# *Algorithms for tracking formant frequencies of a continuous speech with speaker variability*

***A***

***Thesis***

*Submitted in the partial fulfil ment of requirement s for the award of the degree of*

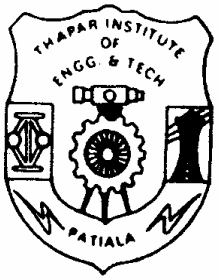
***Master of Engineering***
***In***
***Electronics and Communication Engineering***

*by*


***Poonam jindal***
***Regn. No. 80341 13***


*Under the guidance of*

***Mr. Balwant Singh***
*Lecturer*





**Department of Electronics and Communication Engineering**
**THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY**
**(DEEMED UNIVERSITY)**
**PATIALA (PUNJAB)-147004**
**INDIA**

# CERTIFICATE

I, Poonam Jindal, hereby certify that the work presented in this thesis entitled **"Algorithms For Tracking Formant Frequencies Of A Continuous Speech With Speaker Variability"** by me in partial fulfillment of requirements for the award of degree of **Master of Engineering (Electronics and Communication Engineering)** at **Thapar Institute of Engineering and Technology (Deemed University), Patiala,** is an authentic record of my own work carried under the supervision of Mr. Balwant Singh.

The matter presented in this thesis has not been submitted to any other University / Institute for the award of any degree.

(**Poonam Jindal**)
**Regn. No. 8034113**

This is certified that the above statement made by the candidate is correct to the best of my knowledge

**(Mr. Balwant Singh)**
Lecturer & Supervisor
Department of
Electronics &
Communication Engineering
T.I.E.T, Patiala

**(Dr. R.S Kaler)**
Prof. & Head
Department of Electronics & Communication Engineering
T.I.E.T, Patiala

**(Dr. D.S Bawa)**
Dean of Academic Affairs
T.I.E.T, Patiala

# ACKNOWLEDGEMENT

The real spirit of achieving a goal is through the way of excellence and austerous discipline. I would have never succeeded in completing my task without the cooperation encouragement and help provided to me by various personalities.

With deep sense of gratitude I express my sincere thanks to my esteemed and worthy supervisor, **Mr. Balwant Singh**, Lecturer, Department of Electronics and Communication Engineering for his valuable guidance in carrying out this work under his effective supervision, encouragement, enlightenment and cooperation. I am further indebted to **Dr. R.S Kaler**, Professor and Head, Department of Electronics and Communication Engineering, for his moral support at every step. I am also thankful to **Mr. R.K. Khanna,** the P.G coordinator, Department of Electronics and Communication Engineering for his full cooperation and help.

I take pride of my self in being the daughter of ideal great parents for their everlasting desire, affectionate blessings, support and help without which it would have not been possible for me to complete my studies.

The technical guidance and constant encouragement made it possible to tide over the numerous problems which so ever came up during the study. My greatest thanks to all who wished me success. Above all I render my gratitude to the ALMIGHTY who bestowed self-confidence, ability and strength in me to complete this work.

Place: T.I.E.T Patiala, INDIA (Poonam Jindal)

# ABSTRACT

Exposure to loud sounds can cause damage to the inner ear, leading to degradation of the neural response to speech and to formant frequencies in particular. This may result in decreased intelligibility of speech. An amplification scheme for hearing aids, called Contrast Enhanced Frequency Shaping (CEFS), may improve speech perception for ears with sound-induced hearing damage. CEFS takes into account across-frequency distortions introduced by the impaired ear and requires accurate and robust formant frequency estimates to allow dynamic, speech-spectrum-dependent amplification of speech in hearing aids. Several algorithms have been developed for extracting the formant information from speech signals, however most of these algorithms are either not robust in real-life noise environments or are not suitable for real-time implementation. Two algorithms are discussed in the present work. One is Robust formant tracking algorithm and other is Recursive least square algorithm (RLS). The first algorithm achieves formant extraction from continuous speech by using a time-varying adaptive filter bank to track and estimate individual formant frequencies. The formant tracker incorporates an adaptive voicing detector and a gender detector for robust formant extraction from continuous speech. And the second algorithm is based on recursive least square values. Forgetting factor approach is used for estimating formant frequencies with RLS algorithm. Algorithms are tested for both male and female speakers in the presence of background noise. Thorough testing of the algorithms using various speech sentences has shown promising results over a wide range of signal to noise ratio's (SNR's) for various types of background noises, such as additive white gaussian noise, single and multiple competing background speakers and various other environmental sounds. Results from both algorithms showed that the robust formant tracking algorithm gives very good tracking performance in every environment and at every SNR. RLS algorithm also provide good estimate of formant frequencies in every environment but the estimate is not smooth and tracking of

formant frequencies is very noisy. By observing the limitations of the traditional formant tracking algorithms and the present RLS algorithm it can be seen that the robust formant tracking algorithm is the most accurate algorithm and fulfill all the requirements for accurate formant tracking.

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| CEFS | Contrast Enhanced Frequency Shaping |
| AWGN | Additive White Gaussian noise |
| AZF | All Zero Filter |
| DTF | Dynamic Tracking Filter |
| RLS | Recursive Least Square |
| LMS | Least Mean Square |
| LPC | Linear Predictive Coding |
| SNR | Signal to Noise Ratio |
| RMS | Root Mean Square |
| TIMIT | Joint effort among the Texas Instruments (TI) and Massachusetts Institute of Technology (MIT) for designing a speech data. |

CONTENTS

# Chapter 1

# INTRODUCTION

## 1.1 Introduction

Formants are the resonant frequencies of the vocal tract when vowels are pronounced. Formants can be found where there are large concentrations or peaks of energy in the spectrogram reading of a voiced sample [1]. In order to implement Contrast Enhanced Frequency Shaping (CEFS) [2] amplification in hearing aids for continuous speech, the second formant frequency ($F_2$) needs to be accurately estimated for voiced speech [3] [4] [5]. Accurate formant estimation for continuous speech (in real time noise environments) is a challenge because formant frequencies are not simple to track in such a dynamic environment. The formant estimation algorithm needs to be robust and be able to operate in a wide range of real-time noise scenarios. It must also be able to recover quickly if it encounters any problem and after periods of silence. For estimating formant frequencies two algorithms have been presented. One is robust formant tracking algorithm and another is using RLS algorithm.

## 1.2 Traditional Formant Estimation Techniques

Development of accurate formant estimation algorithms began in the 1950s. Since then numerous techniques have been proposed for formant analysis. Most of the work can be classified as frequency domain techniques (such as picking peaks in the short-time frequency spectrum), parametric techniques (also called “analysis by synthesis”) [6] [7] in which one generates a best match to the incoming signal based on a model of speech production. The traditional approaches to formant frequency estimation are misled by spectral peaks in unvoiced speech and perform very poorly in transient background noise. Also, these traditional algorithms are not robust and are unable to recover quickly after periods of silence. These problems limit the possible use of the traditional techniques for estimation of the second formant frequency ($F_2$) for CEFS amplification.

## 1.3 Formant Tracking Algorithms

The traditional formant tracking algorithms do not track formants accurately. Another formant tracking algorithms whose performance is good as compare to the above mentioned algorithms are: Robust formant tracking algorithm [8], RLS algorithm [9]. These algorithms represent the best known formant analysis techniques and have been implemented in MATLAB in order to test and compare their performance under different conditions. Brief introduction to each of these formant estimation techniques is presented below.

### 1.3.1 Robust Formant Tracking Algorithm

The robust formant tracking algorithm discussed in the present work is the most accurate formant tracking algorithm. This algorithm is robust and accurate in continuous speech and mitigates the effects of speaker variability and different background noises. This allows the algorithm to operate independently and provide reliable formant frequency estimates for contrast enhanced frequency shaping (CEFS) amplification [4] and other applications. Figure 1.1 shows a block diagram of the Robust Formant Tracker [8].

The speech signal is first pre-emphasized using a high-pass filter to equalize the energy and remove the spectral tilt of the speech signal. An approximate, analytic version of the signal is then calculated to increase spectral accuracy for the formant estimates through an approximate Hilbert transformer.

The analytic signal is then filtered into four different bands using a bank of adaptive band-pass filters (called Formant Filters). Each of the four formant filters ($F_1$, $F_2$, $F_3$ and $F_4$) in the filter bank is made up of an All- Zero Filter (AZF) and a Dynamic Tracking Filter (DTF) [10]. The zeros of each of the AZF's are set to the latest estimate of the formant frequencies from the other three bands. The DTF provides the single pole located at the latest estimate of the formant frequency for that band. This cascade arrangement results in each of the filters having a pole around its own formant frequency and zeros at the other formant frequency locations.

Each of the four band-pass filters allows only the signal around the frequency region of the desired formant to pass through and suppresses the other frequency regions. The formant filter bank has a fundamental modification that the $F_1$ filter of the filter bank has

an added zero at the pitch frequency ($F_0$) for further suppression of the region below the $F_1$ frequency (the pitch region). This decreases the effects of the pitch on the $F_1$ estimate.

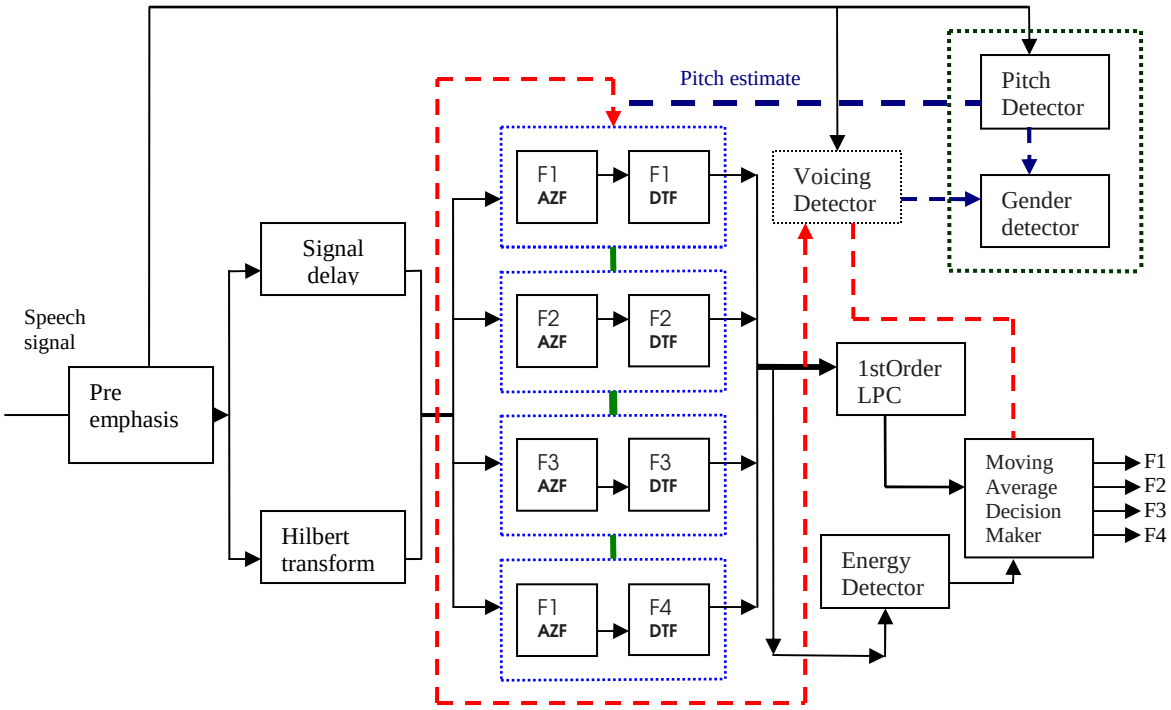



**Fig1.1 Robust Formant Tracker**

A first-order Linear Prediction Coefficient (LPC) is then calculated for the analytic signal in each of the four bands [11]. From each of these coefficients a formant frequency estimate is obtained. As the value of the four formant frequencies vary with time, the formant pre-filters are modified to track them by changing their pole and zero locations. Due to the band-pass pre-filtering of each formant frequency region prior to LPC, the frequency estimates provided by LPC are more accurate and the algorithm is less susceptible to errors due to background noise.

The formant estimation is further refined by adding an adaptive voicing detector to detect a voiced and unvoiced speech segments. LPC estimates for the formant frequencies are only used during the voiced segments of speech [12]. During the unvoiced speech segments or when the signal energy of a particular formant frequency region (determined by the adaptive energy detector) is below a set threshold level [13], the formant

frequency estimates are assigned their moving average value. This approach ensures that the formant tracker is able to recover quickly and with minimum error to the formant estimates, after unvoiced or low-energy speech segments.

The energy detector threshold levels [8] are also made adaptive for each of the formant filters so that they can adjust to long term changes in the energy levels of each formant frequency region. The voicing detector calculates the log ratio between the energy in the lower and higher frequencies of the speech to determine if a speech segment is voiced or unvoiced. If there is more energy in the lower frequencies than the higher ones, the speech segment is classified as being voiced.

The voicing detector also has a threshold with hysteresis [8] to ensure that switching from voiced to unvoiced speech (or vice versa) does not erroneously occur too quickly. Finally, an autocorrelation-based energy test is performed to ensure that voicing is not detected erroneously when there is no actual voicing in the speech but sufficient energy is present in the lower frequencies due to 'colored Gaussian noise' (or other background noises) . The voicing detector provides a sample by sample decision on whether a segment is voiced or unvoiced.

In order for the voicing detector to work properly for both male and female speakers, various parameters of the voicing detector need to be modified. The main purpose of the gender detector is to determine the gender of the speaker and pass this information to the voicing detector so that it is able to modify its parameters. The gender detector uses a pitch [14] based method to classify the gender of the speaker where the pitch is calculated using an autocorrelation based method. The gender detector also provides the pitch estimate to the first formant filter so that an additional zero can be added at the location of the pitch in the AZF of the first formant filter.

Extensive testing of the robust formant tracking algorithm has been done which showed that the formant tracking algorithm is robust to a wide variety of real-time background noise conditions. The algorithm is able to provide reliable formant frequency estimates from continuous speech for both male and female speakers. It recovers quickly and with minimal error when problems do occur and when there is a switch in speakers.

### 1.3.2 Recursive Least-Squares (RLS) algorithm

The formants can be estimated through an adaptive algorithm known as RLS [9] algorithms. Least-square algorithms aim at the minimization of the sum of the squares of the difference between the desired signal and the model filter output. When new samples of the incoming signals are received at each iteration, the solution for the least-squares problem can be computed in recursive form resulting in the recursive least-squares (RLS) algorithms.

An important feature of the RLS algorithm is that it utilizes information contained in the input data, extending back to the instant of time when the algorithm is initiated. This improvement in performance, however, is achieved at the expense of a large increase in computational complexity and some stability problems. Its performance is good as compared to the LMS algorithm [15] [16]. RLS algorithm gives good formant tracking but its performance is not as good as with robust formant tracking algorithm. Although RLS algorithm achieves good formant estimation but due to its noisy tracking response CEFS amplification cannot be achieved.

## 1.4 Applications of Formant Tracking Algorithms

Formant frequencies play a major role in vowel identification and are also important for consonant identification. Formant tracking algorithms estimates the formant frequencies accurately. Accurate formant frequency estimates can be used for a variety of applications.

1.These algorithms provides contrast enhanced frequency shaping amplification for hearing aids.

2. Formant frequencies have been used to make natural sounding computer synthesized speech.

3. Formant frequency estimates can be used for speech recognition.

4. Formant estimates can be used in speech coding.

5. The formant tracking algorithm can also be used for concatenation synthesis of speech

## 1.5 Layout of Thesis

Chapter 1 describes the brief introduction about the formant tracking algorithms and also gives the applications of these algorithms.

Chapter 2 gives the brief discussion about the anatomy of speech production followed by a detailed discussion on what formant frequencies are and their importance in speech perception. Also included in chapter 2 is a detailed look at the formant frequency characteristics in different types of sounds.

Chapter 3 describes the robust formant tracking algorithm and RLS algorithms in detail.

Chapter 4 describes the details of the test cases for which the algorithms was tested and present the results and observations obtained from the robust formant tracking algorithm and RLS algorithm.

Chapter 5 describes the conclusions and future scope of the formant tracking algorithms.

**Chapter 2**

# ACOUSTIC THEORY OF SPEECH PRODUCTION

## 2.1 Introduction

"Speech" refers to the transmission of language orally [17]. It is a process of making definite vocal sounds those form words to express thoughts and ideas. Speech sounds are air pressure vibrations produced by air exhaled from the lungs and modulated and shaped by the vibrations of the glottal cords and the vocal tract as it is pushed out through the lips and nose. [18]

## 2.2. Anatomy of Speech Production

An outline of the anatomy of the human speech production system is shown in Figure 2.1. It consists of the lungs, larynx, vocal tract, lips, nose and the connecting tubes [19]. The combined voice production mechanism produces the variety of vibrations and the spectral-temporal compositions that form different speech sounds. [20]

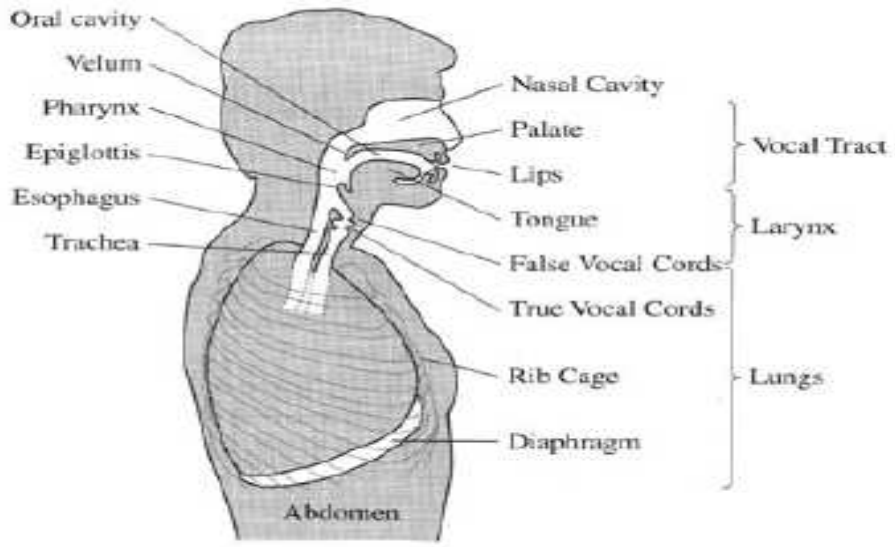



**Fig.2.1 Cross-sectional view of the anatomy of speech production**

The lungs provide the airflow needed for speech production to the larynx [17]. The larynx modulates the continuous airflow from the lungs into either a periodic or noise-like airflow and then passes it into the vocal tract [17]. The vocal tract is made up of the oral, nasal and pharyngeal cavities and provides spectral shaping to the modulated airflow (periodic or noisy) from the larynx. Sound sources can also be produced within the vocal tract itself by constrictions and relaxations creating an impulsive airflow. Following the spectral provided by the vocal tract to all three sound sources, the lips vary the air pressure of the airflow resulting in traveling sound waves that are perceived as speech. This description provides an idealized model of the anatomy of speech production however, in reality the sound sources required to produce most sounds are not ideal (periodic, noisy or impulsive) but usually a mixture of these types and change with the environment.

### 2.2.1. The Lungs

The lungs normally inhale and exhale air in a rhythmic manner for respiration. However, during speaking the lungs override this rhythmic pattern of inhalation and exhalation of air to exhale air more slowly. Usually the period of exhalation roughly coincides with the length of the sentence being spoken. A steady and slow contraction of the rib cage provides a timed exhalation from the lungs during which the air pressure in the lungs is maintained to be roughly constant. This allows the lungs to provide a steady airflow to the larynx for the entire duration of the sentence being spoken. Even though the sound source provides a steady airflow to the larynx, the properties of the larynx and the vocal tract allow the pressure of the airflow being produced to vary.

### 2.2.2. The Larynx

The main purpose of the larynx in the speech production system is to control the vocal folds. The vocal folds are a mass of flesh, ligament and muscle that stretch between the back and the front of the larynx. The glottis is a slit-like opening between the two vocal folds and the size of this slit can be varied. The tension in the vocal folds can also be varied by the muscle and cartilage around them.

The vocal folds (along with the epiglottis) close during eating to prevent food from entering the larynx. The vocal folds have three main states: breathing, voiced and unvoiced. During the breathing state the vocal folds are wide open and the muscles within them are relaxed to allow the air from the lungs to flow through freely. During speaking (for both voiced and unvoiced states) an obstruction to the airflow is provided by the vocal folds. In the voicing state the vocal folds tense up and are brought close together partially closing the glottis leading to self-sustained oscillations as air passes through the glottis. The contraction of the lungs results in air flowing through the glottis. As the airflow velocity increases the local pressure in the glottis decreases and the tension in the vocal folds increases. These two factors lead to an abrupt closing of the glottis. This is followed by an air pressure build-up behind the vocal folds causing them to open slowly and allowing air to flow through. The process is then repeated again resulting in the periodic release of puffs of air into the vocal tract. The time period during which the vocal folds are closed is called the ‘closed phase’. The time period during which there is some airflow before the maxima is reached is called the ‘open phase’. The time between the airflow maxima and the total closure of the vocal folds is called the ‘return phase’. The time duration for one such complete cycle is called the pitch period and its reciprocal is called the pitch frequency or just the pitch (or fundamental frequency). The pitch ranges from about 60 Hz to 400 Hz for most phonemes depending on various factors including the gender of the speaker. Adult males typically have lower pitch than adult females because their vocal cords are longer and larger.

The Fourier transform of the periodic glottal waveform is characterized by harmonics and the spectral envelope of the harmonics has an approximately -12 dB/octave roll off. The exact value of the roll off depends on the speaker and can change slightly. The lower frequencies of the speech spectrum contain more energy than the upper frequencies [18]. This is because the lower frequencies contain the glottal pulsing and radiation from the lips. This causes the speech to have a spectral tilt, which needs to be compensated prior to speech processing, to allow equalization of the energy distribution in the speech spectrum and obtain better spectral estimation in the higher frequency regions.

During the unvoiced state, the shape of the vocal folds is similar to that in the breathing state and there is no vocal fold vibration. However, the vocal folds are closer together

during un-voicing than in the breathing state and this leads to some amount of turbulence being caused at the folds, as the air passes through. This turbulence is called aspiration and sounds produced through aspiration are sometimes called ‘whispered’ sounds because this turbulence is also created during whispering (without any oscillations of the vocal folds). Aspiration can also occur with voicing leading to a ‘breathy’ voice. The vocal folds can also move in a form that does not fall clearly in any of the three states defined above. These includes a ‘creaky’ voiced state where the vocal folds are tense but only a short portion of the vocal folds are actually in oscillations, resulting in a harsh sounding voice with a very high and irregular pitch.

### 2.2.3. The Vocal Tract

The vocal tract is made up of the oral cavity from the larynx to the lips and the nasal passage coupled with the oral passage (through the vellum). The oral tract can take on various different configurations depending upon the shape and movement of the tongue, mouth, teeth, lips, jaw, etc [17]. The vocal tract provides frequency shaping to the output from the larynx and also generates new sources for sound production (impulsive source). Under certain circumstances, the vocal tract can be modeled as a linear filter with resonances.

The resonance frequencies of the vocal tract are called formant frequencies or just formants. The formant frequencies change with different vocal tract configurations [18]. The peaks of the vocal tract response correspond roughly to its formant frequencies. If the vocal tract is modeled as a time-invariant, all-pole linear system, then each of the conjugate pole pairs corresponds to a formant frequency (resonance frequency). Generally, as the length of the vocal tract increases the formant frequencies decrease, so the formant frequencies of adult males are somewhat lower than those of adult females, for the same sound.

When using the time-invariant, all-pole linear system model of the vocal tract, the speech waveform can be obtained through the convolution of the glottal flow with the vocal tract impulse response. It is important to discriminate between the formant frequency and the harmonic frequency. Formant frequencies correspond to the vocal tract frequency response poles while harmonic frequencies arise from the periodicity of the glottal

source. When the vocal tract vellum is lowered, the nasal passage is introduced into the vocal tract and the oral tract closes resulting in the acoustic waves propagating though the nasal cavity, this produces ‘nasal’ sound such as ‘m’. These sounds are often dominated by the lower frequency formants due to the large volume of the nasal cavity. When the vellum is lowered while keeping the oral cavity open the resulting sounds are referred to as ‘nasalized speech’. The effect of the nasal passage on the vocal tract is to broaden the formant bandwidths (due to greater loss of energy in the nasal passage) and to introduce anti-resonances (zeros) into the vocal tract system model due to coupled resonances.

It should be noted that the time-invariant vocal tract model can only be applied when the vocal tract configuration is steady and constant. As mentioned earlier, the vocal tract changes its shape with time so the time-invariant model can only be applied over short time periods or for sounds with a long, repetitive duration, such as sustained vowels, with a temporal windowing heuristic.

### 2.2.4. Voiced and Unvoiced Speech

The broadest way to categorize sounds is by the source to the vocal tract that produces the sound. As described earlier there are three main sound sources: periodic, noisy and impulsive. In a broad sense, sounds produced due to a periodic glottal source are called voiced sounds, and sounds produced otherwise are called unvoiced sounds. Generally, voiced speech [20] has more low-frequency energy and is quasi-periodic (such as steady state vowels) requiring vibration of the vocal cords.

On the other hand, unvoiced speech [20] has more high-frequency energy, is noisy in nature and does not require the vibration of the vocal cords. Figure 2.2 shows an example of a typical waveform for a voiced speech segment displaying the lower frequency and quasi-periodic characteristics of voiced speech sounds.

The waveform is for the vowel sound /a/ (as in ‘father’). There are a variety of unvoiced sounds. Those that are created due to a noise source at an oral constriction are called fricatives because the noise is created by friction of the air moving against the constriction. The sound of ‘th’ in the word ‘thin’ is a fricative with the friction being provided between the tongue and the upper teeth. The waveform for ‘th’ is shown in Figure 2.3 and shows the typical higher frequency and noise-like characteristics of

unvoiced speech. Another unvoiced sound class is the plosives (such as‘t’ and ‘p’) created with an impulsive airflow from the vocal tract as the sound source. When the barrier to the airflow is provided by partially closed vocal folds a new class of unvoiced sounds is produced called whispers (such as ‘h’). Sometimes the sound source is from a combination of voiced and unvoiced sources such as in the case of the sound ‘z’ where there is friction as well as simultaneous voicing; this class of sounds is therefore called voiced fricatives. Similar in concept are voiced plosives which occur due to simultaneous impulsive and voiced sources as in the sound ‘b’.

