is only prominent in case of first formant estimation up to a certain positive value of SNR. The method, proposed in [24], is completely based on [27]. The parameters involved in this method are only adjusted for a particular synthesized sentence of one male speaker. Errors in estimation of higher formants (i.e. other than the first formant) are very high. Recently in [25], peak-picking algorithm is used on the spectral segments containing the formants. Here, a sequential digital resonator model is used for spectral segmentation, while, in [28], a set of parallel digital formant resonators has been proposed for formant frequency estimation. A major advantage of [25] is that it determines the segment boundaries sequentially and avoids the need for dynamic programming as done in [28]. However, the segmentation algorithm requires access to the ACF of the clean signal, which is not available in noisy environment. To determine an estimate of that noise-free ACF, at the beginning, spectral subtraction technique is used which restricts the performance of the method up to 5 dB. Some of the recent formant estimation methods also provide better results in noise free environments, such as vocal tract modeling method proposed in [29] and multi cyclic covariance method [30]. Conventional time domain formant estimation methods, such as Linear Predictive Coding (LPC), exhibit poor performance in noisy environments [31]. Spectral domain peak picking based formant estimation methods suffer from performance degradations under presence of background noise as spurious peaks are introduced. Some methods proposed so far to deal with noisy environments utilize sequential segmentation of spectrum and time domain Adaptive Bandpass Filter Bank (AFB) [32].

1.5.2 Vowel Recognition Using Formants

It has been known for a long time that the frequencies composition in voiced speech can be used to discriminate between vowels [33]. One of the first digit recognition system ever built used the spectral resonances of the vowels from each of the digits [34]. In the early years of speech recognition when computers were not available, spectrum analyzers were built to be used in vowel recognition systems [35]. In order to obtain better recognition accuracies, better estimation accuracies for formants were demanded, and LPC based parameter estimation methods were employed for better estimation of formant frequencies, leading to better vowel detection systems [36]. However, mel frequency cepstral coefficients have become popular due to their more separable inter class characteristics [37]. Even then, formant frequencies continue to receive attention due to their association with the fundamental nature of voiced speech [38, 28]. Formant frequencies have been utilized to derive parameters characterizing gross VT dimensions, which are in turn used for speaker dependent speech recognition [39]. Formant based feature vectors also offer the advantage of dimensionality reduction, facilitating faster processing [40]. Recently, formants as a distinguishing feature of vowels have received renewed attention due to better understanding of the human speech perception mechanism, where formants play a big role [41]. Methods for efficient separation of vowels and consonants from continuous speech are reported that can be utilized for vowel recognition from continuous speech [42].

One aspect of vowel recognition systems that has largely been overlooked is noise robustness. Methods like minimum mean square error (MMSE) and filter bank based methods has been proposed that can improve the recognition performance in moderately high signal to noise ratios [43, 44]. Model based spectral estimation methods have been reported to provide better performance in the presence of noise [44]. Methods that utilize the autocorrelation operation for a better modeling of AR parameters for speech has also been reported to increase the recognition accuracy [45]. Recently spectro temporal methods are being investigated for improved vowel recognition in the presence of noise [46].

1.6 Objective of the Thesis

The objective of the thesis are to

- Develop formant estimation methods that offer significant performance improvements over conventional methods in the presence of noise.
- Evaluate the performance improvements in automatic speech recognition systems incorporating the estimated formants in their feature vector.

1.7 Organization of the Thesis

The major objectives of this thesis are to develop noise robust formant estimation techniques that can perform even at a very low SNR. In the next chapter, first a general explanation of the spectral representation of the vocal tract is presented and the problem is formulated. Then the problems in formant estimation arising due to the presence of environmental noise is discussed, and autocorrelation is presented as an operation that exhibit significant noise reduction. Afterward, a novel spectral matching technique is developed based on the spectrum of autocorrelation of speech. Thereafter the problem of vowel recognition in the presence of noise is presented, and a feature set incorporating the formants estimated from the proposed noise robust method is proposed that can offer significantly better recognition performance under the presence of severe background noise. It is to be noted that extensive experimentation were done to evaluate the performance of the algorithms throughout the thesis on the TIMIT speech corpus, which contains a comprehensive selection of uttered sentences from English speakers. Also variations in the estimation performance by varying the frame lengths is observed and analyzed.

In chapter 3, the similar problem of noise robust formant estimation technique is developed based on a spectral matching technique involving the repeated autocorrelation of speech. Due to the advantages offered by the pole preserving and pole increasing nature of autocorrelation, repeated autocorrelation offers even better noise robustness compared to single autocorrelation operation, and the previously proposed spectral matching

technique is expanded to incorporate double autocorrelation operation. As in previous section, a linear discriminant analysis based classifier is deployed for the task of vowel recognition and in the presence of noise, vowel recognition accuracies show improvements compared to that obtained using conventional mel frequency cepstral coefficients. Extensive experimentation on the TIMIT speech corpus is performed and the obtained results are compared with performances of traditional methods like linear predictive coding and adaptive filter bank methods.

The formant estimation technique involving double autocorrelation is again addressed in Chapter 4, where a band limiting method based on the observed formant frequency zones is presented. Due to the effect of repeated autocorrelation on the spectrum of speech, better estimation performance can be obtained for second and third formant frequencies by band limiting the speech signal first and then performing double autocorrelation. A new spectral model for the repeated autocorrelation for band limited speech signals is presented and used in the proposed spectral matching technique.

In the final chapter, chapter 5, ideas for future improvements are presented and the whole scenario of this literature is summerized with some concluding remarks.

Chapter 2

Spectral Model of Autocorrelation of Speech

Formants are the free resonances of vocal tract, which represent distinguishable characteristics of human voice. Formant frequencies are associated with peaks in the smoothed spectrum of a speech signal [13]. Formants are widely used in many applications, such as speech synthesis, emotion detection [47], and voice disorder detection [48]. In particular, formants of the voiced sound can serve as a unique voice template, which can be used as a fundamental speech property in Automatic Speech Recognition (ASR). Since in the real life speech is corrupted by various types of noise, a formant estimation method with its performance robust to noise is required to be designed. But, estimating formant accurately in the presence of a severe background noise becomes extremely difficult task. Most of the formant frequency estimation methods deal with only noise-free environments. Some of the recent formant estimation methods also provide better results in noise free environments, such as vocal tract modeling method proposed in [29] and 2-D time-frequency transformations proposed in [22]. However, effects of noise on them were not investigated. The multi cyclic covariance method can detect formant frequencies at relatively high signal to noise (SNR) ratios [30]. Conventional time domain formant estimation methods, such as Linear Predictive Coding (LPC), exhibit poor performance in noisy environments [31]. Spectral domain peak picking based formant estimation meth-

ods suffer from performance degradations under presence of background noise as spurious peaks are introduced. Some methods proposed so far to deal with noisy environments utilize sequential segmentation of spectrum and time domain Adaptive Bandpass Filter Bank (AFB) [32]. However, an accurate estimation of formants in the presence of severe noise yet remains a challenging task.

In this chapter, an efficient scheme for estimating the formant frequencies in the presence of noise is presented. In order to overcome the effects of noise in formant estimation, operations that offer similar advantages as strengthening the poles responsible for the formant frequencies are investigated. Autocorrelation operation, which strengthens the dominant poles, and exponentially increases the peak-valley ratio at formant frequencies of the magnitude response, is proposed to be employed with the purpose of canceling out the effects of noise. Formant estimation is carried out in the spectral domain where instead of direct peak-picking from the speech spectrum, a spectral domain model of autocorrelation function (ACF) of speech signal is first proposed considering the vocal tract to comprise of cascaded subsystem responsible for single resonant frequencies. A spectral domain model fitting based algorithm is also developed to extract the model parameters which in turn give the formant. Through the simulation results on standard speech databases, it is shown that the developed method is effective in maintaining a high success rate in formant estimation even in the presence of a significant background noise.

2.1 Background

A typical voiced speech signal is the result of air passing through the human vocal tract, with multiple resonances created due to the structure of the vocal tract. The vocal tract system can be modeled as an autoregressive (AR) filter whose input is a periodic impulse train [49]. These resonant frequencies, known as formants, are evident as peaks in the spectral domain representation of speech. As the ranges for formant frequencies of different vowels are different, this property can be used in ASR systems for vowel

detection. However, real life speech signals are affected by background noise, which alters its frequency spectrum of and can make the detection of formant frequencies difficult.

2.2 Proposed Method

In this section, the composition of the human vocal tract system used to produce voiced sounds is first investigated. Then the effect of everyday noise on voiced speech is demonstrated. Then methods for countering the effect of noise on formant estimation are evaluated and the performance of autocorrelation as a facilitator for better formant detection under noise is demonstrated. Then a model for the human vocal tract is developed, considering the vocal tract to comprise of cascaded subsystems responsible for a single formant frequency. Finally a model matching method is developed for extracting the formants from the autocorrelation of band limited speech.

2.2.1 Spectral Representation of the Vocal Tract System

In order to estimate the formant frequencies from observed speech signal, it is sufficient to restrict the analysis only for the voiced sound. In case of the voiced speech signals, considering the excitation as a periodic impulse-train, the overall vocal tract filter can be represented by a P-th order autoregressive (AR) system with the following transfer function

$$H(z) = \frac{C}{\prod_{i=1}^{P}(1 - p_i z^{-1})} \tag{2.1}$$

where p_i denotes the pole of the AR system and C is the gain factor. As mentioned before, the resonances of the vocal tract correspond to the formant peaks in the speech spectrum [13]. Each pair of complex conjugate poles in the AR system can generate a peak in the frequency response. Hence, the vocal tract system in (2.1) can exhibit $P/2$ formants. However, as far as formant estimation is concerned, only the first three formants are significant and contain a very high portion of the total energy. In this regard, it would be sufficient to consider the vocal tract to be represented as a cascaded network of three

separate subsystems, each causing a resonant peak in the speech spectrum.

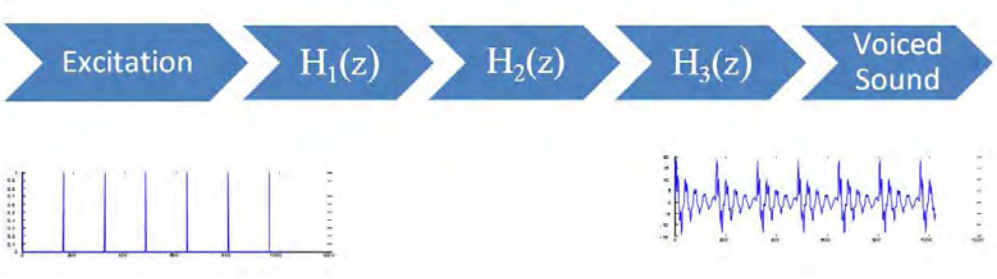


Figure 2.1: Voiced sound generation through a simplified model of vocal tract system

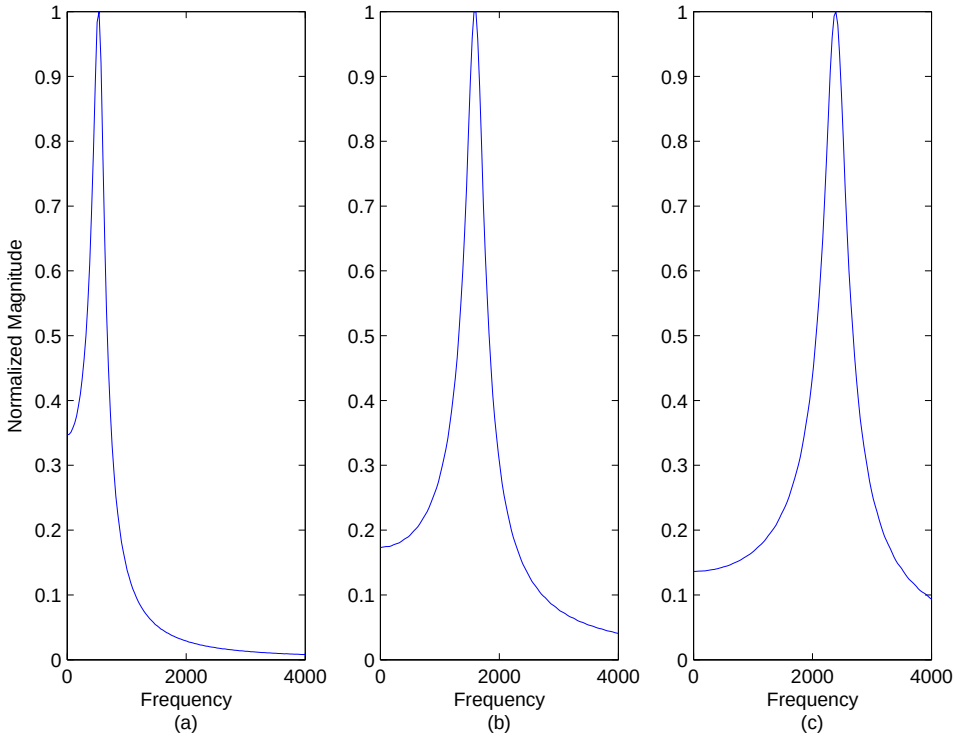


Figure 2.2: Frequency response of the individual subsystems responsible for single formants

In Fig. 2.1, a simplified vocal tract model consisting of three subsystems is shown where a periodic impulse train is used as the excitation, which can be expressed as

$$u_{imp}(n) = \sum_{i=0}^{\lambda-1} \delta(n - iT) \tag{2.2}$$

where T is the period of the impulse train and λ denotes the total number of impulses. Each individual subsystem in Fig. 2.1 can be represented as

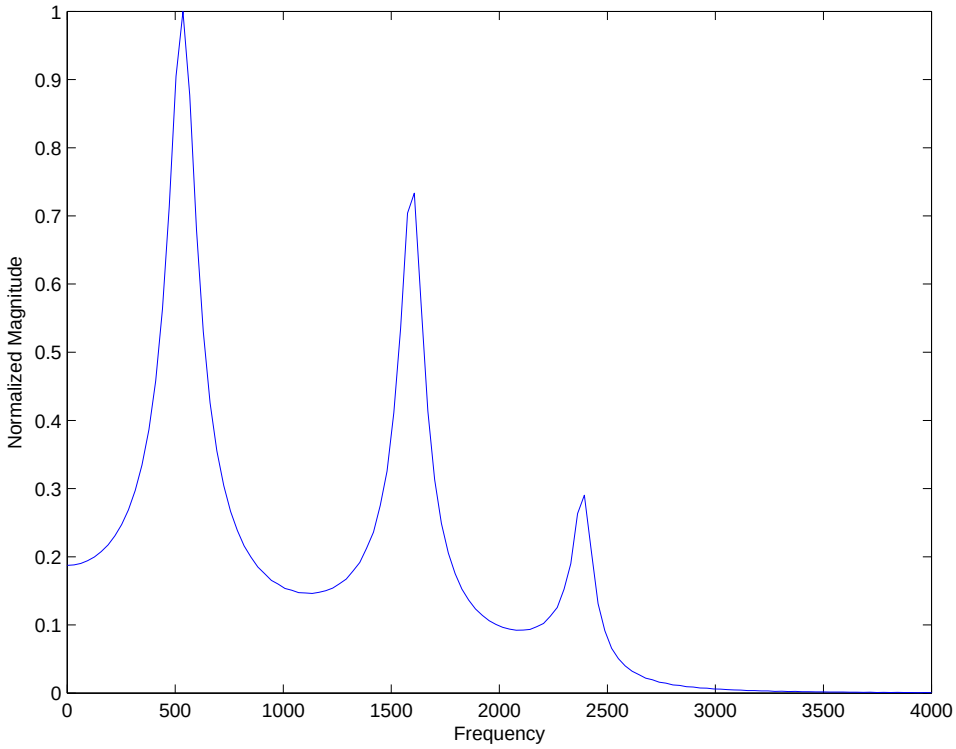


Figure 2.3: Frequency response of the overall vocal tract system

$$H_i(z) = \frac{C_i}{(1 - p_i z^{-1})(1 - p_i^* z^{-1})} \tag{2.3}$$

Here for each pair of complex conjugate poles $p_i = r_i e^{(j\theta_i)}$ the magnitude r_i and angle θ_i are related to a particular formant F_i and the formant bandwidth B_i as

$$r_i = e^{\frac{-\pi B_i}{Fs}} \tag{2.4}$$

$$\theta_i = \frac{2\pi F_i}{Fs} \tag{2.5}$$

where Fs is the sampling frequency. In Fig. 2.2, the frequency responses corresponding to each subsystem shown in Fig. 2.1 are presented, which clearly demonstrates the behavior of spectral peak as discussed above.

In Fig. 2.3, the overall frequency response of the cascaded system considered in Fig. 2.1 is shown. Next a synthetic sound is generated based on the model shown in Fig. 2.1 and the spectrum corresponding to that synthetic sound is shown in Fig. 2.4. It can be

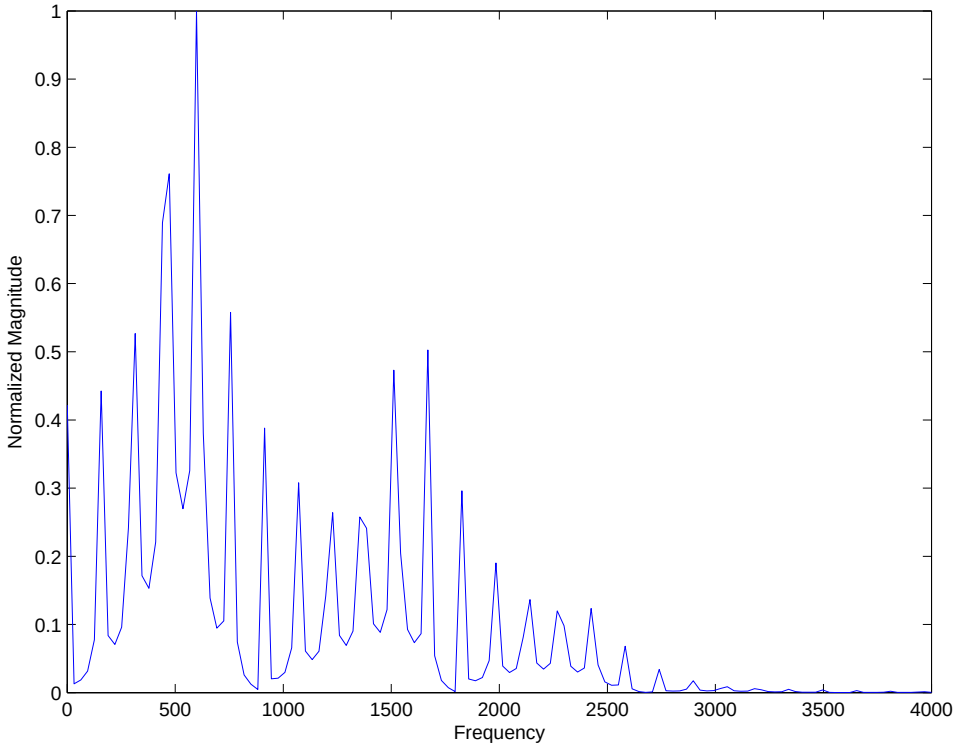


Figure 2.4: Frequency spectrum of a synthetic sound generated by the system whose frequency response was presented in Fig. 2.3

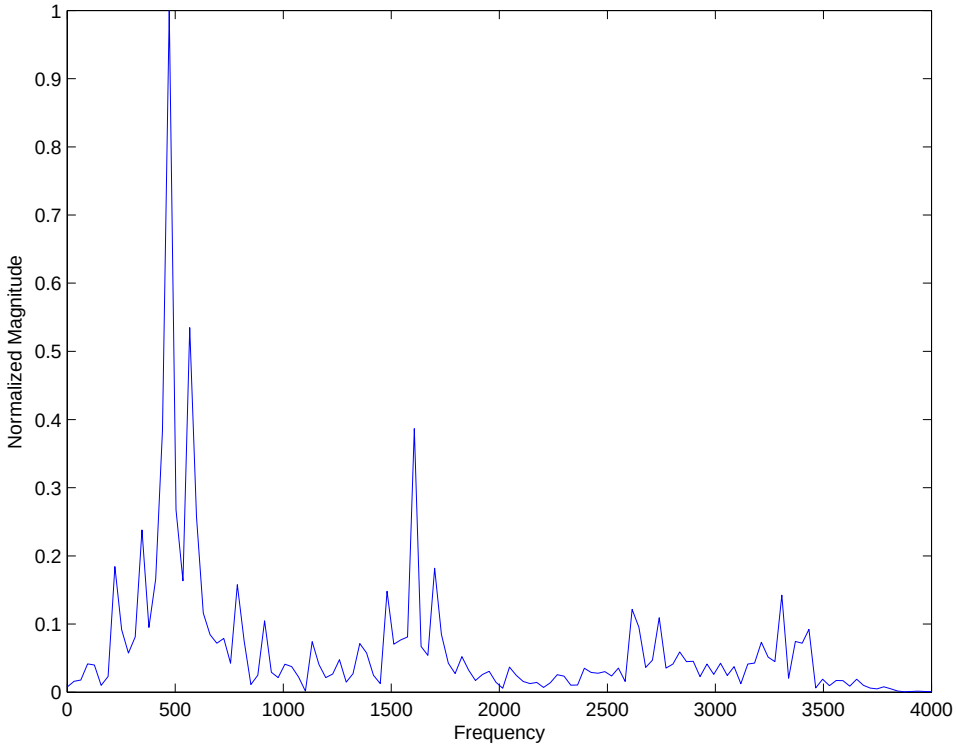


Figure 2.5: Frequency spectrum of natural voiced speech /eh/ in noise free environments

observed that the location of formant peaks are preserved both in Fig. 2.3 and Fig. 2.4. However, the reason behind the lack of smoothness in the spectrum as seen in Fig. 2.4 is mainly because of the nature of the excitation, which in this case is a periodic impulse train. Apart from the synthetic sound, for a more clear understanding, a natural voiced sound /eh/ is taken which contains formants at locations that closely match with the formant locations of the synthetic sound considered before. In Fig. 2.5, the spectrum corresponding to the natural sound is shown. It is observed that the spectra obtained from the natural and synthetic speech signals match closely even though a simplified model is used to generate the synthetic sound. Hence in order to estimate formant frequencies, one can employ the conventional method of spectral peak picking on speech spectrum or can look for a suitable spectral model that fits the speech spectrum. However, due to the effect of the fundamental frequency or pitch of speech and the presence of noise in real life scenario, these approaches may not be able to provide accurate formant estimation.

2.2.2 Formant Estimation in Noise

For noise-free voiced speech signal conventional peak picking formant estimation methods may provide satisfactory results. However, presence of background noise is very common in everyday situations and it affects the accuracy of traditional estimators.

For a voiced sound $x(n)$ in the presence of additive noise $v(n)$ with zero mean and unit variance, the noise corrupted speech $y(n)$ can be written as

$$y(n) = x(n) + v(n) \tag{2.6}$$

In a time domain representation of the noise corrupted speech signal, it is very difficult to distinguish the original speech samples even at a moderate level of noise. The presence of additive noise completely destroys the original speech pattern resulting in a noise like pattern. In order to show the effect of noise in time domain, in Figs. 2.7(a) and 2.7(b), a noise free speech $x(n)$ and corresponding noise corrupted speech $y(n)$ are shown, respectively. Here the natural sound /eh/ is considered and its spectral representation is presented in Fig. 2.6.

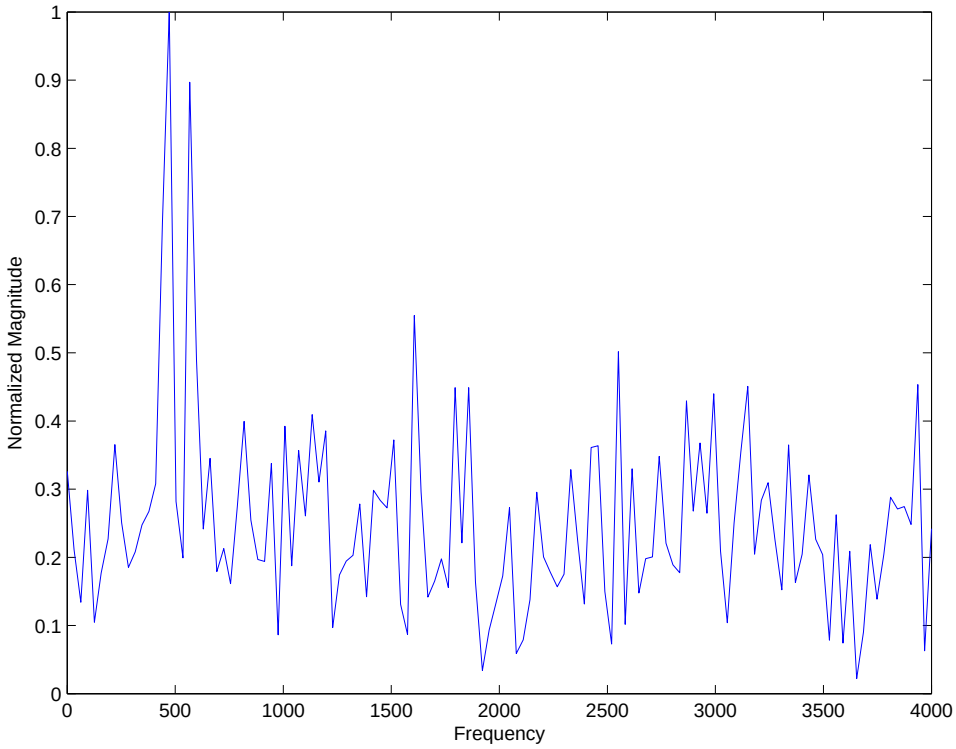


Figure 2.6: Spectrum of natural voiced speech /eh/ under $-5dB$ background noise

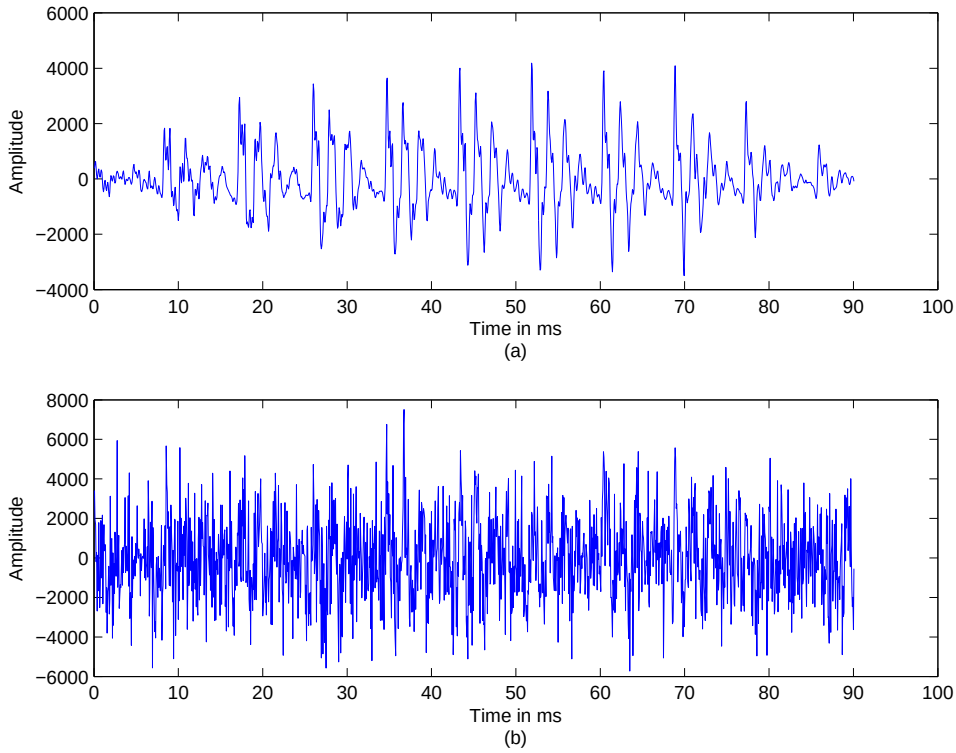


Figure 2.7: (a) Time domain waveform of an utterinance of /eh/ and (b) the same waveform under $-5dB$ background noise

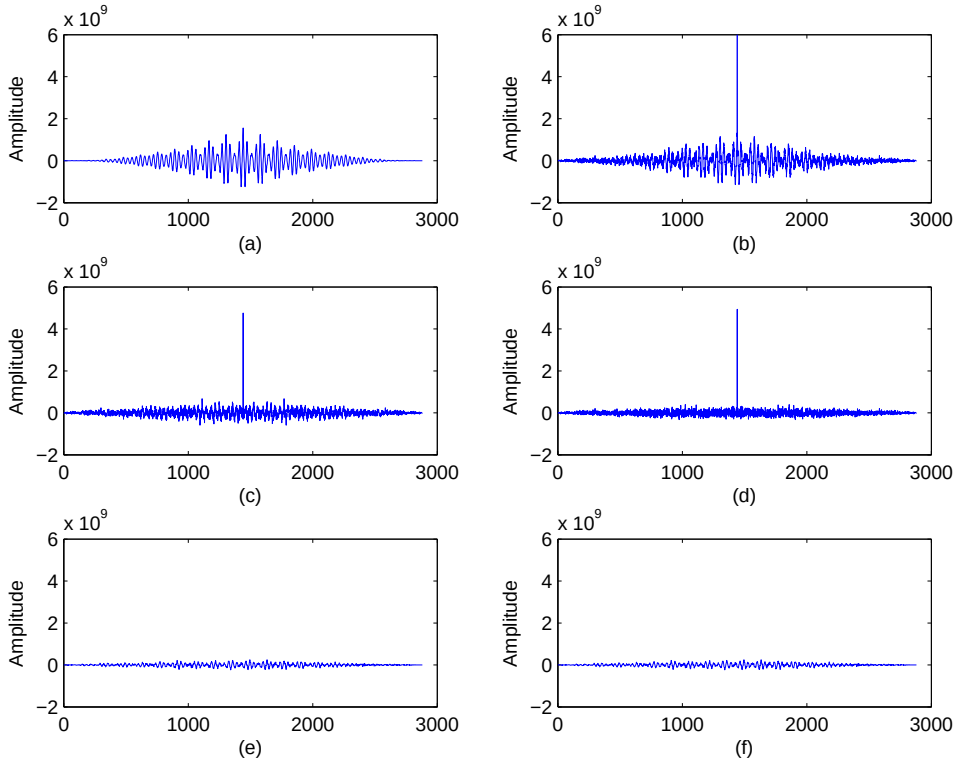


Figure 2.8: Effect of noise in the autocorrelation domain: plot of different autocorrelation functions (a) $r_x(n)$, (b) $r_y(n)$, (c) $r_w(n)$, (d) $r_v(n)$, (e) $r_{xv}(n)$ and (f) $r_{vx}(n)$

For the purpose of comparison, here the same natural sound that is considered in Fig. 2.5 is taken in the presence of additive white Gaussian noise (AWGN) and the SNR is set at $-5dB$. It is clearly observed in Fig. 2.6 that the presence of noise causes several spurious peaks and obscures some of the formant peaks in the frequency spectrum causing error in formant estimators. It is obvious that the formant peaks with lower magnitudes will be heavily affected because of the presence of noise.

As formant frequencies are represented by peaks in speech spectrum, methods that facilitate better detection of peaks are highly desired for a formant estimation system. One possible solution for this can be to increase the strength of the poles, in turn increasing the strength of the related spectral peaks. In order to achieve this goal, one may duplicate the existing poles, by placing additional poles at the positions of original poles. However, as far as the construction of the vocal tract system cannot be changed, this approach would not be feasible in practical situations. This has led to the search for a method that can be used to imitate duplication of the poles and autocorrelation emerges as a prime candidate. In what follows, it is shown that the ACF of an impulse

response $h(n)$ of the AR system can be represented as the impulse response of a system that possesses twice the number of poles of the AR system and among these poles, apart from the original poles, there are some new poles located at conjugate reciprocal positions of the original poles with respect to the unit circle. If the original poles were all inside the unit circle, the additional poles generated by the autocorrelation operation will lie outside the unit circle but at the same angles.