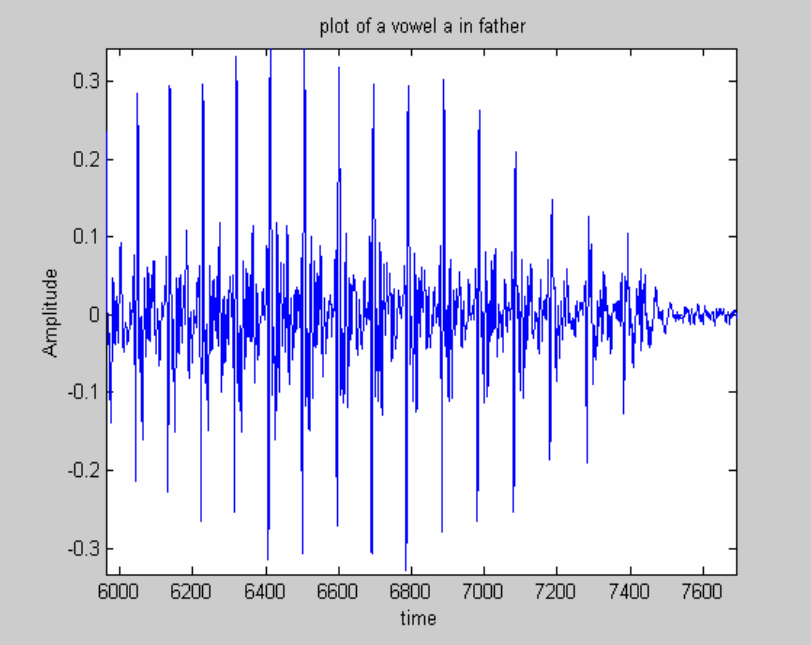



**Fig 2.2– Waveform of a Voiced Speech segment (for /a/ as in ‘father’)**

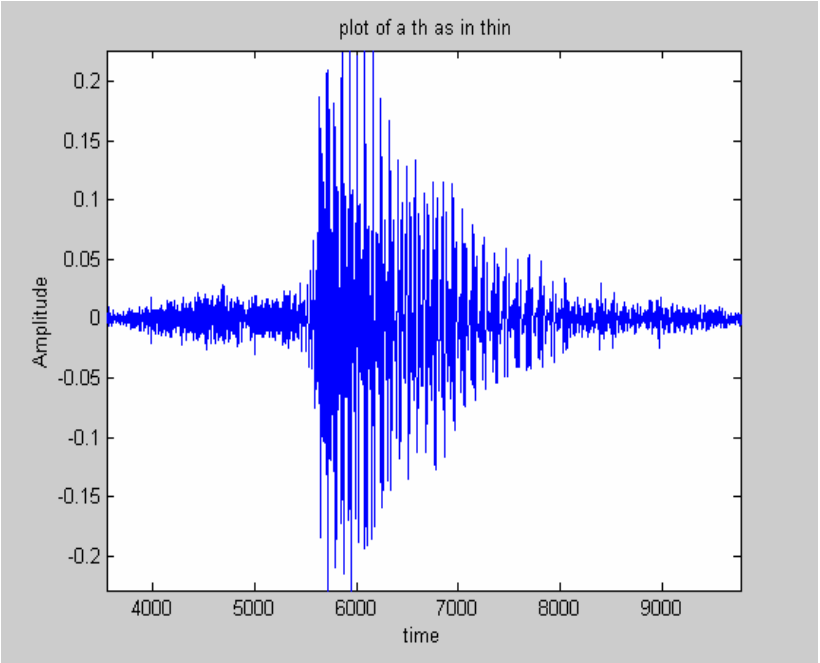



**Fig 2.3– Waveform of a fricative sound (for /th/ as in 'thin')**

### 2.3. Formant Frequencies

As mentioned earlier, the resonance frequencies of the vocal tract are called the formant frequencies or formants. In this section, the origins and characteristics of formants are explored further in terms of their behavior in different types of sounds and the problems that are caused in extracting the formants from these sounds. [1]

#### 2.3.1. Vocal Tract Filtering and Formant Frequencies

Figure 2.4 shows a complete discrete-time speech production model for periodic, noisy and plosive speech [17]. G (z) is the Z-transform of the glottal flow input, Rg(z) is the radiation impedance modeled by a single zero and V(z) is the stable all-pole vocal tract transfer function. $A_v$, $A_n$, and Ai are the gains that controls the loudness of the sound for periodic, noisy and plosive sources respectively. $R_l$ (z), in H (z), models the radiation impedance of the lips.

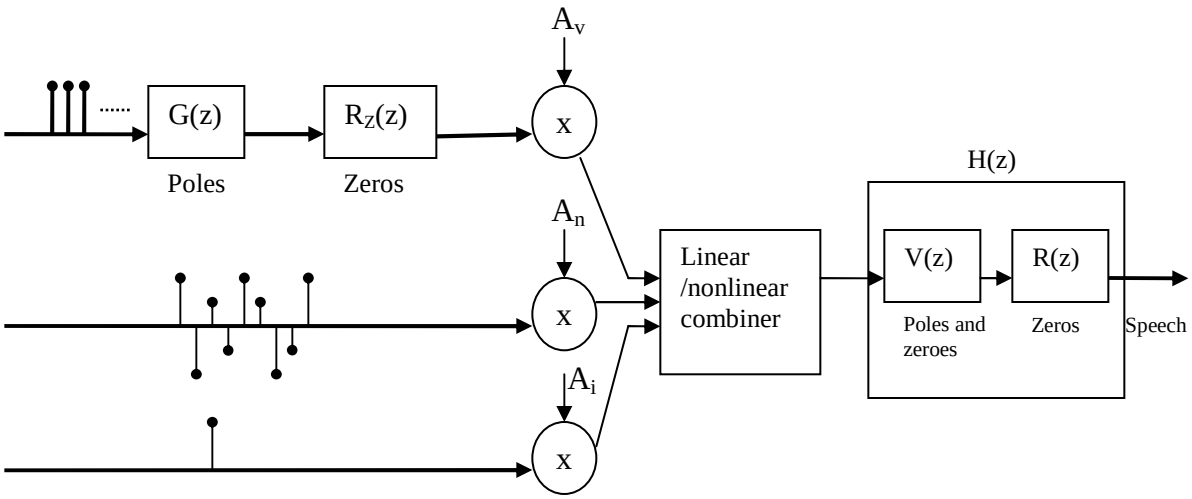



**Figure 2.4– Discrete-Time Model of Speech Production**

The vocal tract transfer function V (z) varies with the type of sound produced and also depends on the speakers and their speaking style. The formant frequencies vary with different vocal tract configurations and therefore, formant frequencies vary in speech with time as the vocal tract changes its shape.

The peaks of the vocal tract response in each configuration correspond roughly to its formant frequencies. The first resonance of the vocal tract is called the first formant frequency (or $F_1$), the second resonance of the vocal tract is called the second formant frequency (or $F_2$), and so on. For perfectly voiced and periodic speech (as in sustained vowels) the vocal tract can be accurately modeled by the stable all-pole model for V (z). However, in order to model other types of sounds, zeros are also added to V (z) in order to model the nasal cavity of the vocal tract. The resonances or peaks of the vocal tract transfer function (poles of the V (z) transfer function) correspond roughly to the formant frequencies of a particular sound. The characteristics and behavior of formant frequencies change in different types of sound and estimating formants in continuous speech is a challenging task.

### 2.3.2. Phonemic Classification of Speech and Formant Behavior in Phonemes

In this section, the behavior and characteristics of formant frequencies in different types of sounds are explored in greater detail to understand the problems associated with

estimating formant frequencies in such cases. In general, formant frequency regions have more energy than the other frequencies in the speech spectrum and can easily be visually identified in spectrograms. Phonemes are the fundamental distinctive units of sound. Each distinct and identifiable sound in a language forms a phoneme. Table 2.1 shows all the different phonemes in American English grouped together by their phoneme class. Formant frequencies for each phoneme vary and will also depend on the speakers and their individual speaking style. However, formant frequencies for a particular phoneme class (for a particular speaker) have similar characteristics and behavior. The formant frequency behavior and characteristics of some of these phoneme classes are discussed below.

**PHONEMES**

| Vowels | | affricatives | Diphthongs | Semi vowels | | Consonants | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Center | Back | | | liquids | glides | nasals | plosives | | fricatives | |
| | | | | | | | voiced | unvoiced | voiced | unvoiced |
| R | a | tS | Y | r | w | m | b | p | v | f |
| A | c | J | W | l | y | n | d | t | D | T |
| | o | | O | | | G | g | k | z | s |
| | U | | JU | | | | | | Z | S |
| | u | | | | | | | | | |

**Table2.1 Classification of phonemes**

### 2.3.2.1 Vowel

Vowels make up the largest class of phonemes [21]. The three types of vowels are grouped according to the tongue hump position required to make the sound (front, central or back). Vowel sounds are produced by quasi-periodic airflow through the glottis that vibrates the vocal folds at a certain fundamental frequency [22]. The nasal tract remains closed in vowel sound production so the vocal tract does not contain the effects of the nasal cavity. The lips can contribute to the vocal tract configuration through their degree of opening and rounding. The position of the tongue (front, centre or back) determines the phoneme produced, e.g. /a/ (‘father’) and /i/ (‘eve’) are differentiated primarily through the position of the tongue hump.

Figures 2-5 and 2-6 show the waveform and spectrogram of the vowel sound /i/ (‘eve’). The quasi-periodic nature of the vowels can be seen from its waveform. From the spectrogram it can be observed that the formant frequency regions have concentrated energy. These features are common to all vowels. The strong energy of the formant regions and the periodic nature of the waveform make it relatively easy to extract formant frequencies of pure and sustained vowel-like sounds. Despite general similarities between different vowel sounds, it is important to remember that the exact formant frequencies of different vowels differ from each other and depend on a wide variety of parameters including the speakers and their speaking style.

#### 2.3.2.2 Fricatives

There are two main types of fricative consonants [21]: unvoiced and voiced. Unvoiced fricatives are generated through turbulence in the airflow being provided at some point in the oral tract without any vocal fold vibration e.g. /f/ (‘father’). The constrictions provided through the hump in the tongue, lips, teeth, etc., help separate the rear and front oral cavity regions. The primary source of spectral shaping is the front of the oral cavity however; anti-resonances that are provided by the rear of the oral cavity also have an effect on the overall spectral shaping provided by the vocal tract. The transfer function of the vocal tract is made up of primarily higher frequency resonances that vary with the location of the vocal tract constrictions. Figures 2-7 and 2-8 show the waveform and the spectrogram of the unvoiced fricative /f/. From these figures it can be seen that unvoiced fricatives have ‘noisy’ waveforms as expected.

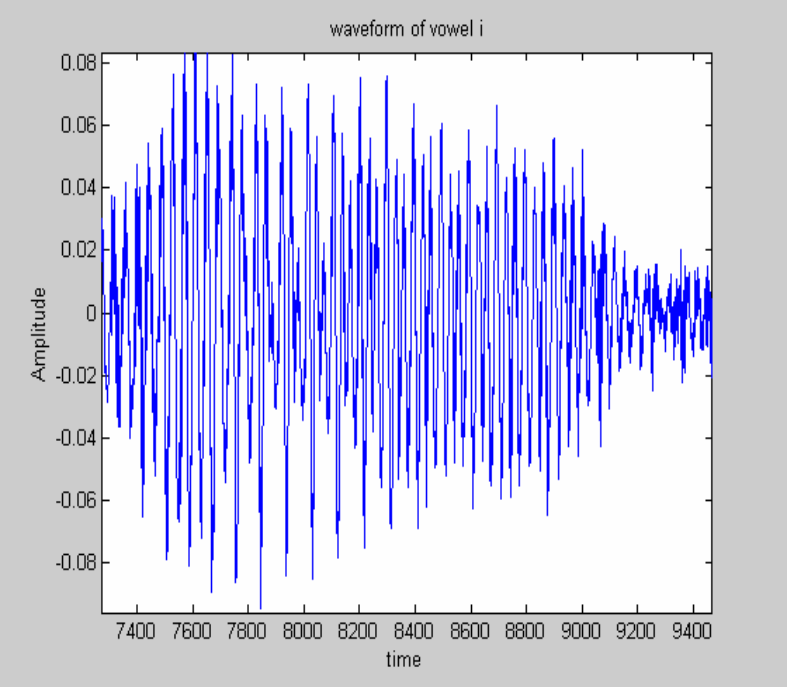



**Fig 2.5– Waveform of vowel /i/ (‘eve’)**

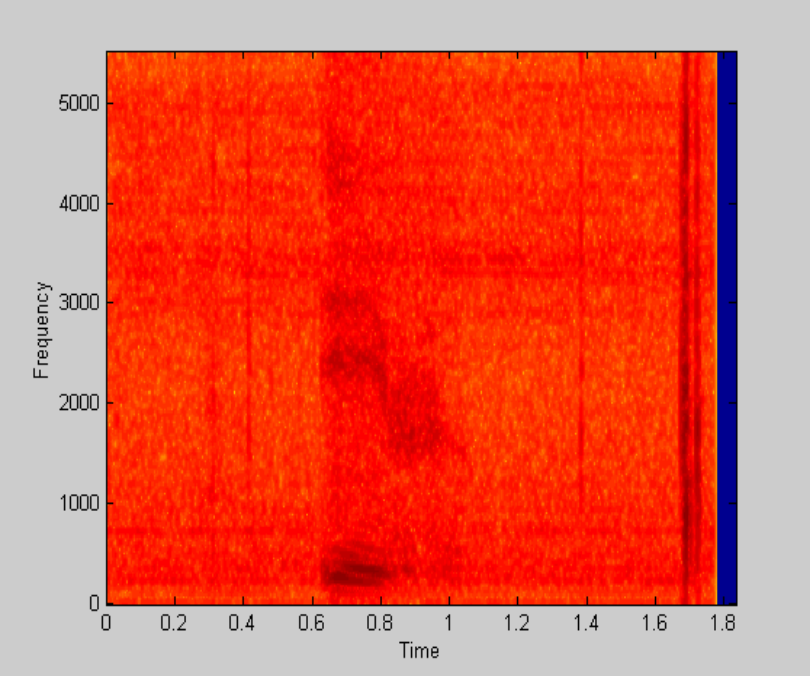



**Fig 2 .6 – Spectrogram of vowel /i/ (‘eve’)**

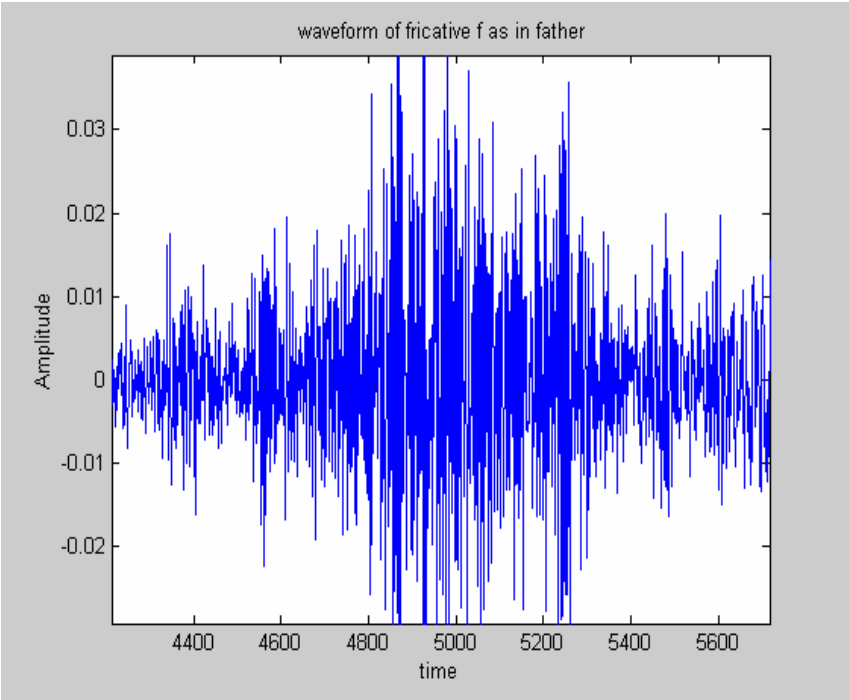



**Fig 2.7 – Waveform of unvoiced fricative /f/ (‘father’)**

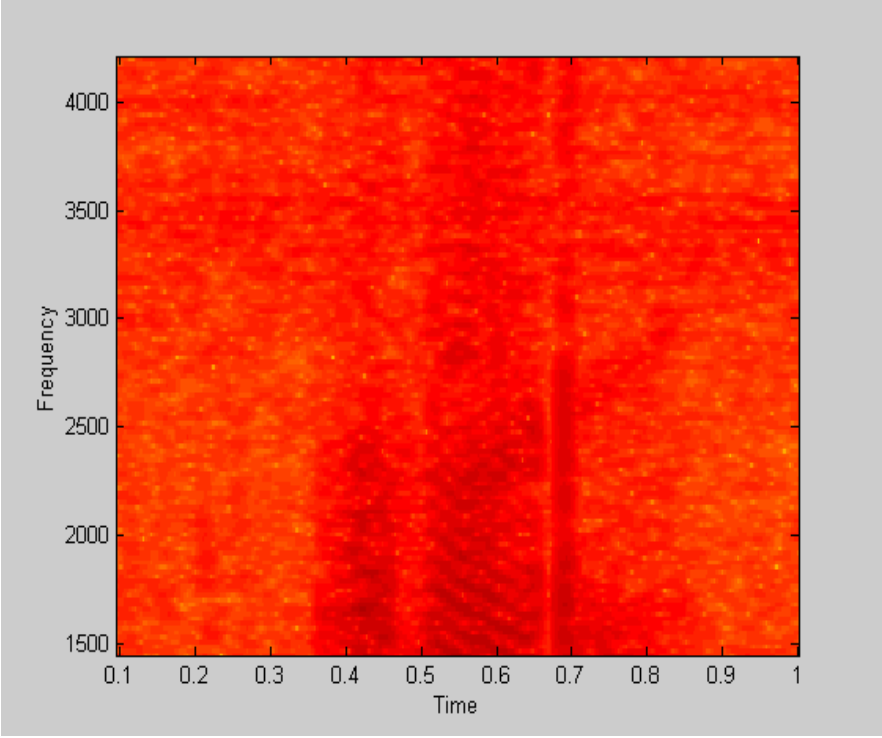


**Fig 2.8 – Waveform of unvoiced fricative /f/ (‘father’)**

Voiced frication [23] also occurs due to turbulence in the airflow provided from within the oral tract but it is often accompanied by some vocal fold vibration as in /v/ (‘vote’). The vibration of the vocal folds means that the airflow in the vocal tract is periodic and frication only takes place when the periodic airflow has reached a certain minimum level. This leads to frication being roughly synchronized with the glottal airflow velocity. Voiced fricatives can be differentiated from unvoiced fricatives through the onset of voicing. The formant transitions from fricatives to vowels also serve as a cue to distinguish between voiced and unvoiced fricatives. In voiced fricatives the voicing occurs sooner in the transitions than for unvoiced fricatives. Figures 2.9 and 2.10 show the waveform and spectrogram for the voiced fricative /v/. The figures 2.9 and 2.10 show that in voiced fricatives the noise in the waveform is super-imposed on a quasi-periodic envelope. The spectrogram shows characteristics of both noisy and periodic signals.

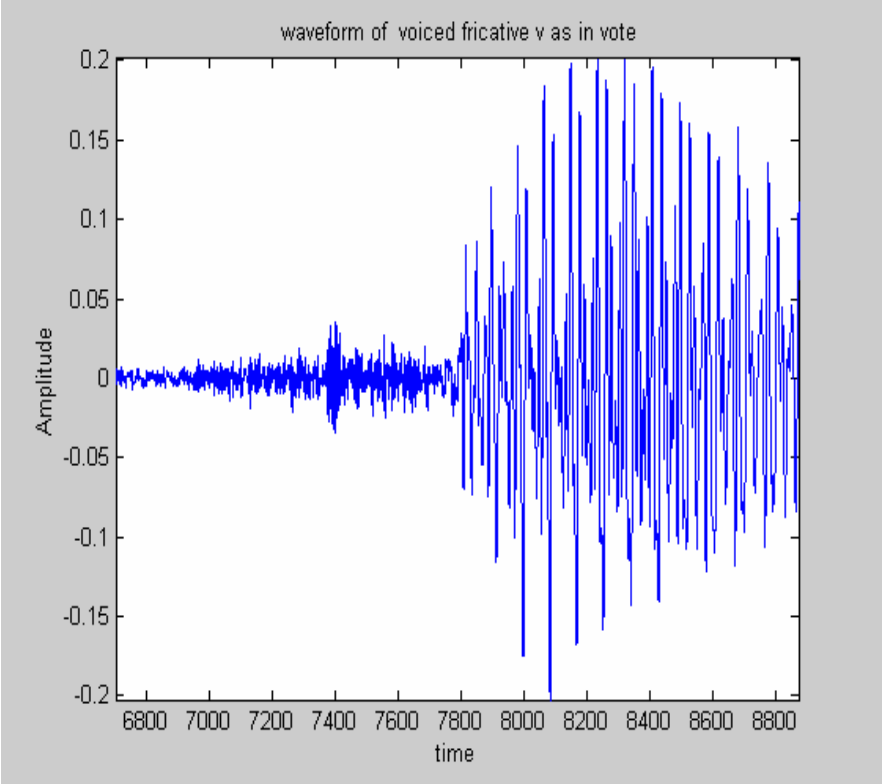



**Fig 2.9– Waveform of voiced fricative /v/ (‘vote’)**

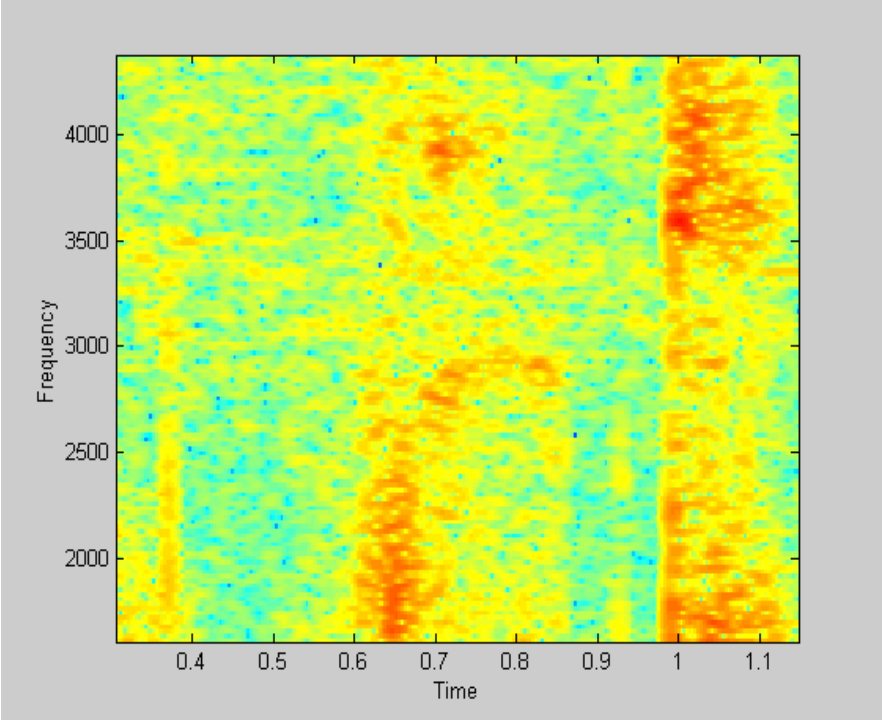



**Fig 2.10– Spectrogram of voiced fricative /v/ (‘vote’)**

### 2.3.2.3 Nasals

Nasal consonants are produced from a source similar to that used for producing vowels – semi-periodic airflow through the vocal tract that vibrates the vocal folds. For nasals, the vellum is lowered and air is mainly radiated through the nostrils because the oral cavity is constricted [17] [18]. Due to the large volume and low resonance of the nasal cavity, nasals are dominated by lower frequency energy with the first formant frequency usually being the most prominent in the spectrogram. The formant transitions that follow the release of the constriction into the steady state vowel position are used to perceptually differentiate between the different nasal consonants.

Figures 2.11 and 2.12 show the waveform and spectrogram for the nasal consonant /m/ (‘mee’). It can be seen from the spectrogram that the nasal sound is dominated by lower frequency energy and the first formant has the highest energy.

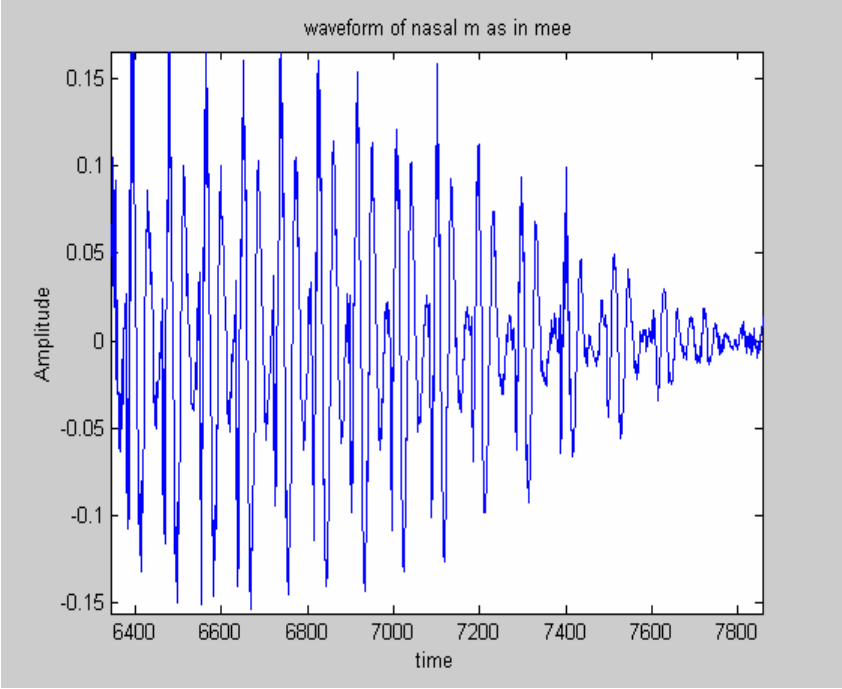



**Fig 2.11 – Waveform of Nasal /m/ (‘more’)**

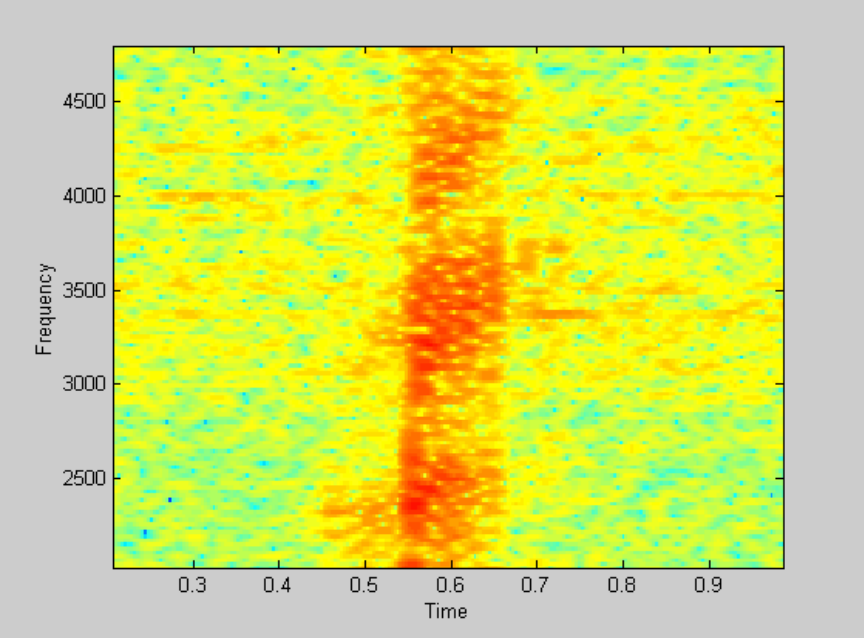


**Fig 2.12 – Spectrogram of Nasal /m/ (‘mee’)**

#### 2.3.2.4 Plosives

Plosives can be both voiced and unvoiced [17]. In unvoiced plosives there is complete closure of the oral tract causing a build-up of air pressure behind the closure followed by the release of air leading to turbulence over a short duration. Then turbulence is generated at the vocal folds and finally a vowel sound is produced. Figures 2.13 and 2.14 show the waveform and the spectrogram for an unvoiced plosive /k/ (‘key’). From the figures 2.13 and 2.14 the main stages that make up the unvoiced plosive can be clearly seen: the silence (as pressure builds up), the burst of air, the aspiration and then the transition of the oral tract from the constricted state leading to vibration of the vocal folds.

Voiced plosives are generated by a mechanism similar to that of unvoiced plosives. However, in voiced plosives there is vocal fold vibration during the pressure build-up stage. This is called the voice bar and it is generated due to the low-frequency vibration of the walls of the throat. Also, after the release of air there is no aspiration and the start of the transition to the vowel occurs much faster than in unvoiced plosives. Figures 2.15 and 2.16 show the waveform and spectrogram of the voiced plosive /g/ (‘go’). The low frequency voice bar can be seen in the spectrogram and the waveform. Most of the energy in both voiced and unvoiced plosives is lower frequency and so the first formant is very strong compared to the other formants. [17]

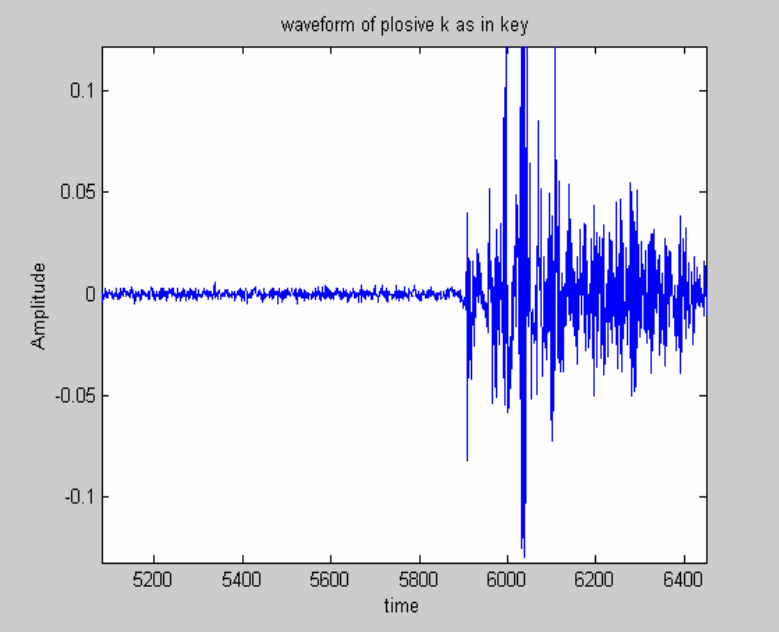



**Figure 2.13– Waveform of unvoiced plosive /k/ (‘key’)**

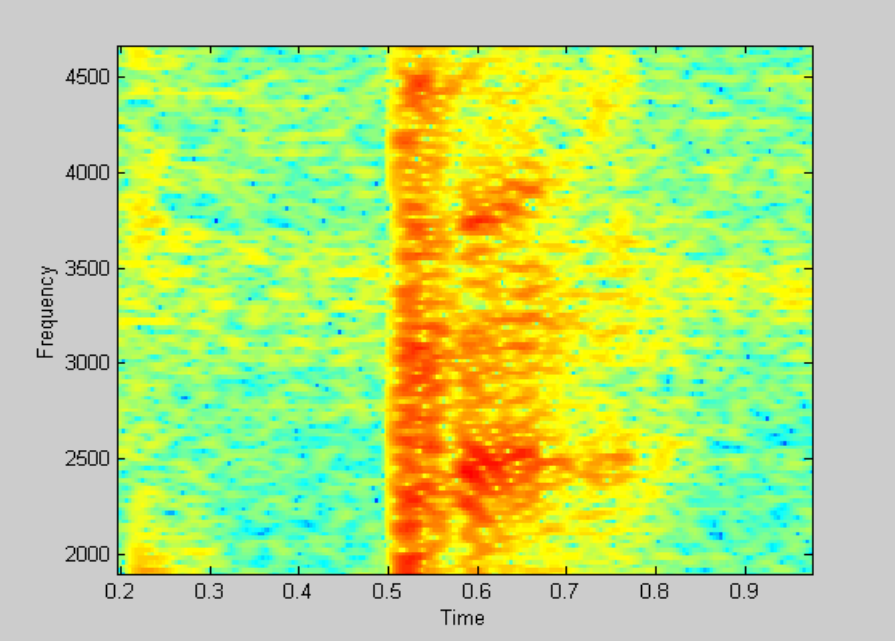



**Figure 2.14 – Spectrogram of unvoiced plosive /k/ (‘key’)**

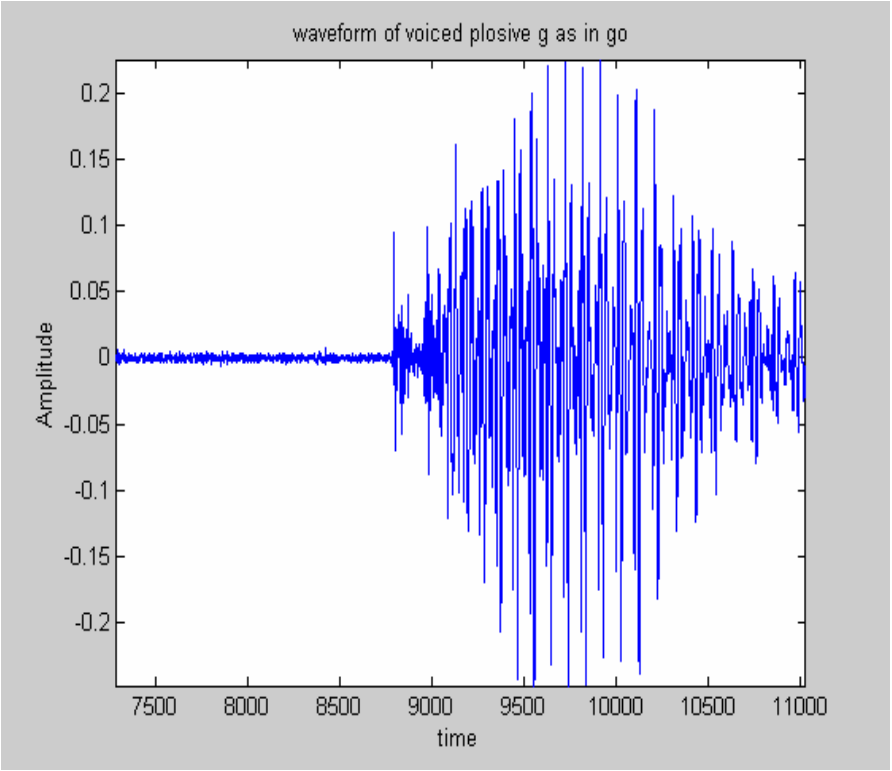



**Fig 2.15 – Waveform of voiced plosive /g/ (‘go’)**

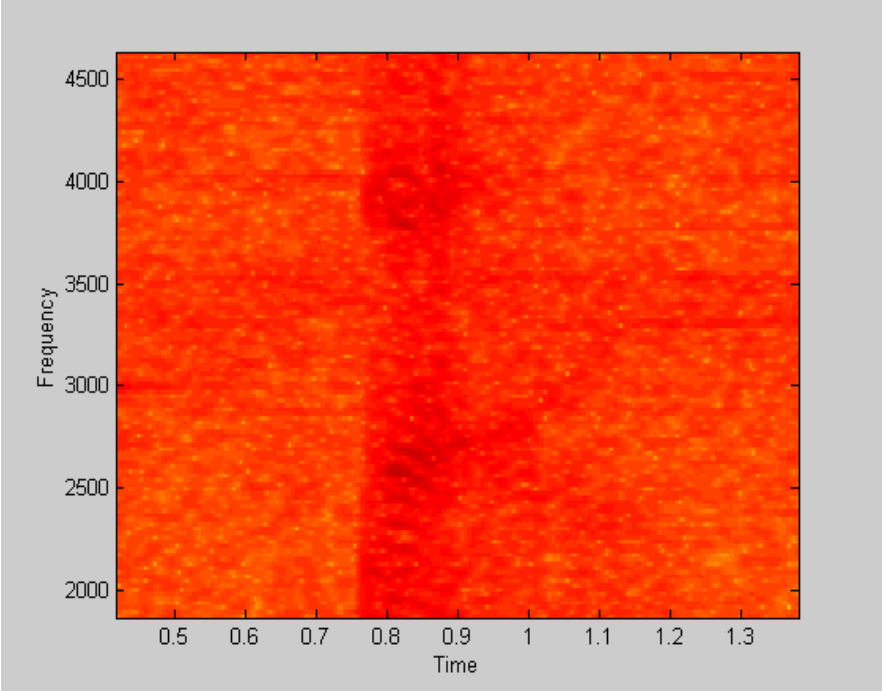



**Figure 2.16 – Spectrogram of voiced plosive /g/ (‘go’)**

### 2.3.3. Importance of Formant Frequencies in Speech Perception

In this section the roles of different features of speech that help differentiate human perception between the phonemes will be discussed. In vowels the primary source of discrimination between the phonemes can be provided by the formant frequencies. Although all phonemes have their own formants, vowel sound formants are usually the easiest to identify [1]. Almost all formants have the trait of waxing and waning in energy in all frequencies, which is caused by the repeated closing and opening of the human vocal tract. On average, this repeated closing and opening occurs at a rate of 125 times per second in an adult male and 250 times per second in an adult female. This rate gives the sensation of pitch (higher frequencies result in higher pitches). Formant values can vary widely from person to person, but the spectrogram reader learns to recognize patterns which are independent of particular frequencies and which identify the various phonemes with a high degree of reliability. For instance, in the vowels, the first formant ($F_1$) can vary from 300 Hz to 1000 Hz. The lower it is, the closer the tongue is to the roof of the mouth. The vowel /i: / as in the word 'beet' has one of the lowest $F_1$ values - about 300 Hz; in contrast, the vowel /A/ as in the word 'bought' (or 'Bob' in speakers who distinguish the vowels in the two words) has the highest $F_1$ value - about 950 Hz. It has been shown that $F_1$ and $F_2$ are highly discriminable features for vowel identification and the higher formants also play a smaller role. It is thought that the formant spacing in vowels and vowel-like phonemes are an essential feature for proper identification of vowels. Nasalization is another important feature for vowel identification; it can be checked by observing the increase in bandwidth of $F_1$. Identification of consonants is a more complex problem than identification of vowels. Among the cues used for consonant identification are formant frequency values, formant frequency transitions into the vowel following the consonant, voicing during the consonant production, and the timing of the onset of the vowel following the consonant.

It is clear from the above discussion that formant frequencies play a major role in vowel identification and are also important for consonant identification.

## 2.4 Contrast Enhanced Frequency Shaping

Sound induced or sensorineural hearing loss causes broadening of the neural response to the first formant frequency, leading to a reduction in speech perception. Simple hearing aid amplification schemes that apply amplification independently across different frequency bands cannot satisfactorily compensate for sound-induced hearing loss. Miller et al. [5] describes a hearing aid amplification technique that can improve the neural response to vowel sounds in sensorineural damaged auditory systems. This scheme, called Contrast Enhanced Frequency Shaping (CEFS) amplification, tries to reverse the effects of the sensorineural hearing loss by compensating for the frequency dependent threshold shift and tries to restore the neural representation to that of a 'normal' ear [1] [4]. CEFS tries to boost the speech signal energy in the regions where the neural thresholds have shifted to higher values. The result of proper CEFS amplification is to restore the 'normal' neural response representation of the formant frequencies to vowel like sounds [4] [5].

# Chapter 3

# FORMANT TRACKING ALGORITHMS

## 3.1 Introduction

Formant can be interpreted as adaptive non-uniform samples of the signal spectrum that are located in the resonance frequencies of the vocal tract. Formant tracking means to track the frequencies of speech signal or to track the variations in the input speech signal. Formant tracking algorithms estimate the formant frequencies of a speech signal and also track the variations in the speech signals. Two formant tracking algorithms have been discussed in the present work:

1. Robust formant tracking algorithm
2. RLS algorithm

## 3.2 Speech Signal and Its Spectrogram

The speech signal is a slow time varying signal in the sense that [18], when examined over a sufficiently short period of time(between 5 and 100 msec), its characteristics are fairly stationary; however over long periods of time(on the order of 1/5 seconds or more) the signal characteristics change to reflect the different speech sounds being spoken.

In the present work formant estimation and tracking has been achieved for a speech sample “five women played basket ball”. The way of characterizing the speech signal and representing the information associated with the sounds is via a spectral representation. Perhaps the most popular representation of this type is the sound spectrogram in which a three–dimensional representation of the speech intensity, in different frequency bands, over time is portrayed. Spectrogram is defined as a graph of the energy content of a signal expressed as function of frequency and time.

The example of speech representation is given in figure 3.1, which shows the wideband spectrogram in first panel [18], a narrowband spectrogram in second panel [18], and a waveform amplitude plot in the third panel, of a spoken version of the utterance” five women played basketball” by a male speaker. The wideband spectrogram corresponds to performing a spectral analysis on 15-msec section of waveform using a broad analyzing

filter (125Hz bandwidth) with the analysis advancing in intervals of 1msec. The spectral intensity at each point in time is indicated by the intensity (darkness) of the plot at a particular analysis frequency. Because of the relatively broad bandwidth of the analysis filters, hence the relatively short duration of the analysis window, the spectral envelope of the individual periods of the speech waveform during voiced sections are resolved and are seen as vertical striations in the spectrogram. The narrowband spectrogram corresponds to performing a spectral analysis on 15-msec section of waveform using a narrow analyzing filter (40Hz bandwidth) with the analysis again advancing in intervals of 1msec. Because of the relatively narrow bandwidth of the analysis filters, individual spectral harmonics corresponding to the pitch of the speech waveform, during voiced regions, are resolved and are seen as almost horizontal lines in the spectrogram.

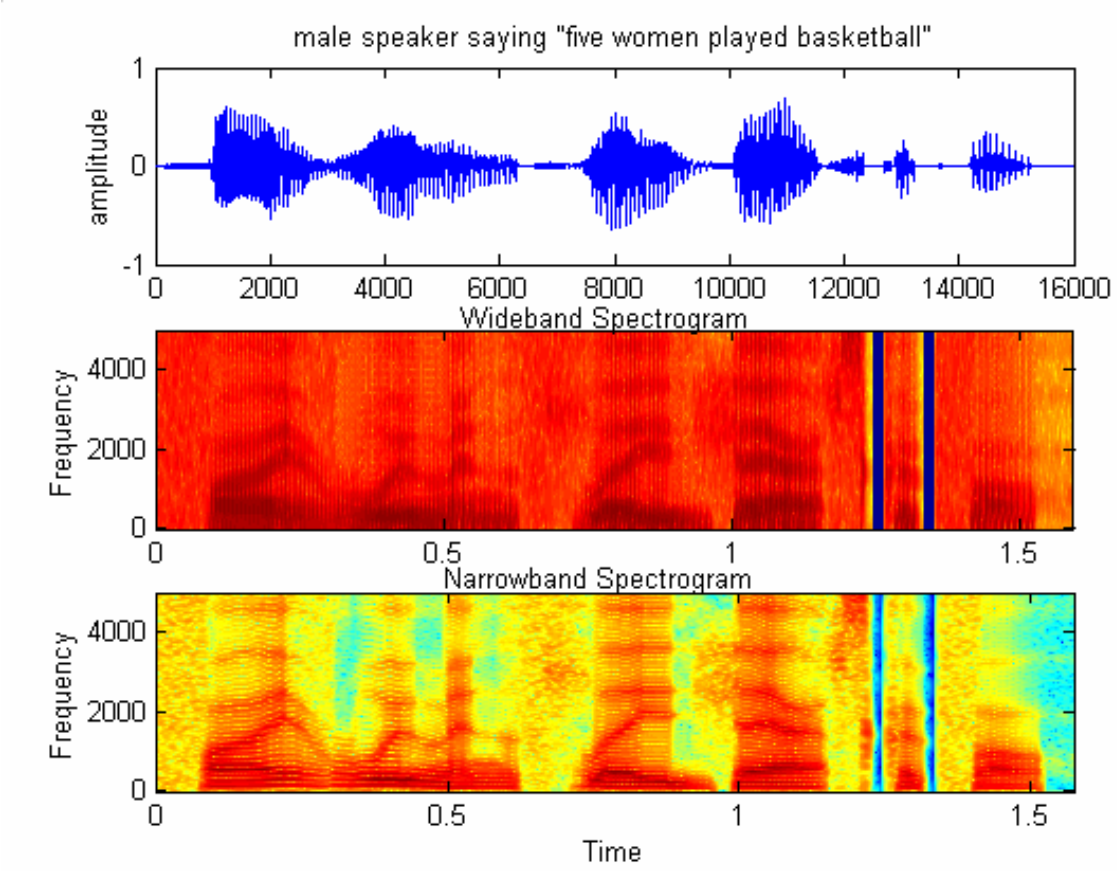



**Fig. 3.1 Speech signal and wideband and narrowband spectrogram of the utterance "five women played basketball"**

## 3.3. Robust formant tracking algorithm

The Robust formant tracking algorithm [8] is the most accurate algorithm for tracking formant frequencies of a speech signal. This can be implemented easily in real time. This algorithm can track accurately the first four formant frequencies in the noisy environment and for both male and female voices. Block diagram of robust formant tracking is shown in chapter one. (See figure 1.1)

### 3.3.1 Pre-Emphasis

A speech sample has been taken saying “five women played basketball” as an input signal for which formant estimation and tracking has to be done. The waveform for speech sample and its spectrograms are shown in figure 3.1. This waveform shows each utterance of the speech sample. After taking speech signal the first step in formant tracking with this algorithm is pre-emphasis.

Pre-emphasis is a system process designed to increase, within a band of frequencies, the magnitude of some (usually higher) frequencies with respect to the magnitude of other (usually lower) frequencies in order to improve the overall signal-to-noise ratio by minimizing the adverse effects of such phenomena as attenuation differences or saturation of recording media in subsequent parts of the system. Voiced speech signals have a natural spectral tilt, with the lower frequencies (below 1 kHz) having greater energy than the higher frequencies [1]. The lower frequencies have more energy because they contain the glottal waveform and the radiation load from the lips. In some speech processing applications it is desirable that this spectral tilt be removed by pre-emphasis or spectral equalization of the signal.

A common method of pre-emphasis is to filter the speech signal using a High-Pass Filter (HPF) [23] that attenuates the lower frequencies. The result of the pre-emphasis is the approximate removal of the contribution of the glottal waveform and the radiation load effect from the lower frequencies of the signal, i.e. the energy in the speech signal is re-distributed to be approximately equal in all frequency regions.

Figure 3.2 shows the frequency response of the FIR pre-emphasis HPF filter that is used in this formant tracking algorithm. Figure 3.3 shows a spectrogram of a speech signal before and after it has been pre-emphasized using the filter from Figure 3.2.

After the signal has been pre-emphasized it is equalized to have a global RMS energy value of 0 dB. This equalization ensures that the energy threshold levels are set properly and to appropriate energy levels.

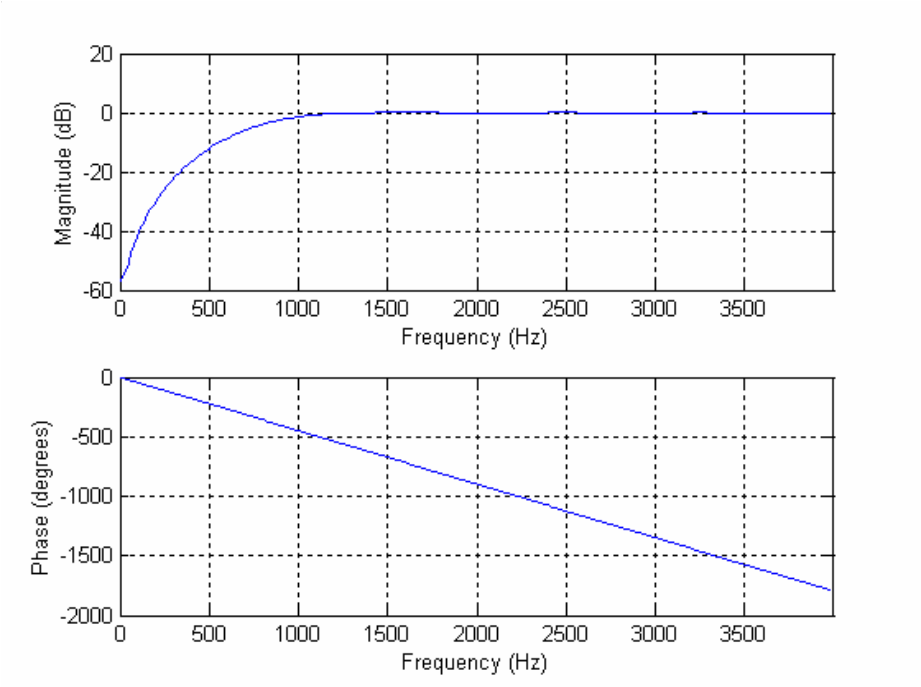



**Fig 3.2– Frequency and phase responses of the FIR pre-emphasis high-pass filter**

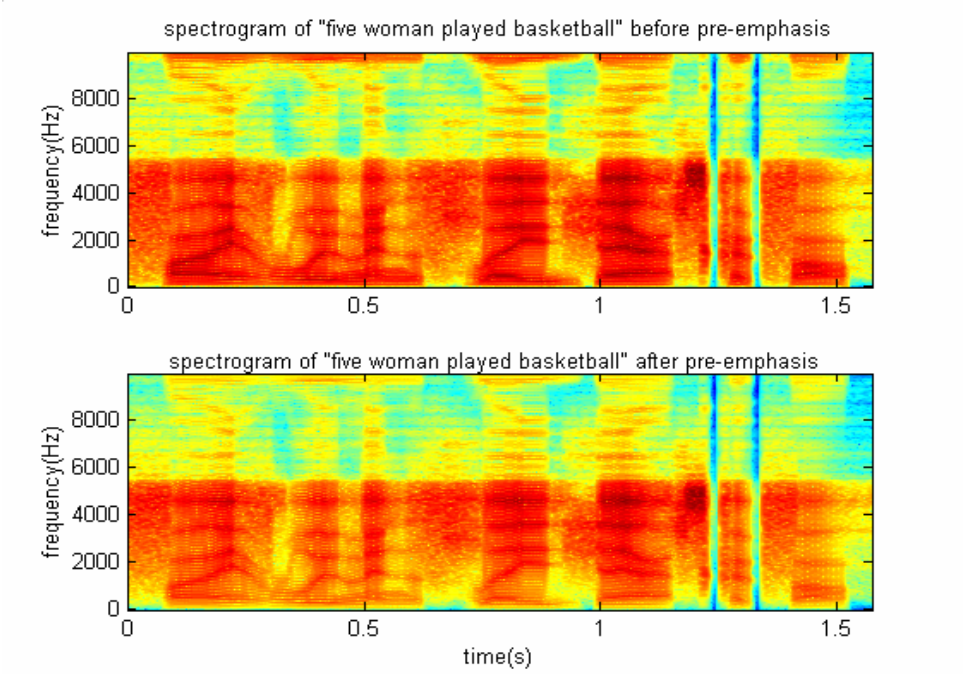



**Figure 3.3 – Spectrogram of the speech signal before and after pre-emphasis**

### 3.3.2 Hilbert Transformer

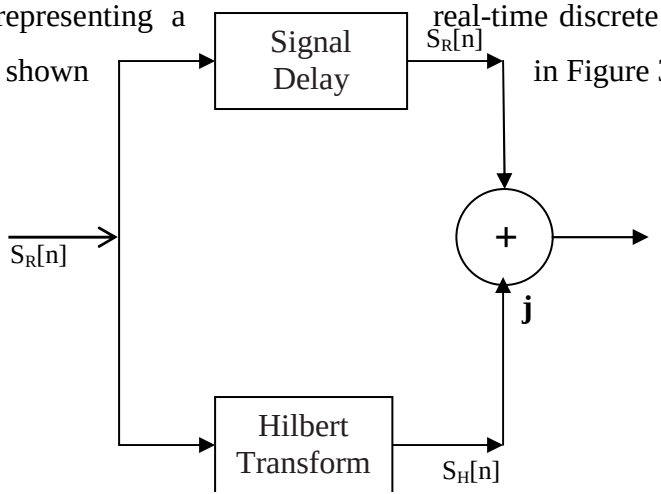
After the speech signal has been pre-emphasized, a complex version of the signal is calculated using an approximate Hilbert transformer [8]. The Hilbert transform of a function is defined as:

$$f(x) \in L_2(IR) \tag{3.1}$$

$$H\{f\}(y) = \frac{1}{\pi} PV \int_{-\infty}^{\infty} \frac{f(x)}{x-y} dx \tag{3.2}$$

Where PV stands for "principal value." The primary reason behind converting the signal into its complex representation is to allow the use of complex filters in the formant filter bank (AZF's and DTF's).

The method of representing a real-time discrete signal as a discrete complex signal is shown in Figure 3.4.



$S_C[n]$

**Fig 3.4 Converting the real-valued signal into its analytic representation**

The real-time discrete signal, $S_R[n]$, can be represented by its complex form, $S_C[n]$, as

$$S_C[n] = S_R[n] + jS_H[n], \qquad (3.3)$$

Where $S_H[n]$ is the Hilbert transform [10] of $S_R[n]$. The particular technique used to implement the Hilbert transformer in the formant tracking algorithm uses an optimum FIR filter.

The Hilbert transformer is implemented with a 20th-order linear-phase FIR filter designed using the Parks-McClellan algorithm (Remez exchange algorithm). The frequency and phase responses of the filter are shown in Figure 3.5. The filter is designed using the Remez exchange algorithm [24] and Chebyshev approximation to have an optimal fit between the desired and actual frequency responses.

The real part of the signal is added back to the Hilbert transformed part after a signal delay to account for the delay in implementing the approximate Hilbert transform (10 samples in this case). The results obtained for the analytic signal using the FIR filter method were found to be approximately the same as those obtained using an ideal Hilbert transform. Figure 3.6 shows the speech signal after taking its Hilbert transform.

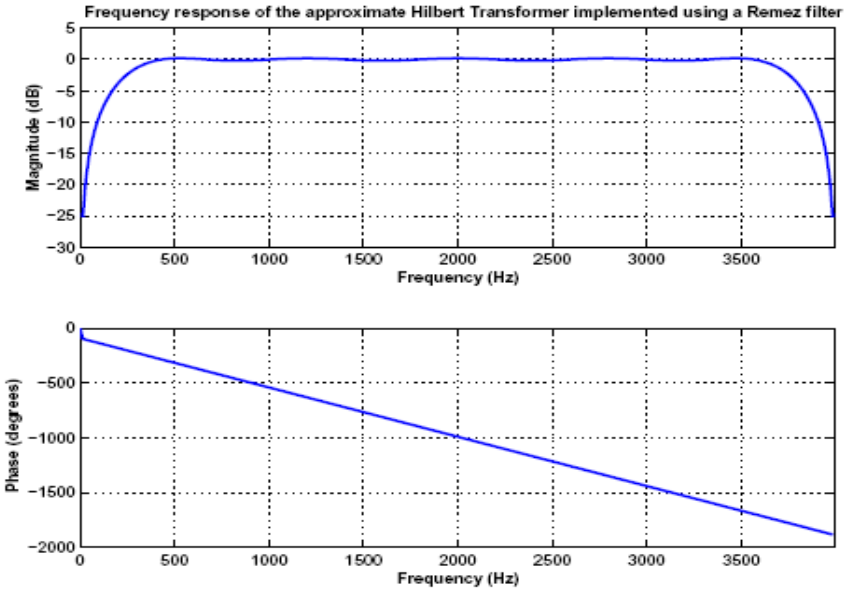



**Figure 3.5 – Frequency response of the Hilbert transform**

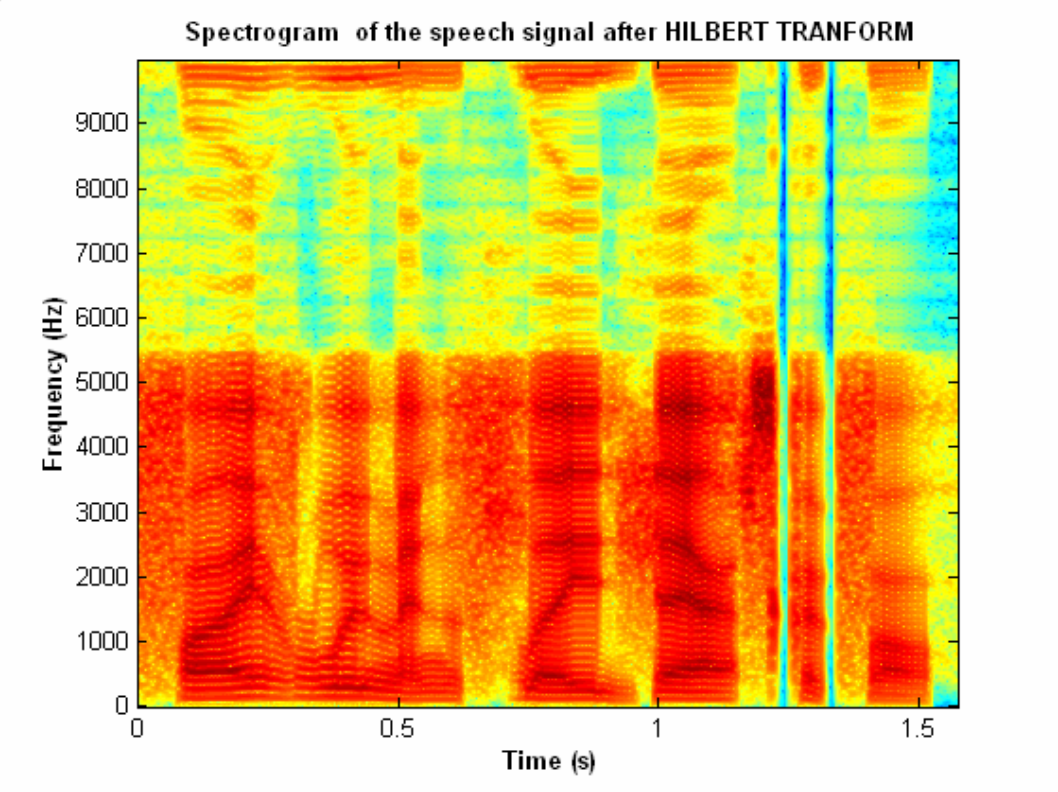



**Fig 3.6 Spectrogram of speech signal after Hilbert transform**

### 3.3.3 The Adaptive Band-Pass Filterbank

The adaptive band-pass filterbank used in the formant tracking algorithm (shown in Figure 3-7) is similar to the one proposed by Rao and Kumaresan [10] but it has a modified first formant filter that removes the effects of the pitch from the first formant filter band.