The autocorrelation function (ACF) of a voiced sound $x(n)$ is defined as

$$r_x(\tau) = E[x(n)x(n-\tau)] \tag{2.7}$$

where τ denotes the lag. ACF is an even function, with the output being symmetric with respect to the amplitude axis. In practical application the ACF of $x(n)$ is computed by using the working formula given below

$$r_x(n) = \frac{1}{N} \sum_{k=0}^{N-1-|n|} x(k)x(k+|n|), n = 0, 1, 2. \ldots\ldots\ldots, M-1 \tag{2.8}$$

Using (2.6) and (2.7), the ACF of noisy speech y(n) can be expressed as

$$\begin{aligned} r_y(n) &= r_x(n) + r_w(n) \\ r_w(n) &= r_v(n) + r_{vx}(n) + r_{xv}(n) \end{aligned} \tag{2.9}$$

Here $r_v(n)$ 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 $r_x(n)$, 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 of all other lags should be very small and ideally should be zero. In Figs. 2.8(a)-2.8(f), different ACFs, namely $r_x(n)$, $r_y(n)$, $r_w(n)$, $r_v(n)$, $r_{xv}(n)$ and $r_{vx}(n)$ are plotted. From Figs. 2.8(e) and 2.8(f), it can be observed that the values of the cross correlation terms are very small as expected. However, as seen in Fig. 2.8(d), although $r_v(n)$ exhibits a very large peak at the zero lag, nonzero small values exist at all other lags because of the finite data length. It is also observed in Fig. 2.8(c) that $r_w(n)$ 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. 2.7, the effect of $r_w(n)$ on $r_x(n)$ is drastically reduced because of the autocorrelation operation. Since the autocorrelation is a pole preserving operation and it exhibits higher noise immunity, it is advantageous to deal with the ACF of $y(n)$ instead of directly using $y(n)$ in spectral domain formant estimation.

Considering $x(n)$ as an output of an LTI system with transfer function $H(z)$, $x(n)$ can be written as

$$x(n) = h(n) * u(n) \tag{2.10}$$

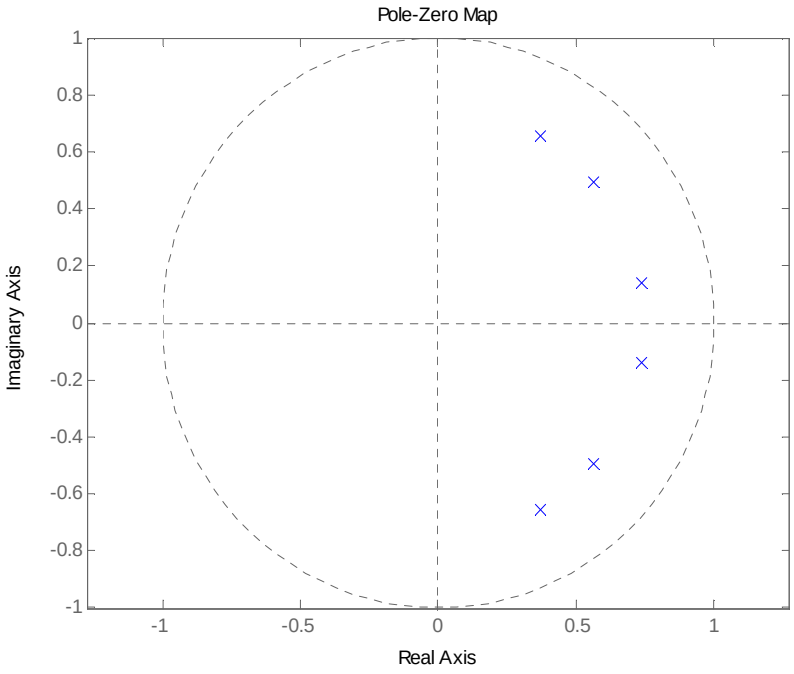
It can be shown that the ACF of $x(n)$ can be expressed as

$$r_x(n) = r_h(n) * r_u(n) \tag{2.11}$$

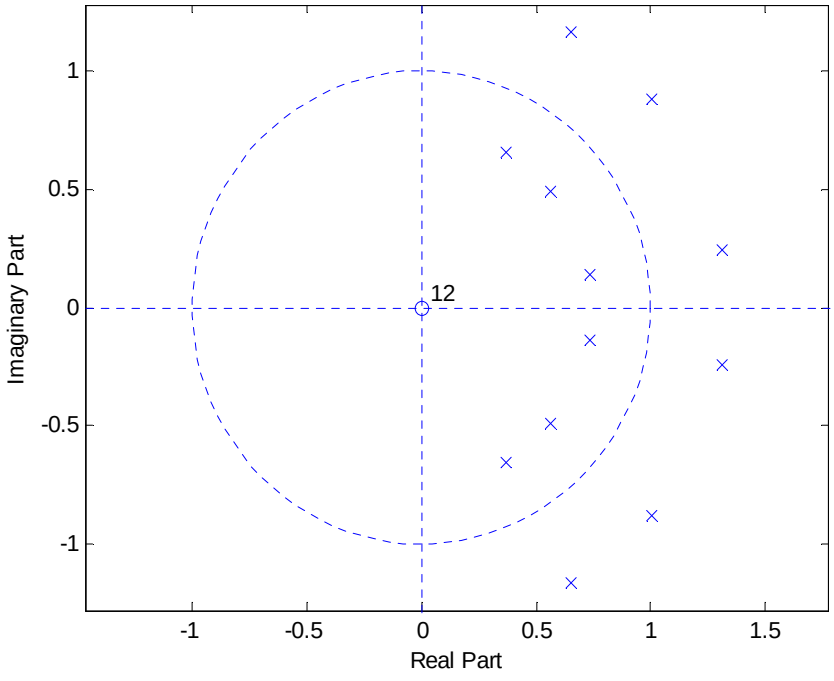
where $r_u(n)$ is the ACF of $u(n)$. For a voiced sound with periodic impulse train excitation, when the length of the period is sufficiently large, the variation of $r_x(n)$ within a period will match closely with that of $r_h(n)$. Next a synthetic signal $x(n)$ is generated from an AR(6) system using a periodic impulse train excitation with a period of $T = 200$ samples. It is expected that in the frequency domain representation of $r_h(n)$ and $r_x(n)$, dominant peak locations will be similar. Hence, in what follows to develop a frequency domain scheme for formant frequency estimation it would be sufficient to consider the detailed analysis of $r_h(n)$ instead of $r_x(n)$. As per the definition of the ACF provided in (2.7), the ACF of $h(n)$ can be written as

$$r_h(n) = h(n) * h(-n) \tag{2.12}$$

In view of analyzing the frequency domain effects, for simplicity, first the Z domain representation is considered. The Z transform of $r_h(n)$, as obtained from (2.12) is given by



(a)



(b)

Figure 2.9: Effect of autocorrelation in z-domain (a) $H(z)$ (b) $R_h(z)$

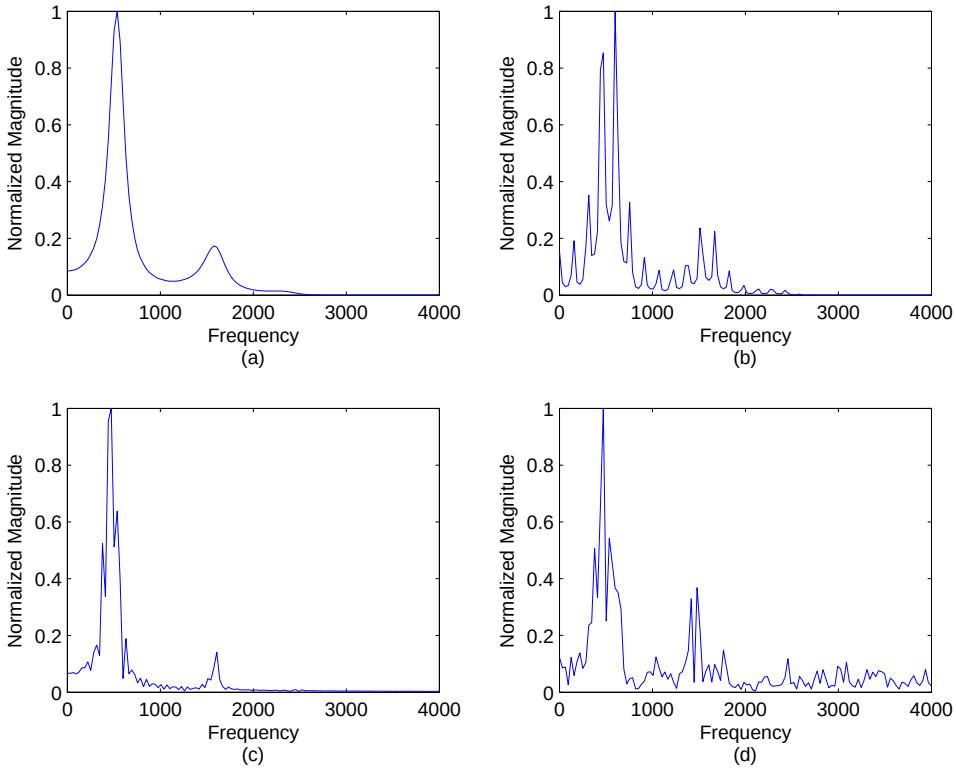


Figure 2.10: Effect of spectral strengthening because of autocorrelation operation. Spectrum of: (a) $r_h(n)$, (b) $r_{synx}(n)$, (c) $r_x(n)$ and (d) $r_y(n)$

$$R_h(z) = H(z)H(z^{-1}) \tag{2.13}$$

It is observed from (2.13) that the ACF operation produces a set of new poles with equal number of the original system poles corresponding to $H(z)$ and the new poles corresponding to $H(z^{-1})$ are located at the conjugate reciprocal location of the original poles. For a clear understanding, a sample z-plane pole representation of $H(z)$ having three pairs of complex conjugate poles and corresponding $R_h(z)$ are shown in Figs. 2.9(a) and 2.9(b) respectively. It is seen that from the figure 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 peak strengthening effect is important as it helps 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 both in noise-free and noisy condition,

in Fig. 2.10, spectra corresponding to $r_h(n)$, $r_{synx}(n)$, $r_x(n)$ and $r_y(n)$ are shown. It is to be mentioned that the synthetic speech considered here to calculate $r_{synx}(n)$ and $r_h(n)$ is the one that is used in Fig. 2.4, while the same natural sound /eh/ as shown in Fig. 2.5 is used to obtain the spectrum of $r_x(n)$. In case of noise for both the synthetic and natural sound, 0dB background noise is used. In comparison to the spectra corresponding to $y(n)$, it is clearly observed that the spectra corresponding to $r_y(n)$ exhibits better noise immunity. Strengthening of dominant peaks is evident as the first formant peak is significantly strengthened. However, in comparison to the increase in the first formant peak, the spectral peaks corresponding to other formants remain very weak. In view of overcoming this problem, one practical solution is to consider the vocal tract to consist of cascaded subsystems, each responsible for a single formant peak, as presented in Fig. 2.1. As the subsystems responsible for the formants are in a cascaded formation, the spectral output of the whole vocal tract system is the product of the separate spectral outputs for the subsystems. 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. Considering only the first three formants, the impulse response $h(n)$ of the whole system can be written as

$$h(n) = h_1(n) * h_2(n) * h_3(n) \tag{2.14}$$

where $h_1(n)$, $h_2(n)$ and $h_3(n)$ are the impulse responses of the individual systems. After performing autocorrelation, the system impulse response becomes

$$r_h(n) = r_{h1}(n) * r_{h2}(n) * r_{h3}(n) \tag{2.15}$$

The Z Transform of $r_h(n)$, as obtained from (2.15) is given by

$$R_h(z) = R_{h1}(z)R_{h2}(z)R_{h3}(z) \tag{2.16}$$

The first formant peak is prominent in the spectrum of ACF presented in Fig. 2.10, indicating that the effect of $R_{h2}(z)$ and $R_{h3}(z)$ are negligible on $R_{h1}(z)$. Using this

property, it can be assumed that the output response closely match $R_{h1}(z)$ 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.

2.2.3 Proposed Spectral Model of ACF of Speech

As seen from the previous section, the spectrum of the vocal tract response within a particular formant band generally exhibits a prominent peak corresponding to the formant. 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. Hence it is sufficient to consider a band limited speech signal corresponding to a particular formant band to analyze the effect of an individual formant. In this regard, according to (2.16) considering the vocal tract system as a cascade of a set of subsystems, each subsystem that is responsible for generating a formant peak is denoted as $H_i(z)$.

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 ACF operation 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 output ACF of the most prominent subsystem, namely $H_1(z)$. 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 ACF of the noise corrupted signal. In this case, the transfer function of the subsystem responsible for the ACF spectrum around the first formant peak as per (2.13) can be represented as

$$R_{h1}(z) = \frac{C_{R1} z^2}{(1 - p_1 z^{-1})(1 - p_1^* z^{-1})(1 - p_1 z)(1 - p_1^* z)} \tag{2.17}$$

where C_{R1} is a constant.

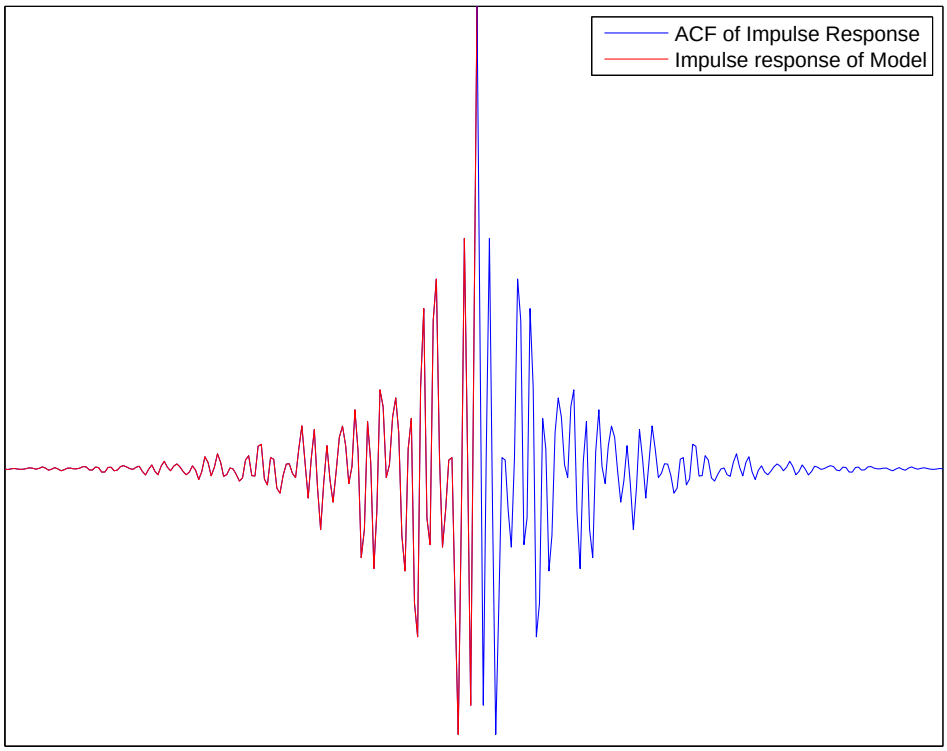


Figure 2.11: Impulse response of $R_{h1}(z)$ and the ACF of the impulse response

With the introduction of each new pole outside the unit circle, a trivial zero is also introduced at the origin. The effect of these zeros is to introduce a delay in the output. Thus a pair of zeros is incorporated in (2.17). If the ACF of an impulse response for a synthetic speech signal is taken, it is expected that this will match with an impulse response obtained from a system which contains new poles in addition to the original ones, as described above. This is evident in Fig. 2.11, where these two signals match perfectly, showing the validity of the proposed model generation approach. If the trivial zeros were not included while constructing the new system for generating an impulse response similar to the ACF, we would have experienced a delay between the signals, which can be observed in Fig. 2.12.

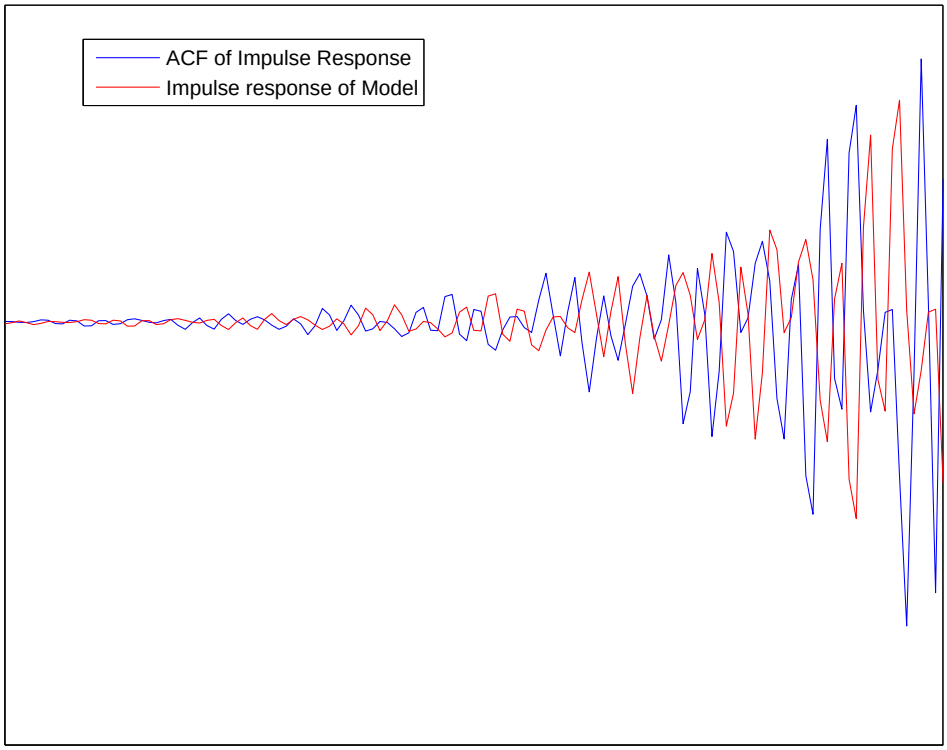


Figure 2.12: Impulse response of $R_{h1}(z)$ and the ACF of the impulse response without using trivial zeros

2.2.4 Proposed Spectral Matching Technique

In the proposed formant estimation method, a spectral model corresponding to the first formant zone of the spectrum of the ACF 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 the z-transform representation given by (2.17), for estimating each formant one such Z-domain model is required and the i-th model within the region of convergence of $r_i < z < \frac{1}{r_i}$ can be represented in the Fourier transform as

$$R_{Mi}(e^{j\omega}) = \frac{C_i e^{j2\omega}}{(1 - p_i e^{-j\omega})(1 - p_i^* e^{-j\omega})(1 - p_i e^{j\omega})(1 - p_i^* e^{j\omega})} \qquad (2.18)$$
$$p_i = r_i e^{j\theta_i}$$

where the spectrum $R_y(e^{j\omega})$ of the ACF of the observed noisy signal $y(n)$ is used in conjunction with the proposed model $R_{Mi}(e^{j\omega})$ to form an objective function and for the first formant with $i = 1$ based on the square of absolute difference of these spectra, namely

$$\begin{aligned} e^i_{min}(r_j, \theta_j) = \quad & min \quad \sum_{\omega=\omega_{lc}}^{\omega_{hc}} \left(|R_{Mi}(e^{j\omega})| - |R^i_y(e^{j\omega})|\right)^2 \\ & r_l < r_i < r_h \\ & \theta_l < \theta_i < \theta_h \end{aligned} \tag{2.19}$$

Note that here the superscript i is introduced to control the step by step algorithm. In particular, at the first step with $i = 1$, in order to obtain $e^1_{min}(r_1, \theta_1)$, one has to consider $R^1_y(e^{j\omega}) = R_y(e^{j\omega})$. However in later part it will be shown that $R^i_y(e^{j\omega})$ will vary for different formants. 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 [30]. Formant frequencies are estimated from the pole angle θ_j that produces the best match between the spectra using (2.5).

Once the first formant frequency $F1$ is obtained, (2.16) is utilized to estimate the second formant frequency $F2$. $R_y(e^{j\omega})$ can be written as the product of $R_{y1}(e^{j\omega})$, $R_{y2}(e^{j\omega})$ and $R_{y3}(e^{j\omega})$ according to (2.16). The magnitude spectrum of $R_y(e^{j\omega})$ is divided by $R_{M1}(e^{j\omega})$ so that the resulting spectrum $R^2_y(e^{j\omega})$ closely resembles the product of $R_{y2}(e^{j\omega})$ and $R_{y3}(e^{j\omega})$. Hence $R^i_y(e^{j\omega})$ in general for estimating second and third formant can be expressed as

$$R^i_y(e^{j\omega}) = R^{i-1}_y(e^{j\omega}).(R_{M(i-1)}(e^{j\omega}))^{-1}, i > 1 \tag{2.20}$$

Then similar to the matching in the first formant zone, matching is performed in the

second formant zone and $F2$ is estimated. Then the magnitude spectrum of $R_y^2(e^{j\omega})$ is divided by $R_{M2}(e^{j\omega})$ to obtain $R_y^3(e^{j\omega})$. According to the simplified modeling of the vocal tract presented above, $R_y^3(e^{j\omega})$ should closely match with $R_{M3}(e^{j\omega})$, leading to a similar approach as described in (2.18) and (2.19) to obtain $F3$.

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 consideration 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/or the formants are very closely spaced.

A simple block diagram representing the major steps involved in the proposed formant frequency estimation scheme is presented in Fig. 2.13. Here it is to be noted that a feedback from the estimated first and second formants is taken in selecting the pass band ranges of the bandpass filters corresponding second and third formants.

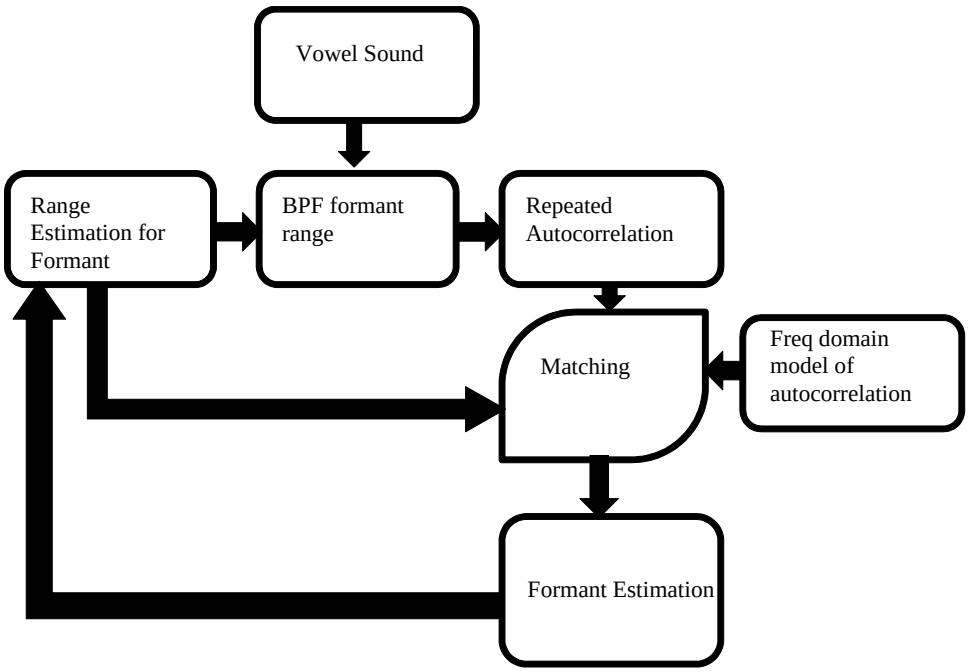


Figure 2.13: Block diagram of the proposed formant estimation system

2.2.5 Vowel Recognition

After estimating formants in this manner, in the proposed scheme they are employed in a vowel recognition systems as potential features along with the commonly used Mel frequency cepstral coefficients (MFCC). For the purpose of recognition two major steps are followed. First given the train data set for different 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 covariance matrix, defined as

$$S = \sum_{i=1}^{N} [x_i - \mu][x_i - \mu]^T \tag{2.21}$$

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

$$S_b = N_+[\mu_+ - \mu][\mu_+ - \mu]^T + N_-[\mu_- - \mu][\mu_- - \mu]^T \tag{2.22}$$

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

$$[\mu_+ - \mu] = \frac{N_-}{N}[\mu_+ - \mu_-] \tag{2.23}$$

and

$$[\mu_- - \mu] = -\frac{N_+}{N}[\mu_+ - \mu_-] \tag{2.24}$$

using the values obtain from (2.23,2.24) in (2.22), the between class scatter matrix is obtained as

$$S_b = \frac{N_- N_+}{N}[\mu_+ - \mu_-][\mu_+ - \mu_-]^T \tag{2.25}$$

in addition, the within class scatter matrix is defined as

$$S_w = \sum[x_i - \mu_+][x_i - \mu_+]^T + \sum[x_i - \mu_-][x_i - \mu_-]^T \tag{2.26}$$

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 [50]

$$J(w) = \frac{w^T S_b w}{w^T S_w w} \tag{2.27}$$

2.3 Simulation Results and Discussion

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) [51]. 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 [52]. The formant estimates reported in [52] 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 analyzed. In VTR database, formant estimates are reported for every 10 ms interval. However, vowel duration is general 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 [53] 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 [32].

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

Table 2.1: Comparison of the estimation performance for synthetic vowels

Vowels			$5dB$			$-5dB$		
			Proposed	LPC	AFB	Proposed	LPC	AFB
Male	/a/	F1	4.86	21.57	43.65	4.95	24.53	47.17
		F2	10.17	7.24	25.74	7.23	99.56	27.25
		F3	12.48	20.49	10.68	17.57	39.35	10.42
	/o/	F1	5.43	61.38	124.73	7.36	73.15	21.63
		F2	14.44	167.49	43.93	17.32	144.60	58.65
		F3	17.90	36.74	12.54	17.90	37.68	11.66
	/u/	F1	5.84	93.53	149.02	9.13	117.36	13.56
		F2	11.13	158.74	46.60	14.45	148.07	63.59
		F3	2.59	69.03	38.05	2.82	72.38	19.40
Female	/a/	F1	5.62	20.24	46.90	5.98	20.46	49.77
		F2	9.27	65.23	32.58	6.98	113.79	30.99
		F3	7.65	17.80	8.45	11.35	34.02	9.84
	/o/	F1	11.03	49.53	128.07	22.16	78.29	18.29
		F2	10.05	138.88	20.42	17.89	133.29	46.61
		F3	4.80	39.93	9.56	7.41	36.28	12.53
	/u/	F1	10.02	72.96	109.00	10.02	98.29	12.98
		F2	9.39	116.33	14.62	13.89	121.92	33.72
		F3	6.74	52.31	11.40	7.64	40.60	13.74

Table 2.2: Number of samples and average duration for different vowels available in the TIMIT database

Vowel	Male		Female	
	No of Occurrences	Average Duration	No of Occurrences	Average Duration
'aa'	129	124.96	93	122.90
'ae'	112	136.16	58	128.62
'ah'	98	88.57	82	80.98
'eh'	194	90.15	134	97.09
'ih'	187	78.02	160	79.31
'ix'	397	48.72	309	50.26
'iy'	262	87.14	179	92.18
'ow'	94	125.74	54	118.15
'uh'	18	77.33	23	79.13
'uw'	30	96.33	14	116.43
'ux'	62	95.32	37	97.03
Total	1583	86.06	1143	84.58

Table 2.3: Comparison of the estimation performance in terms of average error for male speakers

Vowel		$-5dB$			$10dB$		
		Proposed	LPC	AFB	Proposed	LPC	AFB
/ah/	F1	13.54	31.64	24.12	13.43	22.08	15.66
	F2	10.46	57.43	28.88	10.32	23.30	18.84
	F3	12.55	39.21	13.09	8.65	35.47	13.16
/eh/	F1	13.52	27.70	24.62	13.37	14.05	17.07
	F2	14.96	33.30	24.18	8.88	11.95	17.91
	F3	13.22	39.13	13.39	7.56	33.11	11.67
/ih/	F1	14.36	38.52	23.47	14.47	12.40	23.98
	F2	15.90	27.12	25.50	8.32	15.16	20.05
	F3	12.02	39.50	13.45	7.60	32.66	11.34
/ow/	F1	14.87	22.63	35.49	14.40	18.78	36.11
	F2	11.95	47.20	26.03	11.37	34.22	22.37
	F3	12.31	36.68	14.20	10.12	36.92	14.15
/uh/	F1	15.13	20.14	36.49	14.79	19.21	36.59
	F2	11.56	38.02	23.49	11.23	35.66	22.50
	F3	9.95	37.24	13.89	9.82	37.06	14.07
/ux/	F1	13.49	49.04	36.77	13.58	14.28	39.30
	F2	12.27	30.11	22.86	9.81	23.26	21.70
	F3	10.73	41.23	13.63	9.35	36.14	13.48

Table 2.4: Comparison of the estimation performance in terms of average error for female speakers

Vowel		$-15dB$			$0dB$		
		Proposed	LPC	AFB	Proposed	LPC	AFB
/aa/	F1	15.40	48.89	46.25	10.95	15.91	41.89
	F2	21.26	83.33	21.40	11.88	50.37	25.37
	F3	14.25	43.46	14.23	12.89	27.05	12.70
/ah/	F1	21.32	50.77	37.88	12.11	12.32	35.70
	F2	20.81	66.14	21.65	10.54	33.23	19.26
	F3	13.11	34.12	16.12	13.35	22.09	14.33
/eh/	F1	15.97	55.19	31.51	12.13	9.32	24.85
	F2	29.09	31.17	28.87	14.33	11.90	23.45
	F3	14.00	28.41	12.15	9.15	19.62	12.65
/ow/	F1	16.21	76.56	24.15	12.68	10.81	22.83
	F2	26.94	41.29	31.59	14.55	25.46	27.60
	F3	13.29	28.29	14.97	9.73	20.64	14.49
/uh/	F1	15.47	77.76	24.67	13.12	11.46	22.14
	F2	25.12	40.40	32.34	13.36	24.47	26.61
	F3	12.95	28.21	15.12	8.77	20.88	13.60
/uw/	F1	15.45	81.07	24.32	12.96	10.74	23.16
	F2	24.96	40.34	32.41	13.51	23.79	27.00
	F3	12.84	29.35	15.31	9.12	21.40	13.80

Table 2.5: Comparison of the estimation performance in terms of average error for different frame lengths ($Fs = 16kHz$,$SNR = 10dB$)

Vowels		128 samples			512 samples		
		Proposed	AFB	LPC	Proposed	AFB	LPC
Male /ah/	F1	57.68	72.67	54.40	13.43	28.29	22.08
	F2	16.16	24.25	19.60	10.32	29.34	23.30
	F3	66.54	69.12	125.10	8.65	14.56	35.47
Female /ah/	F1	57.19	77.34	57.54	13.69	32.81	11.19
	F2	18.82	25.70	14.67	11.67	29.52	10.60
	F3	49.40	42.63	96.28	9.88	15.78	17.86

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 (2.6), where the noise variance σ_v is appropriately determined according to a specified level of SNR defined as

$$SNR = 10log_{10}\frac{\sum_{n=0}^{N-1} x(n)^2}{\sum_{n=0}^{N-1} v(n)^2} \tag{2.28}$$

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 120 Hz and 220 Hz, respectively, for male and female speakers. 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_O is defined as

$$E = |\frac{f_E - f_O}{f_O}| \times 100\% \tag{2.29}$$

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.

Next the simulation results for TIMIT database is presented. For this analysis, the number of occurrence of each vowel along with the average vowel duration for male and

female speakers available in the TIMIT database is presented in Table 2.2 . Overall the average duration of vowel utterances is 85.44 ms.

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 are presented in Tables 2.3 and 2.4. For each of the Tables 2.3 and 2.4, two different SNR levels are considered, for which the results for a selection of vowels are presented. 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 the 12 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.

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. 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/. However, 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. The standard deviation of the estimated errors is also measured and it is found that the standard deviation is very small, indicating a consistent estimate of the formant under various conditions. Hence the formant estimation obtained by the proposed method is very reliable and accurate.