Each channel of the filter bank consists of an all-zero filter (AZF) cascaded with a single pole dynamic tracking filter (DTF). The combination of the AZF and the DTF is called a formant filter [10] and is responsible for tracking one individual formant frequency. The filters are designed in the complex domain because it is easier to design the unity gain and zero phase lag filters in the complex domain.

Adaptively varying the zeros and pole of each formant filter, allows the suppression of interference from neighboring formant frequencies and from other spectral noise sources, while tracking an individual formant frequency as it varies with time.

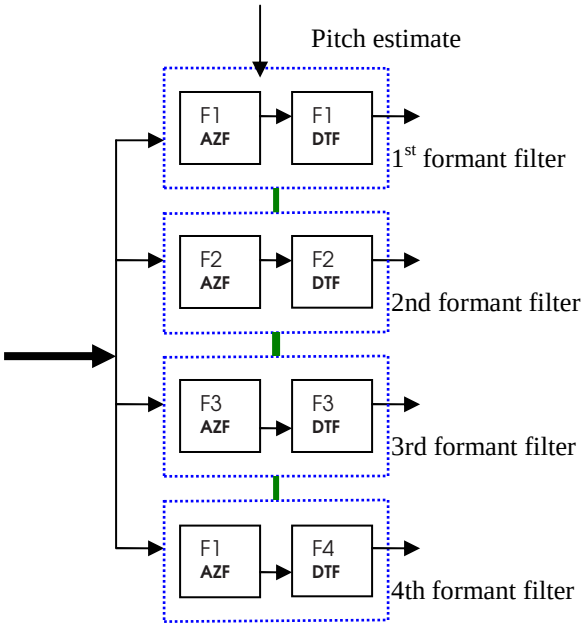



**Figure 3.7 – Adaptive band-pass filterbank**

### 3.3.3.1 All Zero Filters

In Figure 3.7 the box labeled 'AZF' in each formant filter is the adaptive all-zero filter whose three zero locations are always set to the value of the previous formant frequency estimated from the other three formant filters [8]. The transfer function of the kth AZF at time sample index n is

$$H_{AZFK}(n.z) = K_k[n] \times \prod_{\substack{l=1 \\ l \neq k}}^{4} \left(1 - r_z e^{j2\pi f l[n-1]} z^{-1}\right) \tag{3.4}$$

Where

$$K_k[n] = \frac{1}{\prod_{\substack{l=1 \\ l \neq k}}^{4} \left(1 - r_z e^{j2\pi (f_l[n-1] - f_k[n-1])}\right)} \tag{3.5}$$

and $r_z$ is the radius of the zeros on the Z-plane, $f_l[n-1]$ is the formant frequency of the $l^{th}$ filter estimated at time index n–1 and, $f_k[n]$ is the formant frequency of this filter ($k^{th}$ filter) estimated at index n–1. The gain of $K_k[n]$ (Equation 5) ensures that the AZF has unity gain and zero phase lag at the estimated formant frequency of the $k^{th}$ component. A wide range of values for $r_z$ were tested and the best results were obtained (for the range of values tested) for $r_z = 0.98$ [10].

### 3.3.3.2 Dynamic tracking filters

The box labeled 'DTF' in each formant filter [8] in Figure 3.7 is a single-pole dynamic tracking filter. The pole location is always set to the previous estimate of the formant frequency of that formant filter. The transfer function of the kth DTF at index n is

$$H_{DTFk}(n,z) = \frac{1-r_p}{\left(1-r_p e^{j2\pi f_k[n-1]} z^{-1}\right)} \tag{3.6}$$

Where $r_p$ is the radius of the pole and $f_k[n-1]$ is the formant frequency of the $k_{th}$ filter at time index n-1. A wide range of values for $r_p$ were tested and the best results were obtained (for the range of values tested) using $r_p = 0.90$.

### 3.3.3.3 The First Formant Filter

The transfer function of the $1_{st}$ formant AZF is slightly different than that of the other AZFs. The AZF of the first formant filter has an additional zero at the location of the pitch estimate to suppress pitch effects on the first formant estimate. The transfer function of the $1_{st}$ AZF at index n is

$$H_{AZF1(n,z)} = K_k[n] \times \prod_{\substack{l=1 \\ l \neq k}}^{4} \left(1 - r_z e^{j2\pi f_l}[n-1] z^{-1}\right) \tag{3.7}$$

Where

$$K_k[n] = \frac{1}{\prod_{\substack{l=1 \\ l \neq k}}^{4} \left(1 - r_z e^{j2\pi(f_l[n-1] - f_k[n-1])}\right)} \tag{3.8}$$

And $f_0$ [n-1] is the pitch estimate at time index n-1, that is provided to the 1st formant filter by the gender detector [8] [10].

After the placement of the pole and zeros for each of the formant filters, the transfer function and the complex filter coefficients of the four formant filters are calculated. These complex filter coefficients are then used to filter the analytic speech signal into four band-limited spectral regions from which the four formant frequencies are estimated.

### 3.3.3.4 The Frequency Response of Formant Filters

The frequency responses of the four formant filters are set as; pitch ($F_0$) is set to 200 Hz, the first formant frequency ($F_1$) is set to 700 Hz, the second formant frequency ($F_2$) is set to 1500 Hz, the third formant frequency ($F_3$) is set to 2200 Hz and the fourth formant frequency ($F_4$) is set to 3500 Hz.

| | |
|---|---|
| pitch($F_0$) (Hz) | 200 |
| First formant frequency($F_1$) (Hz) | 700 |
| Second formant frequency($F_2$) (Hz) | 1500 |
| third formant frequency($F_3$) (Hz) | 2200 |
| fourth formant frequency($F_4$) (Hz) | 3500 |

**Table 3.1 Positions for first four formant Frequencies**

The position of the pole and the zeros of the filters is updated for each sample. The bandwidth of the formant filters is related to the values of $r_z$ and $r_p$, and is kept constant since the values of $r_z$ and $r_p$ are not changed. All four of the filters have unity gain and zero phase lag at the location of the pole (peak of the band-pass filter that corresponds to the estimated formant frequency).

Figures 3.8, 3.9, 3.10, 3.11, 3.12 show the spectrograms of a speech signal and the spectrograms of the corresponding spectral regions that come out of the first formant filter (used for $F_1$ estimation), second formant filter (used for $F_2$ estimation), third formant filter (used for $F_3$ estimation), and fourth formant filter (used for $F_4$ estimation). As can be seen from the spectrograms, the pitch area is effectively filtered out and the higher formant frequencies are greatly attenuated for the $F_1$ region. The effect of the pitch, the first formant frequency and the upper formant frequencies are all minimized for the $F_2$ region.

**Fig 3.8 Spectrograms of the original speech signal**

**Fig 3.9 Spectrograms of the speech signals from the first formant filter bank**

**Fig 3.10 Spectrograms of the speech signals from the second formant filter bank**

**Fig 3.11 Spectrograms of the speech signals from the third formant filter bank**

**Fig 3.12 Spectrograms of the speech signals from the fourth formant filter bank**

### 3.3.4 Adaptive Energy Detector

After the speech signal has been filtered using the adaptive band-pass filterbank, the RMS energy of the signal over the previous 20 ms in each band is calculated. In order for the algorithm to estimate a particular formant frequency from the spectrum (instead of using the moving average value), the energy calculated in that formant band has to be above a certain 'energy threshold level', in addition to that speech segment being voiced. As mentioned in Section 3.2 the global RMS energy of the speech signal is normalized after pre-emphasis, so that the signal has an RMS of 0 dB. The energy threshold level [10] for each of the formant frequencies is different and is adaptive to long term changes

in the spectral energy of the formant frequency bands. The energy threshold level for each formant frequency is updated at every voiced segment of speech, allowing operation in dynamically changing environments. Equation (3.9) describes how the energy level of each formant frequency is updated during voiced segments of speech:

$$ET_{Fi}(n) = ET_{Fi}(n-1) - (.002 * (ET_{Fi}(n-1) - E_{Fi}(n))) \quad (3.9)$$

where $ET_{Fi}$ (n) is the energy threshold level (in dB) of the $i^{th}$ formant frequency at time index (n), $ET_{Fi}$ (n −1) energy threshold level (in dB) of the $i^{th}$ formant frequency at time index (n-1 ), and $E_{Fi}$ is the RMS energy (in dB) of the previous 20 ms of the speech signal.

The energy in each band is calculated independently of the energy of the other bands. Therefore, it is possible for the energy in some of the bands to be below their threshold level and the energy in other bands to be above the threshold levels concurrently. This scenario results in one or more of the formant frequencies being spectrally estimated, while others revert to their moving average value. Keeping the threshold levels and the energy calculations in each of the frequency bands independent allows accurate formant estimation in at least a few of the formant bands when there is low energy in only some of the frequency bands.

If there are long term changes in the energy of a formant band, the threshold level adapts to these energy changes gradually. Not changing the threshold levels abruptly prevents long term errors to the energy detector and allows the algorithm to recover quickly from brief loud sounds. The threshold levels are measured in decibels and the initial energy threshold levels are set at the start of the algorithm and updated at voiced segments of speech.

Various initial threshold levels were tested and the best results were obtained using the following initial threshold levels:

Initial $F_1$ Energy Threshold Level = –35 dB;

Initial $F_2$ Energy Threshold Level = –40 dB;

Initial $F_3$ Energy Threshold Level = –45 dB;

Initial $F_4$ Energy Threshold Level = –50 dB.

It is important to note that these initial values are calibrated for speech signals whose energy levels have been normalized to have a mean of 0 dB. If the signal energy is not normalized, the algorithm would require some time to adapt to the actual levels of energy present in each formant frequency band, before normal operation of the algorithm can resume.

The variation of the energy threshold levels for the four formant filters throughout an energy normalized speech signal is shown in Figure 3.13. The signal used is a synthesized speech signal for a female speaker saying “five women played basketball”.

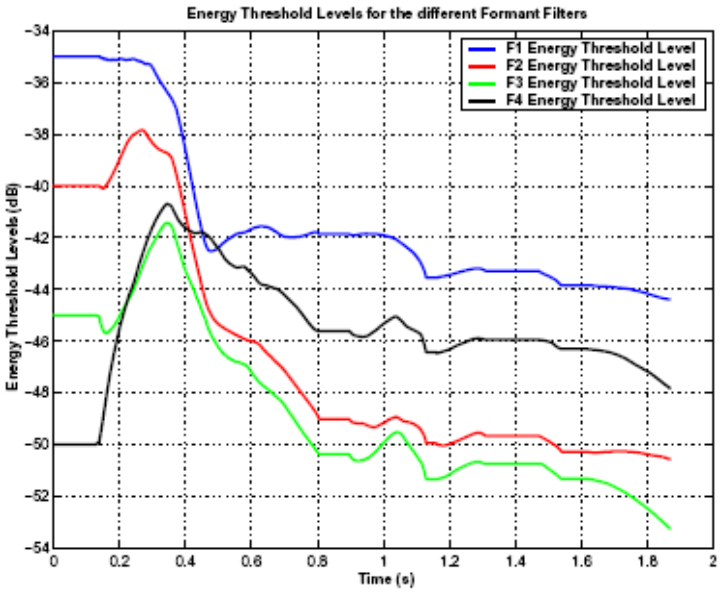



**Figure 3.13 – Variation of the energy threshold levels through time for a female speaker speech signal: “five women playing basketball”**

### 3.3.5 Calculating the Linear Predictor Coefficients

Linear prediction provides a good model of the speech signal. This is especially true for the quasi steady state voiced region of speech in which the all pole model of LPC provides a good approximation to the vocal tract spectral envelope. During unvoiced transient region of speech, the LPC model [14] is less effective than for voiced regions, but it still provides acceptably useful models. The linear prediction method provides a robust, reliable, and accurate method for estimating the parameters that characterize the linear time varying system.

Linear prediction models the human vocal tract as an infinite impulse response (IIR) system that produces the speech signal. For vowel sounds and other voiced regions of speech, which have a resonant structure and high degree of similarity over time shifts that are multiples of their pitch period, this modeling produces an efficient representation of the sound.

The linear prediction [25] problem can be stated as finding the coefficients $a_k$ which results in the best prediction (which minimizes mean-squared prediction error) of the speech sample s [n] in terms of the past samples s [n − k], k = {1. . . P}.

The idea behind linear prediction is to approximate each sample of the speech signal as a linear combination of past samples. By minimizing sum of squared differences between the actual speech samples and the predicted ones, a unique set of predicted coefficients can be determined. A linear predictor of order p is defined as

$$\tilde{S}[n] = \sum_{k=1}^{p} \alpha_k S[n-k] \tag{3.10}$$

Where $\tilde{S}[n]$ is the prediction of S[n] by the sum of p past weighted samples of S[n].

The system function of the $p^{th}$ order predictor is a FIR filter of length p given by:

$$P(z) = \sum_{k=1}^{p} \alpha_k z^{-k} \tag{3.11}$$

And the associated prediction error filter is:

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k Z^{-k} = 1 - P(z) \tag{3.12}$$

And prediction error is defined by:

$$e[n] = S[n] - \tilde{S}[n]$$

$$= S[n] - \sum_{k=1}^{p} \alpha_k S[n-k] \qquad (3.13)$$

The roots of the inverse of the prediction error filter corresponds to the poles placed to model the original signal as closely as possible while minimizing the mean squared error between the estimated and original signals. First order linear prediction (p = 1) obtains one linear predictive coefficient and the corresponding single pole is placed to model the original signal as well as possible. Second order LPC [25] tries to model the original signal using two poles, and so on.

The first four formant frequencies of the speech signal are estimated from the four filterbands of the adaptive bandpass filterbank using first-order LPC. The analytic signal from each of the bands is first windowed using a 20-ms periodic Hamming window and then the linear predictive coefficient (one per band) of the previous 20 ms of the windowed signal from each band is calculated. LPC tries to fit a single pole model to each signal and the location of the pole corresponds roughly to the vocal tract pole (formant frequency) in that band, for voiced segments of speech. The LPCs are only calculated from the bands if the entire previous 20-ms window of the speech signal is voiced (as determined by the voicing detector).

### 3.3.6 Voicing Detector

Figure 3.14 shows a block diagram of the voicing detector [8] that has been designed for use with the formant tracking algorithm. The purpose of the voicing detector is to provide the formant tracking algorithm with a reliable sample by sample decision on whether a signal is voiced or unvoiced. Functionality has been built into the voicing detector to prevent it from switching its decisions spuriously. Parameters of the voicing detector need to be changed to be able to work for both male and female speakers. The gender detector provides regular updates to the voicing detector about the gender of the speaker so that the voicing detector can use the correct set of parameters.

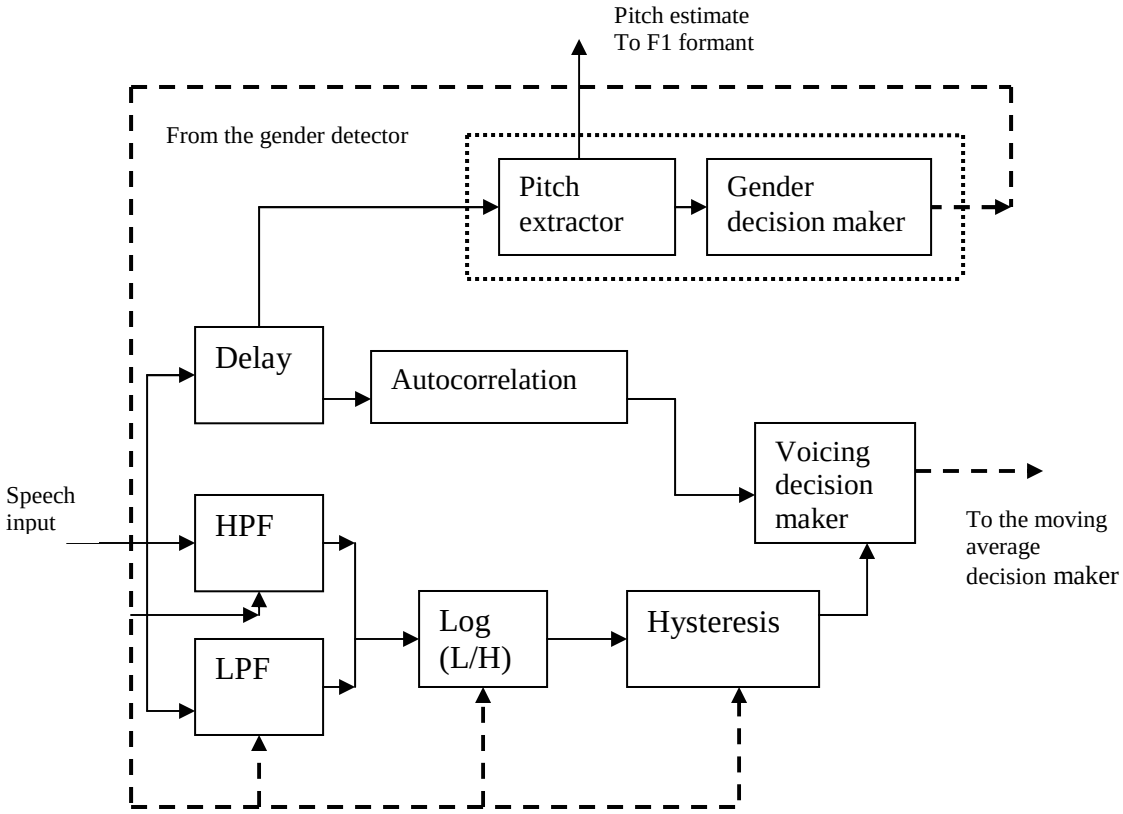



**Fig. 3.14 Block diagram of voicing detector**

### 3.3.6.1 The High Pass Filter and Low Pass Filter of the Voicing Detector

The original speech signal (the real valued signal) is filtered into two different frequency bands by passing it through a High-Pass Filter (HPF) and a Low-Pass Filter (LPF). Figure 3-15 shows the frequency and phase responses of the $20_{th}$-order Butterworth HPF and LPF where the cut-off frequency of the two filters is set to 700 Hz (dotted black line).

Once the signal is filtered into the two frequency bands, the log ratio of the RMS energy for the previous 20 ms of the signal, between the lower and the higher frequency bands is calculated.

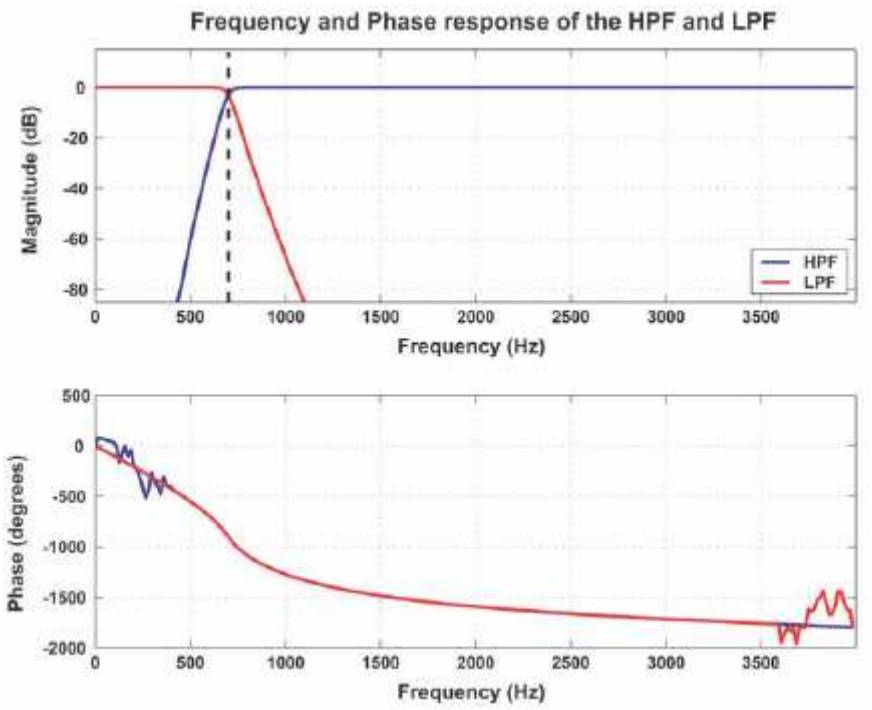



**Fig 3.15 The Frequency and Phase responses of the HPF and LPF**

Voiced speech is made up of lower frequency components than unvoiced speech, so the energy in the lower frequency band is expected to be greater than the energy in the higher frequency band during voiced speech. During voiced speech segments, the log ratio of the low frequency band to high frequency band is positive, indicating that the energy in the lower frequency band is greater than that in the higher frequency band. The log energy ratio is calculated with a sliding window moving sample by sample and a windowed signal is classified as voiced if its log ratio exceeds a set threshold level. This energy ratio measure serves as the primary means of classification for determining if a speech segment is voiced or unvoiced.

The best value of the cut-off frequency of the HPF and the LPF depends on the gender of the speaker. A large number of values were tested for the selection of the cutoff frequency for both genders. For the range of values tested, the best results were obtained when the cut-off frequency was set to 700 Hz for male speech and 1120 Hz for female speech. The voicing detector gets updates every 20 ms about the gender of the speaker from the gender detector and is able to modify the cut-off frequency of the LPF and the

HPF if the gender of the speaker changes. If the cut-off frequency is to be changed, it is slowly increased or decreased so that there are no transient effects, as shown by Equation 3.12. The algorithm is configured to shift the cut-off frequency from 700 Hz to 1120 Hz (from male speaker to female speaker) or vice versa over 40 ms as shown in equations 3.14 and 3.15.

$$F_c[n] = F_c[n-1] + 10, \quad \text{If } G[n] = 0 \text{ and } F_c[n] < 120 \tag{3.14}$$

$$F_c[n] = F_c[n-1] - 10, \quad \text{If } G[n] = 0 \text{ and } F_c[n] > 120 \tag{3.15}$$

Where $F_c[n]$ is the cut-off frequency at time index n and G[n] is the estimated gender at time index n (zero for female and one for male).

**3.3.6.2 Threshold with Hysteresis**

The log energy ratio used to determine if the input is voiced or unvoiced is reliable and accurate only for phonemes whose frequency components do not vary too much over time. The presence of transient frequencies in certain phonemes makes the log energy ratio unreliable on its own for determining voicing in continuous speech. This is because transient frequency components can make the voicing detector results oscillate too quickly between the voiced and unvoiced states. In order to avoid these fast oscillations between the two states, Bruce et al. [26] proposed a threshold with hysteresis. This allows changes in the voicing state (from voiced to unvoiced or vice versa) only if the state of the current sample changes from the previous sample and the current sample has a log ratio greater than a set threshold level. These threshold levels depend on the gender of the speaker and have to be changed as the gender of the speaker changes.

If the previous sample is unvoiced and the current sample has a log ratio greater than a set threshold level (Log_Ratio_Threshold_Voiced), then the current sample is assigned as being voiced, i.e. the switch from unvoiced to voiced state occurs only if the log energy ratio is greater than the proper threshold level.

If the previous sample is voiced and the current sample has a log ratio less than a set threshold level (Log_Ratio_Threshold_Unvoiced), then the current sample is assigned as being unvoiced, i.e. the switch from voiced to unvoiced state occurs only if the log

energy ratio is below the proper threshold level. From the range of values tested, the best results were obtained when the level was set to 0.2 for males and 0.3 for females for Log_Ratio_Threshold_Voiced and 0.1 for males and 0.2 for females for Log_Ratio_Threshold_Unvoiced. The gender of the speaker is checked every 20ms to confirm that, the proper set of parameters are being used. If the gender of the speaker changes, the threshold levels are updated slowly over 40ms, to avoid any transient effects.

#### 3.3.6.3 Autocorrelation Test

The contribution of energy due to additive white Gaussian noise over short time durations may not be 'white', but instead be 'colored' (be randomly concentrated in the lower or upper frequency band). Voicing decisions based solely on the log ratio measure would rely only on the energy distribution of the signal over the previous 20 ms of data. If the short-term energy from AWGN is concentrated in the lower frequency band, the log energy ratio will erroneously detect the signal as being voiced.

In order to avoid the problem of erroneous voicing detection in the presence of AWGN, the voicing detector algorithm performs an autocorrelation based test to check if the energy in the lower frequency band of the signal is due to AWGN or due to some other non-random signal. The autocorrelation of the previous 20 ms of the signal is calculated. The signal is classified as voiced, if the autocorrelation at any lag ($\tau \neq 0$) is greater than the autocorrelation_threshold_multiplier times the autocorrelation at zero ($\tau = 0$) and there is at least one point in the window whose autocorrelation is greater than 0. If the low frequency energy in the signal is determined to be due to AWGN, the autocorrelation of the signal will be very low since random signals have very low or zero autocorrelation values (when $\tau \neq 0$ ), and the above test will fail.

The value of the autocorrelation_threshold_multiplier is different for male and female speakers. Through trial and error the best results were obtained when the autocorrelation_threshold_multiplier was set to 0.25 for female speakers and 0.6 for male speakers. The gender of the speaker is checked every 20 ms and the value of the autocorrelation_threshold_multiplier can be changed if the gender of the speaker changes.

### 3.3.6.4 Voicing Detector Testing and Results

The testing of the voicing detector algorithm was conducted using synthesized sentences. Testing using the synthesized sentences allows quantitative measurements of the performance of the voicing detector for both male and female speakers since the exact time of the onset of voicing is known. Figures 3.16 and 3.17 show the performance of the voicing detector for the male and female synthesized sentence “five women played basketball”. When the lines are at zero (‘low’), it indicates that the speech is unvoiced and when the lines are non-zero (‘high’), it indicates that speech is voiced. The algorithm is also robust and there is very little or no oscillation of the output between voiced and unvoiced states.

**Figure 3.16– Voicing Detector results for a synthesized male speaker**

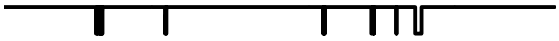


**Figure 3.17– Voicing Detector results for a synthesized female speaker**

### 3.3.7 Gender Detector

The difference in pitch between male and female speakers is sufficient to serve as a discriminating parameter between the two types of speakers. The gender detector calculates the pitch and determines the gender of the speaker. It provides this information to the voicing detector so that the voicing detector can update its parameters to work properly for both male and female speakers. Accurate pitch estimation from continuous speech is a difficult task to accomplish. Several complicated algorithms have been proposed to achieve this task. However, for the purposes of constructing a low-computation and low-delay gender detector, it was deemed sufficient to have an approximate estimate of the pitch as long as there is still clear discriminability between male and female speakers. Therefore, a well known fast and simple approach to pitch estimation [25] is chosen that uses an autocorrelation based approach.

In the gender detector algorithm, pitch is estimated from the real valued speech signal using the short-time autocorrelation of the previous 60 ms of the signal. The 60 ms signal is divided into non-overlapping frames whose length must be greater than at least one pitch period in order to measure the pitch in the frame accurately. The gender detector algorithm segments the signal into non-overlapping 20 ms-frames.

Each frame is lowpass filtered using a fourth-order Butterworth filter (LPF) to reduce the range of spectral estimation. The pitch information is contained within the lower

frequencies of speech (< 1000 Hz) so the higher frequencies contained in the signal can be discarded. The frequency response of the LPF is shown in Figure 3-18

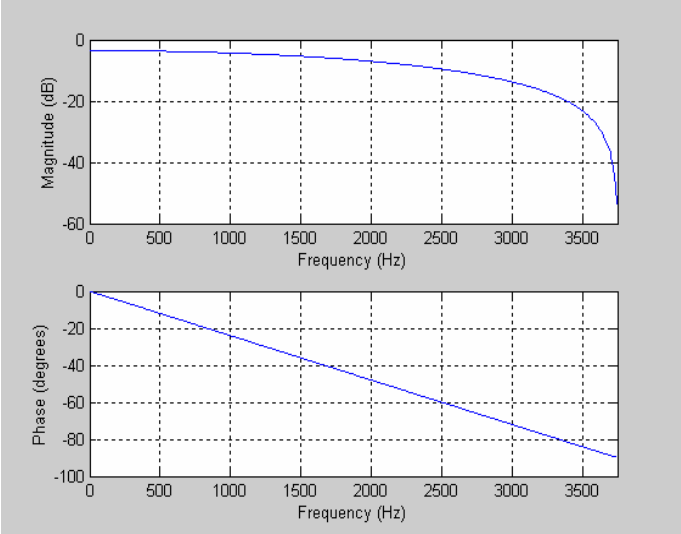



**Figure 3.18– LPF for the Gender Detector**

### 3.3.7.1 Determination of the Average Pitch Period and the Gender of the Speaker

There can be interaction between the pitch frequency and the first formant frequency when the first formant frequency bandwidth is narrow relative to the harmonic spacing. In such cases the autocorrelation function of the signal has higher peaks due to the vocal tract response (first formant frequencies) than due to the vocal excitation (pitch frequency). This makes it difficult to estimate the pitch frequency using short time autocorrelation. To avoid this problem, a nonlinear time-domain technique called centre-clipping is used that makes the periodicity of the speech signal more prominent while suppressing the other features of the speech that contribute to the extra peaks of the autocorrelation function. Center clipping function is shown in figure 3.19.

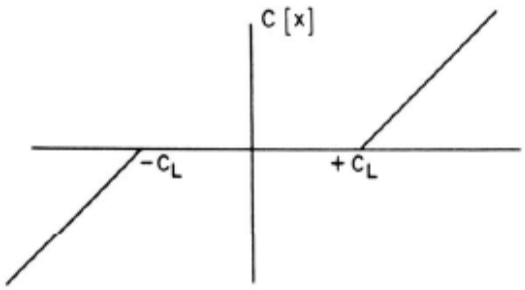


**Fig.3.19 Center clipping function**

Center clipper definition is

$$\text{If } x(n) > C_L,\ y(n) = x(n) - C_L \qquad (3.16)$$

$$\text{If } x(n) \leq C_L,\ y(n) = 0 \qquad (3.17)$$

After the signal has been centre clipped, its autocorrelation, $R_n(p)$, is calculated and the location of the highest peak, p, of the autocorrelation function is located. If Rn(p) is less than 0.4 x Rn(0), then the segment is classified as being unvoiced and its pitch is set to 0 Hz. Otherwise, the pitch period [14] is calculated as being the location of the highest peak of the autocorrelation function.

The range of acceptable values for the pitch frequency is between 60 and 320 Hz, and if the calculated value of the pitch is outside this range then it is set to the moving average value of the pitch in that segment. The value of the pitch frequency in each frame is used to calculate the average pitch frequency of each segment of the signal (of 60 ms duration) passed to the gender detector algorithm.

The average pitch frequency of each segment is sent to the first formant filter to be used for the placement of the additional zero at the pitch frequency location. The gender G[n] of the speaker is considered to be male (‘0’) if the average pitch frequency is below 180 Hz and is set to female (‘1’) if it is above that value.

### 3.3.8 Moving Average Decision Maker

The moving average decision maker [8] has two main purposes:

- To calculate and update the moving average value of each formant frequency and,
- To determine whether to assign the LPC estimated value or the moving average value to each formant frequency.

The moving average decision maker assigns the estimated value to the formant frequencies (from the LPCs) only when the segment is voiced and

the energy of the formant frequency is above its respective threshold level. If the segment is unvoiced or if the energy of a particular formant is below its respective threshold level, then the current value of the formant frequency decays toward the moving average value for that formant frequency according to:

$$F_i[n] = F_i[n-1] - \left(0.002 * \left(F_i[n-1] - F_i^{MA}[n-1]\right)\right) \qquad (3.18)$$

Where $F_i[n]$ is the formant estimate of i[th] formant frequency at time index (n) and $F_i^{MA}[n-1]$ is the previous value of the moving average for the i[th] formant frequency. Equation 3.19 describes the update rule for the moving average value of each formant frequency:

$$F_i^{MA}[n] = \frac{1}{n}\sum_{k=1}^{n} F_i[k] \qquad (3.19)$$

Where $F_i^{MA}$ is the moving average value for the i[th] formant frequency at index n and

$F_i[n]$ is the estimate of the i[th] formant frequency at index n.

### 3.3.9 Limitations on the Proximity of Formant Frequencies

The filter response of the formant filterbank becomes poor when the location of the poles and zeros are very close. Therefore, the formant tracking algorithm limits how close the formant frequencies can come to each other. The algorithm does not allow $F_1$ to be less than 150 Hz from the pitch frequency and any estimate of $F_1$ that is less than 150 Hz from the pitch is set to be pitch+200 Hz. $F_2$ is also limited from being less than 300 Hz from $F_1$. Any $F_2$ values that are less than 300 Hz from $F_1$ are set to be $F_1$+400 Hz. Similarly, $F_3$ is not allowed to be less than 400 Hz from F2. All values of $F_3$ that are less than 400 Hz apart from $F_2$ are set as $F_2$ +400 Hz. Finally, all $F_4$ values that are less than 400 Hz from the $F_3$ values are

set as $F_3$+400 Hz. This limitation on the proximity of the formant frequency values ensures that the poles and zeros of the formant filterbank are never too close to cause problems to the frequency response of the filterbank. Figure 3.20 shows the algorithm used for updating the formant frequencies values when they are too close to each other.

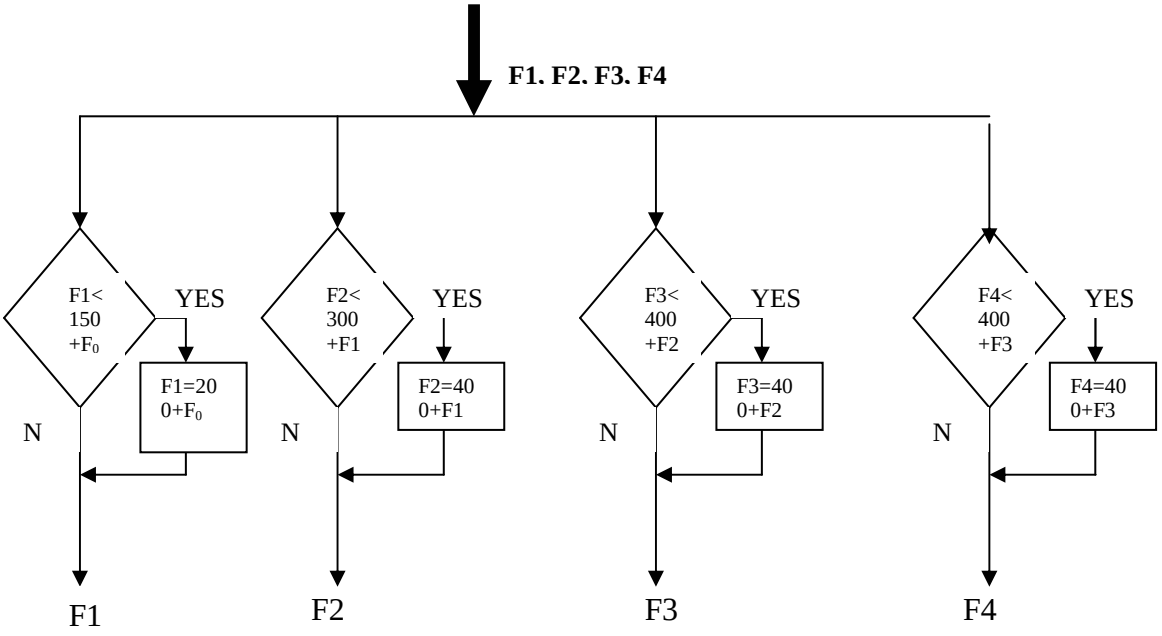



**Fig 3 .20– Update rules for the formant frequency proximity**

## 3.4 Formant Tracking With RLS Algorithm

Several speech processing algorithms assume the signal is stationary during short intervals (approximately 20 to 30 ms). This assumption is valid for several applications, but it is too restrictive in some contexts. This work investigates the application of adaptive signal processing to the problem of estimating the formant frequencies of speech.

Advantages of RLS algorithm [9]:

- Very fast adaptation
- converges very close to the optimum

Disadvantages of RLS algorithm:

- high implementation complexity
- numerical problems
- need for a reference signal or a training sequence

Least-square algorithms aim at the minimization of the sum of the squares of the difference between the desired signal and the model filter output. When new samples of the incoming signals are received at each iteration, the solution for the least-squares problem can be computed in recursive form resulting in the recursive least-squares (RLS) algorithms. An important feature of the RLS algorithm is that it utilizes information contained in the input data, extending back to the instant of time when the algorithm is initiated. The resulting rate of convergence is therefore typically an order of magnitude faster than the simple LMS algorithm.

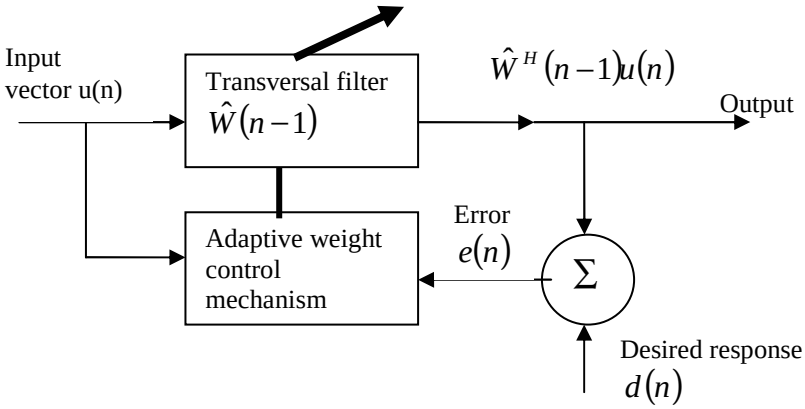



**Fig 3.21 Representation of RLS algorithm**

This improvement in performance, however, is achieved at the expense of a large increase in computational complexity and some stability problems [27]. Figure 3.21 shows the representation of RLS algorithm [9].

The least squares solution manifested by the normal equation operates in a block mode. This means that the process of estimating the weight vector W has to wait until all the

data samples are available. In addition, whenever new samples become available, finding W requires solving the inverse of the input covariance matrix. Doing so, however, is computationally expensive. A recursive procedure that can provide estimates of the parameter vector at every step in time using any new information without having to compute the inverse of $X^T X$, therefore, would be appreciated [29]. Recursive least squares algorithms achieve exactly this goal.

RLS algorithm can be summarized as [9]:
Initialized the algorithm by setting

$$\hat{w}(0) = 0 \tag{3.20}$$

$$P(0) = \delta^{-1} I, \tag{3.21}$$

And

$\delta = \{$ Small positive constant for high SNR
$\{$ Large positive constant for low SNR

For instant of time, n =1, 2… compute

$$\pi(n) = P(n-1)u(n),$$

$$k(n) = \frac{\pi(n)}{\lambda + u^H(n)\pi(n)}, \tag{3.22}$$

$$\xi(n) = d(n) - \hat{w}^H(n-1)u(n),$$

$$\hat{w}(n) = \hat{w}(n-1) + k(n)\xi^*(n),$$

*and* (3.23)

$$P(n) = \lambda^{-1}P(n-1) - \lambda^{-1}k(n)u^H(n)P(n-1),$$

The M-by-M matrix P (n) is referred to as inverse correlation matrix, k (n) is a gain vector, $\lambda$ is a forgetting factor, u(n) is a input vector, w(n) is a weight vector, $\pi(n)$ is a intermediate quantity, $\xi(n)$ is a priori estimation error, $\delta$ is a regularization parameter. The equations shown above are collectively and in that order, constitute the RLS algorithm.

The standard RLS algorithm is known to have optimal properties in stationary environments where the parameter vector is assumed to be time invariant. However, it is unsuitable for tracking the time varying parameters in nonstationary environments as in speech signal. This is due to the property that the adaptation gain in k (n) converges to zero as the parameter estimates approach their true values to reduce the noise effect on the estimates. Consequently, the algorithm loses it ability to track possible parameter changes and, eventually, it gets turned off.

An important requirement for adaptive signal processing and adaptive control is to have a recursive estimator that has the ability to track parameter changes in nonstationary environments. The standard RLS algorithm, however, does not satisfy this requirement. Therefore, considerable research effort has been directed towards the development of modified versions of the algorithm which have good tracking capabilities in time varying environments [28]. The common goal for most of these modification procedures is to prevent the adaptation gain from tending to zero. Numerous parameter-tracking modifications have been proposed in the literature. Most of these modifications, however, follow the forgetting factor approach [29].

The main purpose of such a forgetting factor is to introduce relative weightings on the estimation error contributions to the cost function. Doing so implies that certain data points are more informative than others. In the case of the exponential weighting methods, the most recent data samples are assumed to be more informative than past data samples and hence the past samples are exponentially discarded [30]. Variants of the standard RLS algorithm that are based on the forgetting factor approach minimize a cost function that differs from the cost function in (3.22) by the presence of a weighting or forgetting factor λ. The minimized cost function becomes

$$J = \sum_{i=1}^{n} \lambda^{n-i} e_i^2, \qquad 0 < \lambda \leq 1 \tag{3.24}$$

$$J = \sum_{i=1}^{n} e_i^2 = e^T e \tag{3.25}$$

Notice that using λ = 1 reduces the cost function in (3.24) to the standard cost function described by (3.25). As a consequence of using the modified cost function, the standard RLS algorithm gets transformed into a new algorithm that is capable of tracking possible parameter changes.

The modified RLS algorithm becomes

$$\begin{aligned} W(n) &= W(n-1) + k(n)\xi(n) \\ \xi(n) &= d(n) - x^T(n)W(n-1) \\ k(n) &= \frac{S(n-1)x(n)}{\lambda + x^T(n)S(n-1)x(n)} \\ S(n) &= \lambda^{-1}\left[S(n-1) - k(n)x^T(n)S(n-1)\right] \end{aligned} \tag{3.26}$$

In stationary environments, the parameter vector can be estimated with a forgetting factor λ = 1, thus, emphasizing the fact the all data samples is of equal importance in the estimation process. In fact, setting the forgetting factor to one reduces the modified RLS algorithm in (3.23) to the standard RLS algorithm which is known to have optimal properties in stationary environments. In nonstationary environments (like speech signal), however, λ is required to be smaller than unity to track any parameter changes. This means that only the most recent data is to be used in the estimation algorithm.

The advantage of doing so is that the algorithm is always in a state of alert against possible parameter variations and this improves the tracking capability of the algorithm. The disadvantage, on the other hand, is that the algorithm becomes sensitive to noise

because of the fact that the adaptation gain never converges to zero. One solution that has been proposed to deal with this problem sets the value of $\lambda$ to some number smaller than one when an abrupt parameter change is detected; then it increases $\lambda$ to unity [31] [32]. This procedure allows the estimated parameter vector to converge to the true value. The net result is an algorithm with good parameter tracking capability during the transient stages, and at the same time the adaptation gain is allowed to tend to zero during the stationary stages and hence noise sensitivity is greatly reduced.

Chapter 4

# RESULTS AND OBSERVATIONS FOR FORMANT TRACKING ALGORITHMS

## 4.1 Introduction

The primary goal of formant tracking algorithms is to develop a reliable formant tracking algorithm that is robust in real-time noise scenarios. Different test cases are described and the performance of the algorithms under these conditions has been discussed. Both of the formant tracking algorithms has been tested using synthesized speech signals as well as speech signals from the TIMIT recorded speech database. The TIMIT corpus of read speech has been designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems. TIMIT has resulted from the joint efforts of the Massachusetts Institute of Technology (MIT) and Texas Instruments (TI).

Testing using synthesized sentences allows quantitative analysis of the performance of the formant tracker because the formant frequency values of the synthesized speech signals are known. The TIMIT database speech signals are recorded from actual speakers and therefore sound more natural than the synthesized speech signals. However, the actual formant frequency values of the TIMIT database speech signals are unknown, therefore; only qualitative analysis of the results can be performed through visual inspection. Algorithms are tested for both male and female voices. Table 4.1 shows the range of first four formant frequencies. Generally the estimated formant frequencies vary in these ranges.

| Formants | Minimum frequency (Hz) | Maximum frequency (Hz) |
|---|---|---|
| $F_1$ | 270 | 730 |
| $F_2$ | 840 | 2290 |
| $F_3$ | 1690 | 3010 |
| $F_4$ | 2500 | 4500 |

**Table 4.1 Ranges for formant frequencies**.

## 4.2 Testing With Robust Formant Tracking Algorithm

Robust formant tracking algorithm estimate and track the formant frequencies of a speech signal. The present work shows the testing of robust tracking algorithm for a speech signal in different environmental conditions and for both synthesized and TIMIT data based speakers.

### 4.2.1 Testing with White Noise

Additive White Gaussian Noise (AWGN) may be present in real-time environments from a variety of sources such as fans, air-conditioners, running water, etc. Since the formant tracker is to be implemented and used in a real-time environment, it must be able to operate in AWGN. The operation of the algorithm is tested and analyzed in the presence of background AWGN at various Signal-to-Noise Ratios (SNR's), from 40 dB to -5 dB, for various synthesized and TIMIT database speech signals (for both male and female speakers). AWGN adds wideband spectral noise to each of the four formant bands and the long-time average energy added to each of the bands is roughly equal. Due to the equal energy contribution of AWGN on the formant frequency bands and the nature of the formant tracker, the performance of the formant tracking filters should not be affected greatly in AWGN for voiced segments of speech. However, the performance of the voicing detector and the pitch will both be adversely affected due to AWGN at low SNR's because of the added energy in the lower frequency bands. The voicing detector in particular may erroneously detect voicing during unvoiced segments of speech. The addition of the autocorrelation based testing as well as the adaptive energy thresholds in the voicing detector prevents this from occurring.

Figure 4-1 shows the spectrogram of a male synthesized speaker saying "five women played basketball". Figure 4.2 shows the formant frequencies for female speaker saying "five women played basketball". In these speech samples there is no AWGN. Figure 4.3 shows the formant frequencies for male speaker with added AWGN 40dB. Figure 4.4 shows the formant frequencies for female speaker with added AWGN 40dB. The figure 4.3 and 4.4 also show the original formant frequencies (plotted in black), the estimated formant frequencies (plotted in white) as well as the voicing decisions (plotted in

magenta). It can be seen that at this high SNR level the formant frequencies are estimated accurately and the voicing detector estimates detects voicing accurately. As the formant frequencies change, the formant tracker is able to follow them and capture the formant frequency transitions.

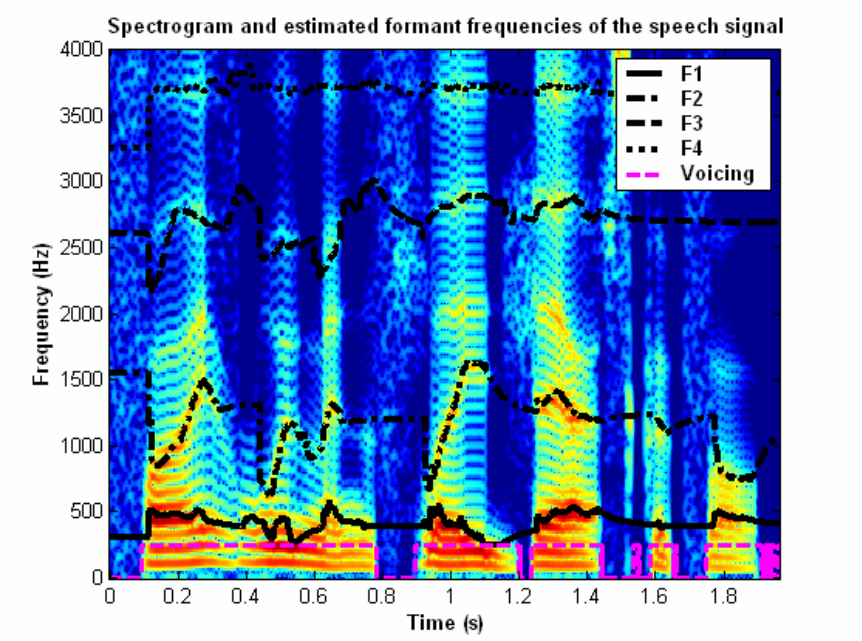



**Fig 4.1 Spectrogram and formant frequencies for a synthesized male speaker "five women played basketball"**

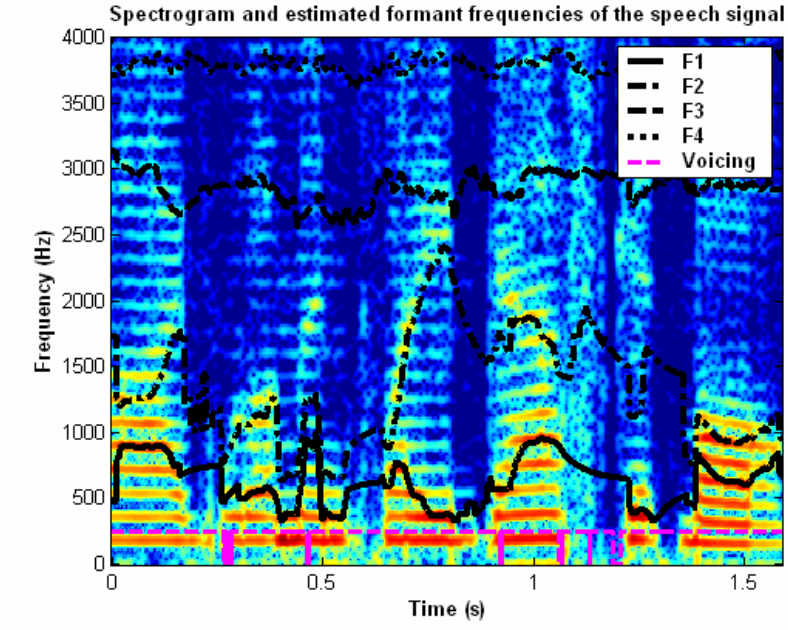


**Fig 4.2 Spectrogram and formant frequencies for a synthesized female speaker "five women played basketball"**

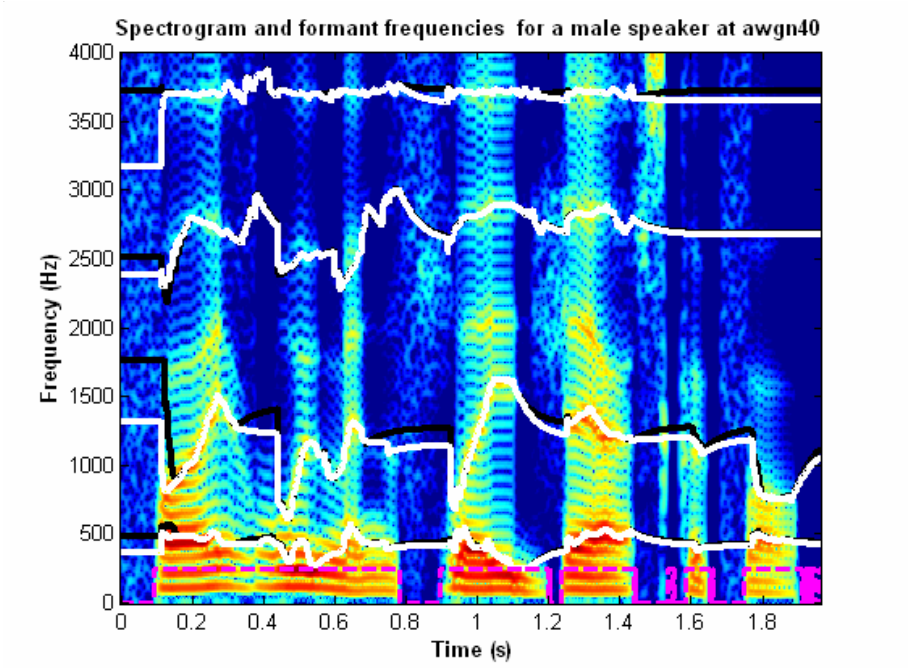



**Fig 4.3 Spectrogram and formant frequencies for a synthesized male speaker with AWGN 40dB**

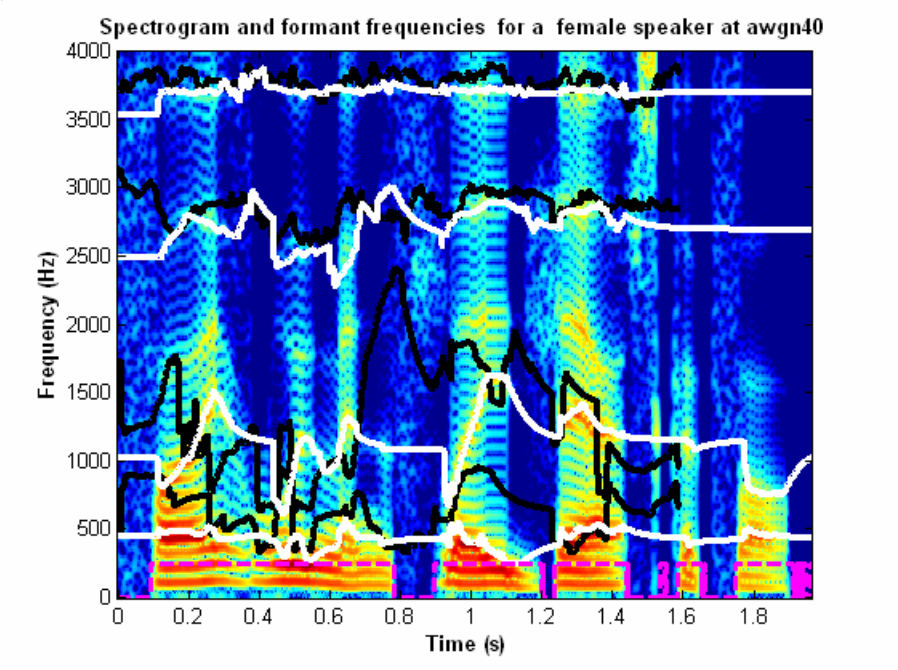



**Fig 4.4 Spectrogram and formant frequencies for a synthesized female speaker with AWGN 40dB**

Figures 4.3 and 4.4 show that at given SNR's the formant tracker estimates the second and third formant frequencies better for the male speaker than for the female speaker. This happens because the second and third formant frequency values of the female speaker have very fast transitions during some phoneme boundaries and the energy of the formant frequency regions drop significantly during these transitions.

The formant tracker is unable to keep track of the formant frequencies during these fast transitions and the algorithm reverts to using the moving average value of the second and third formant frequencies. However, the algorithm recovers quickly and starts tracking the correct formant frequency when as soon as there is sufficient energy present in the formant regions. The algorithm estimates the first formant frequencies quite accurately for both males and females speakers during all voiced speech segments. In totality, the formant frequencies are estimated accurately and the algorithm is robust. The voicing detector performs well in predicting the voiced segments of speech for both male and female speakers.

Figure 4.5 shows the spectrogram and formant frequencies of synthesized male speaker with AWGN 5dB, Figure 4.6 shows the same at AWGN 10dB. Figure 4.7 shows the spectrogram and formant frequencies of synthesized male speaker with AWGN 20dB.