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. In Table 2.5, the

effect of variation in frame length on the estimation accuracy for the vowel /ah/ is shown. A reasonable number of samples are required so that the resolution of the FFT remains good enough. This is also true for the LPC and the AFB methods, as they also require sufficient number of samples to perform time domain estimation. It is observed from the table that with the increase in frame length, estimation errors are significantly reduced for all three methods. Other vowels also show a similar trend. The reason behind the drastic increase in estimation error with decrease in frame length in the proposed method is mainly the finite duration autocorrelation operation which results in a autocorrelation sequence with decreasing tailing lags.

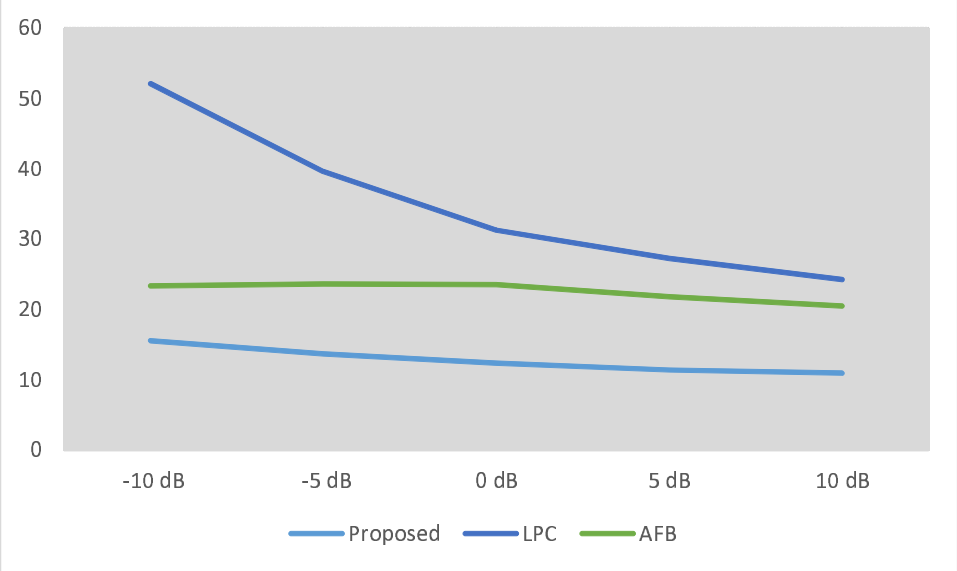


Figure 2.14: First formant estimation performance in terms of percentage error in formant estimation under various noise levels for male speakers

In order to present the overall formant estimation errors over the entire range of SNRs considered in the experimental setup, in Figs. 2.14, 2.15, 2.16 and 2.17, average of estimation error of all vowels for all three formants are shown for the the proposed method and the $LPC-12$ based method considering only male speakers. In this case, the SNR levels considered are ranging from -10 to $+10dB$. In a similar way, in Figs. 2.18, 2.19, 2.20 and 2.21, the average estimation error are shown for the female speakers for a SNR range of -15 to $+5dB$. Finally, in Fig. 2.22, the average estimation error considering both male and female speakers is shown. It is observed that the formant estimation performance obtained by the three methods remains similar in case of high level of SNR.

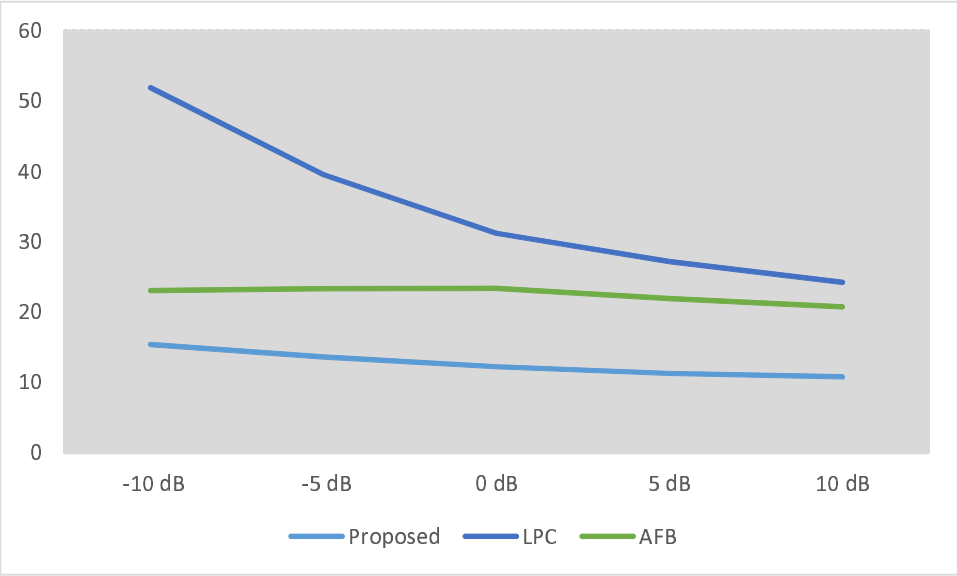


Figure 2.15: Second formant estimation performance in terms of percentage error in formant estimation under various noise levels for male speakers

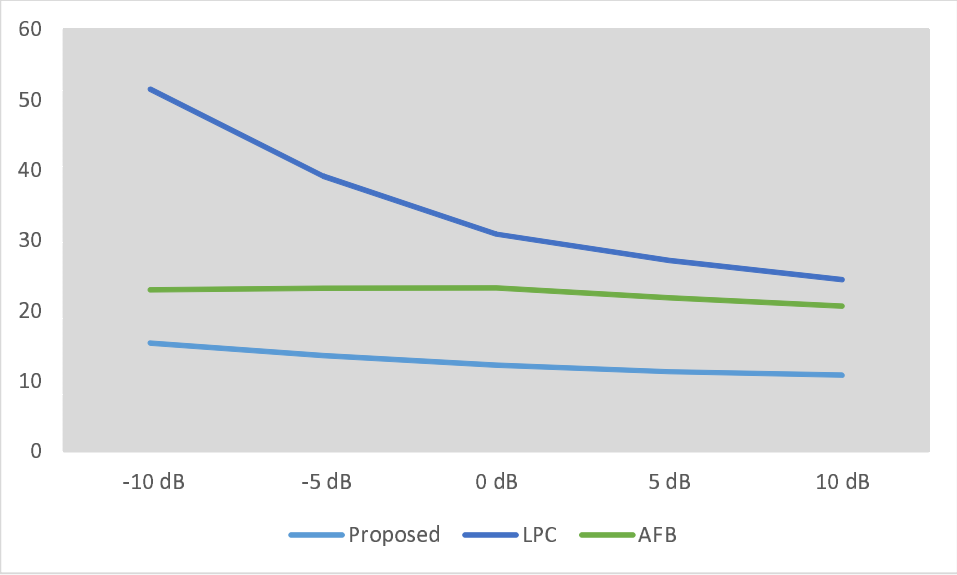


Figure 2.16: Third formant estimation performance in terms of percentage error in formant estimation under various noise levels for male speakers

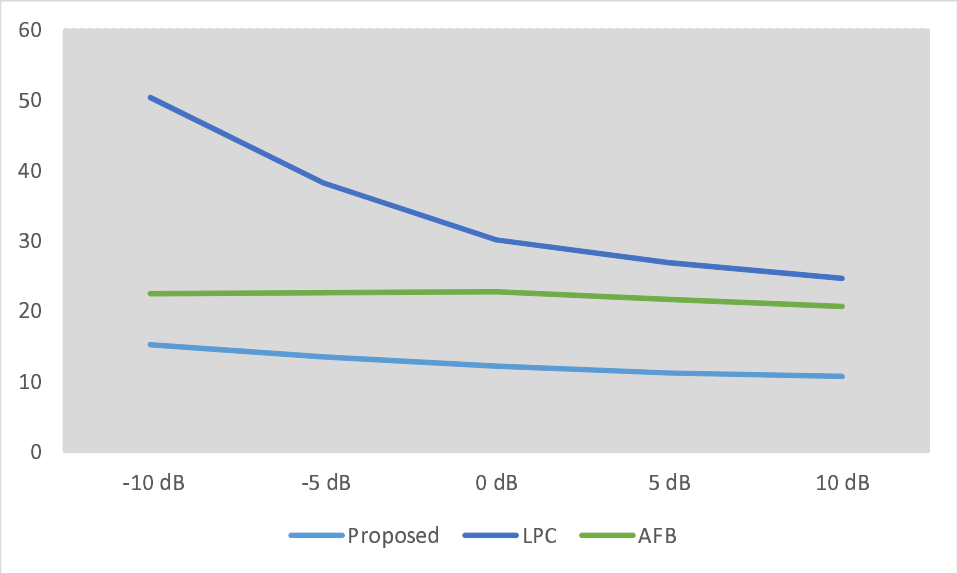


Figure 2.17: Estimation performance in terms of percentage error in formant estimation under various noise levels for male speakers

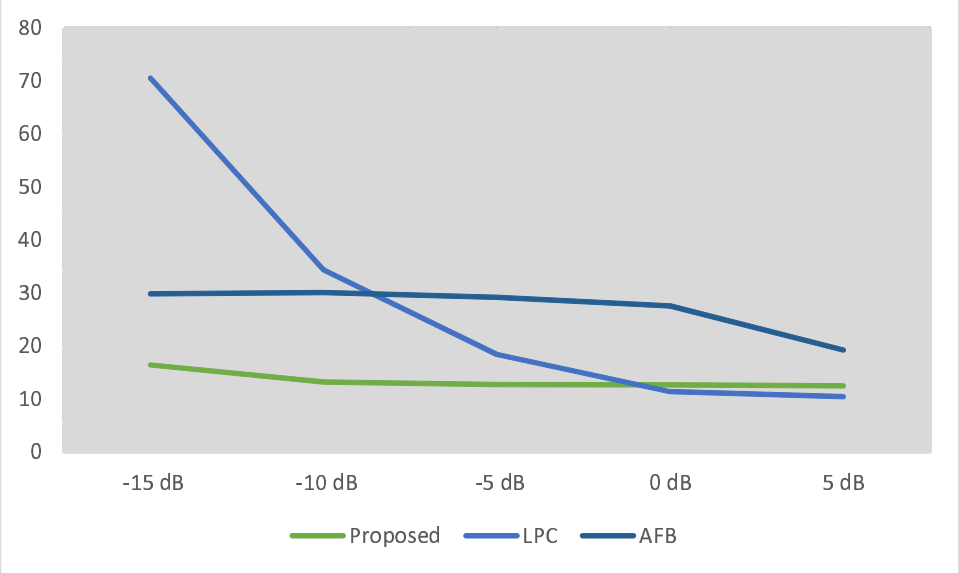


Figure 2.18: First formant estimation performance in terms of percentage error in formant estimation under various noise levels for female speakers

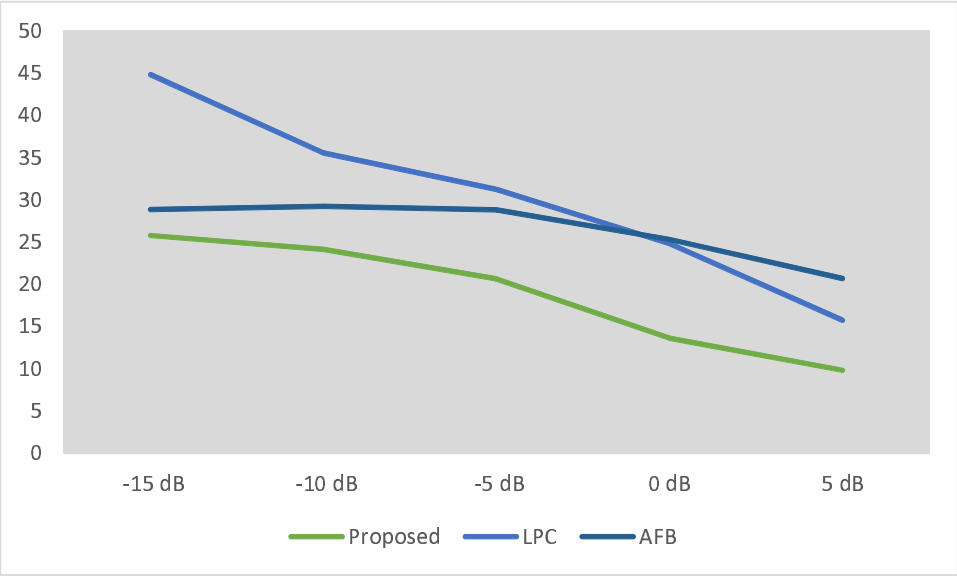


Figure 2.19: Second formant estimation performance in terms of percentage error in formant estimation under various noise levels for female speakers

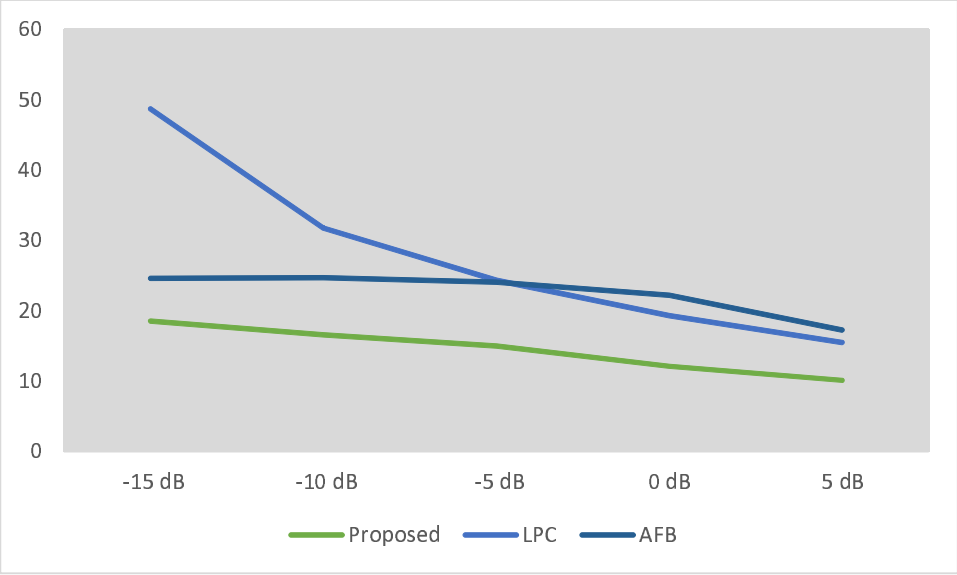


Figure 2.20: Third formant estimation performance in terms of percentage error in formant estimation under various noise levels for female speakers

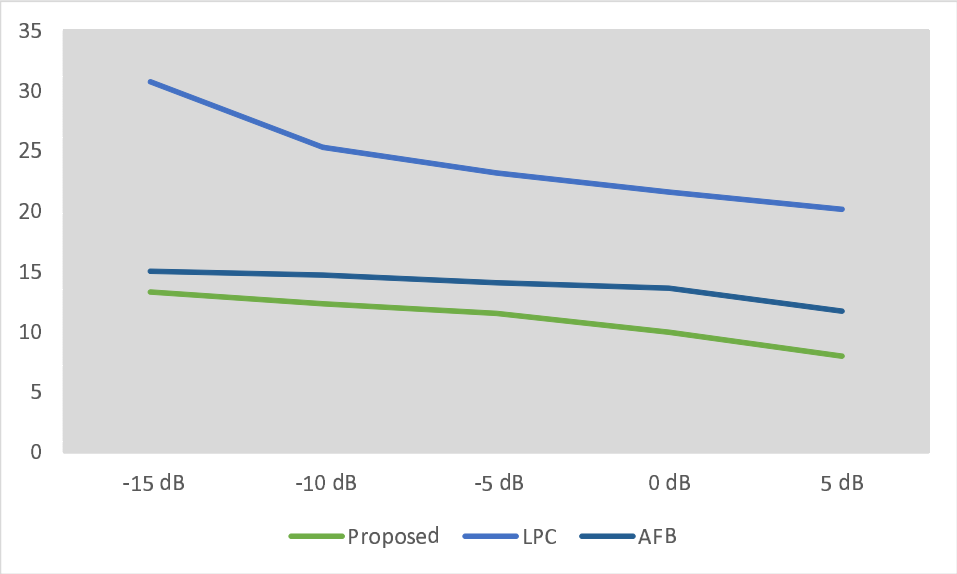


Figure 2.21: Estimation performance in terms of percentage error in formant estimation under various noise levels for female speakers

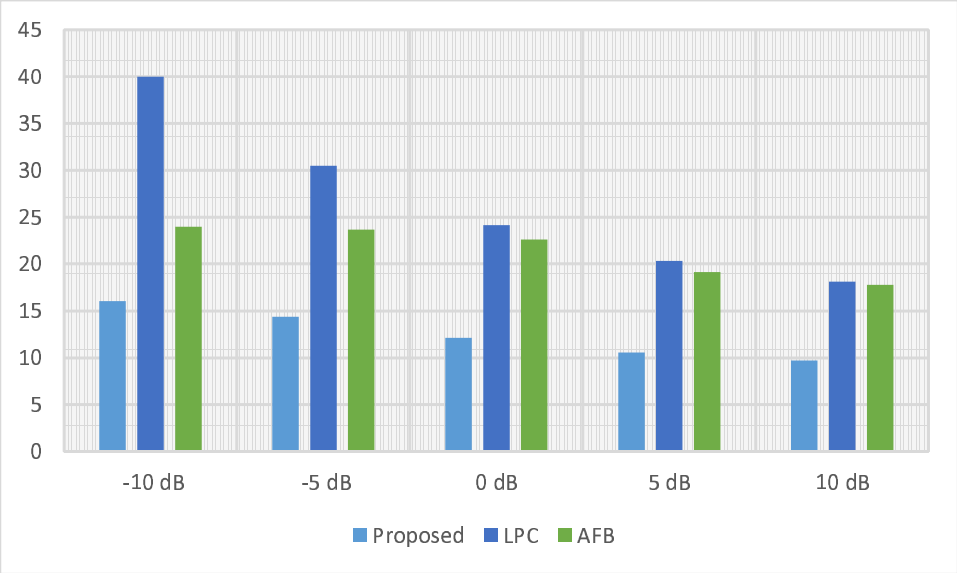


Figure 2.22: Estimation performance in terms of percentage error in formant estimation under various noise levels

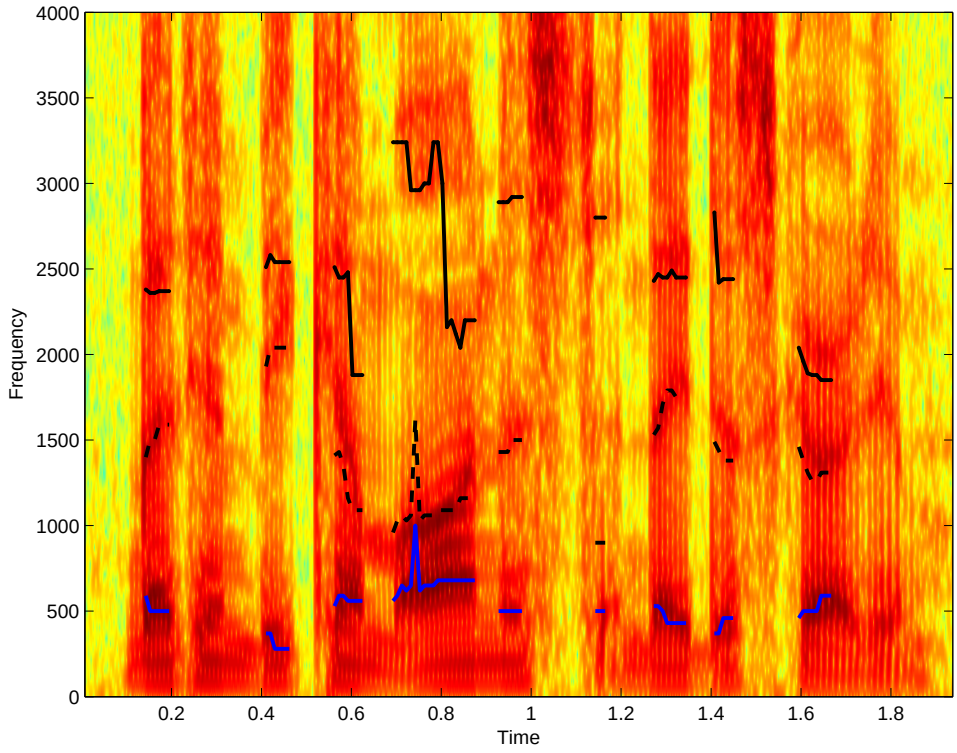


Figure 2.23: Spectrogram of the utterance 'let him become honest and they discard him' , with formant frequencies estimated using the proposed method

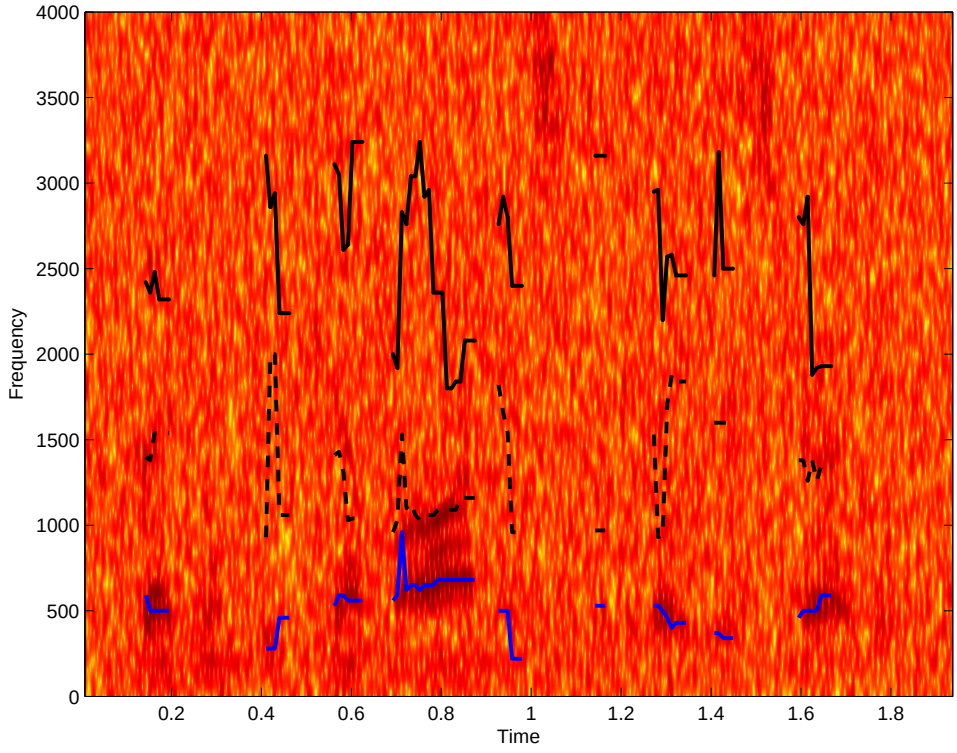


Figure 2.24: Spectrogram of the utterance 'let him become honest and they discard him' , under $-5dB$ of background noise with formant frequencies estimated using the proposed method

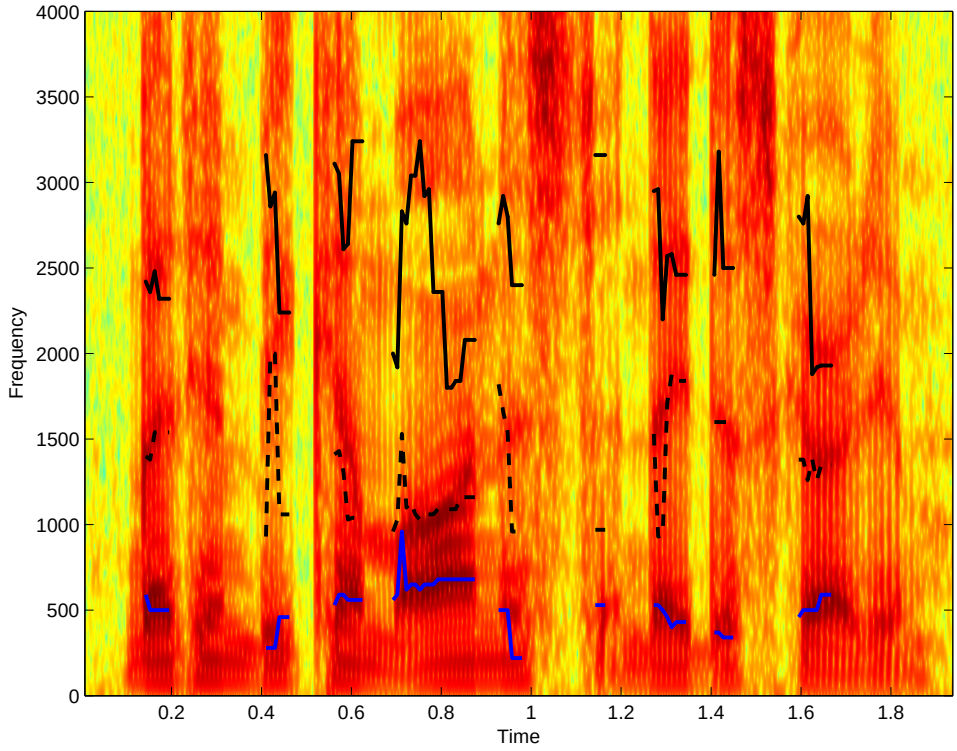


Figure 2.25: Spectrogram of the utterance 'let him become honest and they discard him' , with formant frequencies estimated under $-5dB$ of Background noise using the proposed method

However, with the decrease in SNR level, the estimation performance of the other two methods deteriorates significantly 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 $-15dB$. 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.

Table 2.6: Vowel recognition accuracy

Feature Vector	$10dB$	$-5dB$
MFCC + Proposed Method	93.33	90.00
MFCC + LPC-12	93.33	81.66
MFCC + TIMIT reference	93.33	90.00
MFCC	93.33	81.66

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 features as a feature vector, vowel recognition was performed for the vowels /aa/, /ux/ and /ow/ from the TIMIT database. For the purpose of further comparison, vowel recognition accuracies obtained by incorporating the noise free reference values of the formants are incorporated in the feature vector along with MFCC features obtained from noisy speech. It can be observed that up to $-10dB$, the performance of the proposed feature vector is comparable even to the performance of feature vectors incorporating noise free formant estimations. 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 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 2.6. 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 analysis, 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 'let him become honest and they discard him' uttered by a male speaker taken from the TIMIT database is shown in Fig. 2.23. The formant frequencies estimated at different frames using the proposed method are shown over the spectrogram. 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. For the purpose of comparison, the same sentence, under influence of $-5dB$ background noise, is utilized to obtain the spectrogram present in Fig. 2.24. Here the presence of noise has completely obscured the energy bands, but still the proposed method can successfully track the formant frequencies. With the purpose of gaining a better insight, the formant frequencies obtained from the $-5dB$ noise corrupted speech are overlayed on the spectrogram for

noise free speech, which is shown in Fig. 2.25. The resulting tracking lines obtained by the proposed method is a clear indication of its high level of consistency as well as the accuracy even in heavy noisy condition.

2.4 Conclusion

In this chapter, an effective method for formant frequency estimation of noise corrupted voiced human speech using spectral model of autocorrelation of speech is deployed that can find out the band of successive formant frequencies for pre-processed voiced speech signals. Then autocorrelation is then performed on the speech signal, which strengthens the dominant poles, and exponentially increases the peak-valley ratio at formant frequencies of the magnitude response, canceling out the effects of noise. Instead of using conventional peak picking to find formants from the spectrum of the ACF, a spectral model of autocorrelated speech signal for a single formant is developed and model fitting is employed to find out model parameters which lead to formant estimation. Natural vowels as well as some naturally spoken sentences in noisy environments are tested. Through the simulation results on standard speech databases , it is shown that the developed method is effective in maintaining a high success rate in formant estimation even in the presence of a significant background noise.

Chapter 3

Spectral Model of Repeated Autocorrelation of Speech

In this chapter, the scheme for estimating the formant frequencies is further developed using repeated autocorrelation. Repeated autocorrelation operation, which significantly strengthens the dominant poles, and exponentially increases the peak-valley ratio at formant frequencies of the magnitude response, is proposed to be employed with the purpose of canceling out the effects of noise. Formant estimation is carried out in the spectral domain where instead of direct peak-picking from the speech spectrum, a spectral domain model of repeated ACF of speech signal is first proposed considering the vocal tract to comprise of cascaded subsystem responsible for single resonant frequencies. A spectral domain model fitting based algorithm is also developed to extract the model parameters which in turn give the formant. Through the simulation results on standard speech databases , it is shown that the developed method is effective in maintaining a high success rate in formant estimation even in the presence of a significant background noise.

3.1 Background

In order to estimate the formant frequencies from observed speech signal, it is sufficient to restrict the analysis only for the voiced sound. In case of the voiced speech signals,

considering the excitation as a periodic impulse-train, the overall vocal tract filter can be represented by a P-th order autoregressive (AR) system with the following transfer function

$$H(z) = \frac{C}{\prod_{i=1}^{P}(1 - p_i z^{-1})} \tag{3.1}$$

where p_i denotes the pole of the AR system and C is the gain factor. The vocal tract system in (3.1) can exhibit $P/2$ formants. However, as far as formant estimation is concerned, only the first three formants are significant and contain a very high portion of the total energy. In this regard, it would be sufficient to consider the vocal tract to be represented as a cascaded network of three separate subsystems, each causing a resonant peak in the speech spectrum.

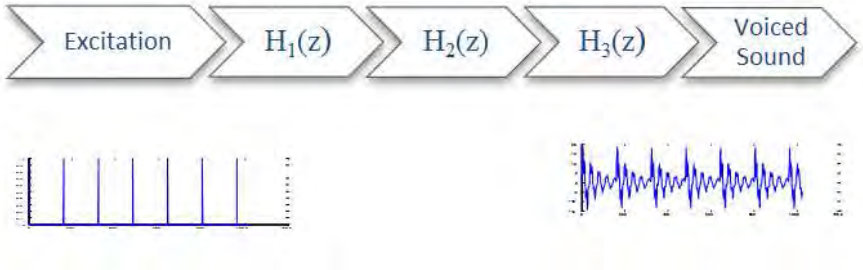


Figure 3.1: Voiced sound generation through a simplified model of vocal tract system

In Fig. 3.1, a simplified vocal tract model consisting of three subsystems is shown.Each individual subsystem in Fig. 3.1 can be represented as

$$H_i(z) = \frac{C_i}{(1 - p_i z^{-1})(1 - p_i^* z^{-1})} \tag{3.2}$$

Here for each pair of complex conjugate poles $p_i = r_i e^{(j\theta_i)}$ the magnitude r_i and angle θ_i are related to a particular formant F_i and the formant bandwidth B_i as

$$r_i = e^{\frac{-\pi B_i}{Fs}} \tag{3.3}$$

$$\theta_i = \frac{2\pi F_i}{Fs} \tag{3.4}$$

where Fs is the sampling frequency.

3.2 Proposed Formant Estimation Scheme

In this section, the effect of everyday noise on voiced speech is demonstrated. Then methods for countering the effect of noise on formant estimation are evaluated and the performance of repeated autocorrelation as a facilitator for better formant detection under noise is demonstrated. It is shown that repeated autocorrelation works better than single autocorrelation for the purpose of noise removal. Finally a model matching method is developed for extracting the formants from the autocorrelation of band limited speech.

3.2.1 Effect of Repeated ACF in Noise

For noise-free voiced speech signal conventional peak picking formant estimation methods may provide satisfactory results. However, presence of background noise is very common in everyday situations and it affects the accuracy of traditional estimators.

For a voiced sound $x(n)$ in the presence of additive noise $v(n)$ with zero mean and unit variance, the noise corrupted speech $y(n)$ can be written as

$$y(n) = x(n) + v(n) \tag{3.5}$$

In a time domain representation of the noise corrupted speech signal, it is very difficult to distinguish the original speech samples even at a moderate level of noise. The presence of additive noise completely destroys the original speech pattern resulting in a noise like pattern. In order to show the effect of noise in time domain, in Fig. 3.2(a) and 3.2(b), a noise free speech $x(n)$ and corresponding noise corrupted speech $y(n)$ are shown, respectively. Here the natural sound /iy/ is considered and its spectral representation is presented in Fig. 3.3(a) for the noise free speech and in Fig.3.3(b) for the noisy speech.

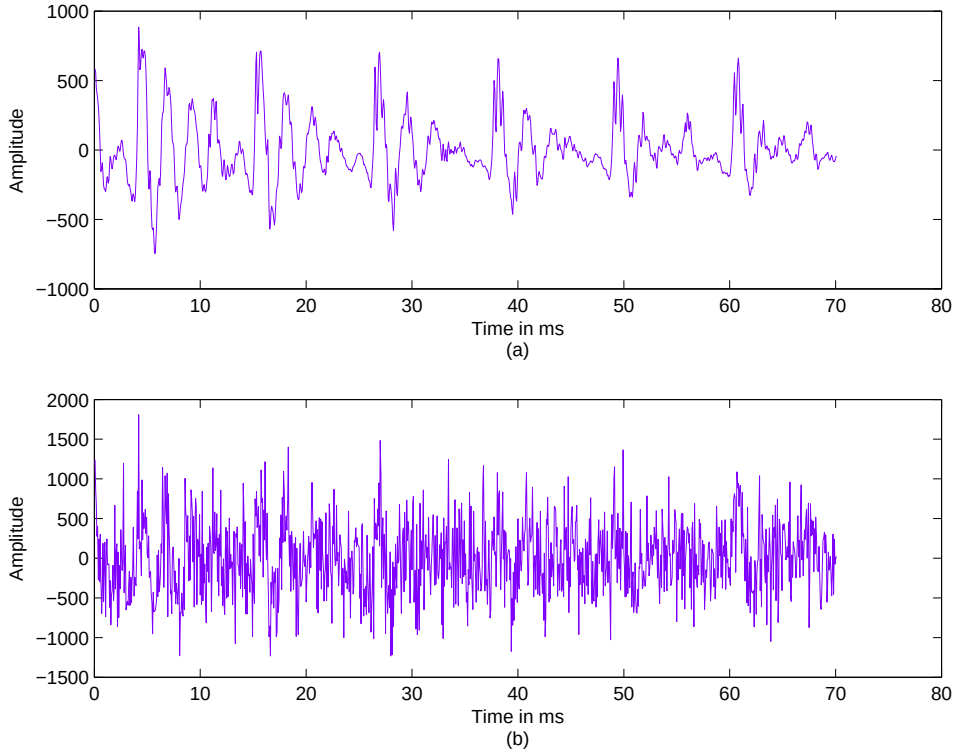


Figure 3.2: (a) Time domain waveform of an utterance /iy/ and (b) the same waveform under $-5dB$ background noise

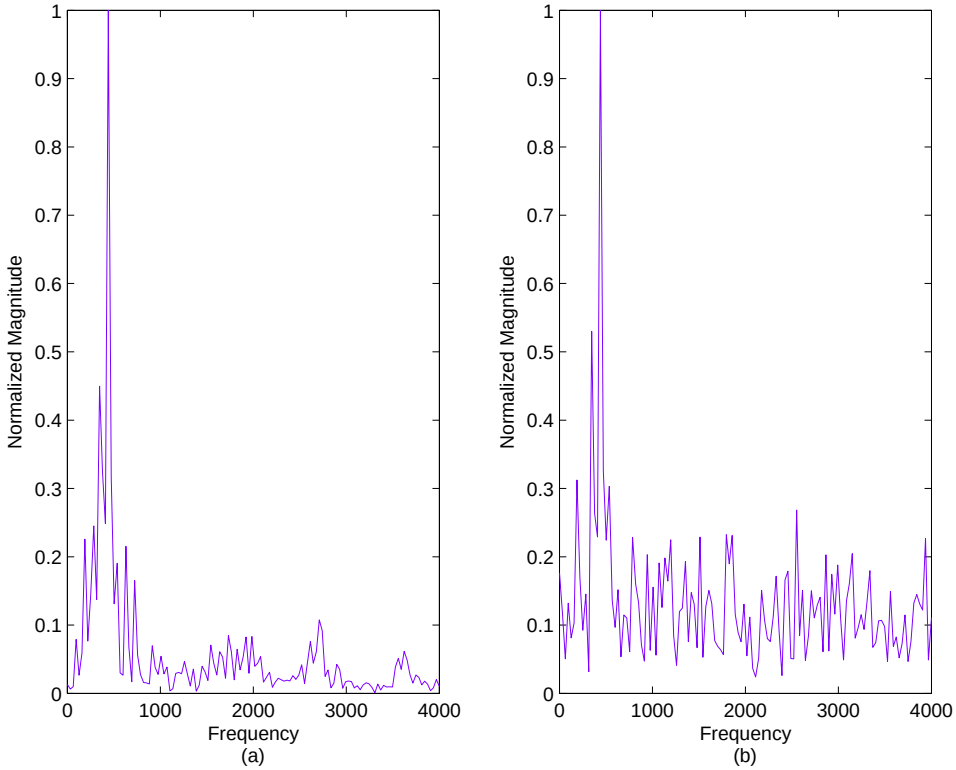


Figure 3.3: Spectrum of natural /iy/ voiced speech, (a) under noise free conditions and (b) under $-5dB$ background noise

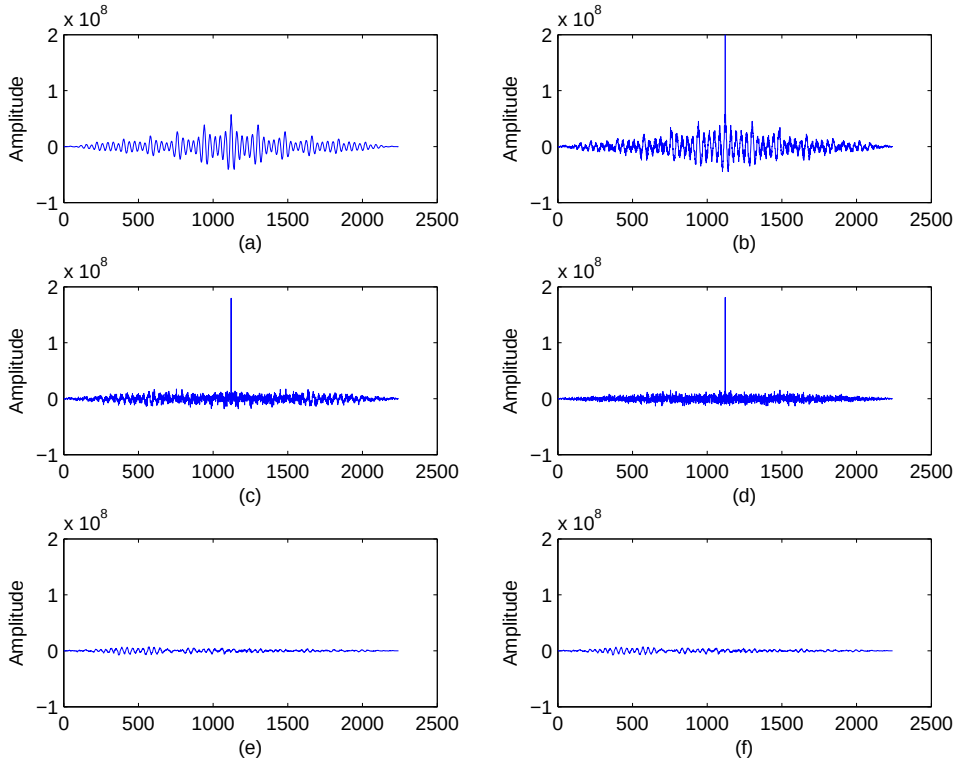


Figure 3.4: Effect of noise in the autocorrelation domain: plot of different autocorrelation functions (a) $r_y(n)$, (b) $r_x(n)$, (c) $r_w(n)$, (d) $r_v(n)$, (e) $r_{xv}(n)$ and (f) $r_{vx}(n)$

The autocorrelation function (ACF) of a voiced sound $x(n)$ is defined as

$$r_x(\tau) = E[x(n)x(n-\tau)] \tag{3.6}$$

where τ denotes the lag. ACF is an even function, with the output being symmetric with respect to the amplitude axis. In practical application the ACF of $x(n)$ is computed by using the working formula given below

$$r_x(n) = \frac{1}{N}\sum_{k=0}^{N-1-|n|} x(k)x(k+|n|), n = 0, 1, 2.\ldots\ldots\ldots M-1 \tag{3.7}$$

Using (3.5) and (3.6), the ACF of noisy speech y(n) can be expressed as

$$\begin{aligned} r_y(n) &= r_x(n) + r_w(n) \\ r_w(n) &= r_v(n) + r_{vx}(n) + r_{xv}(n) \end{aligned} \tag{3.8}$$

Here $r_v(n)$ is the ACF of noise $v(n)$ and $r_{vx}(n)$ and $r_{xv}(n)$ are the cross correlation terms. In Figs. 8(a)-8(f), different ACFs, namely $r_y(n)$, $r_x(n)$, $r_w(n)$, $r_v(n)$, $r_{xv}(n)$ and $r_{vx}(n)$

are plotted. From these figures, it can be concluded that in comparison to the effect of $v(n)$ on $x(n)$ as shown in Fig. 3.2, the effect of $r_w(n)$ on $r_x(n)$ is drastically reduced because of the autocorrelation operation.

Considering $x(n)$ as an output of an LTI system with transfer function $H(z)$, $x(n)$ can be written as

$$x(n) = h(n) * u(n) \tag{3.9}$$

It can be shown that the ACF of $x(n)$ can be expressed as

$$r_x(n) = r_h(n) * r_u(n) \tag{3.10}$$

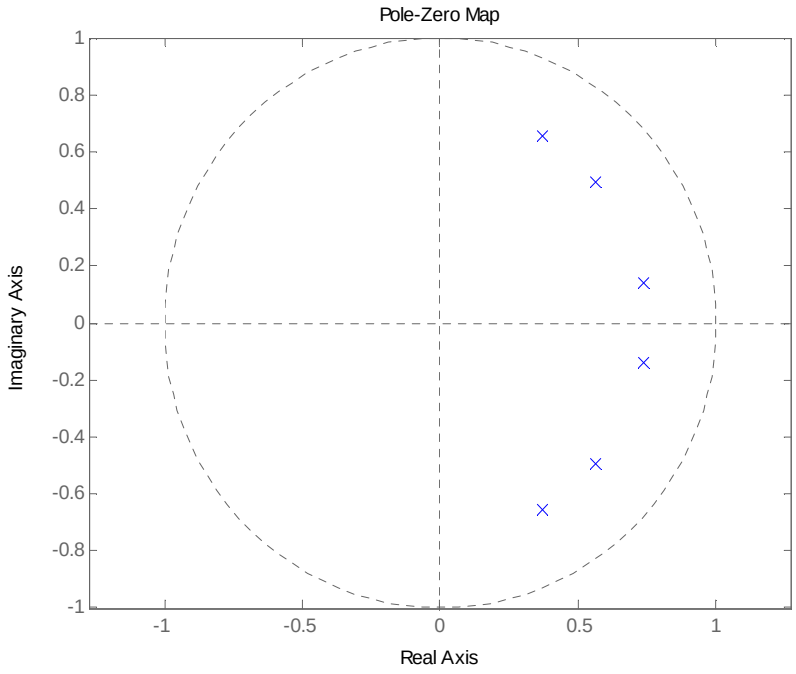
where $r_u(n)$ is the ACF of $u(n)$. As per the definition of the ACF provided in (3.6), the ACF of $h(n)$ can be written as

$$r_h(n) = h(n) * h(-n) \tag{3.11}$$

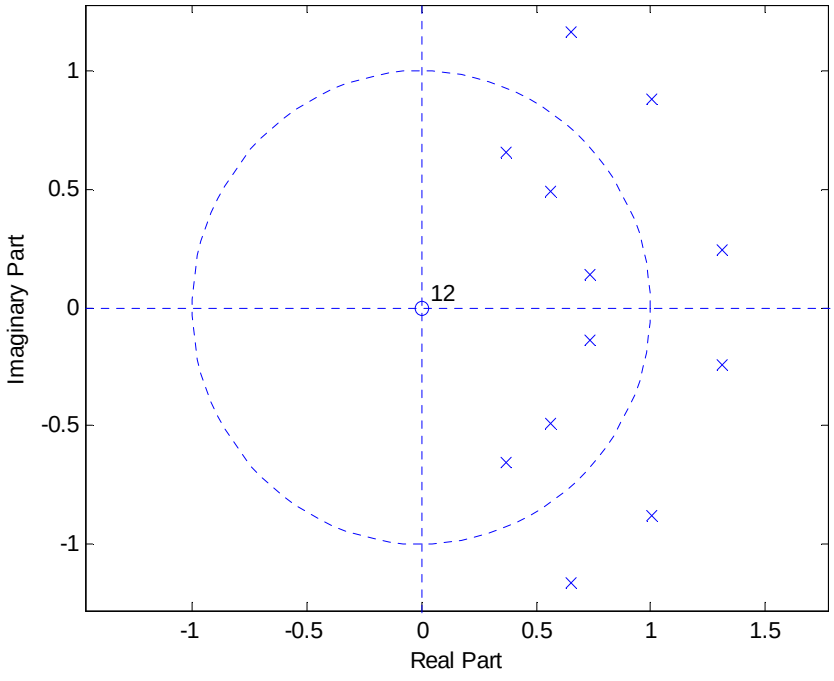
In view of analyzing the frequency domain effects, for simplicity, first the Z domain representation is considered. The Z transform of $r_h(n)$, as obtained from (3.11) is given by

$$R_h(z) = H(z)H(z^{-1}) \tag{3.12}$$

It is observed from (3.12) that the ACF operation produces a set of new poles with equal number of the original system poles corresponding to $H(z)$ and the new poles corresponding to $H(z^{-1})$ are located at the conjugate reciprocal location of the original poles. For a clear understanding, a sample z-plane pole representation of $H(z)$ having three pairs of complex conjugate poles and corresponding $R_h(z)$ are shown in Figs. 3.5(a) and 3.5(b) respectively. It is seen that from the figure that at each angular position of the original poles, one new pole is generated outside the unit circle. In order to present the effect of spectral peak strengthening both in noise-free and noisy condition, in Fig. 3.6, spectra



(a)



(b)

Figure 3.5: Effect of autocorrelation in z-domain (a) $H(z)$ (b) $R_h(z)$

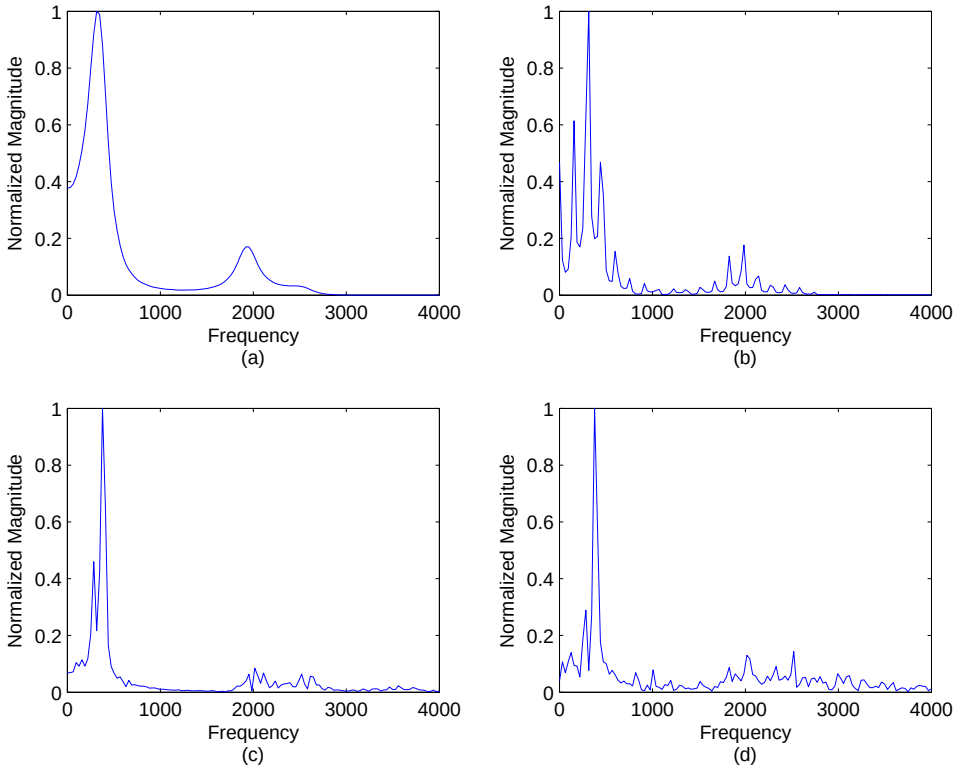


Figure 3.6: Effect of spectral strengthening because of autocorrelation operation. Spectrum of: (a) $r_h(n)$, (b) $r_{synx}(n)$, (c) $r_x(n)$ and (d) $r_y(n)$

corresponding to $r_h(n)$, $r_{synx}(n)$, $r_x(n)$ and $r_y(n)$ are shown. Strengthening of dominant peaks is evident as the first formant peak is significantly strengthened. However, spurious peaks are still present, and this may poses challenges under severely noisy conditions.

Realizing the effect of spectral peak strengthening, in this chapter, 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. Considering the same noisy natural sound /iy/ as shown in Fig. 3.2, the spectral domain effect of Double ACF (DACF) on this speech signal is shown in Fig. 3.7. It is observed from this figure that because of the repeated ACF the resulting

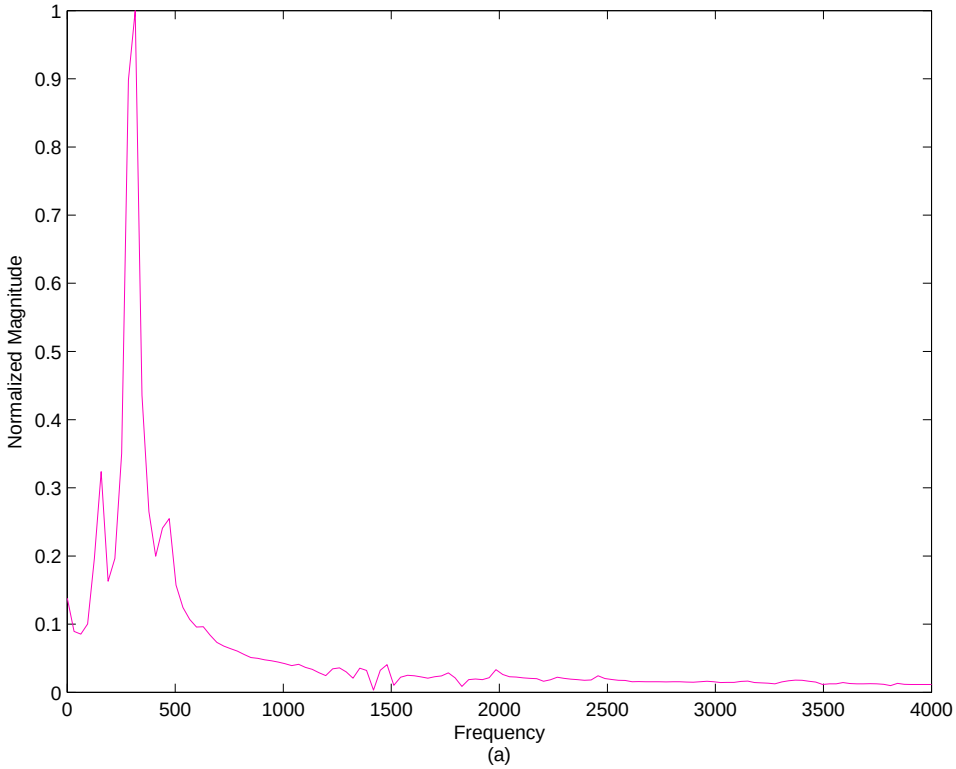


Figure 3.7: Spectrum for the double ACF of $-5dB$ noise corrupted voiced /iy/ sound

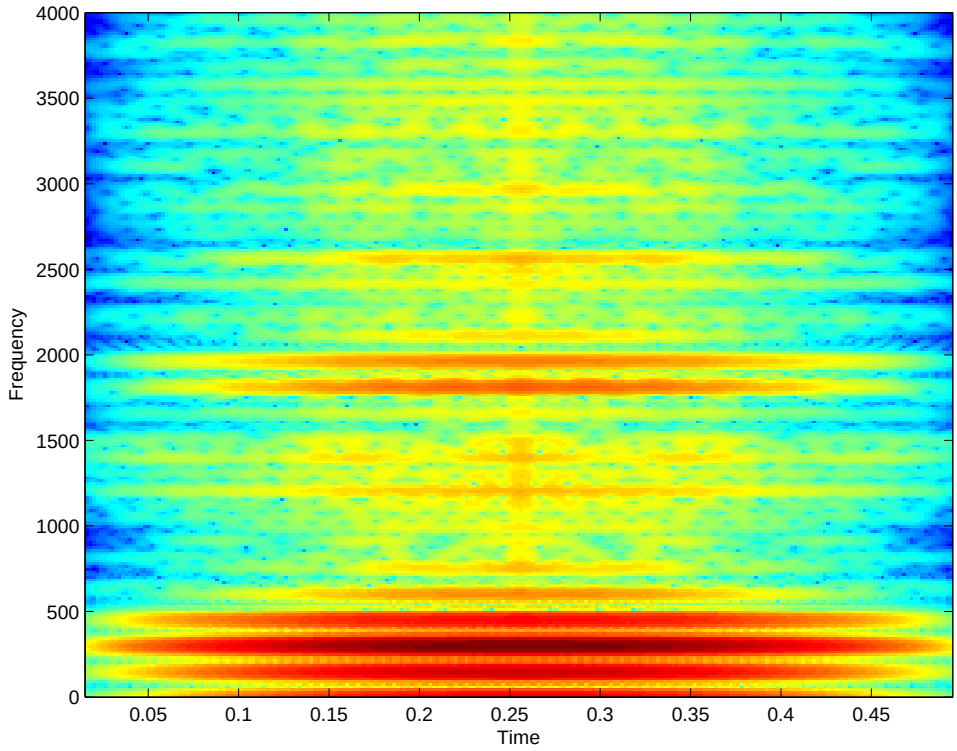


Figure 3.8: Spectrogram for the double ACF of $-5dB$ noise corrupted voiced /iy/ sound

spectrum becomes almost free from spurious peaks around the enhanced first formant peak. The enhancement of the first formant is quite prominent in Fig. 3.7. However, enhancement also occurs for other two formant peaks which can be visible from the enlarged figures shown inside Fig. 3.7. In view of clear understanding of such spectral peak enhancement, corresponding spectrogram is shown in Fig. 3.8. It can be clearly seen from the spectrogram representation that all three formant locations indicate a very high spectral energy (dark red color in the figure). Hence, use of the spectrum corresponding to the double correlated signal, instead that corresponding to the noisy signal, would be much convenient for formant estimation. According to the definition of the ACF mentioned in (3.6), the ACF of $r_x(n)$, namely the repeated ACF of $x(n)$ can be expressed as

$$\rho_x(n) = r_x(n) * r_x(-n) \tag{3.13}$$

using the definiton of $r_x(n)$ from (3.10), it can be shown that

$$\rho_x(n) = \rho_h(n) * \rho_u(n) \tag{3.14}$$

As discussed before, it would be sufficient to consider the detailed analysis of $\rho_h(n)$ instead of $\rho_x(n)$. Using the definition in (3.11), the Z Transform of $\rho_h(n)$ can be written as

$$P_h(z) = P_h(z)P_h(z^{-1}) \tag{3.15}$$

It is observed from (3.15) that the repeated ACF operation produces a set of new poles with equal number of the poles corresponding to $R_h(z)$ and the new poles corresponding to $R_h(z^{-1})$ are located at the conjugate reciprocal location of the poles corresponding to $R_h(z)$. In a similar fashion as the effect of single autocorrelation operation on the system poles is shown in Fig. 3.6(b), the effect of repeated autocorrelation operation on system poles is shown in Fig. 3.6(c). It is clearly observed that in the location of each pole in Fig. 3.6(b), instead of one, there exist two poles in Fig. 3.6(c) resulting from the double

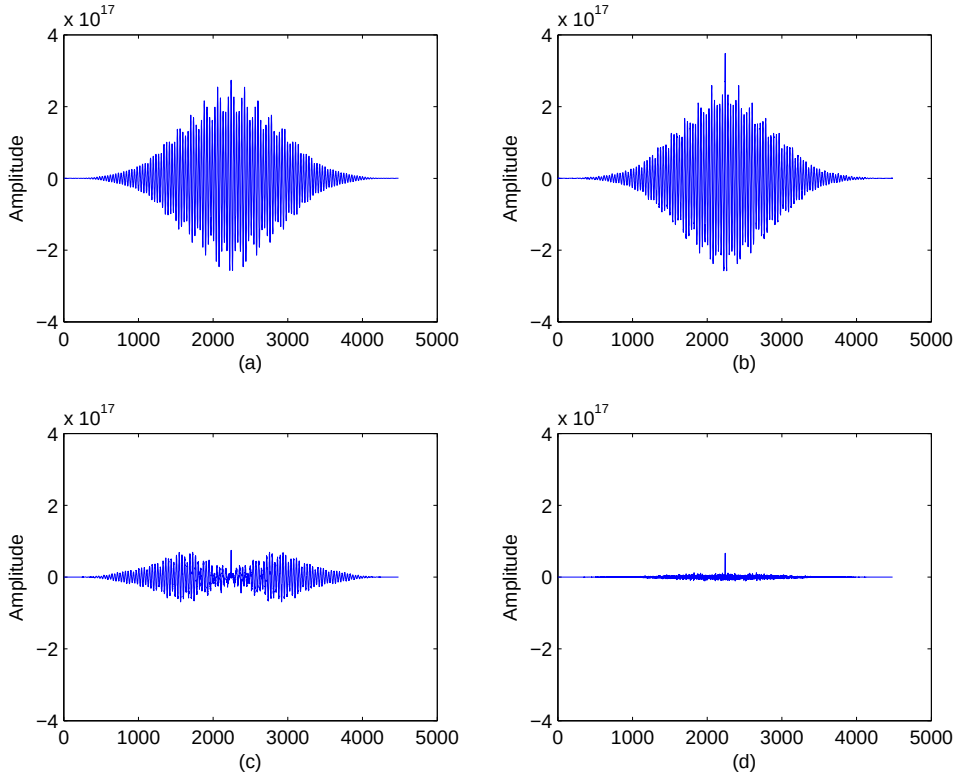


Figure 3.9: Time domain waveforms for (a) $\rho_x(n)$,(b) $\rho_y(n)$, (c) $\rho_c(n)$, and (d) $\rho_w(n)$

autocorrelation operation. As a result, in each formant frequency location, instead of one, now there will be four poles, resulting in huge spectral energy.

It is to be mentioned that in the computation of the DACF, the double sided ACF signal is provided as input. This is done in view of overcoming the adverse spectral domain effect of conventional windowing to be used to obtain the single sided ACF from the given two sided version. The advantage of using the DAC operation can also be demonstrated in time domain as explained before in case of SAC operation. Further application of ACF on the noise corrupted signal $r_y(n)$ produces $\rho_y(n)$ which can be expressed as

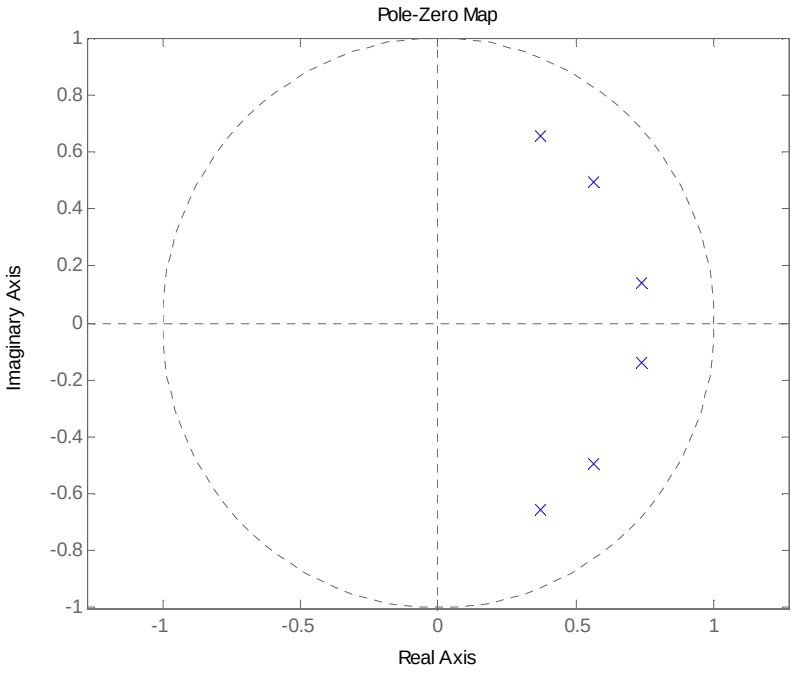
$$\begin{aligned} \rho_y(n) &= \rho_x(n) + \rho_c(n) \\ \rho_c(n) &= \rho_w(n) + \rho_{xw}(n) + \rho_{wx}(n) \end{aligned} \tag{3.16}$$

where$\rho_x(n)$ and $\rho_w(n)$ are the ACF of $r_x(n)$ and $r_w(n)$ and $\rho_{xw}(n)$ and $\rho_{wx}(n)$ are cross correlation terms. It is expected that the effect of $\rho_c(n)$ on $\rho_x(n)$ is very negligible, as there exists very little correlation between $r_x(n)$ and $r_w(n)$, and $r_w(n)$ is quite insignificant at points other than the zero lag. In Figs. 3.9(a) - 3.9(d) , the DACFs $\rho_x(n)$, $\rho_y(n)$, $\rho_c(n)$

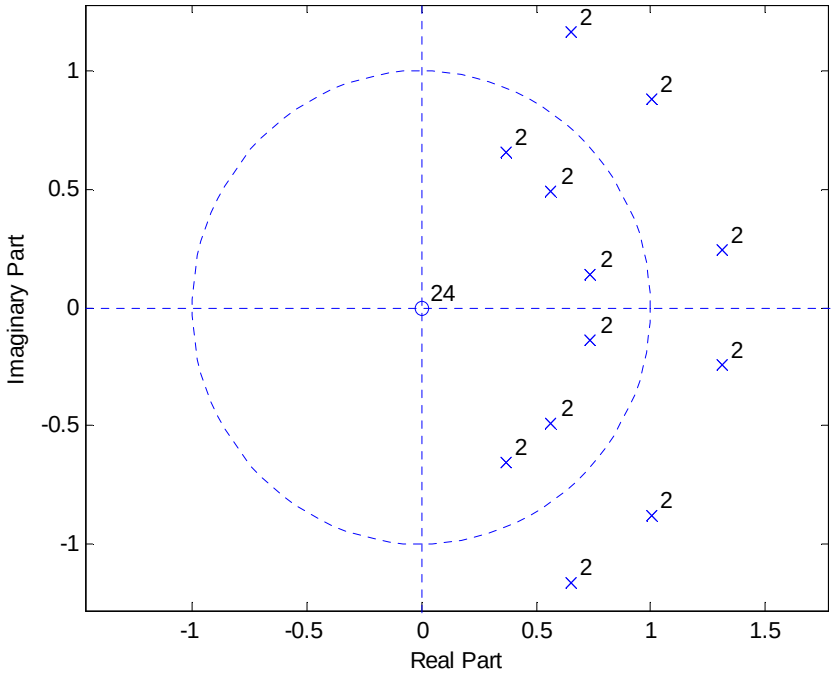
as well as $\rho_w(n)$ are shown. It is clearly observed that the values of $\rho_c(n)$ are extremely small in comparison to that of $\rho_x(n)$ as expected. From these figures, it can be concluded that in comparison to the effect of $r_w(n)$ on $r_x(n)$ as shown in Fig. 3.4, the effect of $\rho_c(n)$ on $\rho_x(n)$ is significantly reduced because of the repeated autocorrelation operation. Hence, it is advantageous to utilize $\rho_y(n)$ instead of $r_y(n)$ in spectral domain formant estimation.