Figure 4.8 shows the spectrogram and formant frequencies of synthesized male speaker with AWGN 30dB, Figure 4.9 shows the spectrogram and formant frequencies of synthesized female speaker with AWGN 5dB. Figure 4.10 shows the spectrogram and formant frequencies of synthesized female speaker with AWGN 10dB. Figure 4.11 shows the spectrogram and formant frequencies of synthesized female speaker with AWGN 20dB. Figure 4.12 shows the spectrogram and formant frequencies of synthesized female speaker with AWGN 30dB. From figures 4.5-4.12 it is clear that even at low SNR's this algorithm track formants accurately. Table 4.2 and Table 4.3 give the estimated formant frequencies for both male and female speaker with added AWGN at different SNR's respectively. Actual frequencies are also shown in these tables. So both estimated and actual frequencies can be easily compared.

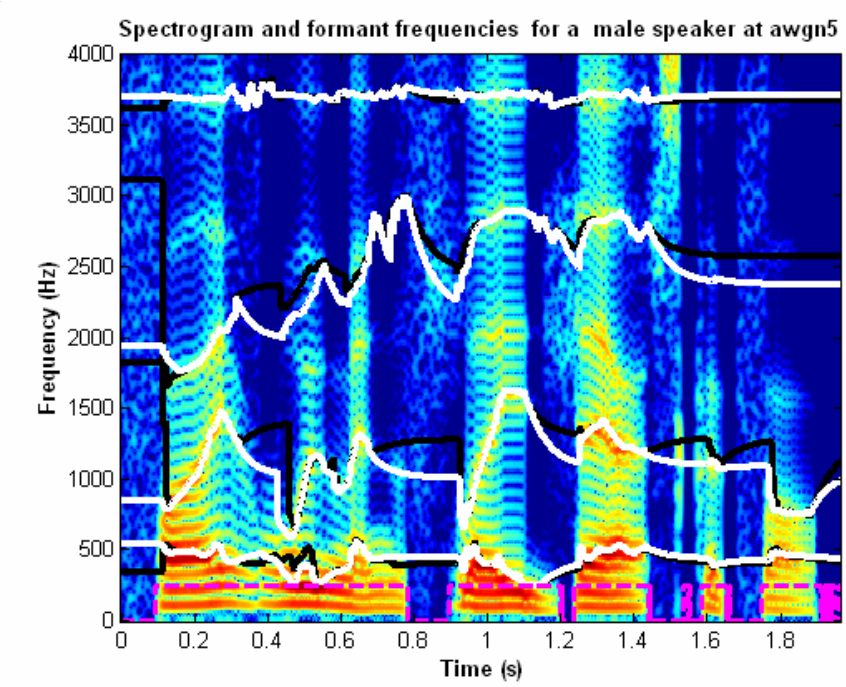



**Fig 4.5 Spectrogram and formant frequencies for a synthesized male speaker with AWGN 5dB**

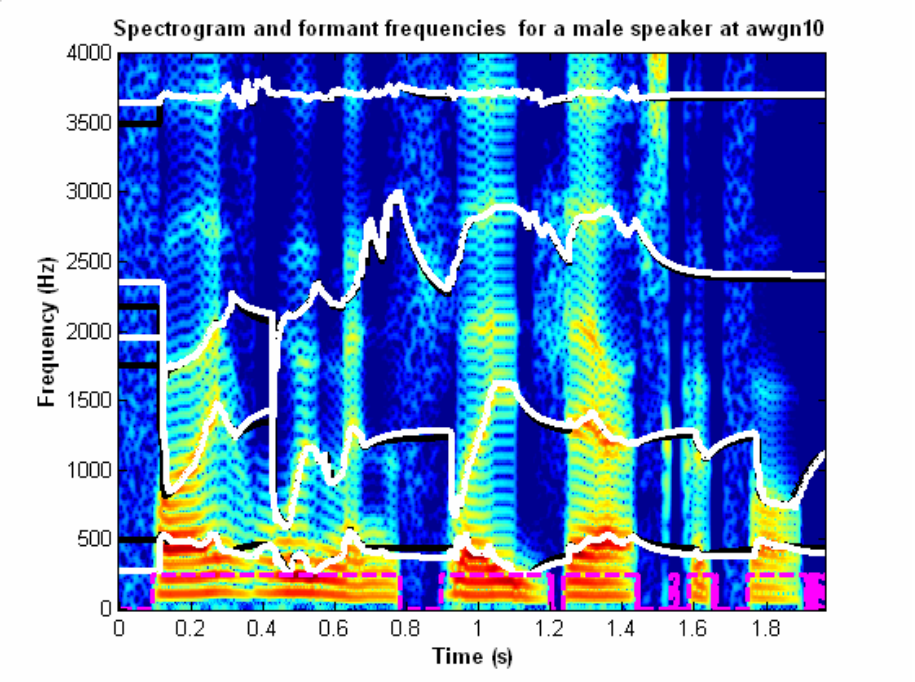



**Fig 4.6 Spectrogram and formant frequencies for a synthesized male speaker with AWGN 10dB**

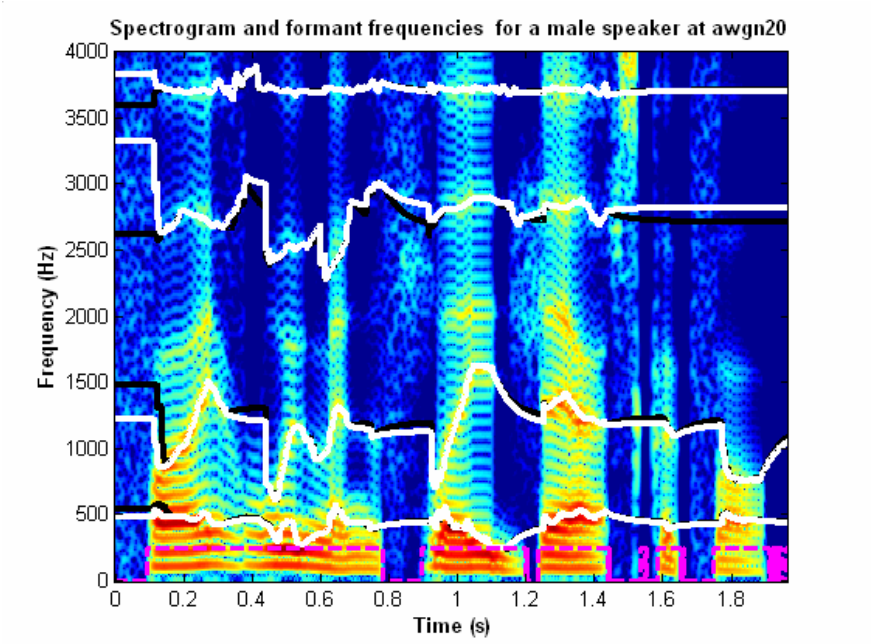



**Fig 4.7 Spectrogram and formant frequencies for a synthesized male speaker with AWGN 20dB**

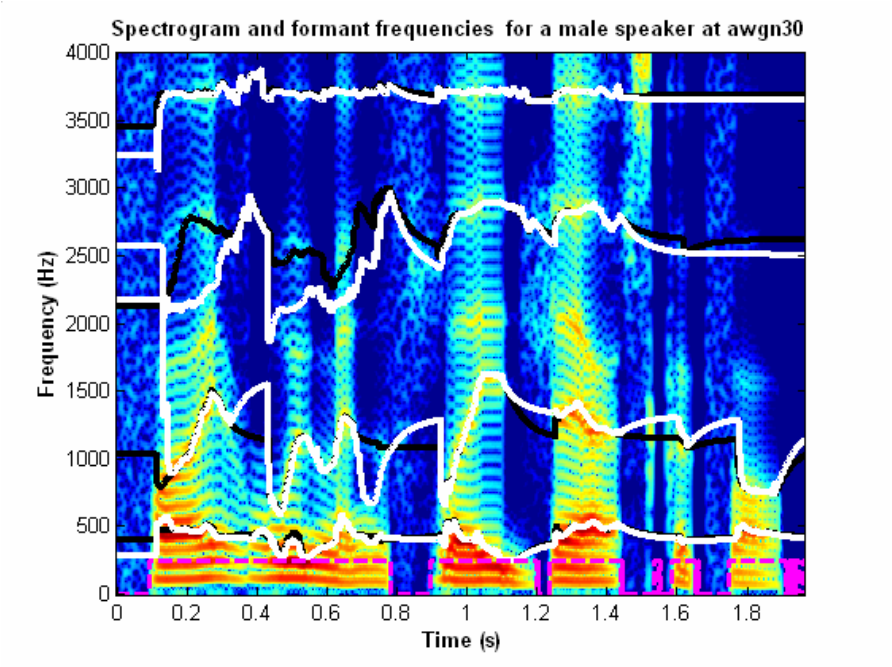



**Fig 4.8 Spectrogram and formant frequencies for a synthesized male speaker with AWGN 30dB**

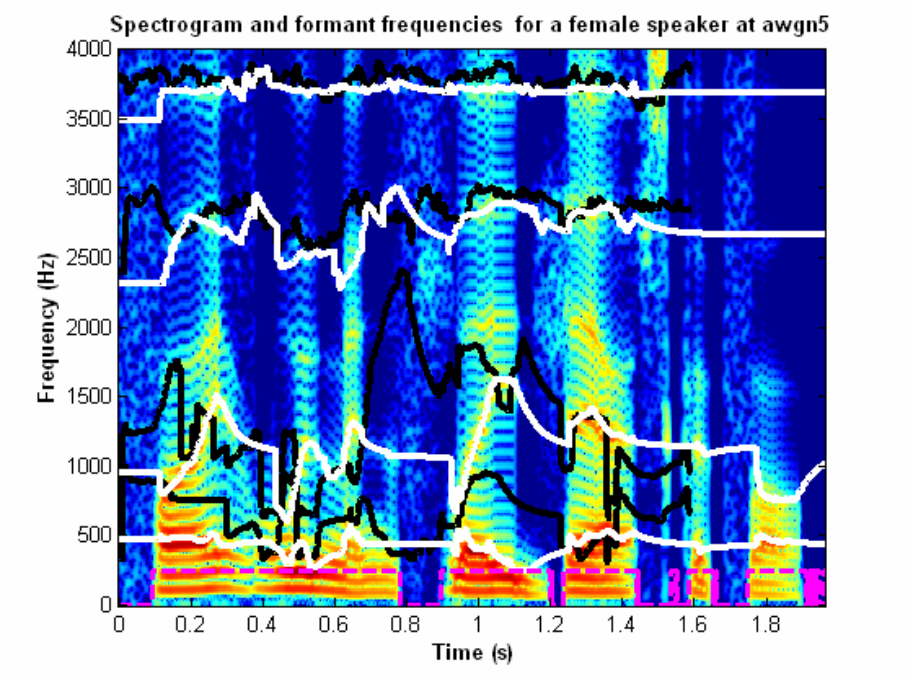



**Fig 4.9 Spectrogram and formant frequencies for a synthesized female speaker with AWGN 5dB**

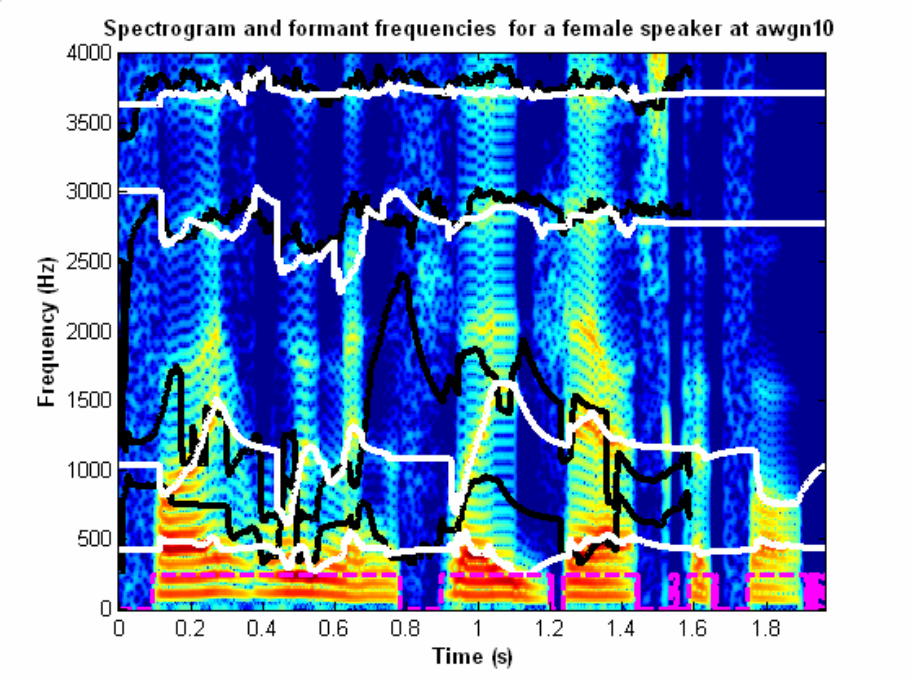



**Fig 4.10 Spectrogram and formant frequencies for a synthesized female speaker with AWGN 10dB**

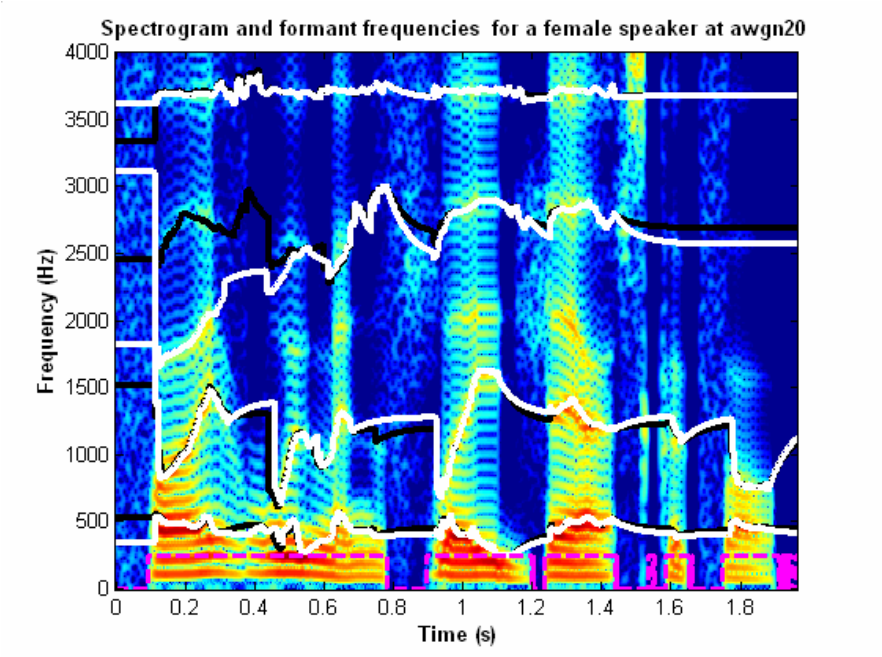



**Fig 4.11 Spectrogram and formant frequencies for a synthesized female speaker with AWGN 20dB**

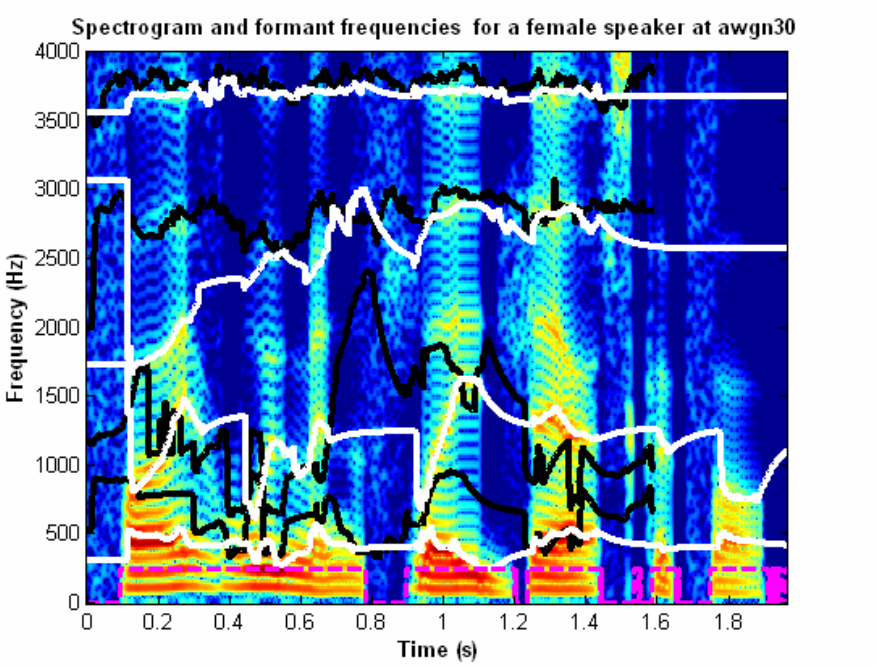



**Fig 4.12 Spectrogram and formant frequencies for a synthesized female speaker with AWGN 30dB**

| Frequencies To be Estimated (Hz) | Value of Actual frequencies (Hz) | Estimated frequencies with a tracking algorithm (Hz) | | | | |
|---|---|---|---|---|---|---|
| | | With SNR 40dB | With SNR 30dB | With SNR 20dB | With SNR 10dB | With SNR 5dB |
| $F_1$ | 700 | 400 | 300 | 500 | 400 | 600 |
| $F_2$ | 1500 | 1400 | 2200 | 1400 | 2000 | 1000 |
| $F_3$ | 2200 | 2400 | 2500 | 2600 | 2400 | 2000 |
| $F_4$ | 3500 | 3200 | 3300 | 3800 | 3600 | 3700 |

**Table 4.2 Table for estimated formant frequencies for male speaker at different SNR's "five woman played basketball"**

| Frequencies To be Estimated (Hz) | Value of Actual frequencies (Hz) | Estimated frequencies with a tracking algorithm (Hz) | | | | |
|---|---|---|---|---|---|---|
| | | With SNR 40dB | With SNR 30dB | With SNR 20dB | With SNR 10dB | With SNR 5dB |
| $F_1$ | 700 | 500 | 400 | 400 | 500 | 500 |
| $F_2$ | 1500 | 1000 | 1100 | 1800 | 1700 | 1000 |
| $F_3$ | 2200 | 2500 | 3100 | 3100 | 3000 | 2400 |
| $F_4$ | 3500 | 3500 | 3500 | 3600 | 3600 | 3500 |

**Table 4.3 Table for estimated formant frequencies for female speaker at different SNR's "five woman played basketball"**

Tables 4.2 and 4.3 give the estimated formant frequencies for both male and female speakers at different SNR's obtained from the robust formant tracking algorithm. By comparing the estimated formant values with the actual formant values it can be seen that frequencies are estimated accurately. Although they differ from actual frequencies, but are within the frequency range as shown in table 4.1.

The algorithm was also tested using recorded natural speech for both male and female speakers from the TIMIT database. Figure 4.13 shows the spectrogram and the estimated formant frequencies for a natural male speaker from the TIMIT database saying "fifth yard contains big juicy peaches". From the spectrogram it can be visually observed the formant tracker is able to detect and track the formant frequencies relatively well and also makes good voicing decisions. The formant frequency transitions are also captured well by the algorithm.

Figure 4.14 shows the spectrogram and the estimated formant frequencies for a natural male speaker from the TIMIT database saying "fifth yard contains big juicy peaches" with AWGN at SNR of 25dB. Figure 4.15 shows the spectrogram and the estimated formant frequencies for a natural male speaker from the TIMIT database saying "a book of scholars". Figure 4.16 shows the spectrogram and the estimated formant frequencies for a natural male speaker in the presence of background AWGN at a SNR of 30 dB.

From the spectrograms of the following figures it can be visually observed the formant tracker is able to detect and track the formant frequencies well and also makes good voicing decisions. The formant frequency transitions are also captured well by the algorithm.

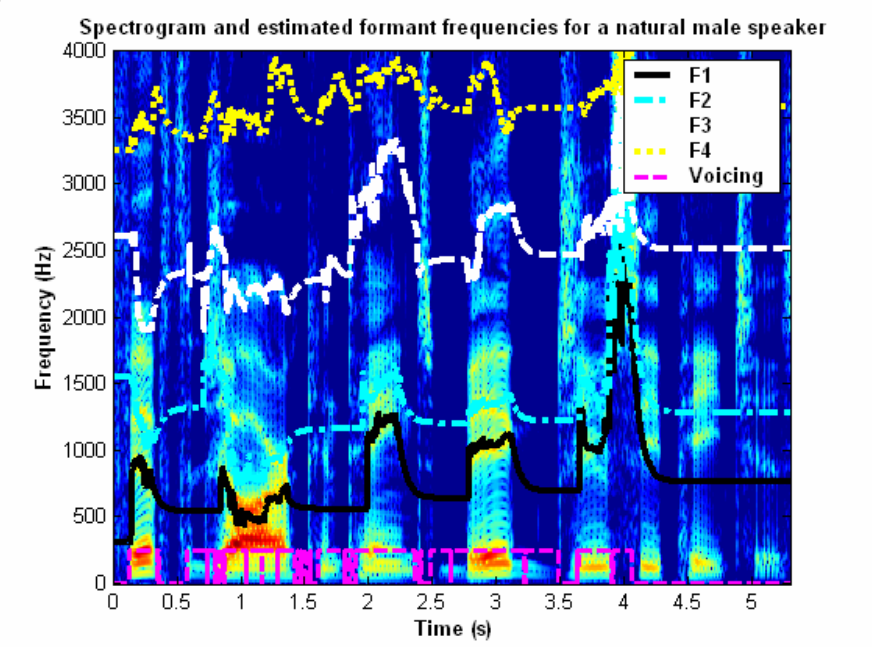



**Fig 4.13 Spectrogram and formant frequencies for a natural male speaker "fifth yard contains big juicy peaches"**

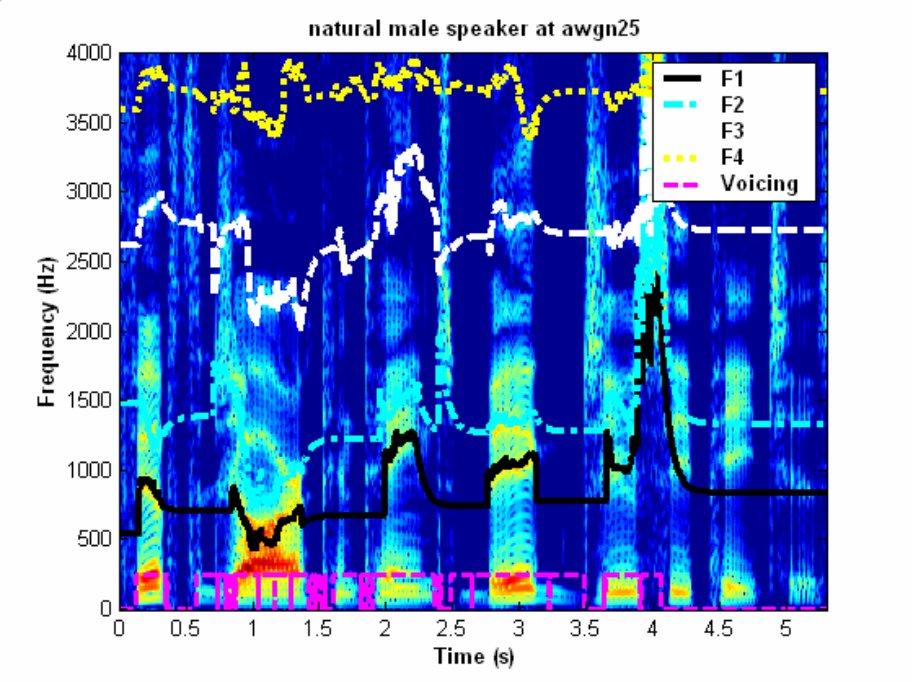



**Fig 4.14 Spectrogram and formant frequencies for a natural male speaker at AWGN 25 dB**

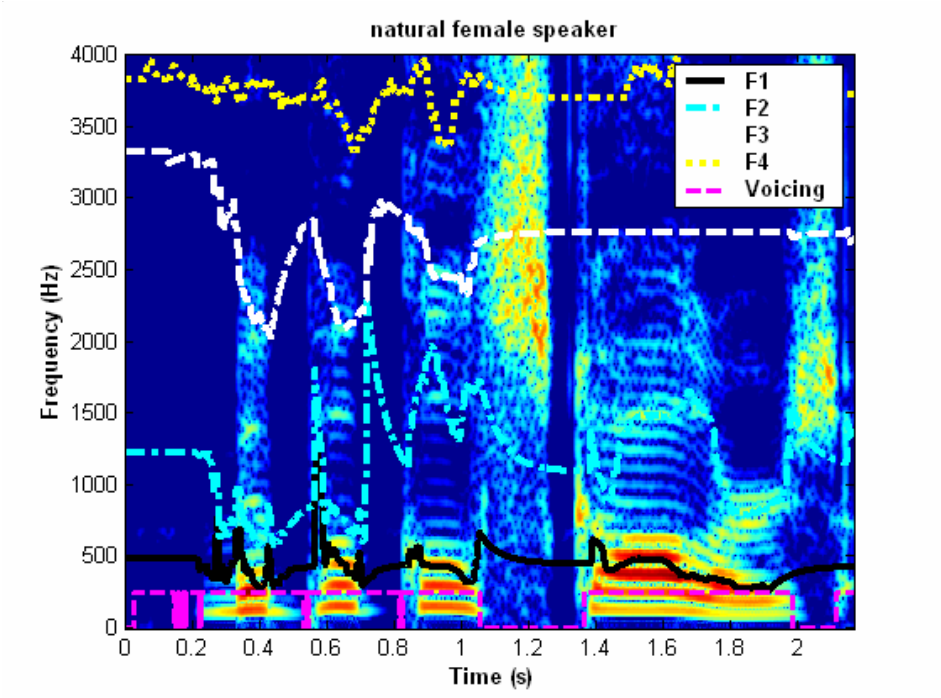



**Fig 4.15 Spectrogram and formant frequencies for a natural female speaker "a book of scholars"**

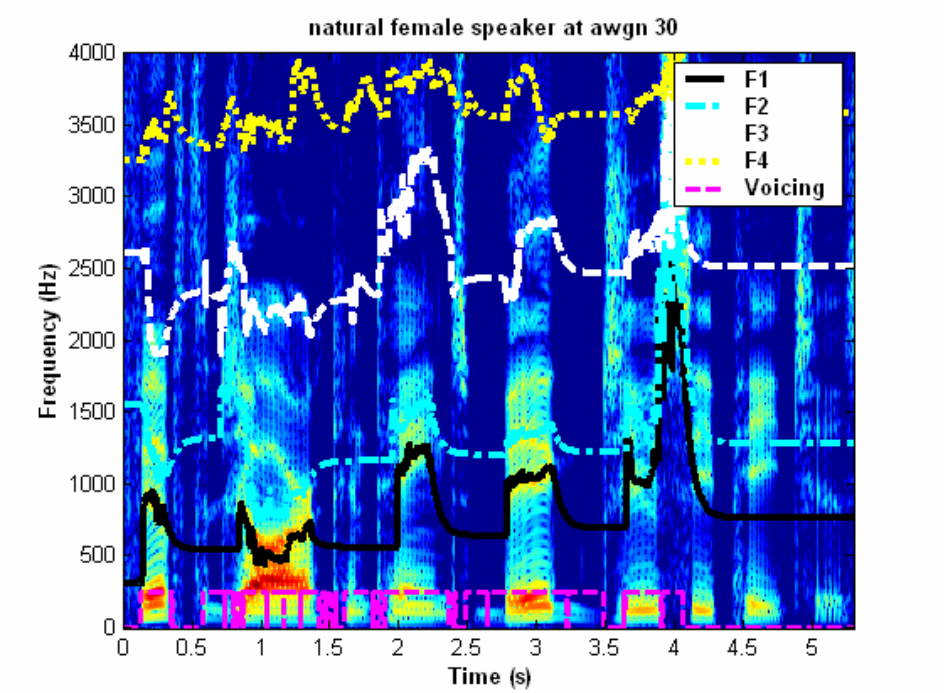



**Fig 4.16 Spectrogram and formant frequencies for a natural female speaker at AWGN 30dB**

### 4.2.2 Testing in the presence of a background speaker

It has been observed that in real-life, there is often more than just one speaker present in an environment. The algorithm was also tested for the environment in which background speaker is present by estimating formant frequencies for the dominant speaker in the presence the background speakers.

Different cases are considered here. Testing is done with the female background speaker, with male background speaker, with multiple background speakers. Here the background speaker serves as the 'noise source'. The loudness of the background speaker often varies in real-life and therefore the algorithm is tested at varying SNR's. In some cases for a particular time period the background speaker may contribute significant energy to the formant frequency regions of the primary speaker, especially at lower SNR's. This will cause the algorithm to start tracking the formant frequencies of the background speaker instead of those of the primary (more dominant) speaker.

There are short moments of silence during the speech of any speaker while the speaker inhales, or exhales, and during phoneme transitions etc. Another source of concern when there are background speakers present is that if the background speaker says something during the brief moments of silence of the primary speaker, the formant tracking algorithm may start to track the formant frequencies of the background speaker. In this case the formant frequencies estimated will switch back and forth between those of the primary and the background speakers. Another point to keep in mind is that the 'noise source' is a female speaker and this can lead to one of two scenarios when the primary speaker is male. The formant frequencies of the background female speaker are higher than those of the primary male speaker. This may lead the overall performance of the algorithm to be better for male speakers, because there will be less energy contribution to the male speakers' formant frequency regions.

#### 4.2.2.1 Testing in the presence of a female single background speaker

First the algorithm has been tested in the presence of female single background speaker. Figure 4.17 shows the spectrogram of a synthesized male speaker saying "five women played basketball" in the presence of a female single background speaker saying "what are you doing" at an SNR of 30dB. It can be seen that at this high SNR the formant frequencies are estimated fairly accurately for most of the speech signal. Figure 4.18 shows the same at low SNR of 5dB

Figure 4.19 shows the spectrogram of a synthesized female speaker saying "five women played basketball" in the presence of a female single background speaker saying "what are you doing" at SNR 20dB. Figure 4.20 shows the same at low SNR of 5dB. It can be seen that at both low and high SNR's, formant frequencies are estimated fairly accurately.

Table 4.4 gives the estimated formant frequencies for both male and female speaker with single background female speaker. From the table 4.4 it is clear that the algorithm estimates the formant frequencies well in the presence of single background female speaker. Values of formant frequencies deviate very less from their actual values. Formant frequency values are also within the standard frequency range as shown in table 4.1.

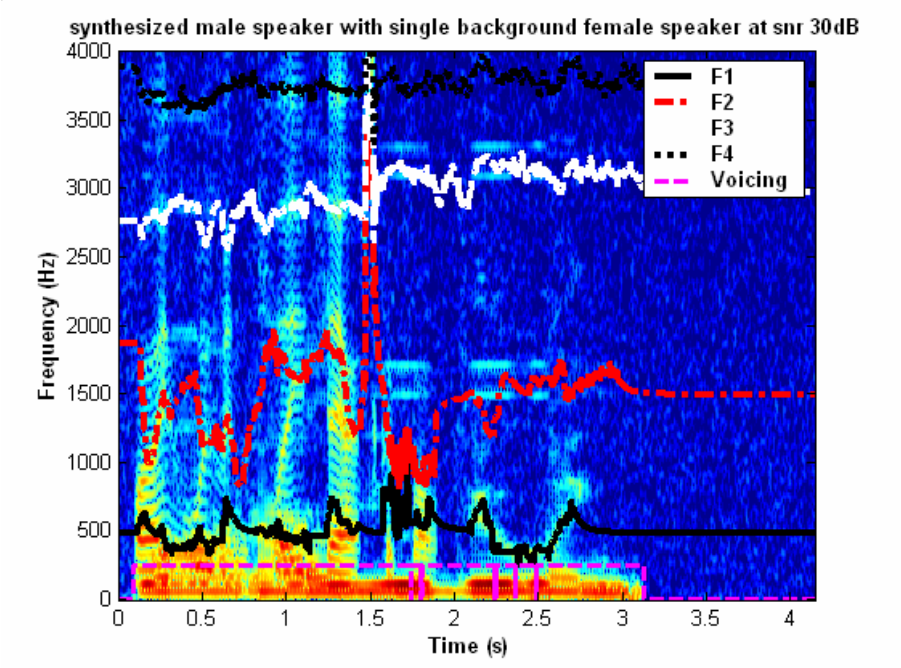



**Fig.4.17 Spectrogram of a synthesized male speaker in the presence of female single background speaker at 30 dB SNR**

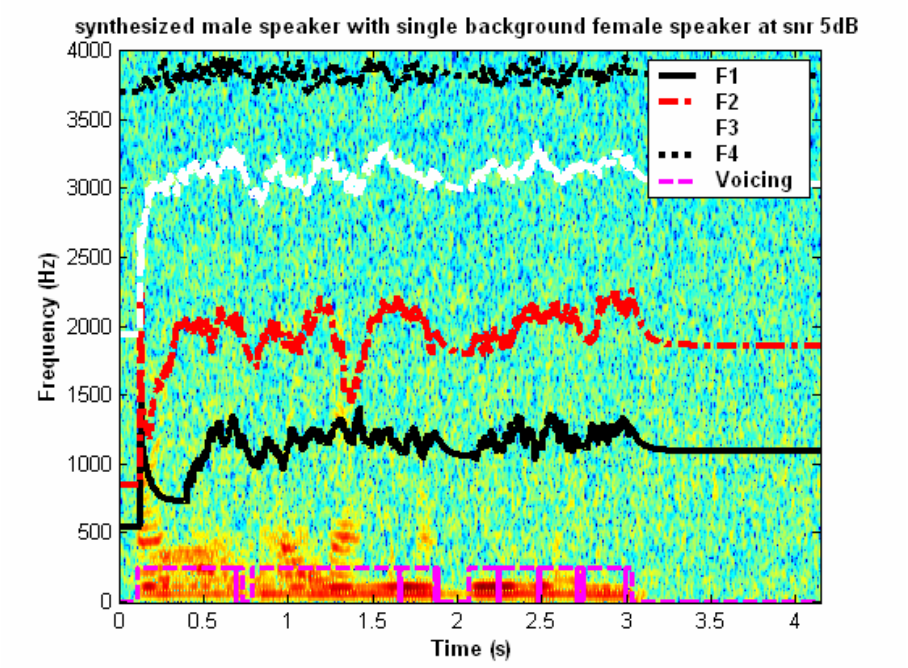


**Fig.4.18 Spectrogram of a synthesized male speaker in the presence of female single background speaker at 5 dB SNR**

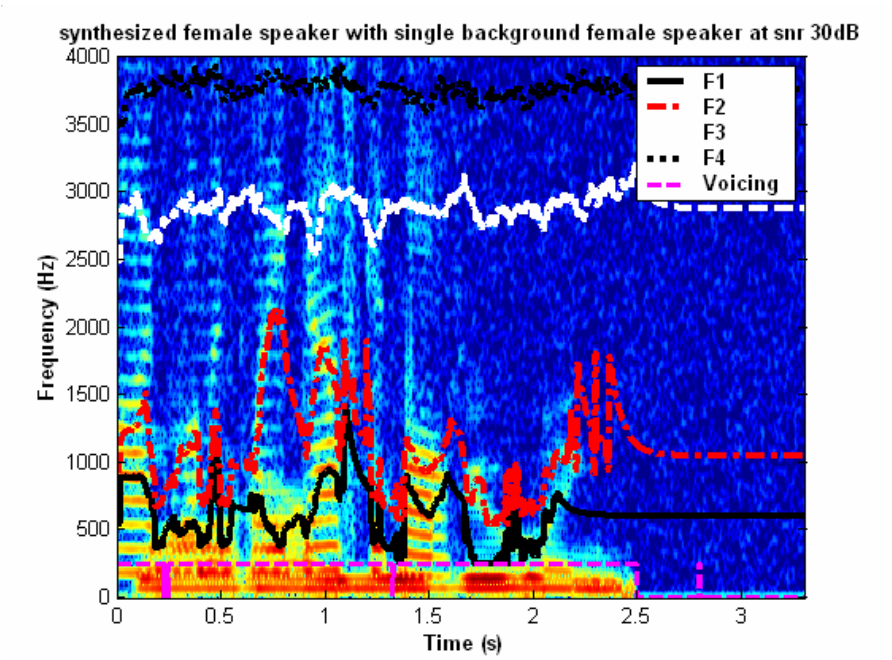



**Fig 4.19 Spectrogram of a synthesized female speaker in the presence of female single background speaker at SNR 30dB**

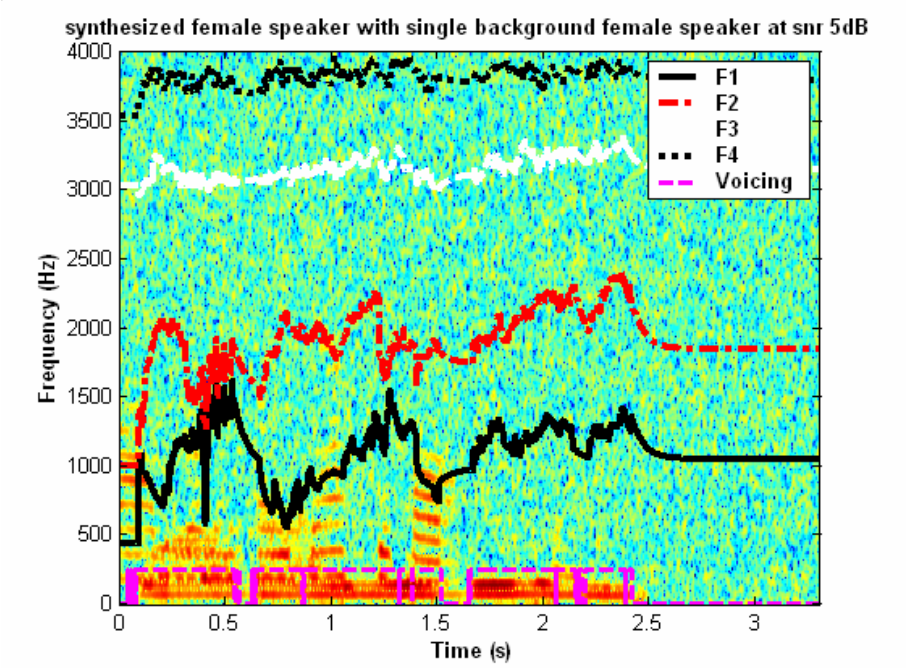


**Fig 4.20 Spectrogram of a synthesized female speaker in the presence of female single background speaker at SNR 5dB**

| Frequencies To be estimated (Hz) | Estimated frequencies for a male speaker (Hz) | | Estimated frequencies for a female speaker (Hz) | |
|---|---|---|---|---|
| | With SNR 30dB | With SNR 5dB | With SNR 30dB | With SNR 5dB |
| $F_1$ | 500 | 500 | 500 | 400 |
| $F_2$ | 1800 | 900 | 1000 | 1000 |
| $F_3$ | 2700 | 2000 | 2500 | 3000 |
| $F_4$ | 3700 | 3800 | 3500 | 3500 |

**Table 4.4 Table for estimated formant frequencies for male and female speakers with single female background speaker at different SNR's**

The algorithm is also tested using more natural sounding speech (for both male and female speakers) from the TIMIT database at various SNR's. A figure 4.21 shows the portion of the spectrogram for natural female speakers saying "a book of scholars" in the presence of a female single background speaker "what are you doing" at 20 dB. Figure 4.22 shows the portion of the spectrogram for natural male speaker saying "fifth yard contains big juicy peaches" in the presence of a female single background speaker saying "what are you doing" at 10 dB.

From visual inspection of these spectrograms, it can be seen that the algorithm performs well for both genders despite the relatively low SNR's. The algorithm is also able to track formant frequencies as the speech switches between voiced and unvoiced segments and provides smooth formant frequency estimates. Testing has been done at different values of SNR ranging from 40dB to -5dB. But results for few values are presented here. Algorithm tracks frequencies accurately for entire range of SNR's.

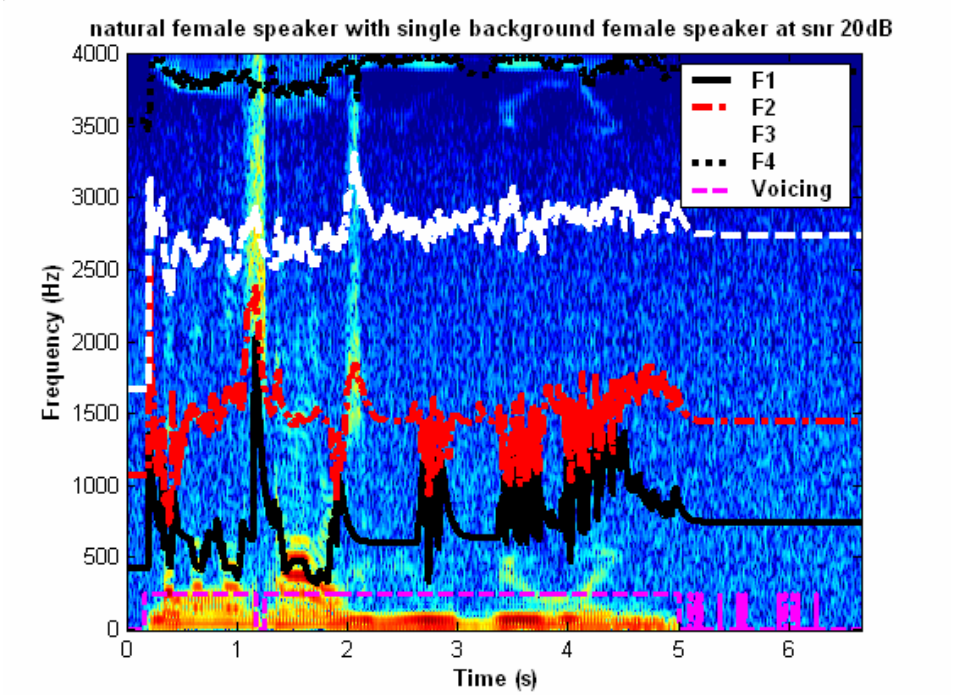



**Fig 4.21 Spectrogram of a natural female speaker in the presence of female single background speaker at SNR 20dB**

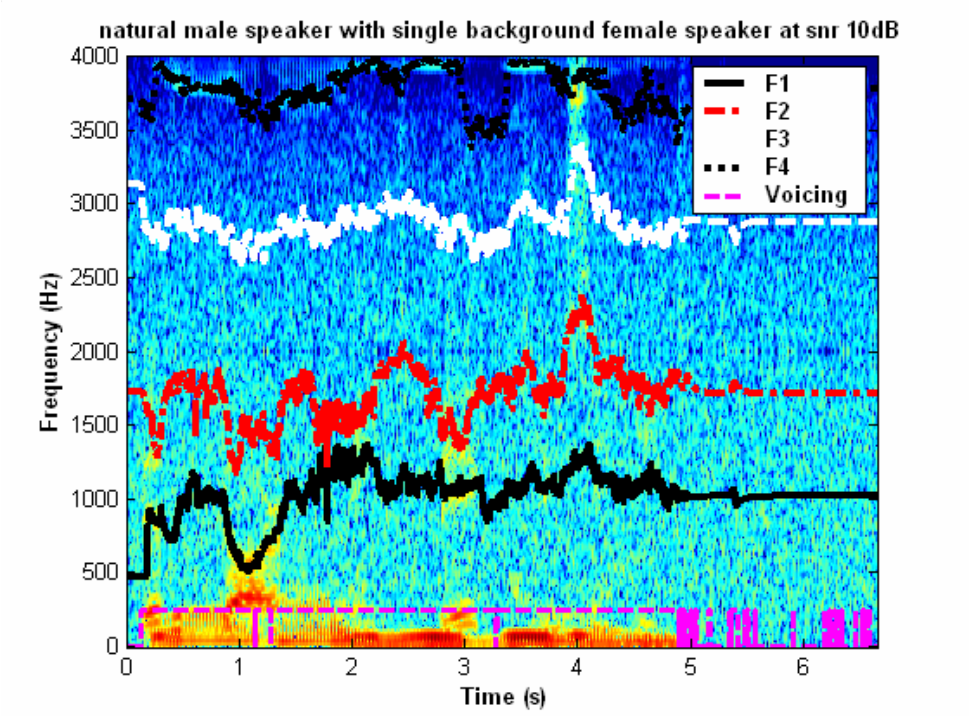


**Fig 4.22 Spectrogram of a natural male speaker in the presence of female single background speaker at SNR 10dB**

#### 4.2.2.2 Testing in the presence of a male single background speaker

Algorithm has also been tested in the presence of a male single background speaker at varying SNR's (from 40 dB to -5 dB). Algorithm again tracks very well for the whole range of SNR's. But the results for few values have been presented here. Concerns still remain regarding the algorithm starting to track the formant frequencies of the background speaker instead of the primary speaker at low SNR's. Similar to the female single background speaker case, the estimated formant frequencies can still switch back and forth between those of the primary speaker and the background speaker due to the noise contributions from the background speaker during momentary periods of silence of the primary speaker.

Figure 4.23 shows the spectrogram and estimated formant frequencies for a male speaker saying "five woman played basketball" in the presence of single background male speaker saying "what are you doing" at SNR of 15dB. Figure 4.24 shows the spectrogram and estimated formant frequencies for a male speaker saying "five woman played basketball" in the presence of single background male speaker saying "what are you doing" at SNR of 35dB.

Figure 4.25 shows the spectrogram and estimated formant frequencies for a female speaker saying "five woman played basketball" in the presence of single background male speaker saying "what are you doing" at SNR of 15dB. Figure 4.26 shows the spectrogram and estimated formant frequencies for a male speaker saying "five woman played basketball" in the presence of single background male speaker saying "what are you doing" at SNR of 15dB. Table 4.5 gives the estimated formant frequencies for both male and female speaker with single background male speaker.

Algorithm has also been tested for TIMIT database. Again it tracks frequencies accurately. Figure 4.27 shows the spectrogram and formant

frequencies for natural male speaker saying “fifth yard contains big juicy peaches” in the presence of single background speaker saying “what are you doing” at SNR of 25dB. Figure 4.28 shows the same for natural female speaker saying “a book of scholars” in the presence of single background speaker saying “what are you doing” at SNR of 25dB.

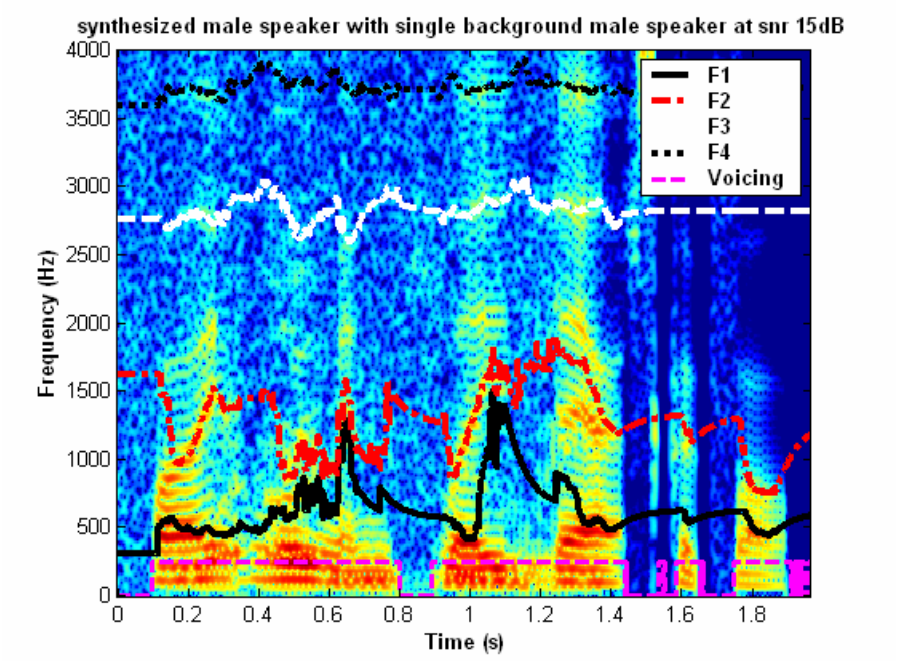



**Fig 4.23 Spectrogram of a synthesized male speaker in the presence of male single background speaker at SNR 15dB**

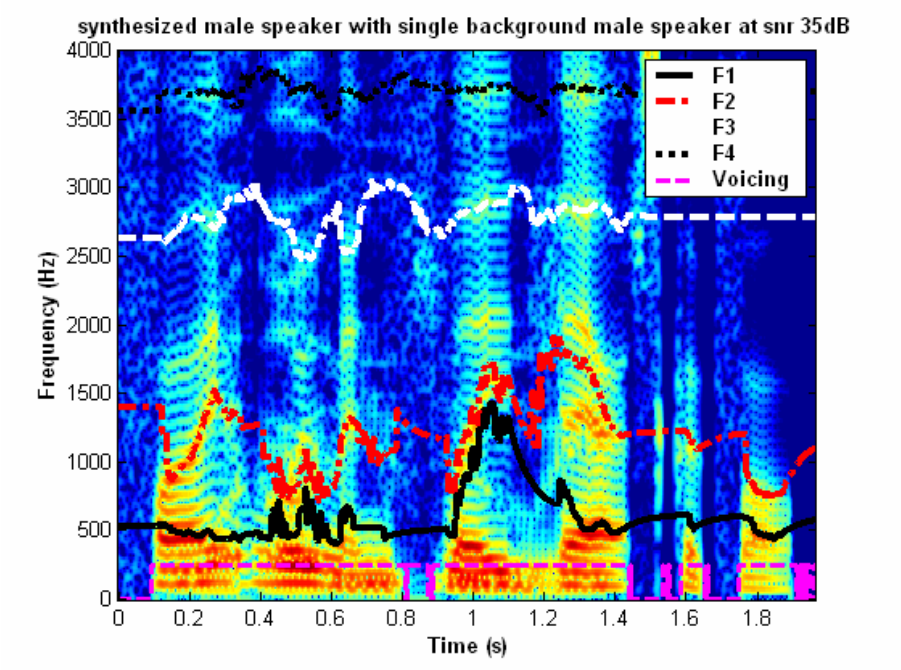



**Fig 4.24 Spectrogram of a synthesized male speaker in the presence of male single background speaker at SNR 35dB**

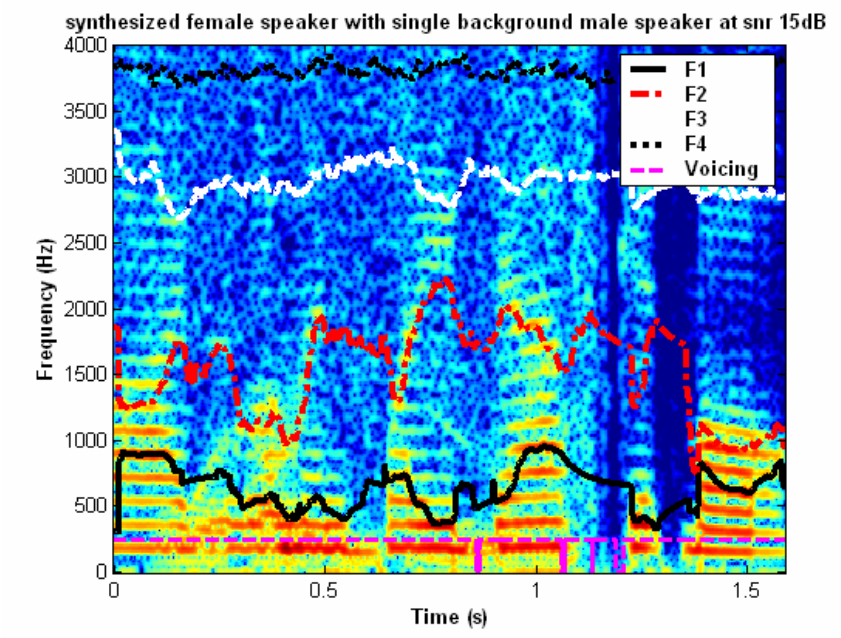



**Fig 4.25 Spectrogram of a synthesized female speaker in the presence of male single background speaker at SNR 15dB**

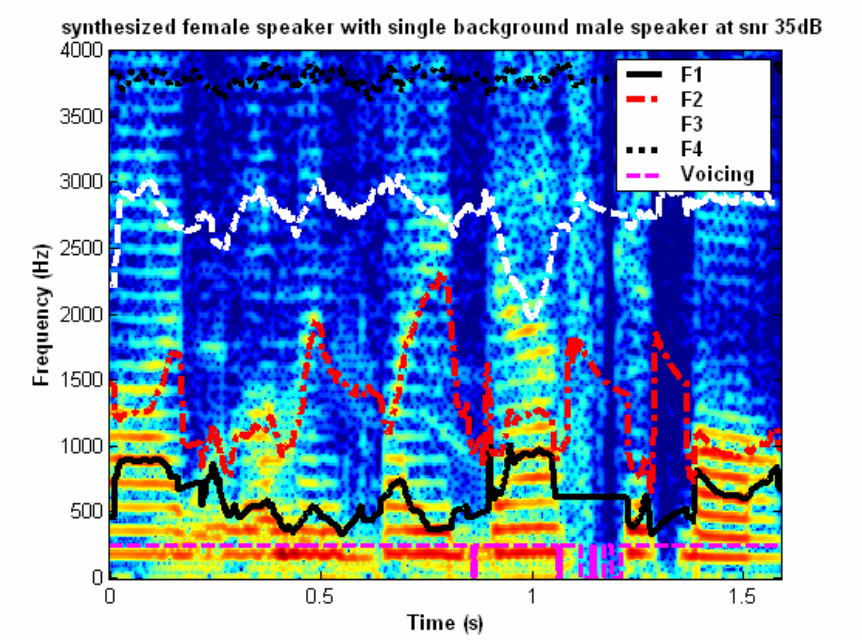



**Fig 4.26 Spectrogram of a synthesized female speaker in the presence of male single background speaker at SNR 35dB**

| Frequencies To be estimated (Hz) | Estimated frequencies for a male speaker (Hz) | | Estimated frequencies for a female speaker (Hz) | |
|---|---|---|---|---|
| | With SNR 35dB | With SNR 15dB | With SNR 35dB | With SNR 15dB |
| $F_1$ | 500 | 300 | 500 | 300 |
| $F_2$ | 1400 | 1600 | 1500 | 1800 |
| $F_3$ | 2600 | 2700 | 2200 | 3400 |
| $F_4$ | 3500 | 3600 | 3800 | 3800 |

**Table 4.5 Table for estimated formant frequencies for male and female speakers with single male background speaker at different SNR's**

Table 4.5 gives the estimated formant frequencies for both male and female speaker with single background male speaker. From the table 4.5 it is clear that the algorithm estimates the formant frequencies well in the presence of single background male speaker. Second

formant frequency is estimated very accurately. Values of formant frequencies deviate very less from their actual values. Formant frequency values are also within the standard frequency range as shown in table 4.1.