In the z domain, the ACF operation creates new poles at conjugate reciprocal positions. In the case of DACF, one major advantage is that new poles are created in the original positions, in addition to the poles produced outside the unit circle after the first autocorrelation operation. z-plane pole representation of $H(z)$ having three pairs of complex conjugate poles and corresponding $P_h(z)$ are shown in Figs. 3.10(a) and 3.10(b) respectively. It is seen that from the figure that at the position of each of the original poles, one new pole is generated, while two other poles corresponding to the same angular frequency are present outside the unit circle. In spectral domain, the peak strengthening effect, as discussed before for the ACF, is more prominent in case of DACF, resulting in a spectral smoothing in the region of unwanted noise peaks. In view of demonstrating the effect DACF on spectral peak strengthening both in noise-free and noisy conditions, similar to the case of ACF as shown in Fig. 3.6, and in Fig. 3.11, spectra corresponding to $\rho_h(n)$, $\rho_{synx}(n)$, $\rho_x(n)$ and $\rho_y(n)$ are shown. In comparison to the spectra corresponding to $y(n)$ presented at Fig. 3.2, It is clearly observed that the first peak in the spectra corresponding to $\rho_y(n)$ exhibits an extremely large peak in comparison to other peaks and significant spectral smoothing is observed in other zones of the normalized spectrum. The pole strengthening effect is much more prominent here compared to single autocorrelation function. It is to be noted that there are no spurious peaks around the dominant first formant peak. This is to be expected as after DACF, the poles responsible for the first formant peak are in effect doubled.

Similar to the spectral matching scheme described for the single autocorrelation function, to overcome the problems posed by extremely dominant first formant peak, we can again consider the vocal tract to consist of separate subsystems each responsible for a



(a)



(b)

Figure 3.10: Effect of double ACF in z-domain (a) $H(z)$ (b) $P_h(z)$

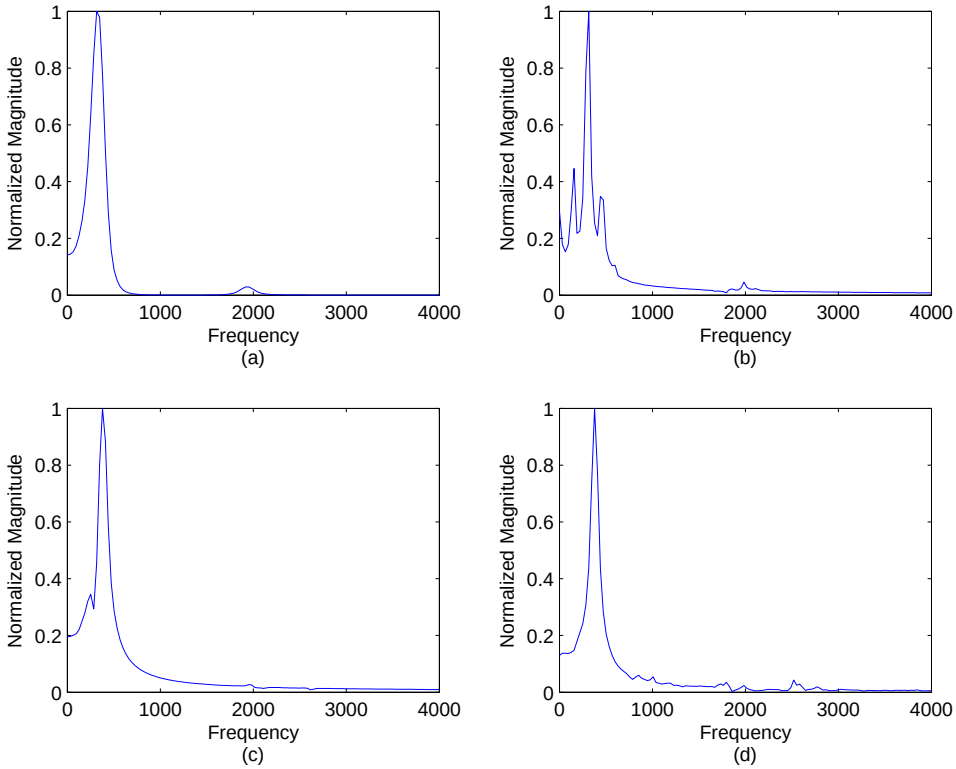


Figure 3.11: Effect of spectral strengthening because of autocorrelation operation. Spectrum of: (a) $\rho_h(n)$, (b) $\rho(n)$, (c) $\rho_x(n)$ and (d)$\rho(n)$

single formant frequency. As higher formants become increasingly weak due to their low energy concentration and the tilt caused by the lip radiation, it is sufficient to consider only the first three formants. Then the impulse response $h(n)$ of the whole system can be written as

$$h(n) = h_1(n) * h_2(n) * h_3(n) \tag{3.17}$$

where $h_1(n)$, $h_2(n)$ and $h_3(n)$ are the impulse responses of the individual systems. After performing double autocorrelation, the system impulse response becomes

$$\rho_h(n) = \rho_{h1}(n) * \rho_{h2}(n) * \rho_{h3}(n) \tag{3.18}$$

The Z Transform of $\rho_h(n)$, as obtained from (3.18) is given by

$$P_h(z) = P_{h1}(z)P_{h2}(z)P_{h3}(z) \tag{3.19}$$

The first formant peak is even more prominent in the spectrum of DACF presented in 3.11, indicating that the effect of $P_{h2}(z)$ and $P_{h3}(z)$ are negligible on $P_{h1}(z)$. Using this property, it can be assumed that the output response to closely match $P_{h1}(z)$ around the first formant peak. Thus instead of conventional peak picking, in this chapter, the task of formant estimation is carried out through model fitting, which ensures that both the frequency and bandwidth of formant peaks are matched.

3.2.2 Proposed Spectral Model of Repeated ACF

As seen from the previous section, the spectrum of the vocal tract response within a particular formant band generally exhibits a prominent peak corresponding to the formant. 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. Hence it is sufficient to consider a band limited speech signal corresponding to a particular formant band to analyze the effect of an individual formant. In this regard, considering the vocal tract system as a cascade of a set of subsystems, each subsystem that is responsible for generating a formant peak is denoted as $H_i(z)$.

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 ACF operation 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 output ACF of the most prominent subsystem, namely $H_1(z)$. 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 ACF of the noise corrupted signal.

In this case, the transfer function of the subsystem responsible for the ACF spectrum around the i-th formant peak as per (3.12) can be represented as

$$R_{Mi}(z) = \frac{C_{Ri}z^2}{(1-p_i z^{-1})(1-p_i^* z^{-1})(1-p_i z)(1-p_i^* z)} \tag{3.20}$$

where C_{Ri} is a constant.

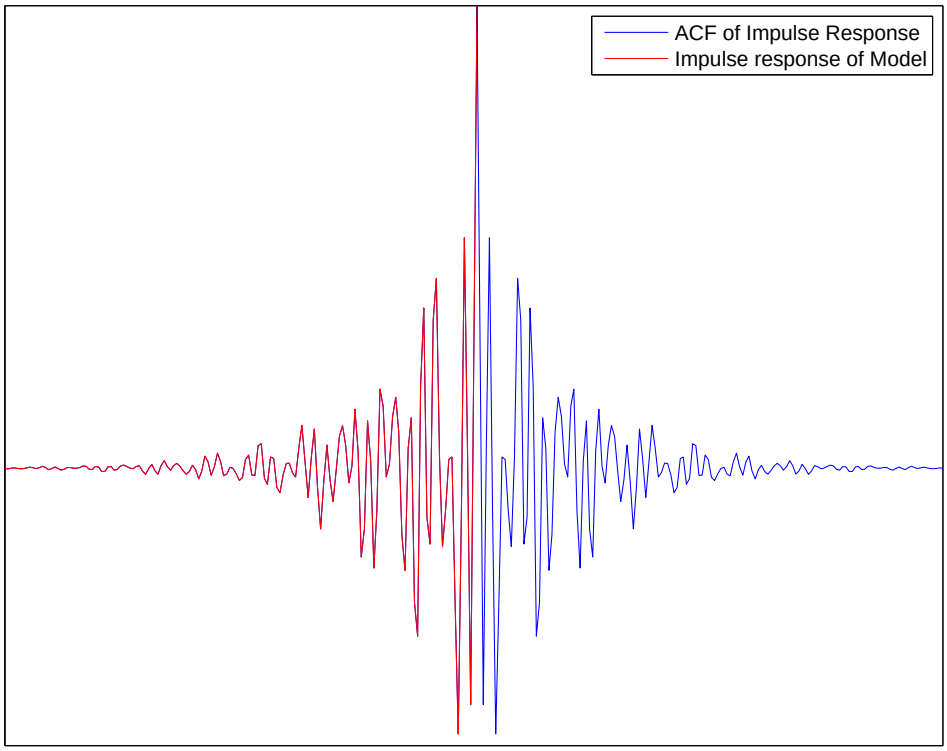


Figure 3.12: Impulse response of the proposed model $R_{h1}(z)$ and the ACF of the impulse response

With the introduction of each new pole outside the unit circle, a trivial zero is also introduced at the origin. The effect of these zeros is to introduce a delay in the output. Thus a pair of zeros is incorporated in (3.20). If the ACF of an impulse response for a synthetic speech signal is taken, it is expected that this will match with an impulse response obtained from a system which contains new poles in addition to the original ones, as described above. This is evident in Fig. 3.12, where these two signals match perfectly. Again, if the trivial zeros were not included while constructing the new system for generating an impulse response similar to the ACF, we would have experienced a delay between the signals, which can be observed at Fig. 3.13. In a similar fashion, if instead

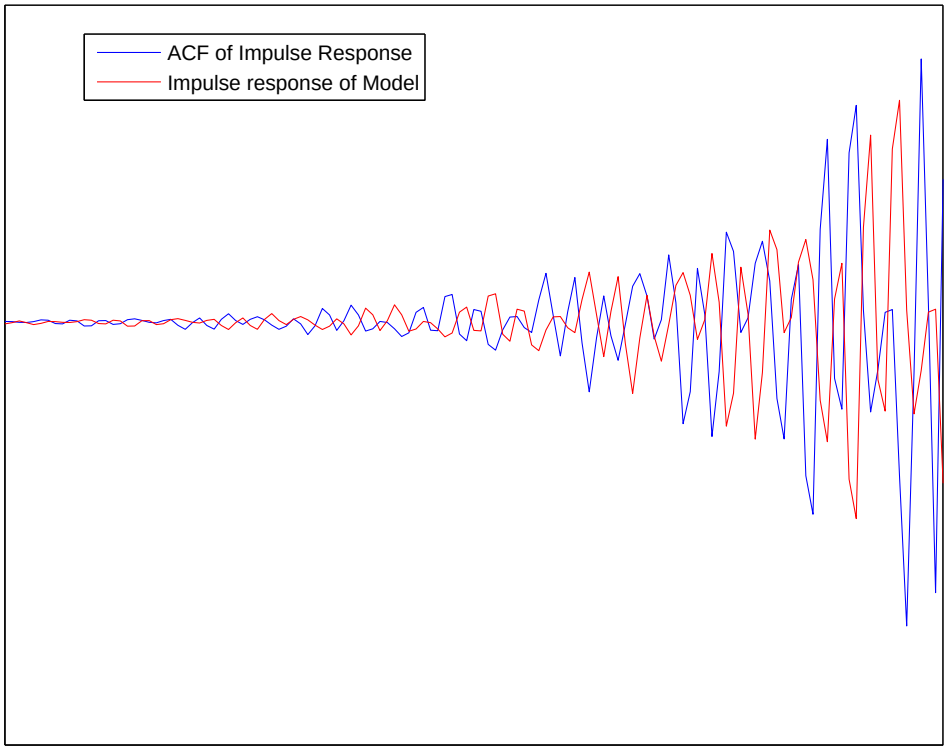


Figure 3.13: ACF data and model impulse response without any trivial zeros used

of one, double ACF is used, as explained before, the effect of noise will be further reduced and the transfer function corresponding to the DACF for the i-th subsystem according to (3.15) can be represented as

$$P_{hi}(z) = \frac{C_{Pi} z^6}{\{(1 - p_i z^{-1})(1 - p_i^* z^{-1})(1 - p_i z)(1 - p_i^* z)\}^2} \tag{3.21}$$

where C_{P1} is a constant.

Similar to the ACF, impulse responses obtained from the model in (3.21) should show an exact match with the DACF of the impulse response of a single pole pair system. This is confirmed in Fig. 3.14, where these two signals are shown to have a perfect match, showing the validity of the proposed model generation approach.

3.2.3 Formant Estimation using Spectral Matching

In the proposed formant estimation method, a spectral model corresponding to the first formant zone of the spectrum of the DACF of the speech signal is introduced, which is

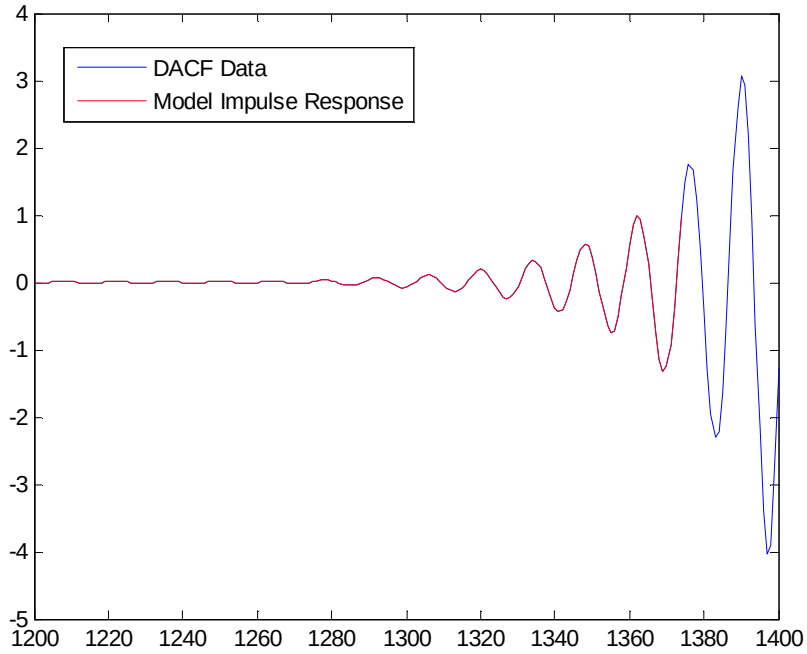


Figure 3.14: DACF data and model impulse response for a single formant band

utilized in a model matching technique to find out the model parameters that in turn will provide the desired formant frequencies. The DACF of each band limited speech frame $y_i(n)$ is computed and used in the proposed model matching technique. The z-transform representation of the DACF of $y_i(n)$ in (3.21) can have a Fourier transform representation for the region of convergence $r_i < z < \frac{1}{r_i}$, which is chosen as the model function of the proposed method and given by

$$P_{Mi}(e^{j\omega}) = \frac{C_i e^{j6\omega}}{\{(1-p_i e^{-j\omega})(1-p_i^* e^{-j\omega})(1-p_i e^{j\omega})(1-p_i^* e^{j\omega})\}^2} \qquad (3.22)$$
$$p_i = r_i e^{j\theta_i}$$

The spectrum $P_{yi}(e^{j\omega})$ of the DACF of the band limited observed noisy signal $y_i(n)$ is used in conjunction with the proposed model $P_{Mi}(e^{j\omega})$ to form an objective function based on the square of absolute difference of these spectra, namely

$$e^i_{min}(r_j, \theta_j) = \quad min \quad \sum_{\omega=\omega_{lc}}^{\omega_{hc}} \left(|P_{Mi}(e^{j\omega})| - |P^i_y(e^{j\omega})|\right)^2$$
$$r_l < r_i < r_h \tag{3.23}$$
$$\theta_l < \theta_i < \theta_h$$

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 [30]. Formant frequencies are estimated from the pole angle θ_j that produces the best match between the spectra using (3.4).

Similar to the process described in the first chapter, once the first formant frequency $F1$ is obtained, (3.19) is utilized to estimate the second formant frequency $F2$. As $P_y(e^{j\omega})$ can be taken as the product of $P_{y1}(e^{j\omega})$, $P_{y2}(e^{j\omega})$ and $P_{y3}(e^{j\omega})$, the magnitude spectrum of $P_y(e^{j\omega})$ is divided by $P_{M1}(e^{j\omega})$ so that the resulting spectrum $P^1_y(e^{j\omega})$ closely resembles the product of $P_{y2}(e^{j\omega})$ and $P_{y3}(e^{j\omega})$. Then similar to the matching in the first formant zone, matching is performed in the second formant zone and $F2$ is estimated. Then the magnitude spectrum of $P^1_y(e^{j\omega})$ is divided by $P_{y2}(e^{j\omega})$ to obtain $P^2_y(e^{j\omega})$. According to the simplified modeling of the vocal tract presented above, $P^2_y(e^{j\omega})$ should closely match with $P_{y3}(e^{j\omega})$, leading to a similar approach as described in (3.22) and (3.23) to obtain $F3$.

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 consideration 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/or the formants

are very closely spaced.

3.2.4 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. 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. 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

$$J(w) = \frac{w^T S_b w}{w^T S_w w} \tag{3.24}$$

where S_w and S_b are within and between-class scatter matrices, respectively[50] .

3.3 Simulation Results and Discussion

In order to evaluate the recognition performance of the proposed methods, experiments have been conducted on the same collection of utterances from the TIMIT acoustic-phonetic continuous speech corpus, introduced in the previous chapter. For the purpose of performance comparison, first the most widely used LPC based formant estimation method [53] 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 [32].

Table 3.1: Comparison of the estimation performance for synthetic vowels

Vowels			$5dB$			$-5dB$		
			Proposed	LPC	AFB	Proposed	LPC	AFB
Male	/a/	F1	4.92	21.57	43.65	6.04	24.53	47.17
		F2	9.74	7.24	25.74	7.47	99.56	27.25
		F3	12.48	20.49	10.68	17.57	39.35	10.42
	/o/	F1	4.56	61.38	124.73	4.56	73.15	21.63
		F2	14.16	167.49	43.93	17.23	144.60	58.65
		F3	17.90	36.74	12.54	17.90	37.68	11.66
	/u/	F1	5.80	93.53	149.02	6.55	117.36	13.56
		F2	10.48	158.74	46.60	14.26	148.07	63.59
		F3	2.59	69.03	38.05	2.82	72.38	19.40
Female	/a/	F1	5.98	20.24	46.90	5.98	20.46	49.77
		F2	8.14	65.23	32.58	7.41	113.79	30.99
		F3	7.78	17.80	8.45	11.35	34.02	9.84
	/o/	F1	10.75	49.53	128.07	10.75	78.29	18.29
		F2	10.05	138.88	20.42	15.79	133.29	46.61
		F3	4.80	39.93	9.56	7.37	36.28	12.53
	/u/	F1	9.43	72.96	109.00	9.52	98.29	12.98
		F2	9.39	116.33	14.62	13.89	121.92	33.72
		F3	6.74	52.31	11.40	7.64	40.60	13.74

In the proposed model fitting approach, the model parameter limits are set according to the general behavior of the vocal tract. The parameter r which determines the bandwidth of the resulting formant has a limit of $[0.8, 0.99]$ for the purpose of our simulation. The search range for θ is set according to the determined formant band. Search resolutions of $\triangle r = 0.01$ and $\triangle\theta = 0.001\pi$ are used for r and θ, respectively.

At first results for three synthetic vowels /a/, /o/ and /u/ are presented in Table 3.1.

Table 3.2: Comparison of the estimation performance in terms of average error for male speakers

Vowel		$-5dB$			$5dB$		
		Proposed	LPC	AFB	Proposed	LPC	AFB
/aa/	F1	14.90	30.53	30.88	14.07	26.48	17.74
	F2	12.00	82.19	36.42	11.47	45.44	21.87
	F3	17.59	43.35	15.47	12.11	39.80	17.07
/ah/	F1	15.36	31.64	24.12	14.83	24.65	16.31
	F2	10.42	57.43	28.88	10.23	35.57	24.41
	F3	13.37	39.21	13.09	9.78	37.72	11.61
/eh/	F1	14.69	27.70	24.62	14.31	15.56	18.17
	F2	17.80	33.30	24.18	10.98	18.03	18.17
	F3	13.57	39.13	13.39	10.14	35.55	13.05
/ix/	F1	14.15	37.19	24.84	13.14	12.22	28.69
	F2	16.41	31.84	24.89	10.74	21.27	23.10
	F3	10.91	39.08	14.50	10.36	35.36	15.29
/ow/	F1	16.05	22.63	35.49	15.28	19.72	37.77
	F2	13.77	47.20	26.03	12.64	41.67	24.65
	F3	14.24	36.68	14.20	12.62	37.74	14.00
/uw/	F1	15.64	29.66	36.72	15.58	20.09	39.58
	F2	13.10	40.36	23.14	12.77	36.45	22.49
	F3	12.39	39.48	14.53	11.53	38.25	14.50

Table 3.3: Comparison of the estimation performance in terms of average error for female speakers

Vowel		$-15dB$			$0dB$		
		Proposed	LPC	AFB	Proposed	LPC	AFB
/aa/	F1	15.89	48.89	46.25	11.72	15.91	41.89
	F2	20.74	83.33	21.40	11.74	50.37	25.37
	F3	13.89	43.46	14.23	13.11	27.05	12.70
/ah/	F1	22.93	50.77	37.88	13.16	12.32	35.70
	F2	19.35	66.14	21.65	9.93	33.23	19.26
	F3	13.28	34.12	16.12	13.35	22.09	14.33
/eh/	F1	16.85	55.19	31.51	11.06	9.32	24.85
	F2	28.91	31.17	28.87	15.18	11.90	23.45
	F3	13.74	28.41	12.15	9.79	19.62	12.65
/ow/	F1	18.50	76.56	24.15	12.44	10.81	22.83
	F2	27.01	41.29	31.59	14.56	25.46	27.60
	F3	13.47	28.29	14.97	9.41	20.64	14.49
/uh/	F1	17.17	77.76	24.67	12.88	11.46	22.14
	F2	25.27	40.40	32.34	13.33	24.47	26.61
	F3	13.04	28.21	15.12	8.50	20.88	13.60
/uw/	F1	17.16	81.07	24.32	12.62	10.74	23.16
	F2	25.09	40.34	32.41	13.50	23.79	27.00
	F3	12.93	29.35	15.31	8.86	21.40	13.80

Vowels with duration of 80 ms are synthesized using the Klatt synthesizer considering the pitch values of 120 Hz and 220 Hz, respectively, for male and female speakers. 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.

The estimation errors for utterances from the TIMIT database 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 are presented in Tables 3.2 and 3.3. The estimation errors obtained by the proposed method and that by the other two methods are presented under the influence of various levels of white gaussian noise conditions for male and female speakers, respectively for a selection of vowels. 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 12 order LPC and 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.

Similar to the first chapter, the proposed method offers very good estimation accuracies even for the third formant, for which estimation is generally quite difficult due to the low level of energy in the spectrum. However, for the female vowels like /iy/ with closely spaced second and third formants, the level of estimation accuracy is low for the second formant. However, 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. As in previous chapter, formant estimation is carried out frame by frame with a frame length of 512 samples and 10 ms overlap between the successive frames.

In order to present the overall formant estimation errors over the entire range of SNRs considered in the experimental setup, in Figs. 3.15, 3.16, 3.17 and 3.18, average of

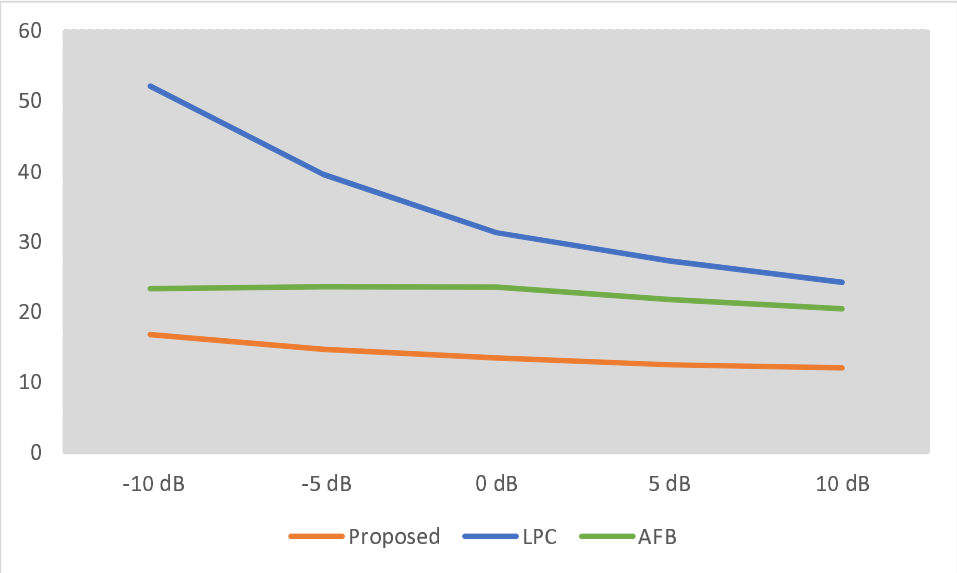


Figure 3.15: First formant estimation performance in terms of percentage error in formant estimation under various noise levels for male speakers

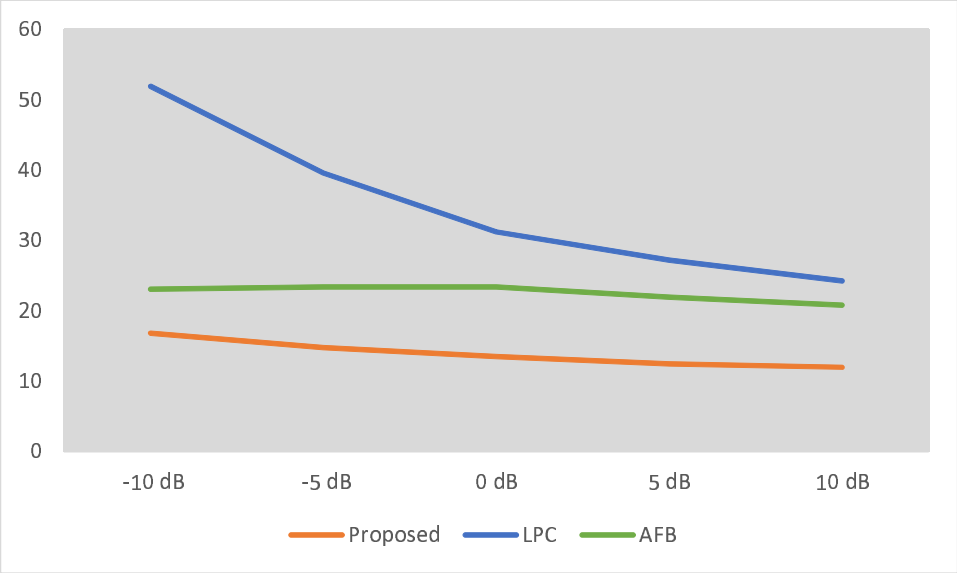


Figure 3.16: Second formant estimation performance in terms of percentage error in formant estimation under various noise levels for male speakers

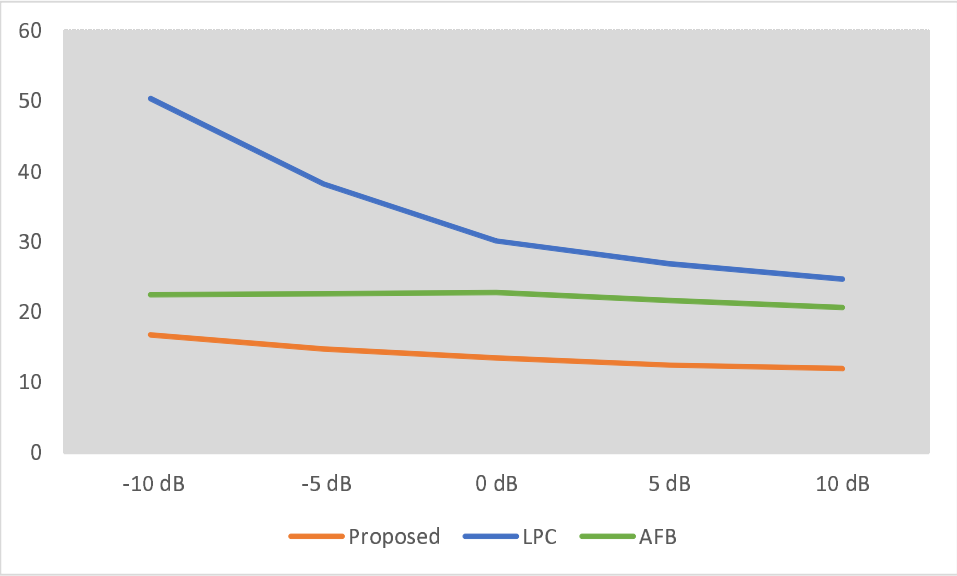


Figure 3.17: Third formsnt estimation performance in terms of percentage error in formant estimation under various noise levels for male speakers

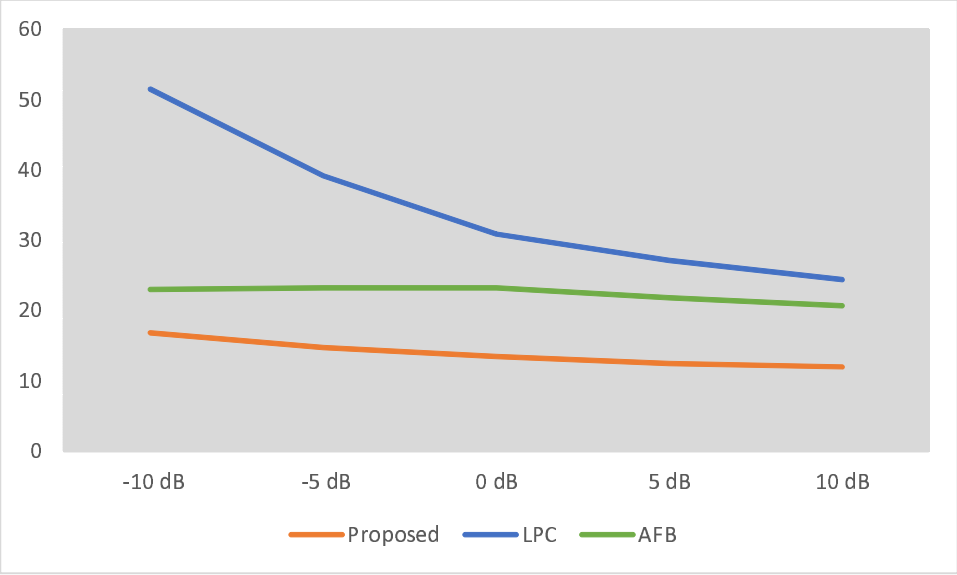


Figure 3.18: Estimation performance in terms of percentage error in formant estimation under various noise levels for male speakers

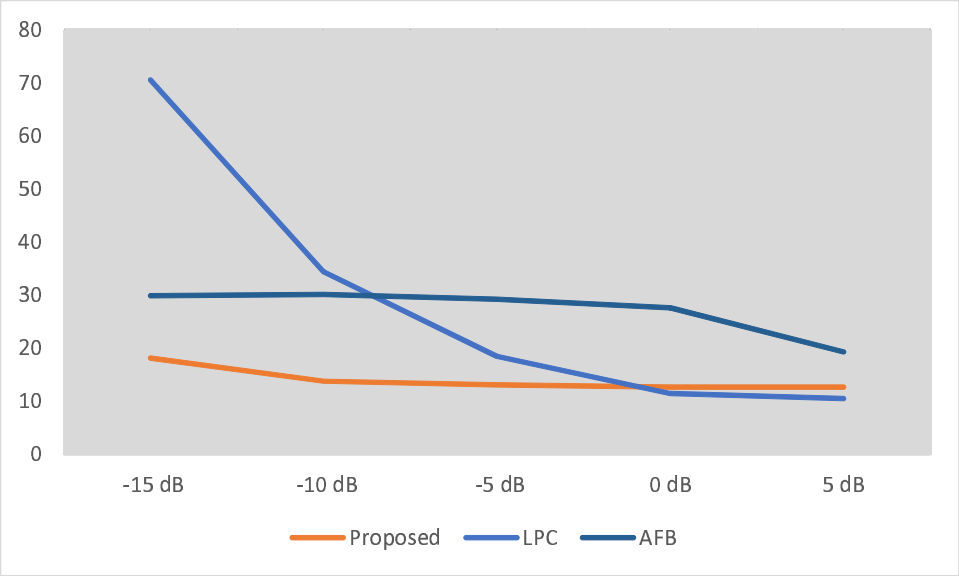


Figure 3.19: First formant estimation performance in terms of percentage error in formant estimation under various noise levels for female speakers

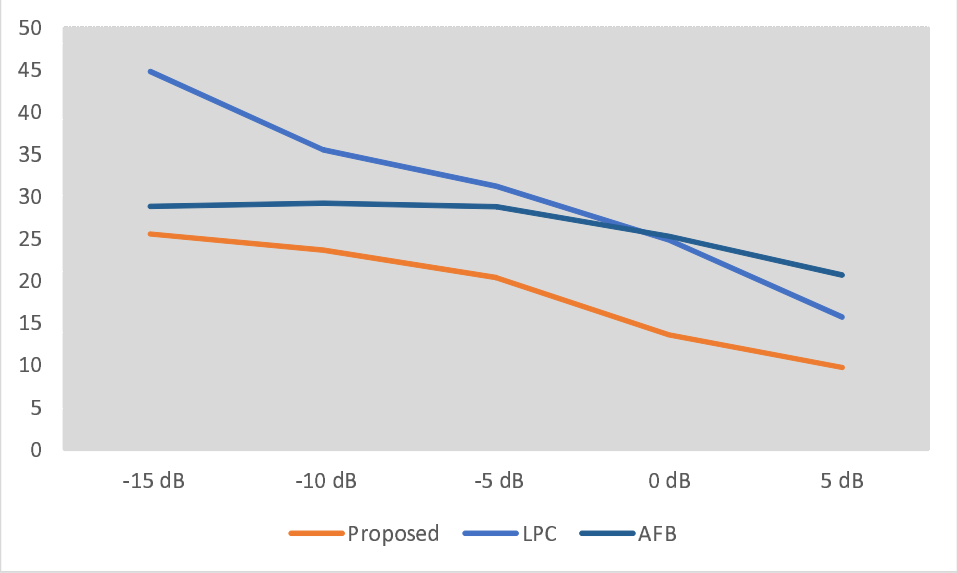


Figure 3.20: Second formant estimation performance in terms of percentage error in formant estimation under various noise levels for female speakers

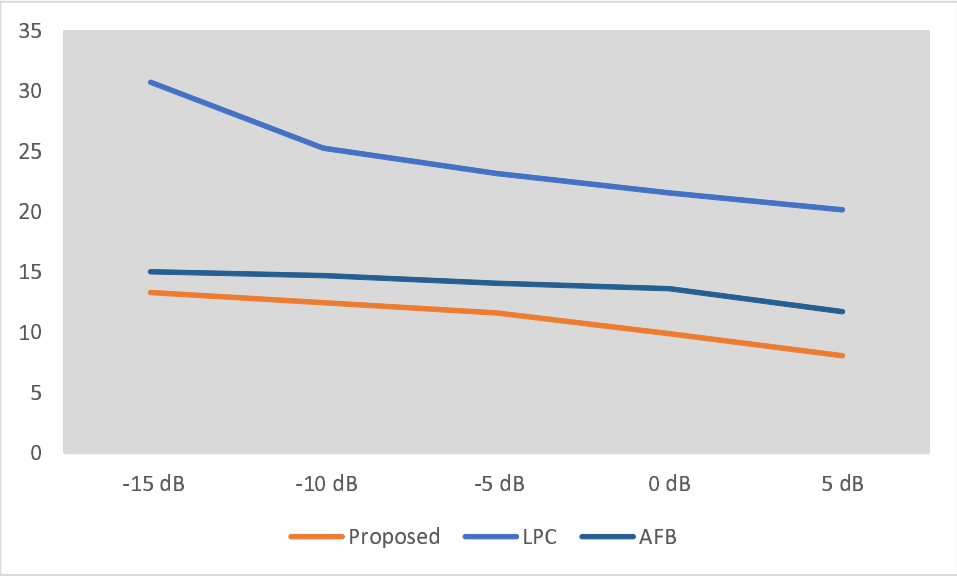


Figure 3.21: Estimation performance in terms of percentage error in formant estimation under various noise levels for female speakers

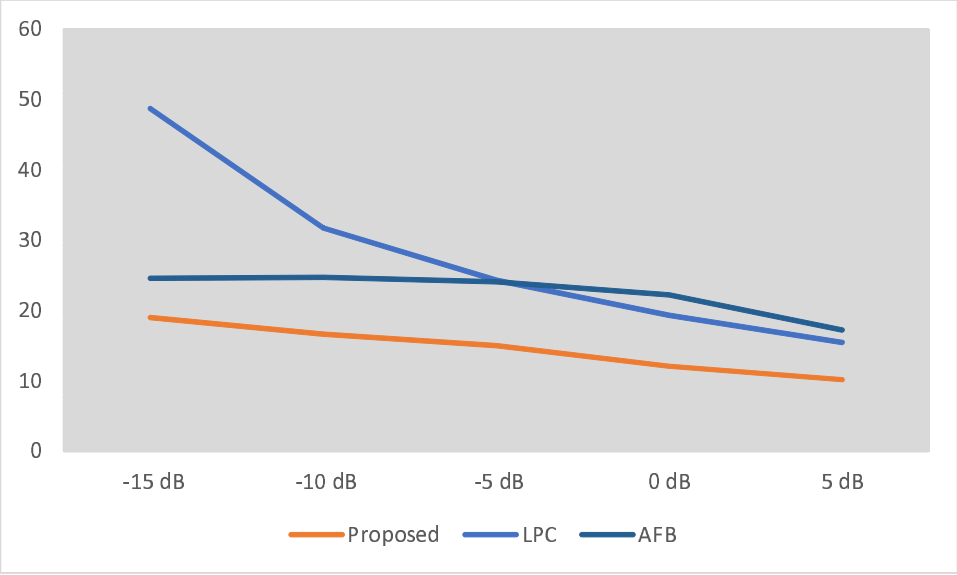


Figure 3.22: Estimation performance in terms of percentage error in formant estimation under various noise levels for female speakers

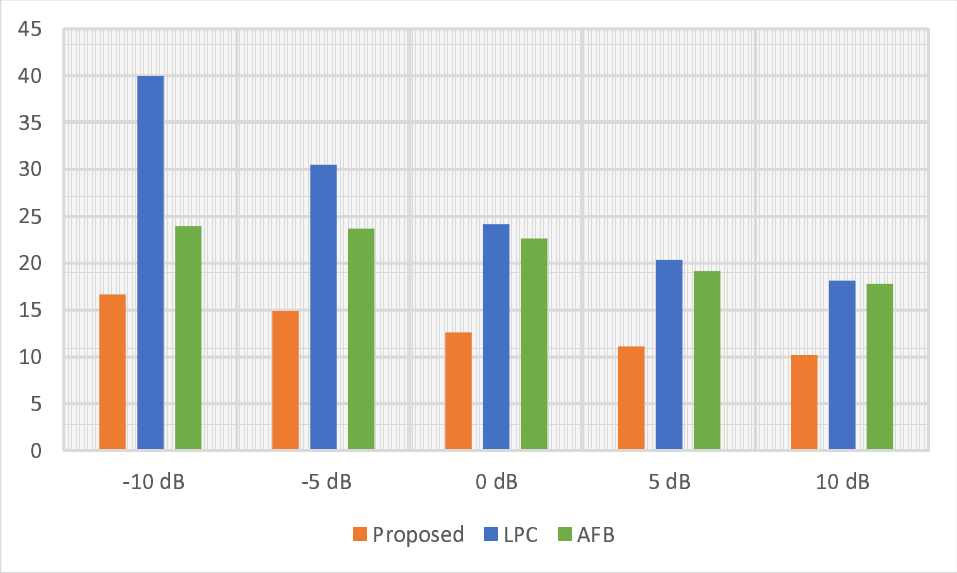


Figure 3.23: Estimation performance in terms of percentage error in formant estimation under various noise levels

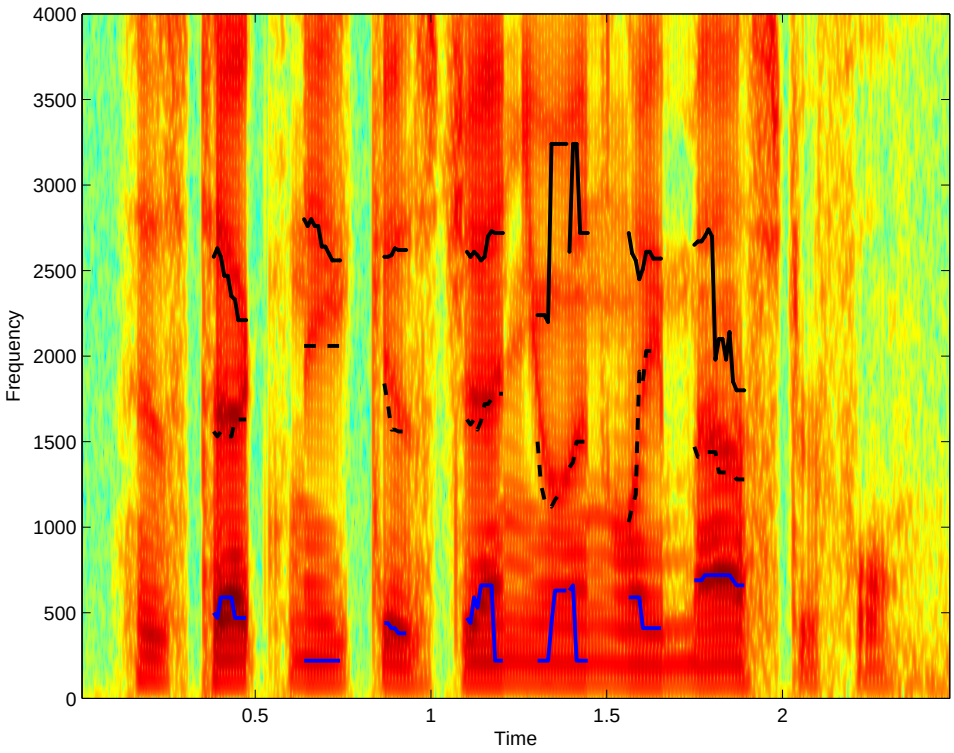


Figure 3.24: Spectrogram of the utterance 'His technique is genuinely masterful' , with formant frequencies estimated using the proposed method

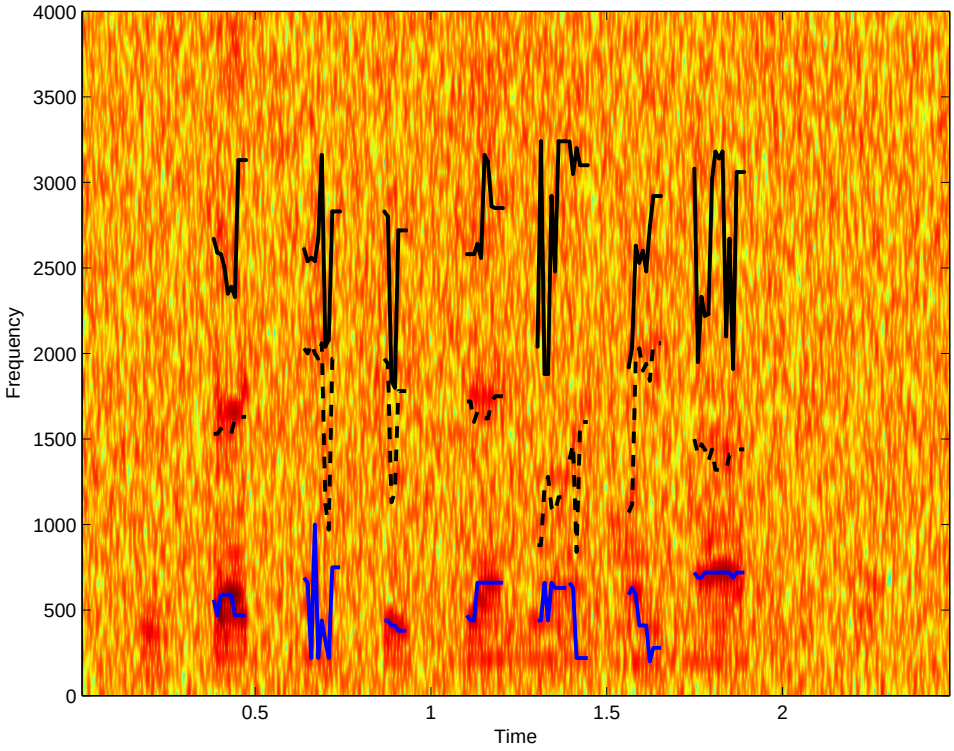


Figure 3.25: Spectrogram of the utterance 'His technique is genuinely masterful' , under $-5dB$ of background noise with formant frequencies estimated using the proposed method

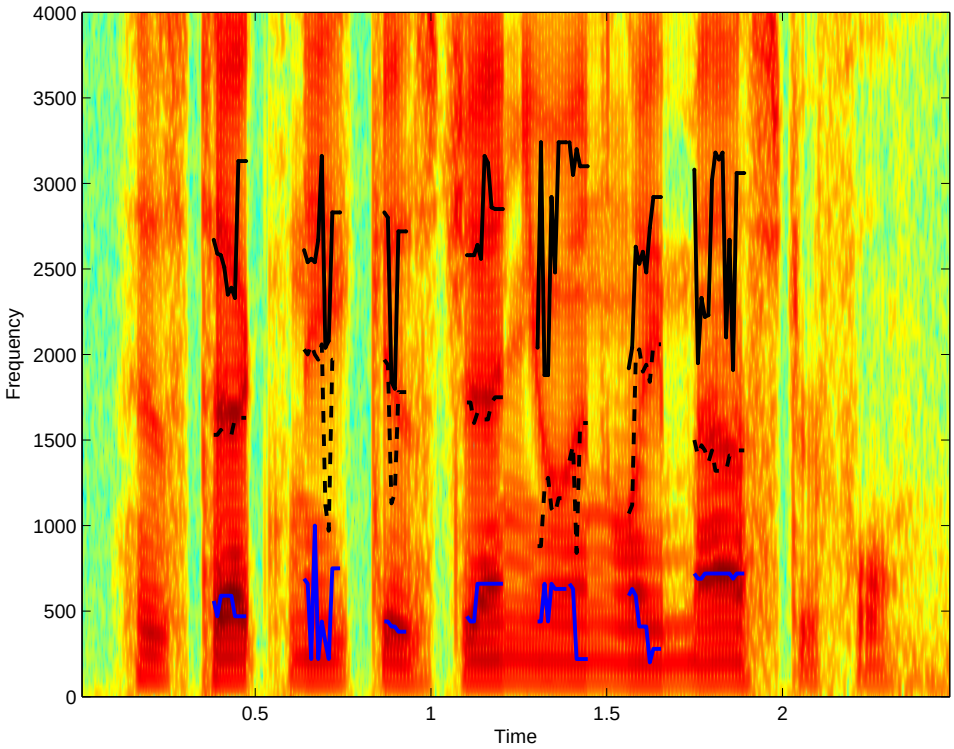


Figure 3.26: Spectrogram of the utterance 'His technique is genuinely masterful' , with formant frequencies estimated under $-5dB$ of Background noise using the proposed method

estimation error of all vowels for all three formants are shown for the the proposed method and the $LPC-12$ based method considering only male speakers. In this case, the SNR levels considered are ranging from -10 to $+10dB$. In a similar way, in Figs. 3.19, 3.20, 3.21 and 3.22, the average estimation error are shown for the female speakers for a SNR range of -15 to $+5dB$. Finally, in Fig. 3.23, the average estimation error considering both male and female speakers is 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 significantly 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 $-15dB$. 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.

Table 3.4: Vowel recognition accuracy

Feature Vector	$10dB$	$-5dB$
MFCC + Proposed Method	93.33	88.33
MFCC + LPC-12	93.33	81.66
MFCC + TIMIT reference	93.33	90.00
MFCC	93.33	81.66

By incorporating the estimated formants in a feature vector along with traditional MFCC, vowel recognition accuracies marginally improved 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 features as a feature vector, vowel recognition was performed for the vowels /aa/, /ow/ and /ux/ from the TIMIT database. For the purpose of further comparison, vowel recognition accuracies obtained by incorporating the noise free reference values of the formants are incorporated in the feature vector along with MFCC features obtained from noisy speech. It can be observed that up to $-10dB$, the performance of the proposed feature vector is comparable even to the performance of feature vectors incorporating noise free formant estimations. The recognition accuracies for different vowels is presented in Table 3.4.

As seen from these analysis, 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 technique is genuinely masterful' uttered by a male speaker taken from the TIMIT database is shown in Fig. 3.24. The formant frequencies estimated at different frames using the proposed method are shown over the spectrogram. 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. For the purpose of comparison, the same sentence, under influence of $-5dB$ background noise, is utilized to obtain the spectrogram present in Fig. 3.25. Here the presence of noise has completely obscured the energy bands, but still the proposed method can successfully track the formant frequencies. With the purpose of gaining a better insight, the formant frequencies obtained from the $-5dB$ noise corrupted speech are overlayed on the spectrogram for noise free speech, which is shown in Fig. 3.26. The resulting tracking lines obtained by the proposed method is a clear indication of its high level of consistency as well as the accuracy even in heavy noisy condition.

3.4 Conclusion

In this chapter, a formant estimation scheme based on frequency domain modeling of repeatedly autocorrelated speech. An adaptive band recognition system is deployed that can find out the band of successive formant frequencies for pre-processed voiced speech signals. The speech signal is then passed through an adaptive filter designed to separate the responses of different formants. Repeated autocorrelation, which strengthens the dominant poles, and exponentially increases the peak-valley ratio at formant frequencies of the magnitude response, canceling out the effects of noise, is then performed on the filtered speech signals. Formant estimation is carried out in the spectral domain where instead of direct peak-picking from the speech spectrum, a spectral domain model of ACF of speech signal is first proposed considering the resonances of the vocal tract. A

spectral domain model fitting based algorithm is also developed to extract the model parameters which in turn give the formant. Through the simulation results on standard speech databases , it is shown that the developed method is effective in maintaining a high success rate in formant estimation even in the presence of a significant background noise.

Chapter 4

Spectral Model of Repeated Autocorrelation of Band Limited Speech

In this chapter, an adaptive band recognition system is deployed that can find out the band of successive formant frequencies for pre-processed voiced speech signals. The speech signal is then passed through an adaptive filter designed to separate the responses of different formants. As seen in the previous chapters, repeated autocorrelation strengthens the dominant poles and exponentially increases the peak-valley ratio at formant frequencies of the magnitude response. In order to exploit this characteristic, a formants estimation method involving the repeated autocorrelation of band limited speech is developed. Formant estimation is carried out in the spectral domain where instead of direct peak-picking from the speech spectrum, a spectral domain model of ACF of speech signal is first proposed considering the resonances of the vocal tract. A spectral domain model fitting based algorithm is also developed to extract the model parameters which in turn give the formant. Through the simulation results on standard speech databases as well as synthetic speech signals, it is shown that the developed method is effective in maintaining a high success rate in formant estimation even in the presence of a significant background noise.

4.1 Background

In order to estimate the formant frequencies from observed speech signal, it is sufficient to restrict the analysis only for the voiced sound, as described in the previous chapters. In case of the voiced speech signals, considering the excitation as a periodic impulse-train, the overall vocal tract filter can be represented by a P-th order autoregressive (AR) system comprising of with the following transfer function

$$H(z) = \frac{C}{\prod_{i=1}^{P}(1 - p_i z^{-1})} \tag{4.1}$$

where p_i denotes the pole of the AR system and C is the gain factor. This system can be further subdivided into individual cascaded subsystems whose transfer function can be presented as

$$H_i(z) = \frac{C_i}{(1 - p_i z^{-1})(1 - p_i^* z^{-1})} \tag{4.2}$$

Here the magnitude r_i and angle θ_i of each pair of complex conjugate poles $p_i = r_i e^{(j\theta_i)}$ are related to a particular formant F_i and the formant bandwidth B_i as

$$r_i = e^{\frac{-\pi B_i}{Fs}} \tag{4.3}$$

$$\theta_i = \frac{2\pi F_i}{Fs} \tag{4.4}$$

where Fs is the sampling frequency.

For a voiced sound$x(n)$in the presence of additive noise $v(n)$ with zero mean and unit variance, the noise corrupted speech $y(n)$ can be written as

$$y(n) = x(n) + v(n) \tag{4.5}$$

In a time domain representation of the noise corrupted speech signal, it is very difficult to distinguish the original speech samples even at a moderate level of noise. The presence

of additive noise completely destroys the original speech pattern resulting in a noise like pattern. As described before, the autocorrelation operation on the noisy signal may reduce the effect of noise. The autocorrelation function of a voiced sound $x(n)$ is defined as

$$R_x(\tau) = E[x(n)x(n-\tau)] \tag{4.6}$$

where τ denotes the lag. ACF is an even function, with the output being symmetric with respect to the amplitude axis.

The ACF of noisy speech y(n) can be expressed as

$$\begin{aligned} r_y(n) &= r_x(n) + r_w(n) \\ where \quad r_w(n) &= r_v(n) + r_v x(n) + r_x v(n) \end{aligned} \tag{4.7}$$

Here $r_v(n)$ is the ACF of noise $v(n)$ and $r_{vx}(n)$ and $r_{xv}(n)$ are the cross correlation terms. Since the autocorrelation is a pole preserving operation and it exhibits higher noise immunity, it is advantageous to deal with the ACF of $y(n)$ instead of directly using $y(n)$ in spectral domain formant estimation.

Considering $x(n)$ as an output of an LTI system with transfer function $H(z)$, $x(n)$ can be written as

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

And it can be shown that the ACF of $x(n)$ can be expressed as

$$r_x(n) = r_h(n) * r_u(n) \tag{4.9}$$

where $r_u(n)$ is the ACF of $u(n)$. As per the definition of the ACF provided in (4.6), the ACF of $h(n)$ can be written as

$$r_h(n) = h(n) * h(-n) \tag{4.10}$$

In view of analyzing the frequency domain effects, for simplicity, first the Z domain representation is considered. The Z Transform of $r_h(n)$, as obtained from (4.10) is given by

$$R_H(z) = H(z)H(z^{-1}) \tag{4.11}$$

According to the definition of the ACF mentioned in (4.6), the ACF of $r_x(n)$, namely the repeated ACF of $x(n)$ can be expressed as

$$\rho_x(n) = r_x(n) * r_x(-n) = \rho_h(n) * \rho_u(n) \tag{4.12}$$

As discussed before, it would be sufficient to consider the detailed analysis of $\rho_h(n)$ instead of $\rho_x(n)$. Using the definition in (4.10), the Z Transform of $\rho_h(n)$ can be written as

$$P_h(z) = R_h(z)R_h(z^{-1}) \tag{4.13}$$

Further application of ACF on the noise corrupted signal $r_y(n)$ produces $\rho_y(n)$ which can be expressed as

$$\begin{aligned} \rho_y(n) &= \rho_x(n) + \rho_c(n) \\ \rho_c(n) &= \rho_w(n) + \rho_{xw}(n) + \rho_{wx}(n) \end{aligned} \tag{4.14}$$

where$\rho_x(n)$ and $\rho_w(n)$ are the ACF of $r_x(n)$ and $r_w(n)$ and $\rho_{xw}(n)$ and $\rho_{wx}(n)$ are cross correlation terms. As per discussions from the previous chapter, the effect of $\rho_c(n)$ on $\rho_x(n)$ is significantly reduced because of the repeated autocorrelation operation. Hence, it is advantageous to utilize $\rho_y(n)$ instead of $r_y(n)$ in spectral domain formant estimation.

In comparison to the spectra corresponding to $y(n)$, It is clearly observed that the first peak in the spectra corresponding to $\rho_y(n)$ exhibits an extremely large peak in comparison to other peaks and significant spectral smoothing is observed in other zones of the spectrum. 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.

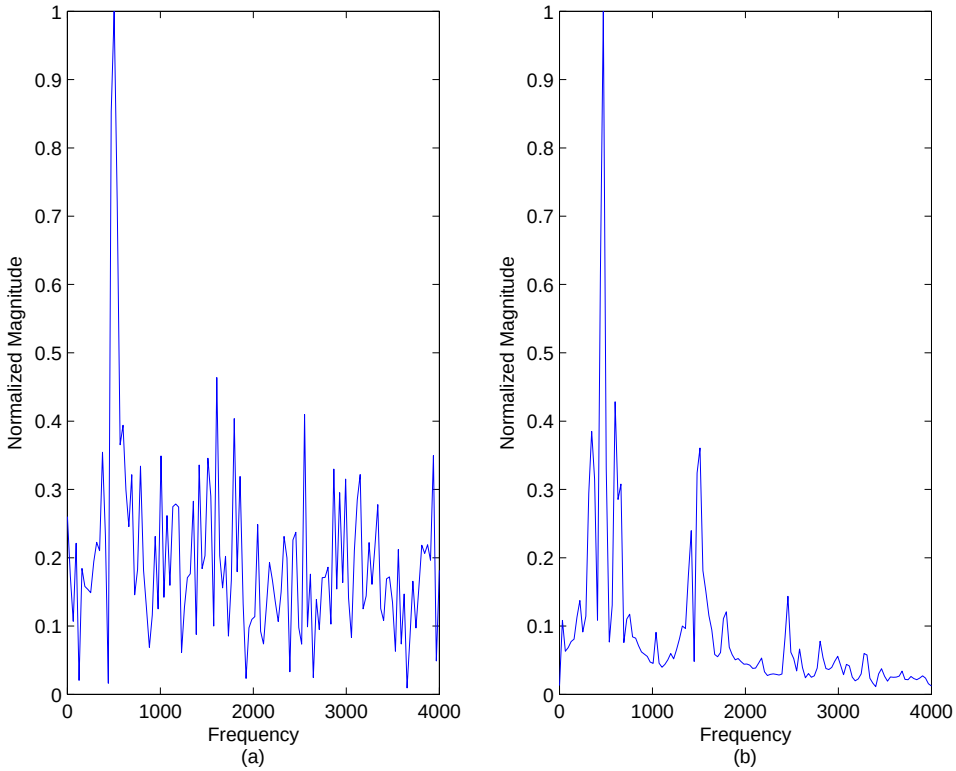


Figure 4.1: Spectrum of (a) natural utterance /eh/ under the influence of $-5dB$ white noise and (b) the DACF of the same utterance

This phenomenon is also observed in Fig. 4.1. 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.

4.2 Effect of Bandlimiting on Repeated ACF of Speech

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 pre-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 passband 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 bandpass 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

$$X_i(z) = X(z)B_i(z) \tag{4.15}$$

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

$$B_i(z) = B_{ih}(z)B_{il}(z) \tag{4.16}$$

where $B_{ih}(z)$ and $B_{il}(z)$ correspond to the transfer function of the highpass and lowpass 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-th filter produces $H_i(z)$. It is to be mentioned that the DACF operation which offers more peak-strengthening effect in comparison to the ACF, 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 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.

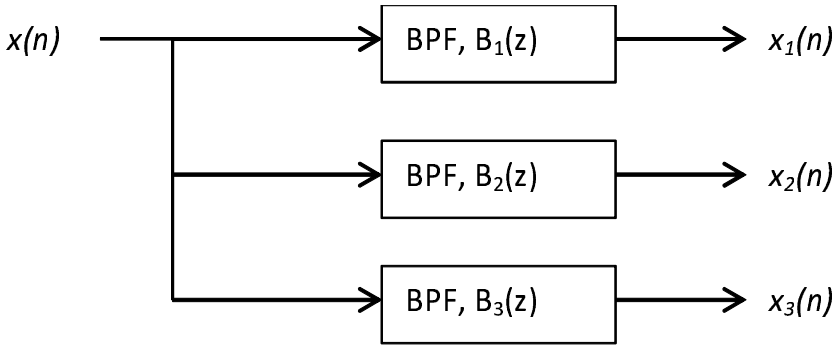


Figure 4.2: Banding using filters

In order to demonstrate the banding operation explained above, in Fig. 4.2 the process for bandlimited signal generation using three bandpass filters is shown. Considering a noise-free voiced speech /eh/ uttered by a male speaker taken from the TIMIT database,

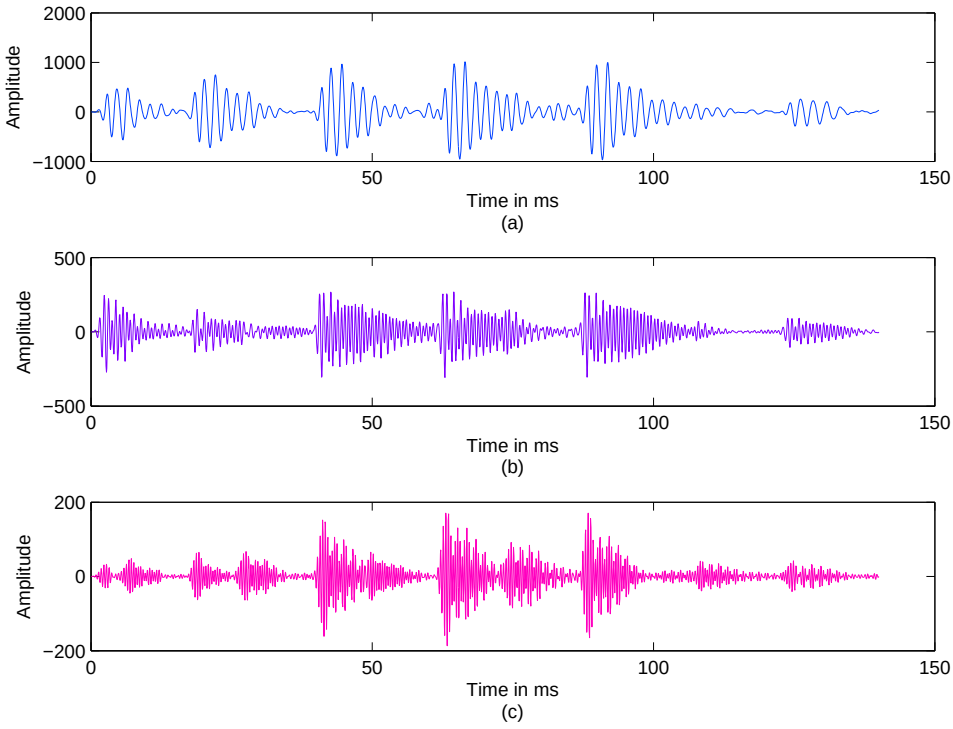


Figure 4.3: Time domain waveforms of bandlimited signal (a)$x_1(n)$, (b)$x_2(n)$, (c)$x_3(n)$

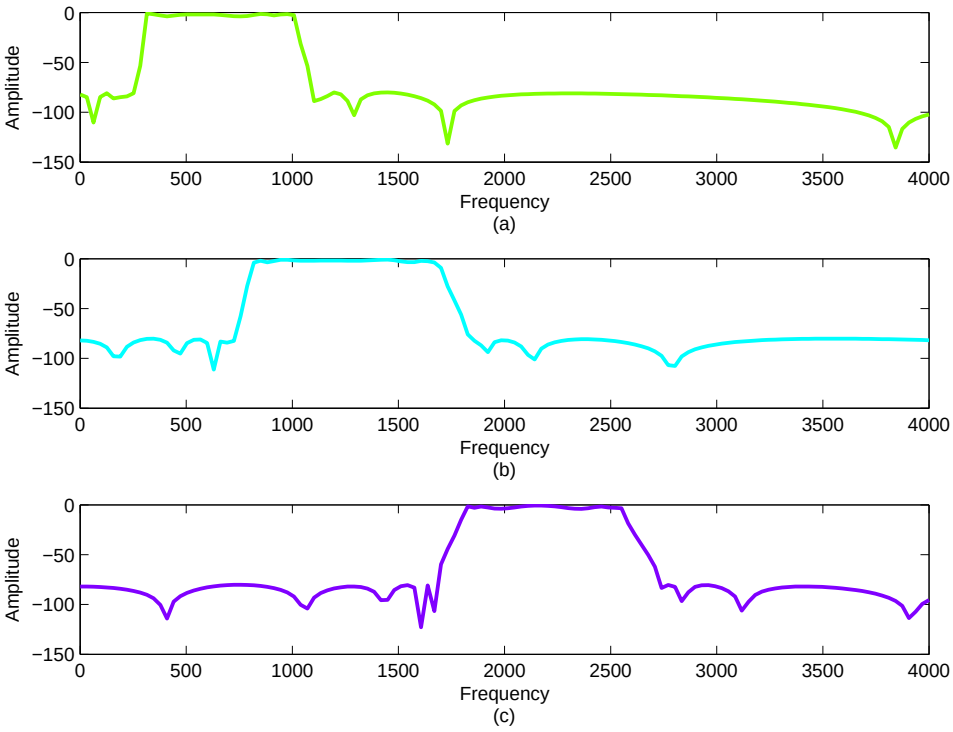


Figure 4.4: Frequency zones for the three bandpass filters and the spectra for the bandpass filter corresponding to (a) the first formant zone, (b) second formant zone and (c) The third formant zone

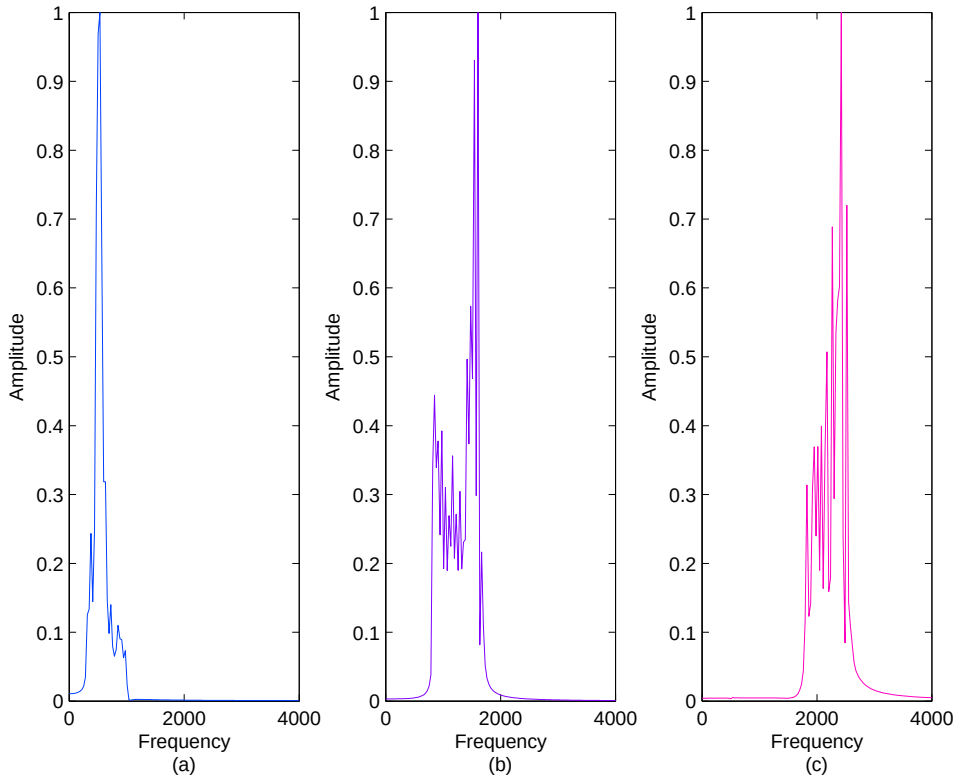


Figure 4.5: Spectrum corresponding to the output of (a) the first bandpass filter, (b) second bandpass filter and (c) The third bandpass filter

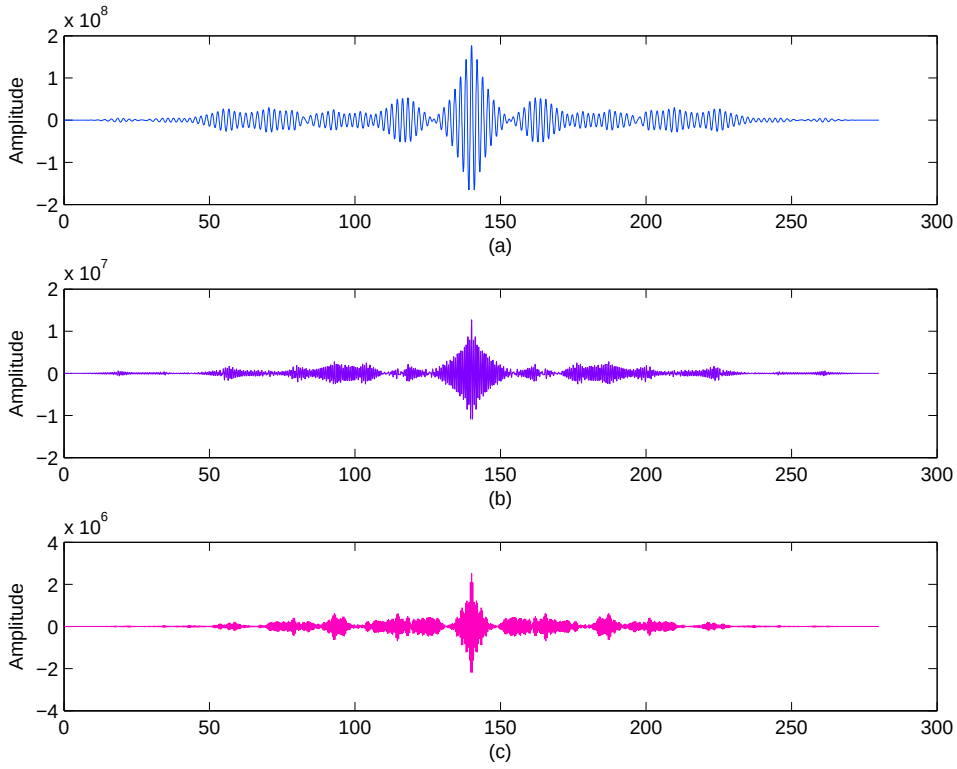


Figure 4.6: ACF of the output waveforms of the three bandpass filters

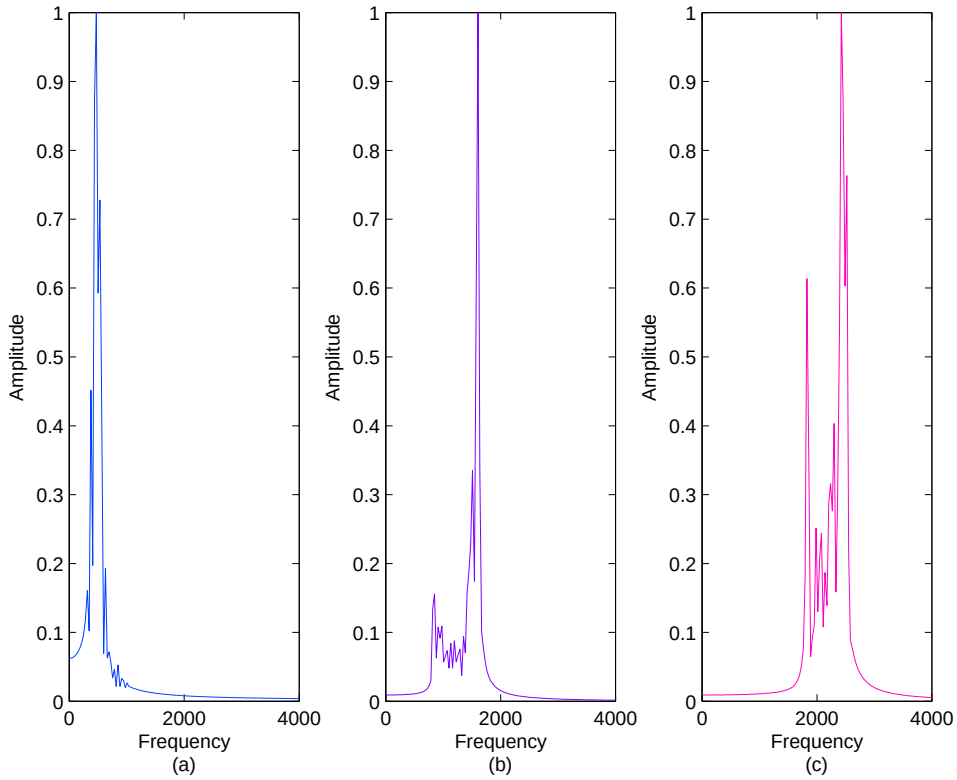


Figure 4.7: Spectrum corresponding to the ACFs of the three bandlimited signals

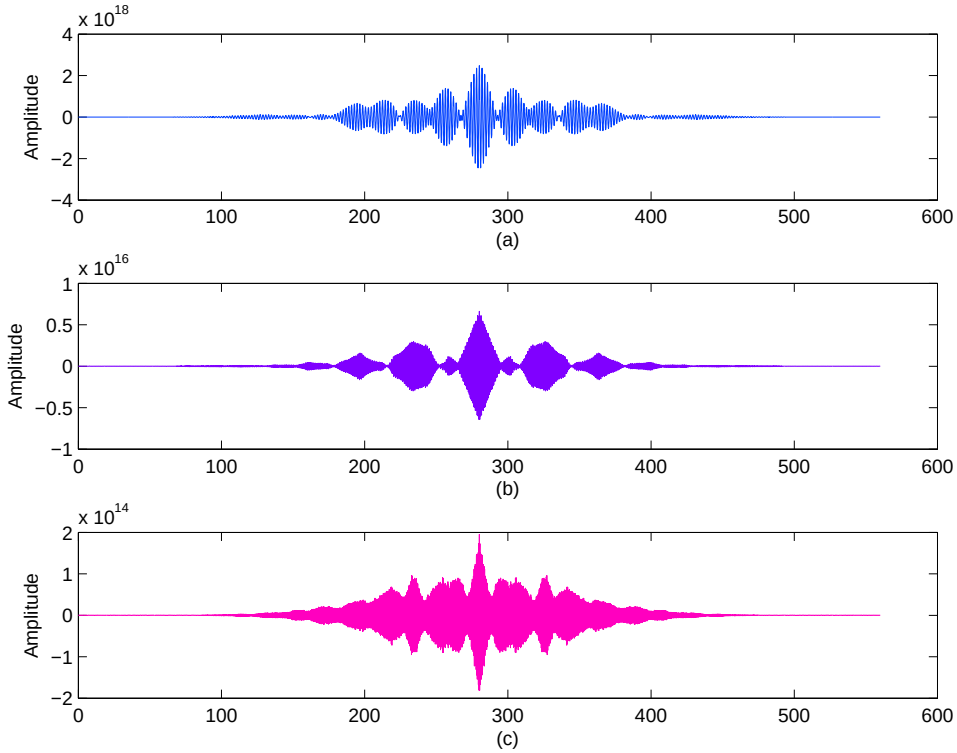


Figure 4.8: Waveforms for the DACF of the output of the three bandpass filters

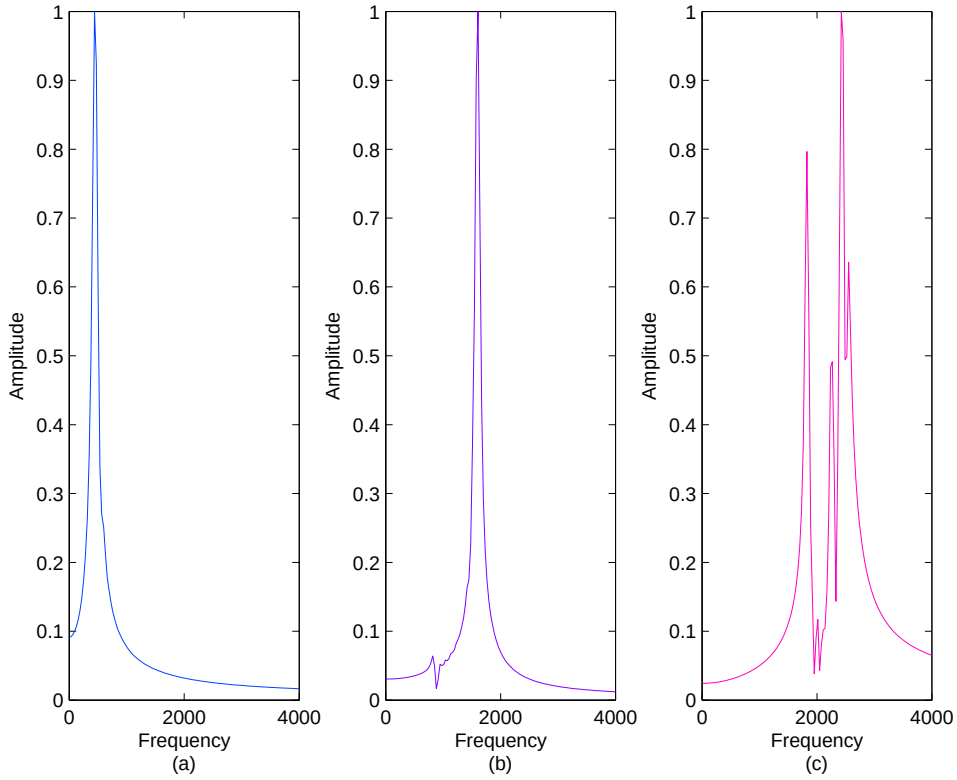


Figure 4.9: Spectrum for the DACF of the output of the three bandpass filters

three bandlimited outputs corresponding to Fig. 4.2, namely $x_1(n)$,$x_2(n)$,$x_3(n)$ is presented in Fig. 4.3.The typical frequency responses of three bandpass filters used in Fig. 4.2 tuned to the formant zones are shown in Fig. 4.4. The pass band for each filter is chosen in such a way that it covers the conventional formant bands. The spectra corresponding to the output of each bandpass filter are shown in Fig. 4.5. It is clearly observed that in the spectrum of a band limited signal, even for the third formant, the formant peak is clearly distinguishable for a noise-free signal. Next the effect of performing the autocorrelation operation on the bandlimited signal is demonstrated in time and spectral domain. In Fig. 4.6, the ACFs $r_{x1}(n)$,$r_{x2}(n)$,$r_{x3}(n)$, corresponding to the bandlimited signals are presented. It is observed that the autocorrelation function of each bandlimited signal preserves the corresponding signal property as expected. In Fig. 4.7, the spectra corresponding to the ACF of bandlimited signal are shown. In comparison to the spectrum of the ACF of signal $x(n)$ as shown in Fig. 2.10, in the spectrum of the ACF of the bandlimited signal, the formant peaks, especially the second and third formants are significantly enhanced. In a similar fashion, in Figs. 4.8 and 4.9 the effect of DACF is

demonstrated in time and spectral domain, respectively. Similar to Fig. 4.6, in Fig. 4.8 it is observed that the DACF operation on the bandlimited signal also preserves the signal properties. The peak strengthening obtained by the DACF operation is quite prominent as seen in Fig. 4.9.

In what follows, a frequency domain model is going to be developed which will be used as a target function in a model matching approach where the DACF of the bandlimited signal will be utilized.

4.3 Proposed Spectral Model

As seen from the previous section, the spectrum of the vocal tract response within a particular formant band generally exhibits a prominent peak corresponding to the formant. 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. Hence it is sufficient to consider a band limited speech signal corresponding to a particular formant band to analyze the effect of an individual formant. In this regard, considering the vocal tract system as a cascade of a set of subsystems, each subsystem that is responsible for generating a formant peak is denoted as $H_i(z)$.

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 DACF operation will definitely exhibit significant noise reduction. However, band limiting should be performed before the DACF operation as it would prevent the dominant formant peak from overshadowing other formant peaks. 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 ACF

of the band limited signal. The spectrum corresponding to that transfer function can then be used in a spectral matching technique along with the spectrum obtained from the ACF of band limited noise corrupted signal. For the band limited case, the transfer function corresponding to the ACF as per (4.11) can be represented as

$$R_H(z) = \frac{C_{Ri} z^2}{(1 - p_i z^{-1})(1 - p_i^* z^{-1})(1 - p_i z)(1 - p_i^* z)} \tag{4.17}$$

where C_{Ri} is a constant.

With the introduction of each new pole outside the unit circle, a trivial zero is also introduced at the origin. The effect of these zeros is to introduce a delay in the output. Thus a pair of zeros is incorporated in (4.17). If the ACF of an impulse response for a synthetic speech signal is taken, it is expected that this will match with an impulse response obtained from a system which contains new poles in addition to the original ones, as described above. In a similar fashion, if instead of one, double ACF is used, as explained before, the effect of noise will be further reduced and the transfer function corresponding to the DACF according to (4.13) can be represented as

$$P_H(z) = \frac{C_{Pi} z^6}{\{(1 - p_i z^{-1})(1 - p_i^* z^{-1})(1 - p_i z)(1 - p_i^* z)\}^2} \tag{4.18}$$

where C_{Pi} is a constant.

Similar to the ACF, impulse responses obtained from the model in (4.18) should show an exact match with the DACF of the impulse response of a single pole pair system.

4.3.1 Proposed Model Matching Scheme

In the proposed formant estimation method, a spectral model corresponding to the spectrum of the DACF of the bandlimited speech signal is introduced, which is utilized in a model matching technique to find out the model parameters that in turn will provide the desired formant frequencies. First, the given noise corrupted voiced frame of speech signal $y(n)$ is filtered out using the BPFs so that in the speech spectrum, only one formant range is present. Next, the DACF of each band limited speech frame $y_i(n)$ is computed

and used in the proposed model matching technique. The z-transform representation of the DACF of $y_i(n)$ in (4.18) can have a Fourier transform representation for the region of convergence $r_i < z < \frac{1}{r_i}$, which is chosen as the model function of the proposed method and given by

$$P_M(e^{j\omega}) = \frac{C_i e^{j6\omega}}{\{(1 - r_i e^{j\theta_i} e^{-j\omega})(1 - r_i e^{-j\theta_i} e^{-j\omega})(1 - r_i e^{j\theta_i} e^{j\omega})(1 - r_i e^{-j\theta_i} e^{j\omega})\}^2} \tag{4.19}$$

The spectrum $P_{yi}(e^{j\omega})$ of the DACF of the band limited observed noisy signal $y_i(n)$ is used in conjunction with the proposed model $P_M(e^{j\omega})$ to form an objective function based on the square of absolute difference of these spectra, namely

$$e_{min}(r_j, \theta_j) = \min_{\substack{r_l < r_i < r_h \\ \theta_l < \theta_i < \theta_h}} \left(\sum_{\omega = \omega_{lc}}^{\omega_{hc}} \left(\left(|P_M(e^{j\omega})| - |P_{yi}(e^{j\omega})| \right)^2 \right) \right) \tag{4.20}$$

As in the case of spectral matching without band limiting, 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 band obtained during the band limiting operation described in previous sections. 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 [30]. Formant frequencies are estimated from the pole angle θ_j that produces the best match between the spectra using (4.4).

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 consideration 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/or the formants are very closely spaced.

4.3.2 Vowel Recognition

For 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. 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

$$J(w) = \frac{w^T S_b w}{w^T S_w w} \tag{4.21}$$

where S_w and S_b are within and between-class scatter matrices, respectively[50] .

4.4 Simulation Results and Discussion

Similar to the previous chapters, formant estimation errors are reported for the 2726 utterances of phonemes used from the VTR subset of the TIMIT database. Results are also reported for LPC based formant estimation method [53] is chosen, where the order of the LPC is chosen as 12 and for the adaptive filter bank (AFB) method [32].

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

Table 4.1: Comparison of the estimation performance for synthetic vowels

Vowels			$5dB$			$-5dB$		
			Proposed	LPC	AFB	Proposed	LPC	AFB
Male	/a/	F1	4.70	21.57	43.65	4.70	24.53	47.17
		F2	5.18	7.24	25.74	2.74	99.56	27.25
		F3	6.33	20.49	10.68	6.32	39.35	10.42
	/o/	F1	4.56	61.38	124.73	5.82	73.15	21.63
		F2	3.72	167.49	43.93	5.49	144.60	58.65
		F3	6.71	36.74	12.54	6.71	37.68	11.66
	/u/	F1	5.80	93.53	149.02	7.86	117.36	13.56
		F2	3.00	158.74	46.60	3.92	148.07	63.59
		F3	9.43	69.03	38.05	13.56	72.38	19.40
Female	/a/	F1	4.56	20.24	46.90	4.56	20.46	49.77
		F2	4.09	65.23	32.58	3.43	113.79	30.99
		F3	2.58	17.80	8.45	6.95	34.02	9.84
	/o/	F1	11.85	49.53	128.07	12.05	78.29	18.29
		F2	6.39	138.88	20.42	11.83	133.29	46.61
		F3	6.70	39.93	9.56	9.48	36.28	12.53
	/u/	F1	10.40	72.96	109.00	12.79	98.29	12.98
		F2	6.35	116.33	14.62	8.43	121.92	33.72
		F3	9.57	52.31	11.40	11.78	40.60	13.74

Table 4.2: Comparison of the estimation performance in terms of average error for male speakers

Vowel		$-5dB$			$10dB$		
		Proposed	LPC	AFB	Proposed	LPC	AFB
/ae/	F1	11.66	19.84	27.67	11.04	13.22	12.33
	F2	12.22	29.64	23.60	7.59	13.06	15.36
	F3	13.76	35.94	12.25	7.24	29.24	11.61
/eh/	F1	13.30	27.70	24.62	12.94	14.05	17.07
	F2	14.45	33.30	24.18	8.78	11.95	17.91
	F3	13.04	39.13	13.39	7.53	33.11	11.67
/ih/	F1	14.83	38.52	23.47	14.64	12.40	23.98
	F2	15.64	27.12	25.50	7.97	15.16	20.05
	F3	11.55	39.50	13.45	7.48	32.66	11.34
/ix/	F1	13.77	37.19	24.84	13.40	11.37	27.44
	F2	14.28	31.84	24.89	8.79	17.13	22.73
	F3	10.66	39.08	14.50	7.63	34.03	13.79
/ow/	F1	15.12	22.63	35.49	14.99	18.78	36.11
	F2	12.39	47.20	26.03	11.62	34.22	22.37
	F3	12.92	36.68	14.20	10.56	36.92	14.15
/uh/	F1	15.50	20.14	36.49	15.40	19.21	36.59
	F2	11.76	38.02	23.49	11.50	35.66	22.50
	F3	10.39	37.24	13.89	10.26	37.06	14.07

Table 4.3: Comparison of the estimation performance in terms of average error for female speakers

Vowel		$-15dB$			$0dB$		
		Proposed	LPC	AFB	Proposed	LPC	AFB
/ah/	F1	19.44	50.77	37.88	12.47	12.32	35.70
	F2	16.88	66.14	21.65	15.04	33.23	19.26
	F3	12.30	34.12	16.12	11.00	22.09	14.33
/eh/	F1	13.60	55.19	31.51	11.10	9.32	24.85
	F2	27.00	31.17	28.87	19.01	11.90	23.45
	F3	8.78	28.41	12.15	8.17	19.62	12.65
/ow/	F1	15.33	76.56	24.15	12.47	10.81	22.83
	F2	27.48	41.29	31.59	15.85	25.46	27.60
	F3	11.88	28.29	14.97	10.23	20.64	14.49
/uh/	F1	15.45	77.76	24.67	13.40	11.46	22.14
	F2	27.36	40.40	32.34	16.24	24.47	26.61
	F3	12.07	28.21	15.12	9.89	20.88	13.60
/uw/	F1	15.34	81.07	24.32	13.07	10.74	23.16
	F2	27.77	40.34	32.41	16.38	23.79	27.00
	F3	12.30	29.35	15.31	10.18	21.40	13.80

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.

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_O is defined as

$$E = |\frac{f_E - f_O}{f_O}| \times 100\% \tag{4.22}$$

it can be observed that the proposed method offers far superior performance in the presence of noise for both male and female synthetic vowels.

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 . Similar to the analysis from previous chapters, the analysis is performed for different noise levels. 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.

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. 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/. However, 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.

Table 4.4: Comparison of the estimation performance for synthetic vowels

Vowels			$5dB$			$-5dB$		
			Proposed	LPC	AFB	Proposed	LPC	AFB
Male	/a/	F1	4.95	21.57	43.65	4.95	24.53	47.17
		F2	11.66	7.24	25.74	9.10	99.56	27.25
		F3	7.47	20.49	10.68	12.28	39.35	10.42
	/o/	F1	5.43	61.38	124.73	7.36	73.15	21.63
		F2	4.41	167.49	43.93	4.41	144.60	58.65
		F3	14.92	36.74	12.54	4.53	37.68	11.66
	/u/	F1	5.84	93.53	149.02	10.08	117.36	13.56
		F2	3.03	158.74	46.60	3.41	148.07	63.59
		F3	5.24	69.03	38.05	10.91	72.38	19.40
Female	/a/	F1	4.56	20.24	46.90	4.56	20.46	49.77
		F2	3.85	65.23	32.58	3.46	113.79	30.99
		F3	4.13	17.80	8.45	6.59	34.02	9.84
	/o/	F1	11.85	49.53	128.07	12.05	78.29	18.29
		F2	4.66	138.88	20.42	9.64	133.29	46.61
		F3	7.80	39.93	9.56	9.11	36.28	12.53
	/u/	F1	10.40	72.96	109.00	12.79	98.29	12.98
		F2	5.60	116.33	14.62	7.86	121.92	33.72
		F3	9.41	52.31	11.40	11.40	40.60	13.74

The proposed band limited approach is also applied to once performed ACF. in that case, the model (2.18) from chapter 1 is utilized for a matching approach with the ACF of the band limited speech. The results for this simulation are presented for synthetic

Table 4.5: Comparison of the estimation performance in terms of average error for male speakers

Vowel		$-5dB$			$10dB$		
		Proposed	LPC	AFB	Proposed	LPC	AFB
/ae/	F1	11.15	19.84	27.67	10.71	13.22	12.33
	F2	18.89	29.64	23.60	13.77	13.06	15.36
	F3	14.00	35.94	12.25	10.87	29.24	11.61
/eh/	F1	14.60	27.70	24.62	14.44	14.05	17.07
	F2	17.15	33.30	24.18	10.18	11.95	17.91
	F3	13.99	39.13	13.39	10.34	33.11	11.67
/ih/	F1	14.27	38.52	23.47	13.45	12.40	23.98
	F2	19.08	27.12	25.50	7.81	15.16	20.05
	F3	13.39	39.50	13.45	11.00	32.66	11.34
/ix/	F1	14.04	37.19	24.84	13.28	11.37	27.44
	F2	15.35	31.84	24.89	8.52	17.13	22.73
	F3	12.04	39.08	14.50	11.04	34.03	13.79
/ow/	F1	17.09	22.63	35.49	16.37	18.78	36.11
	F2	12.80	47.20	26.03	11.88	34.22	22.37
	F3	15.44	36.68	14.20	12.47	36.92	14.15
/uh/	F1	17.10	20.14	36.49	16.81	19.21	36.59
	F2	12.08	38.02	23.49	11.93	35.66	22.50
	F3	12.55	37.24	13.89	12.15	37.06	14.07

Table 4.6: Comparison of the estimation performance in terms of average error for female speakers

Vowel		$-15dB$			$0dB$		
		Proposed	LPC	AFB	Proposed	LPC	AFB
/ah/	F1	19.16	50.77	37.88	13.12	12.32	35.70
	F2	17.61	66.14	21.65	14.70	33.23	19.26
	F3	13.19	34.12	16.12	11.68	22.09	14.33
/eh/	F1	15.13	55.19	31.51	10.78	9.32	24.85
	F2	26.72	31.17	28.87	18.14	11.90	23.45
	F3	9.33	28.41	12.15	8.26	19.62	12.65
/ow/	F1	16.06	76.56	24.15	13.23	10.81	22.83
	F2	26.64	41.29	31.59	15.24	25.46	27.60
	F3	13.40	28.29	14.97	10.88	20.64	14.49
/uh/	F1	16.17	77.76	24.67	14.14	11.46	22.14
	F2	26.35	40.40	32.34	15.67	24.47	26.61
	F3	13.88	28.21	15.12	10.61	20.88	13.60
/uw/	F1	16.08	81.07	24.32	13.82	10.74	23.16
	F2	26.79	40.34	32.41	15.74	23.79	27.00
	F3	14.09	29.35	15.31	10.96	21.40	13.80

vowels is presented in Table 4.4. In the case of male speakers, the results are presented in Tables 4.5 and for female speakers, in Table 4.5. It can be observed from the results that this method also exhibits significantly better performance in comparison to the LPC and AFB based methods. However, in the presence of severe noise, the DACF based method offers improved performance.

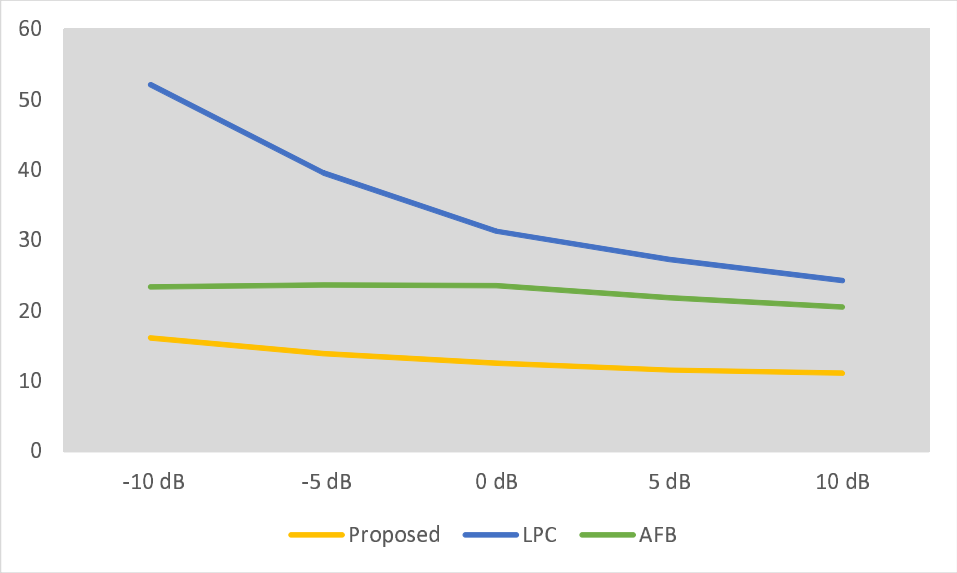


Figure 4.10: First formant estimation performance in terms of percentage error in formant estimation under various noise levels for male speakers

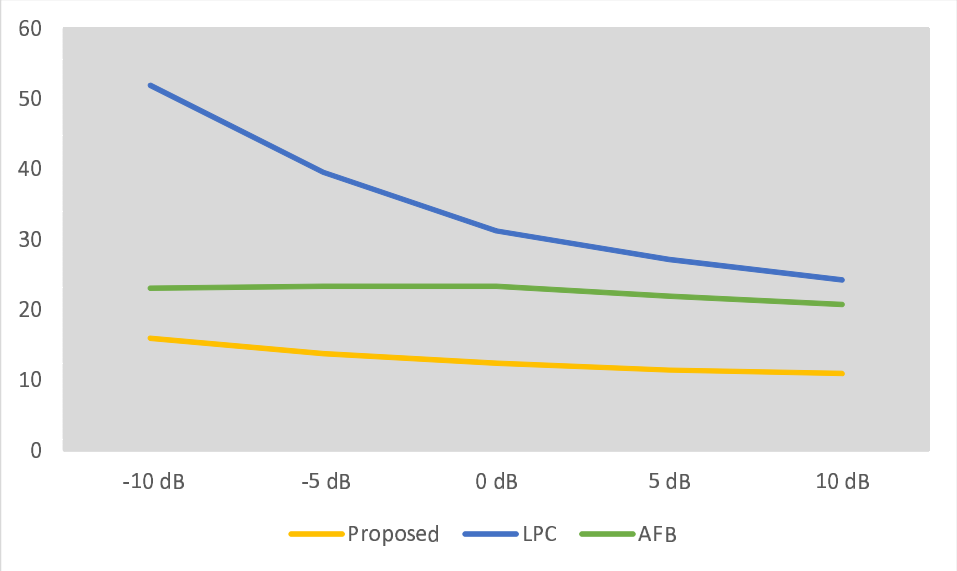


Figure 4.11: Second formant estimation performance in terms of percentage error in formant estimation under various noise levels for male speakers

In order to present the overall formant estimation errors over the entire range of

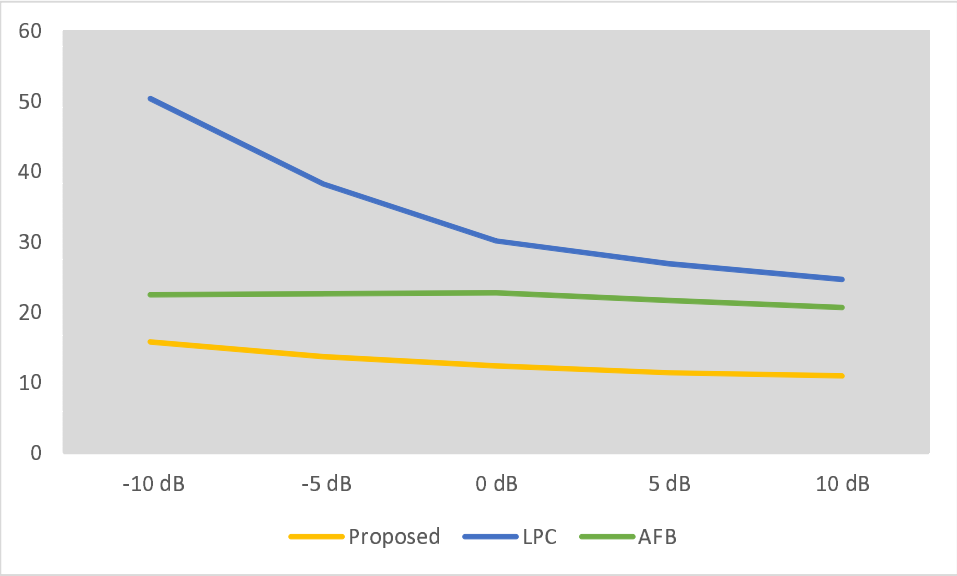


Figure 4.12: Third formant estimation performance in terms of percentage error in formant estimation under various noise levels for male speakers

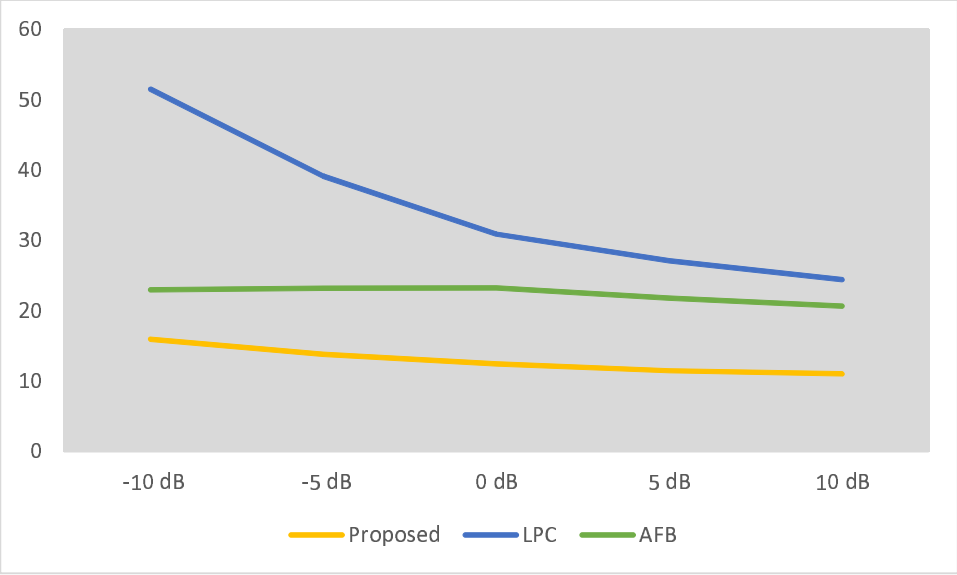


Figure 4.13: Estimation performance in terms of percentage error in formant estimation under various noise levels for male speakers

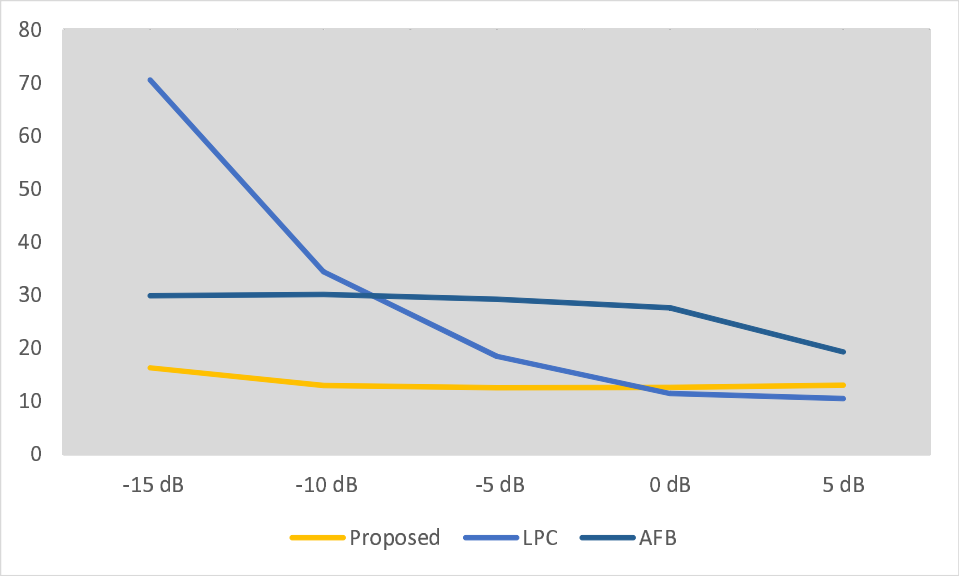


Figure 4.14: First formant estimation performance in terms of percentage error in formant estimation under various noise levels for female speakers

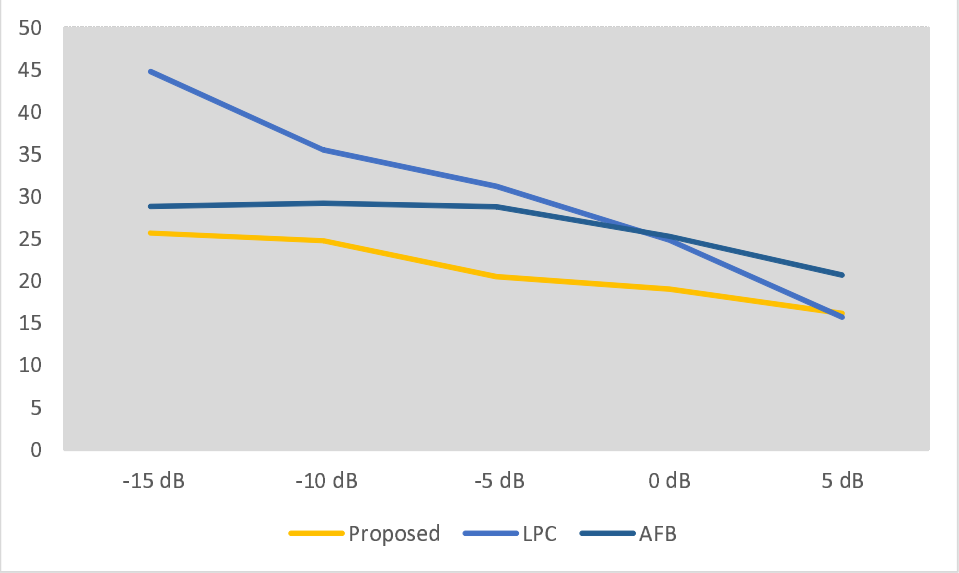


Figure 4.15: Second formant estimation performance in terms of percentage error in formant estimation under various noise levels for female speakers

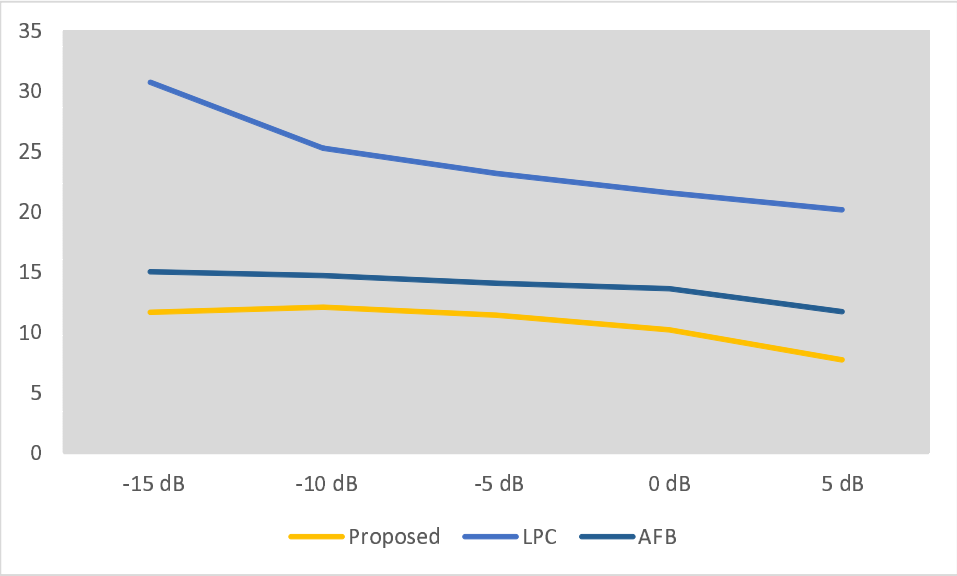


Figure 4.16: Third formant estimation performance in terms of percentage error in formant estimation under various noise levels for female speakers

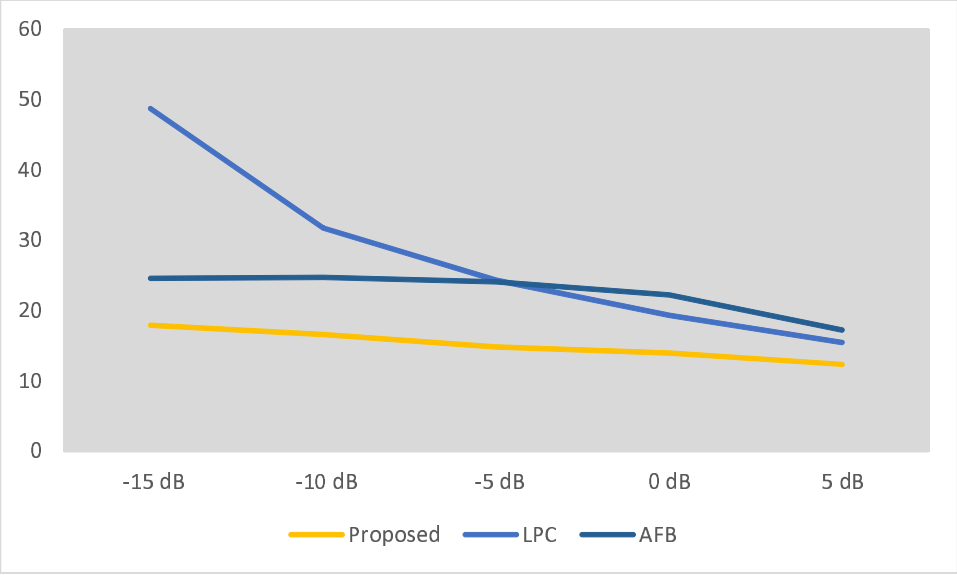


Figure 4.17: Estimation performance in terms of percentage error in formant estimation under various noise levels for female speakers

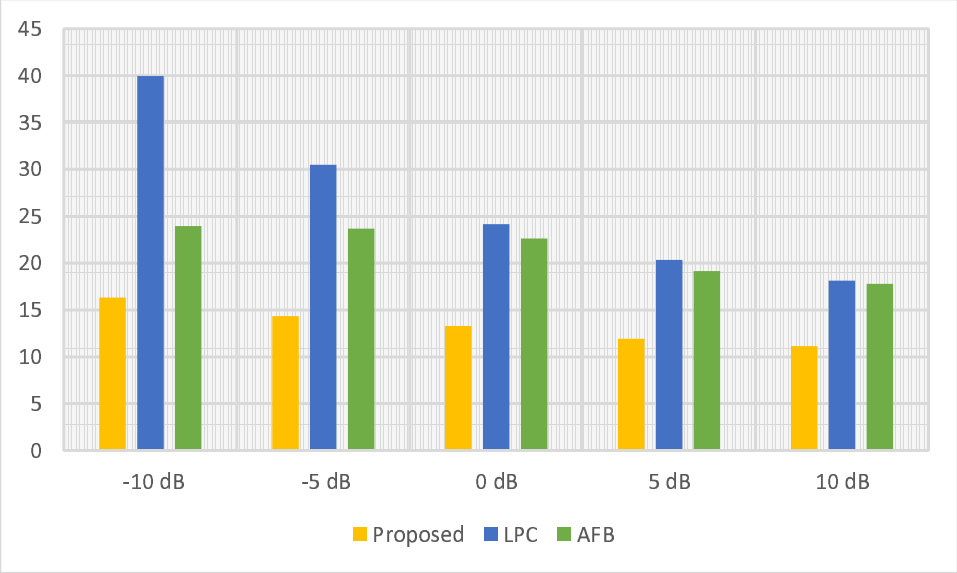


Figure 4.18: Estimation performance in terms of percentage error in formant estimation under various noise levels

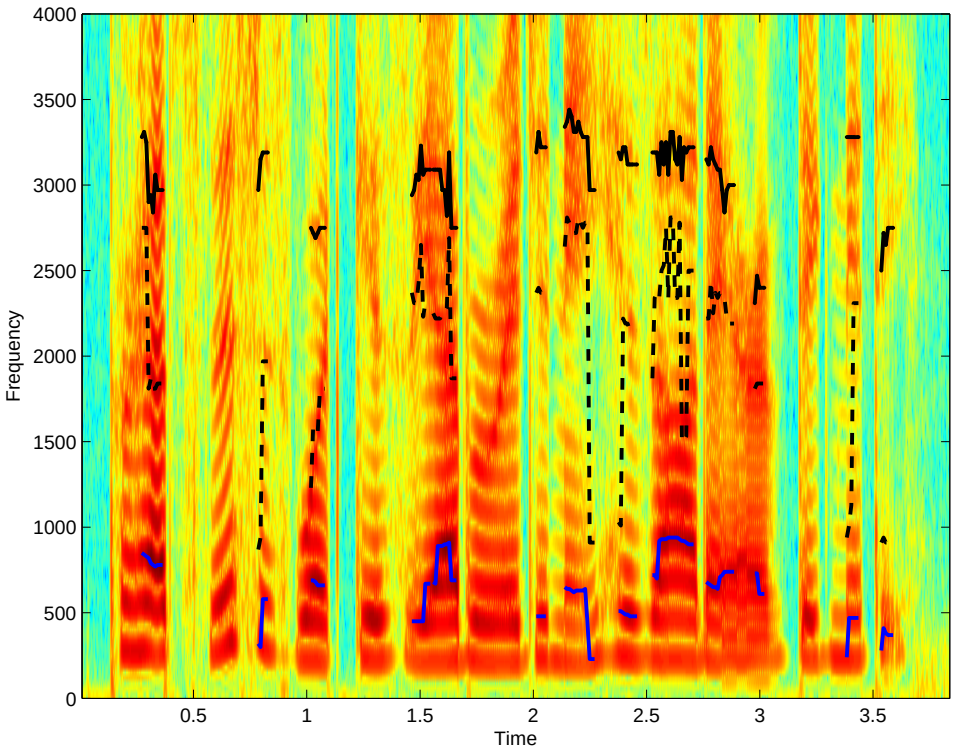


Figure 4.19: Spectrogram of the utterance 'Perhaps this is what gives the aborigine his odd air of dignity' , with formant frequencies estimated using the proposed method

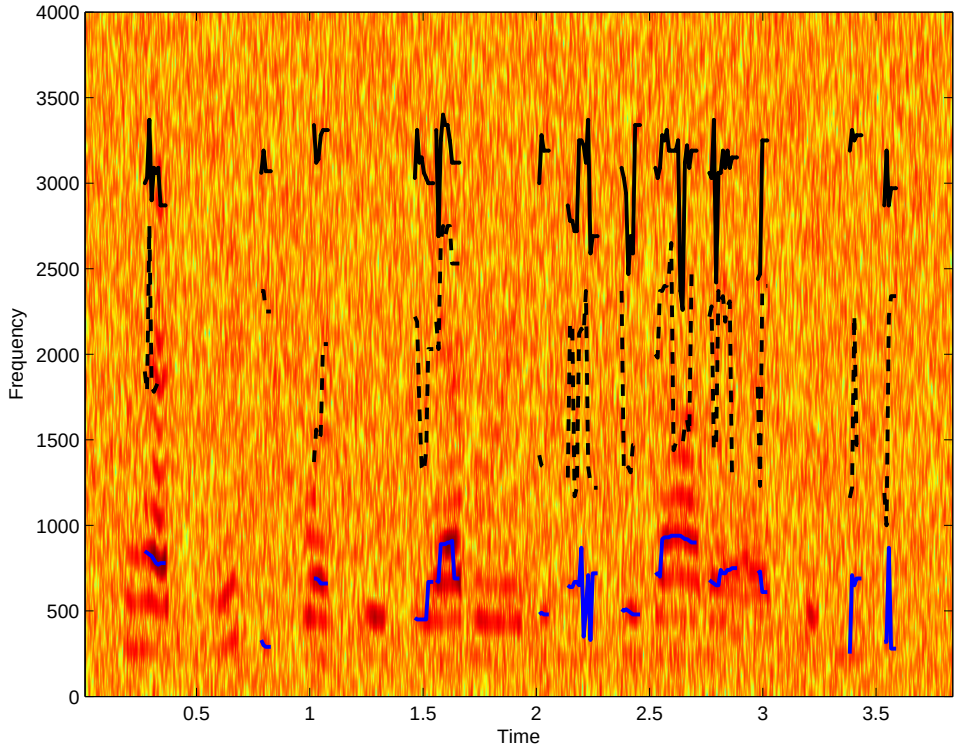


Figure 4.20: Spectrogram of the utterance 'Perhaps this is what gives the aborigine his odd air of dignity' , under $-5dB$ of background noise with formant frequencies estimated using the proposed method

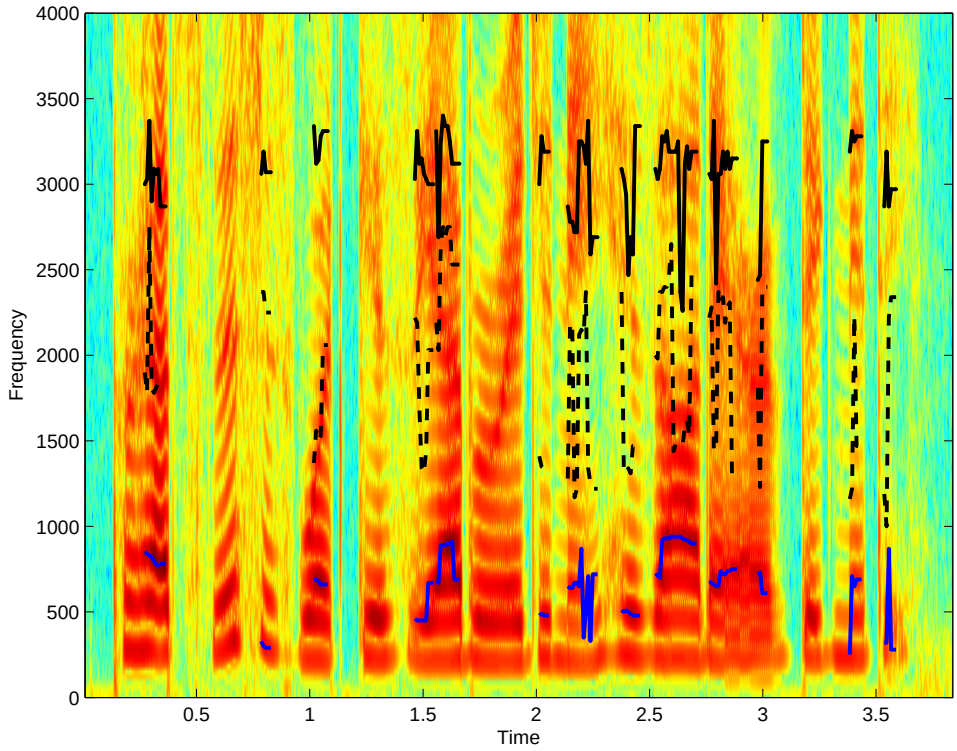


Figure 4.21: Spectrogram of the utterance 'Perhaps this is what gives the aborigine his odd air of dignity' , with formant frequencies estimated under $-5dB$ of Background noise using the proposed method

SNRs considered in the experimental setup, in Figs. 4.10, 4.11, 4.12 and 4.13, average of estimation error of all vowels for all three formants are shown for the the proposed method and the $LPC-12$ based method considering only male speakers. In this case, the SNR levels considered are ranging from -10 to $+10dB$. In a similar way, in Figs. 4.14, 4.15, 4.16 and 4.17, the average estimation error are shown for the female speakers for a SNR range of -15 to $+5dB$. Finally, in Fig. 4.18, the average estimation error considering both male and female speakers is 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 significantly 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 $-15dB$. 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.

Table 4.7: Vowel recognition accuracy

Feature Vector	$10dB$	$-5dB$
MFCC + Proposed Method	93.33	86.66
MFCC + LPC-12	93.33	81.66
MFCC + TIMIT reference	93.33	90.00
MFCC	93.33	81.66

Finally, the recognition accuracies for different vowels is presented in Table 4.7. It can be concluded from the figure 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 analysis, 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 'Perhaps this is what gives the aborigine his odd air of dignity' uttered by a female speaker taken from the TIMIT database is shown in Fig. 4.20. The formant frequencies estimated at different frames using the proposed method are shown over the spectrogram.

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. For the purpose of comparison, the same sentence, under influence of $-5dB$ background noise, is utilized to obtain the spectrogram present in Fig. 4.20. Here the presence of noise has completely obscured the energy bands, but still the proposed method can successfully track the formant frequencies. With the purpose of gaining a better insight, the formant frequencies obtained from the $-5dB$ noise corrupted speech are overlayed on the spectrogram for noise free speech, which is shown in Fig. 4.21. The resulting tracking lines obtained by the proposed method is a clear indication of its high level of consistency as well as the accuracy even in heavy noisy condition.

4.5 Conclusion

In this paper, a formant estimation scheme based on frequency domain modeling of repeatedly autocorrelated speech is presented. An adaptive band recognition system is deployed that can find out the band of successive formant frequencies for pre-processed voiced speech signals. The speech signal is then passed through an adaptive filter designed to separate the responses of different formants. Repeated autocorrelation, which strengthens the dominant poles, and exponentially increases the peak-valley ratio at formant frequencies of the magnitude response, canceling out the effects of noise, is then performed on the filtered speech signals. Formant estimation is carried out in the spectral domain where instead of direct peak-picking from the speech spectrum, a spectral domain model of ACF of speech signal is first proposed considering the resonances of the vocal tract. A spectral domain model fitting based algorithm is also developed to extract the model parameters which in turn give the formant. Even in the presence of a significant background noise, the developed method is effective in maintaining a high success rate in formant estimation.

Chapter 5

Conclusion

5.1 Contribution of the Thesis

- The main objective of this thesis work is to develop a formant frequency estimation method that can provide robust performance even in the presence of severe background noise. In order to achieve this target, one important property of the autocorrelation operation that it can strengthen the dominant formant peaks in the spectrum of speech, removing the effect of spurious noise peaks is utilized. The spectrum of the autocorrelation function offers better noise immunity in identifying the formant peaks. The reason behind is that the autocorrelation operation on a signal offers a similar advantage of increasing the poles in the transfer function generating the signal.

- First, formant estimation is carried out in the spectral domain where instead of direct peak-picking from the speech spectrum, a spectral domain model of autocorrelation function of speech signal is introduced. In this case the vocal tract is considered to be comprised of cascaded subsystems where each subsystem is responsible for single resonant frequency.

- For the purpose of finding out the formant frequencies, instead of employing an exhaustive search method, practical knowledge of formant ranges is utilized. An iterative spectral domain model fitting algorithm is developed to extract the model

parameters which in turn give the formants.

- Since the autocorrelation operation causes replication of original poles resulting in spectral peak strengthening, instead of single autocorrelation, the idea of double autocorrelation is introduced for further spectral peak enhancement, especially in severely noisy condition. An iterative spectral model fitting is employed for formant estimation.

- Although double autocorrelation works well in providing noise robust performance for the purpose of formant estimation, the significant enhancement of the first formant peak in some cases affects the estimation performance regarding other formants with relatively lower spectral energy. In order to overcome such a problem, a band limiting technique is presented that offers significantly improved estimation performance with regard to the iterative approach. As repeated autocorrelation significantly increases the strength of the most dominant peak, because of the band limiting operation each formant corresponding to the dominant peak of a band, thereby the estimation accuracies for the other formants are significantly improved.

- A filter bank based on the developed band limiting algorithm is employed to separate the speech signal into frequency bands containing single formants. After performing repeated autocorrelation on the signal, a spectral model of repeatedly autocorrelated filtered speech signal with only a single formant frequency is developed and model fitting is employed to find out the model parameters required for formant estimation.

- In order to demonstrate the estimation performance, a very large number of voiced utterances are used which are taken from the most widely used TIMIT naturally spoken continuous speech corpus. The experimental results demonstrate a far superior performance obtained by the proposed scheme in comparison to some of the existing methods at low levels of signal-to-noise ratio.

- As an application of the proposed formant estimation scheme, vowel recognition scheme is also developed based on linear discriminant analysis. It is observed that

because of the increased accuracy obtained by the proposed formant estimation scheme in comparison to some of the existing methods, a better vowel recognition accuracy is achieved in noisy environment.

5.2 Scope & Future Work

- One potential future work could be to develop a separate spectral domain noise subtraction block prior to formant estimation, which may further increase the formant estimation accuracy. Such a preprocessing is especially very useful in view of handling different types of environmental noises.

- The effect of noise whitening could also be investigated in case of dealing with different types of practical noises.

Bibliography

[1] D. O'Shaughnessy, *Speech Communications Human and Machine*, 2nd ed. NY: IEEE Press, 2000.

[2] D. Y. Wong, J. D. Markel, and A. H. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, no. 4, pp. 350–355, Aug. 1979.

[3] A. K. Krishnamurthy and D. G. Childers, "Two-channel speech analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, no. 4, pp. 730–743, 1986.

[4] D. G. Childers and C. K. Lee, "Voice quality factors: Analysis synthesis and perception," *J. Acoust. Soc. Amer.*, vol. 90, pp. 2394–2410, 1991.

[5] J. D. Markel and J. A. H. Gray, *Linear Prediction of Speech.* New York: Springer-Verlag, 1976.

[6] 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.

[7] I. Konvalinka and M. Matausek, "Simultaneous estimation of poles and zeros in speech analysis and itif-iterative inverse filtering algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, no. 5, pp. 485–492, Oct. 1979.

[8] 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.

[9] 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.

[10] S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 22, no. 2, pp. 135–141, Apr. 1974.

[11] M. Lee, J. V. Santen, B. Mobius, and J. Olive, "Formant tracking using context-dependent phonemic information," *IEEE Trans. Speech Audio Processing*, vol. 13, no. 5, pp. 741–750, Sept. 2005.

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

[13] L. Deng, A. Acero, and I. Bazzi, "Tracking vocal tract resonances using a quantized nonlinear function embedded in a temporal constraint," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 2, pp. 425–434, Mar. 2006.

[14] I. Bazzi, A. Acero, and L. Deng, "An expectation maximization approach for formant tracking using a parameter-free non-linear predictor," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, vol. 1, Apr. 2003, pp. 464–467.

[15] J. Darch, B. Milner, X. Shao, S. Vaseghi, and Q. Yan, "Predicting formant frequencies from mfcc vectors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, vol. 1, Mar. 2005, pp. 941–944.

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

[17] J. Malkin, X. Li, and J. Bilmes, "A graphical model for formant tracking," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, vol. 1, Philadelphia, PA, Mar. 2005, pp. 913–916.

[18] Y. Zheng and M. H. Johnson, "Formant tracking by mixture state particle filter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, vol. 1, Montreal, Canada, Mar. 2004, pp. 565–568.

[19] 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.

[20] Y. Shi and E. Chang, "Spectrogram-based formant tracking via particle filters," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, vol. 1, Hong Kong, Apr. 2003, pp. 168–171.

[21] Z. Yan and H. Zhao, "Formant estimation algorithm based on digital waveguide models," in *Information Engineering and Computer Science (ICIECS), 2010 2nd International Conference on.* IEEE, 2010, pp. 1–4.

[22] 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.

[23] T. V. Sreenivas and R. J. Niederjohn, "Zero-crossing based spectral analysis and svd spectral analysis for formant frequency estimation in noise," *IEEE Trans. Speech Audio Processing*, vol. 40, no. 2, pp. 282–293, Feb. 1992.

[24] I. C. Bruce, N. V. Karkhanis, E. D. Young, and M. B. Sachs, "Robust formant tracking in noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, vol. 1, May 2002, pp. 281–284.

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

[26] B. S. Atal, "A model of lpc excitation in terms of eigenvectors of the autocorrelation matrix of the impulse response of the lpc filter," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, vol. 1, May 1989, pp. 45–48.

[27] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 240–254, May 2000.

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] W. Tiffany, "Vowel recognition as a function of duration, frequency modulation and phonetic context," *Journal of speech and hearing disorders*, vol. 18, no. 3, p. 289, 1953.

[34] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," *J. Acoust. Soc. Am.*, vol. 24, no. 6, pp. 637–42, 1952.

[35] J. Suzuki and K. Nakata, "Recognition of japanese vowels preliminary to the recognition of speech," *J.Radio Res.Lab*, vol. 37, no. 8, pp. 193–212, 1961.

[36] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 23, no. 1, pp. 67–72, 1975.

[37] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, 1980.

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

[39] M. Naito, L. Deng, and Y. Sagisaka, "Speaker clustering for speech recognition using vocal tract parameters," *Speech Communication*, vol. 36, no. 3, pp. 305–315, 2002.

[40] X. Wang and K. Paliwal, "Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition," *Pattern recognition*, vol. 36, no. 10, pp. 2429–2439, 2003.

[41] R. Summers, P. Bailey, and B. Roberts, "Effects of the rate of formant-frequency variation on the grouping of formants in speech perception," *JARO-Journal of the Association for Research in Otolaryngology*, pp. 1–12, 2011.

[42] A. Vuppala, J. Yadav, S. Chakrabarti, and K. Rao, "Vowel onset point detection for low bit rate coded speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 6, pp. 1894–1903, 2012.

[43] A. Waibel, *Readings in speech recognition*. Morgan Kaufmann, 1990.

[44] A. Erell and M. Weintraub, "Energy conditioned spectral estimation for recognition of noisy speech," *Speech and Audio Processing, IEEE Transactions on*, vol. 1, no. 1, pp. 84–89, 1993.

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

[46] B. T. Meyer and B. Kollmeier, "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition," *Speech Communication*, vol. 53, no. 5, pp. 753–767, 2011.

[47] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals," *IEEE Trans. Multimedia*, vol. 10, no. 10, pp. 936–946, Aug. 2008.

[48] G. Muhammad, M. Alsulaiman, A. Mahmood, and Z. Ali, "Automatic voice disorder classification using vowel formants," in *Proc. IEEE Int. Conf. Multimedia and Expo (ICME)*, July 2011, pp. 1–6.

[49] S. M. Kay and S. L. Marple, "Spectrum analysis a modern perspective," *Proceedings of The IEEE*, vol. 69, pp. 1380–1419, 1981.

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

[51] 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.

[52] 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.

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