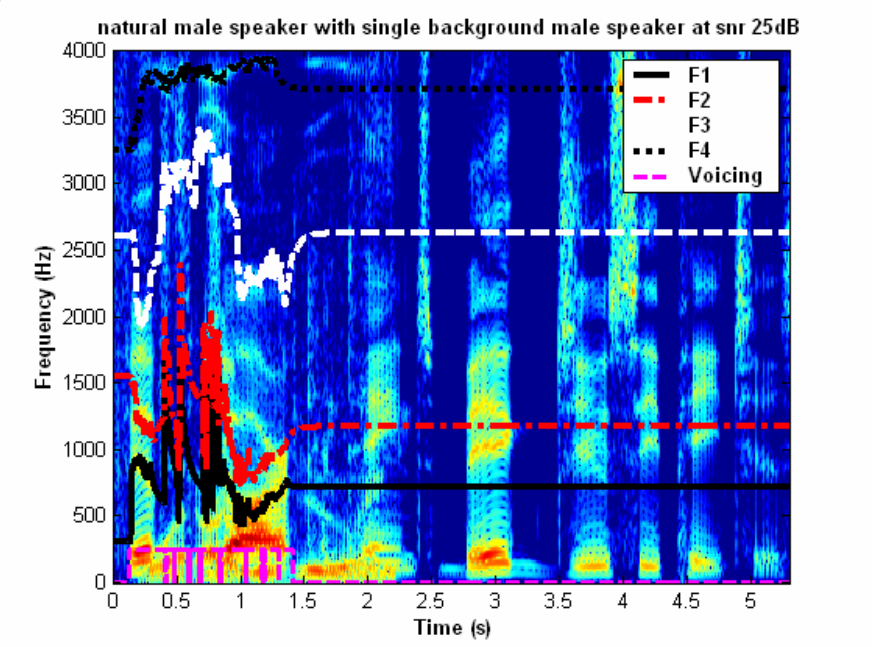



**Fig 4.27 Spectrogram of a natural male speaker in the presence of male single background speaker at SNR 25dB**

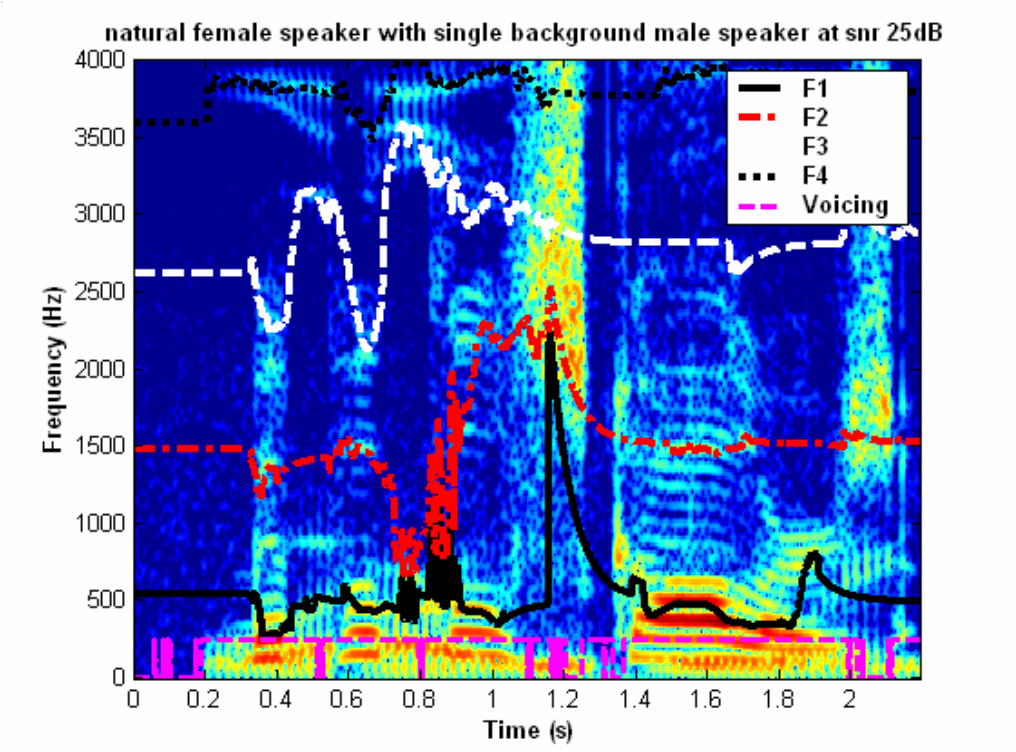



**Fig 4.28 Spectrogram of a natural female speaker in the presence of male single background speaker at SNR 25dB**

#### 4.2.2.3 Testing in the presence of multiple background speakers

In this test case the algorithm is tested using synthesized and natural male and female speakers in the presence of multiple background speakers to analyze the algorithm's behavior in a real-time environment where there are often more than just one or two speakers present in the background. Again algorithm has been tested for a wide range of SNR's but few results are presented here.

Figure 4.29 shows the spectrogram of a synthesized male speaker saying "Five women played basketball" in the presence of multiple background speakers at a SNR of 10 dB. As can be seen from the spectrogram, the algorithm estimates the formant frequencies quite well. Figure 4.30 shows the spectrogram and formant frequencies for a female speaker at SNR with multiple background speakers at SNR 20dB. Table 4.6 gives the estimated formant frequencies for both male and female speaker with multiple background speakers. From the table 4.6 it is clear that the algorithm estimates the formant frequencies accurately.

The algorithm was also tested for this test case using natural speech from the TIMIT database for both male and female speakers. The performance was similar to that in synthesized speech and the algorithm was able to track formant well.

Figure 4.31 and 4.32 show spectrograms of male and female speakers from the TIMIT database at a SNR of 20 dB respectively in the presence of multiple background speakers. The spectrograms show the formant tracker is able to estimate the first and second formant frequencies reasonably accurately for both the female and male speakers in these SNR levels.

| Frequencies To be estimated (Hz) | Estimated frequencies for a male speaker (Hz) | Estimated frequencies for a female speaker (Hz) |
|---|---|---|
| | With SNR 10dB | With SNR 20dB |
| $F_1$ | 300 | 500 |
| $F_2$ | 1800 | 1500 |
| $F_3$ | 2700 | 3000 |
| $F_4$ | 3700 | 3500 |

**Table 4.6 Table for estimated formant frequencies for male and female speakers with multiple background speakers**

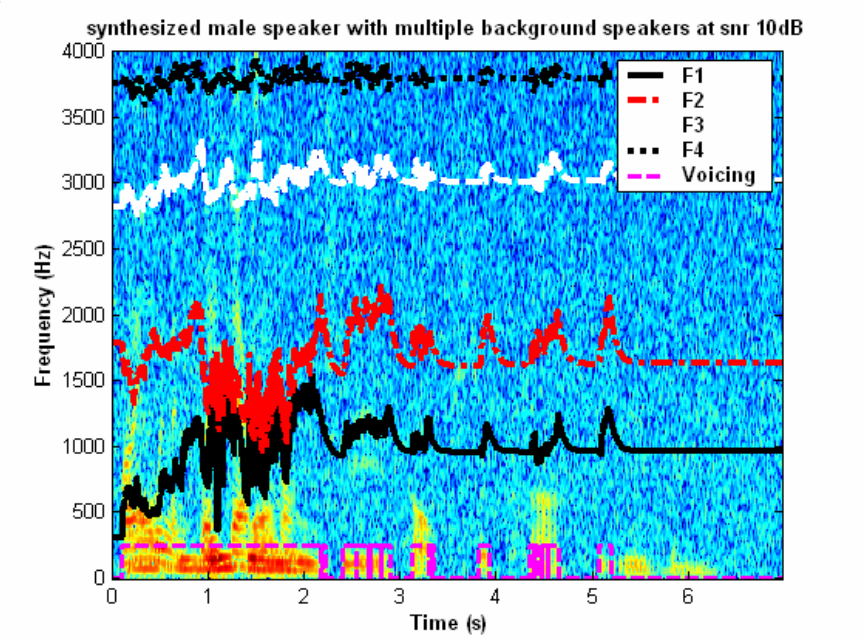



**Fig 4.29 Spectrogram of a synthesized male speaker in the presence of multiple background speakers at SNR 10dB**

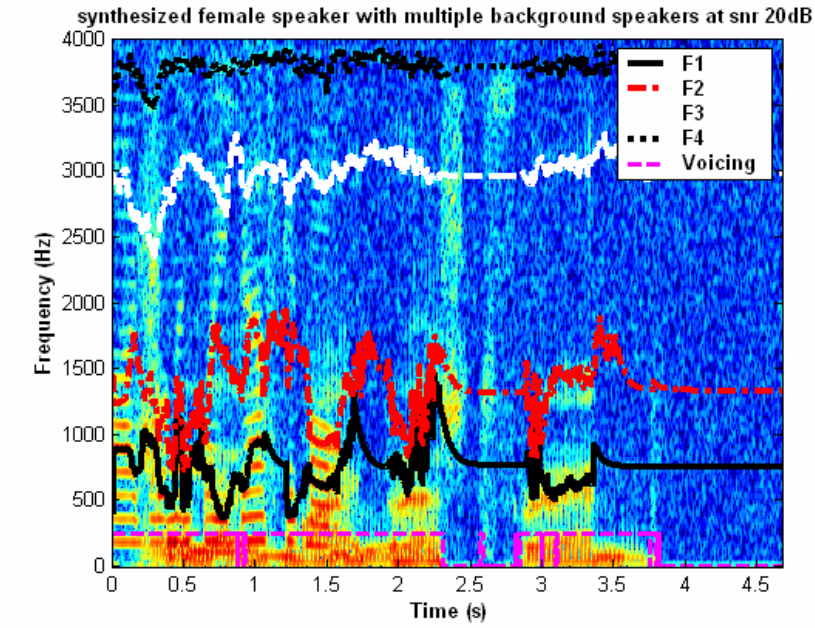



**Fig 4.30 Spectrogram of a synthesized female speaker in the presence of multiple background speakers at SNR 20dB**

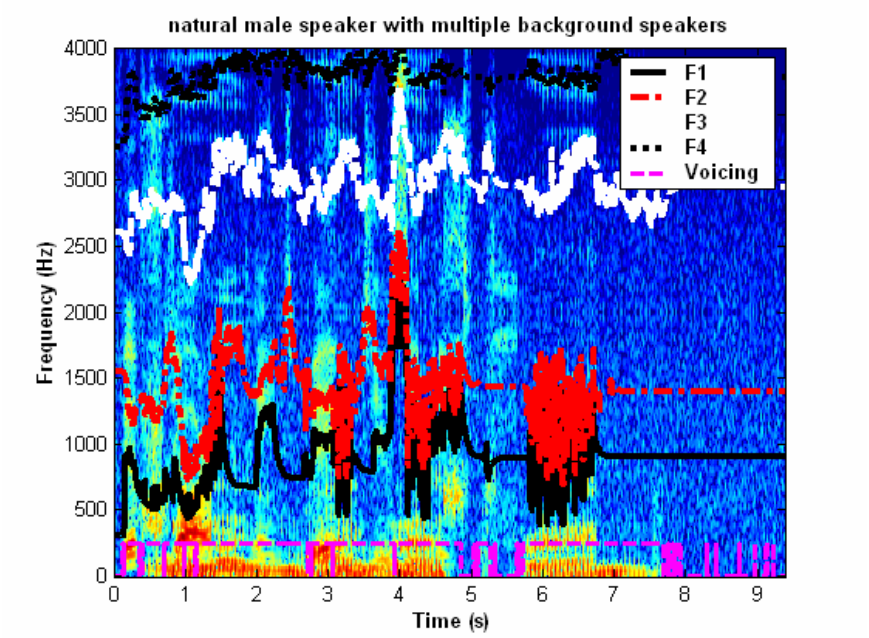



**Fig 4.31 Spectrogram of a natural male speaker in the presence of multiple background speakers at SNR 20dB**

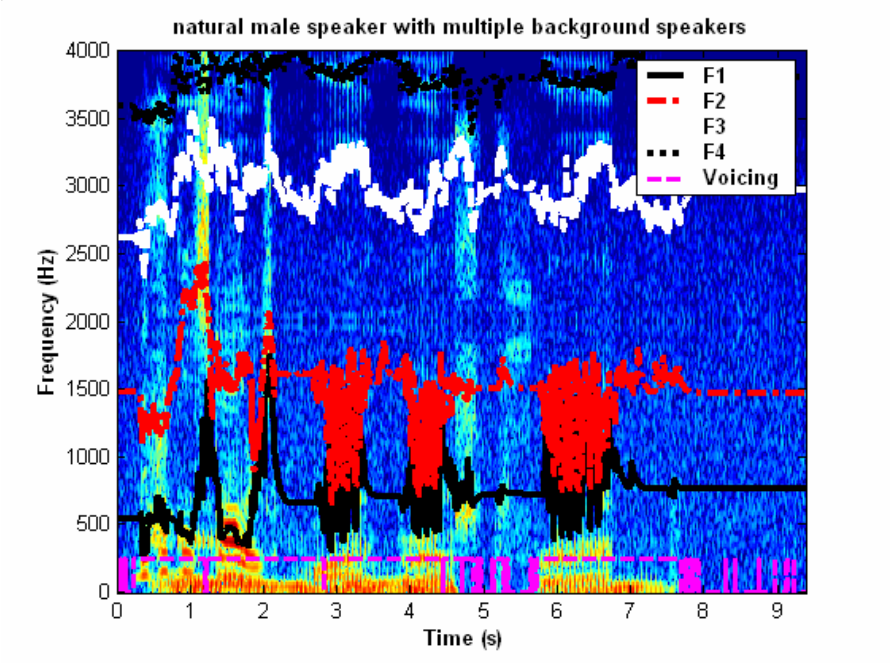



**Fig 4.32 Spectrogram of a natural female speaker in the presence of multiple background speakers at SNR 20dB**

### 4.2.3 Testing in the presence of a Echo

Algorithm has also been tested for both male and female speakers in the presence of echo. Figure 4.33 shows the spectrogram and estimated formant frequencies for a synthesized male speaker saying “five woman played basketball” in the presence of echo. Figure 4.34 shows the spectrogram and estimated formant frequencies for a synthesized female speaker saying “five woman played basketball” in the presence of echo.

Table 4.7 gives the estimated formant frequencies for both synthesized male and female speakers in the presence of echo. Estimated formant frequency values in table 4.7 shows that algorithm tracks formant frequencies accurately even in the presence of echo. Estimated formant frequency values are within the standard frequency range shown in table 4.1 and also estimated frequencies are very close to actual formant frequencies.

| Frequencies To be Estimated (Hz) | Value of Actual frequencies (Hz) | Estimated frequencies with a tracking algorithm (Hz) | |
|---|---|---|---|
| | | Echo present in male speaker | Echo present in female speaker |
| $F_1$ | 700 | 400 | 300 |
| $F_2$ | 1500 | 1600 | 1500 |
| $F_3$ | 2200 | 2600 | 2500 |
| $F_4$ | 3500 | 3300 | 3300 |

**Table 4.7 Estimated formant frequencies for both synthesized male and female speakers in the presence of echo.**

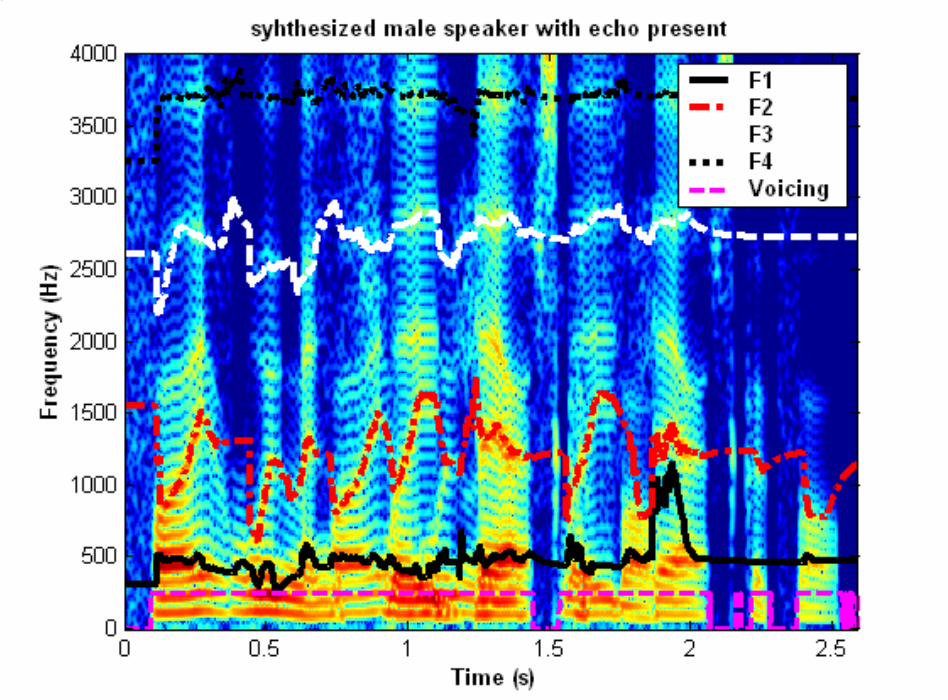



**Fig 4.33 Spectrogram of a synthesized male speaker in the presence of echo**

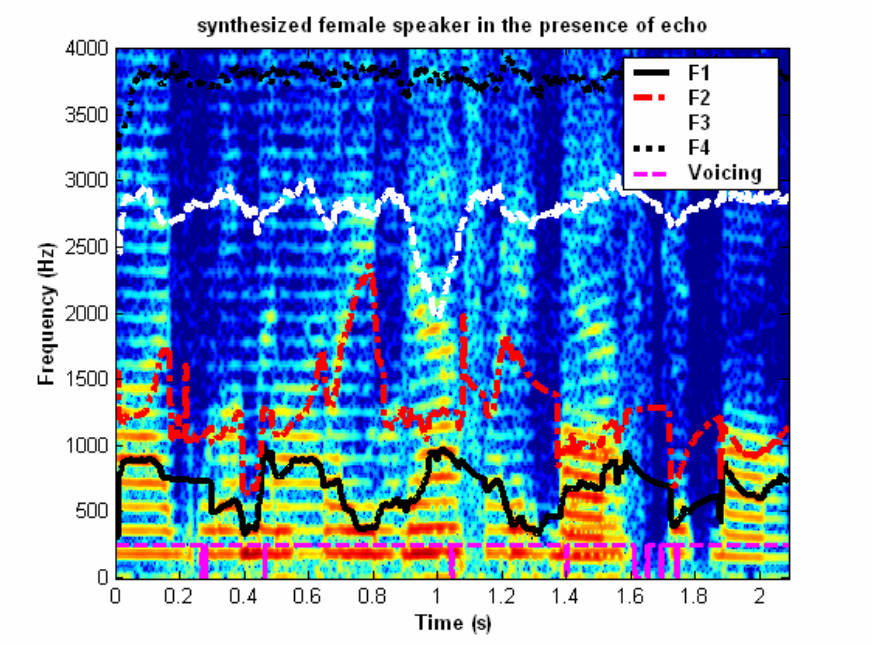


**Fig 4.34 Spectrogram of a synthesized female speaker in the presence of echo**

### 4.2.4 Testing the algorithm for fading speech

In this test case the algorithm is tested to observe the effect of a speaker whose speech is fading 'in and out' on the algorithm's ability to track formant frequencies. Again the algorithm has been tested for both male and female speakers saying "five woman played basket ball".

Figure 4.35 shows the spectrogram and estimated formant frequencies for male speaker when speech is fading 'in'. Figure 4.36 shows the spectrogram and estimated formant frequencies for male speaker when speech is fading 'out'.

Figure 4.37 shows the spectrogram and estimated formant frequencies for female speaker when speech is fading 'in'. Figure 4.38 shows the spectrogram and estimated formant frequencies for female speaker when speech is fading 'out'.

Table 4.8 gives the estimated formant frequencies for both male and female speakers for fading speech. Values in the table 4.8 shows that the formant frequencies are estimated accurately even in the fading environment. Again formant frequency values are very close to actual frequencies.

| Frequencies To be estimated (Hz) | Estimated frequencies for a male speaker (Hz) | | Estimated frequencies for a female speaker (Hz) | |
|---|---|---|---|---|
| | Fading 'in' | Fading 'out' | Fading 'in' | Fading 'out' |
| $F_1$ | 400 | 500 | 500 | 500 |
| $F_2$ | 1500 | 1500 | 1200 | 1500 |

| | | | | |
|---|---|---|---|---|
| $F_3$ | 2600 | 2600 | 3100 | 2500 |
| $F_4$ | 3200 | 3600 | 3800 | 3400 |

**Table 4.8 Estimated formant frequencies for both synthesized male and female speakers with fading speech.**

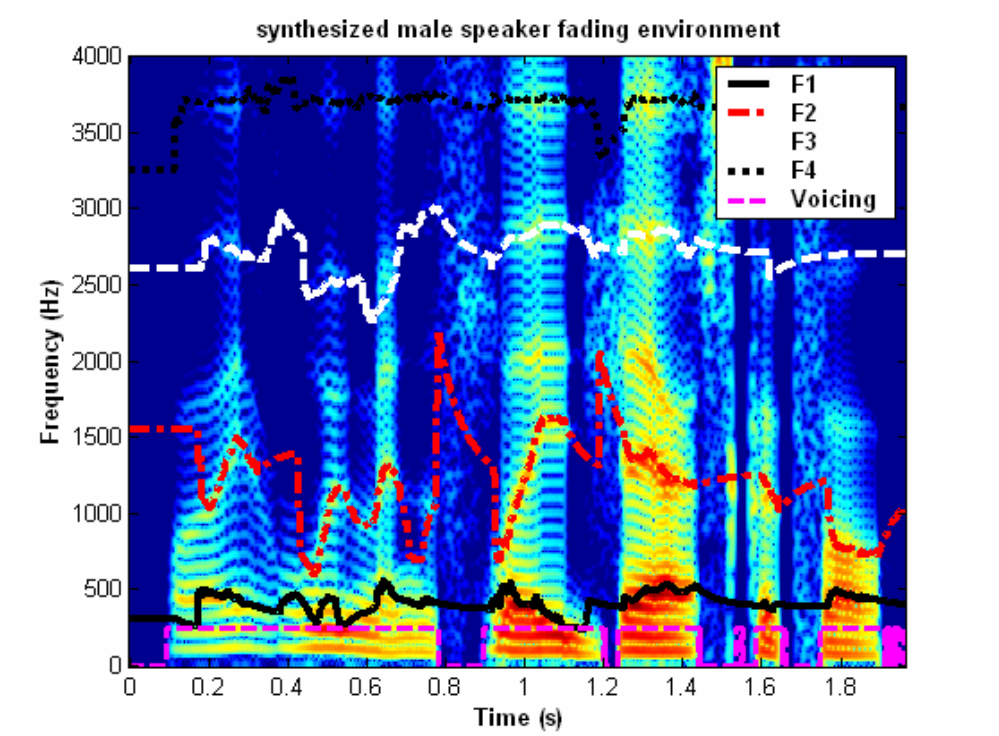



**Fig 4.35 Spectrogram of a synthesized male speaker when signal is fading in**

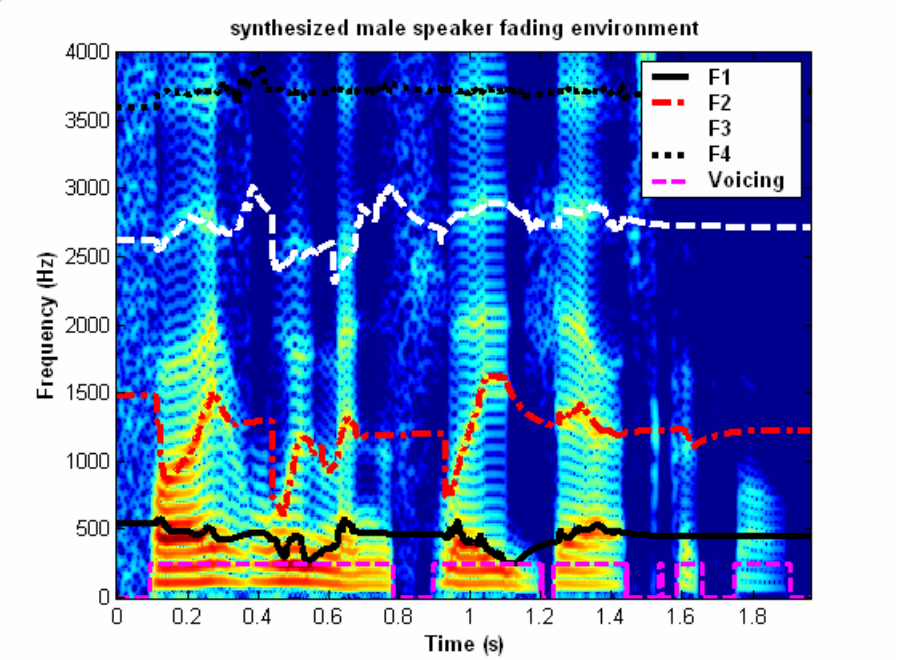



**Fig 4.36 Spectrogram of a synthesized male speaker when signal is fading out**

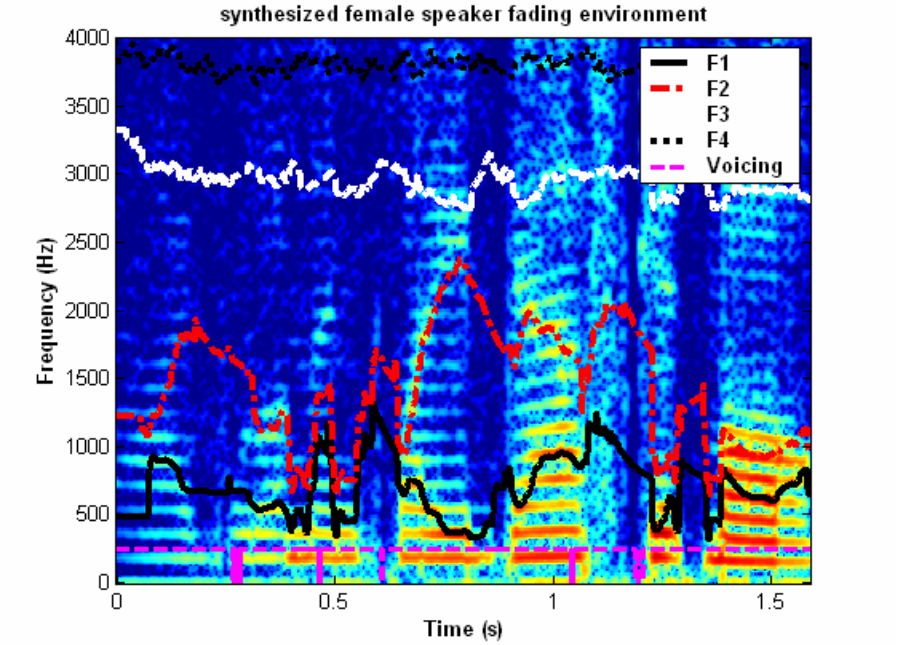



**Fig 4.37 Spectrogram of a synthesized female speaker when signal is fading in**

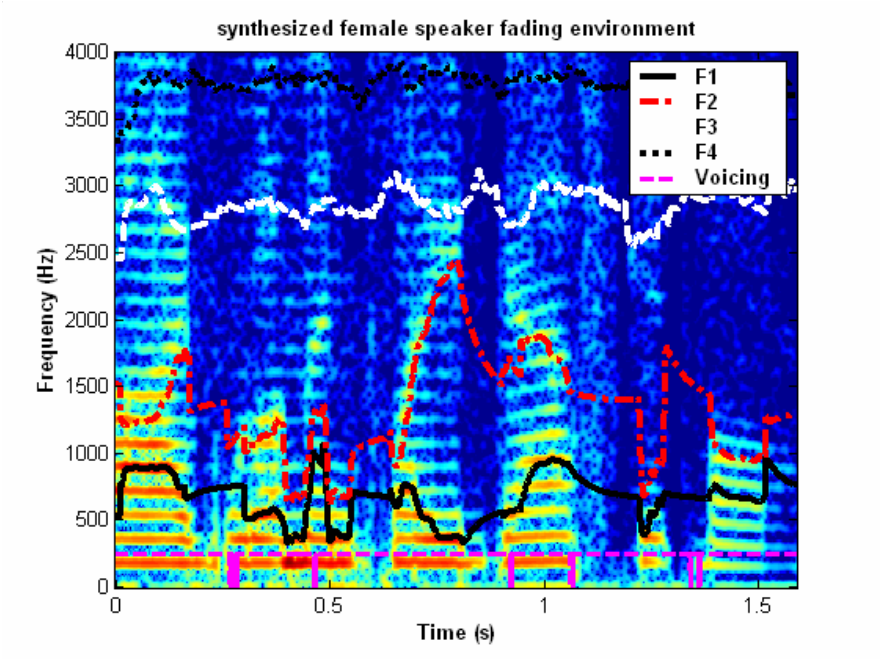



**Fig 4.38 Spectrogram of a synthesized female speaker when signal is fading out**

## 4.3 Testing With Recursive Least Square Algorithm

Recursive formant tracking algorithm estimates and tracks the formant frequencies of a speech signal. This algorithm has been tested again for both male and female speakers. Although formant tracking with RLS algorithm gives good formant estimates but exact values can not be estimated. The estimated formant frequencies are within the standard formant frequency range. Algorithm has been tested for number of coefficients eight and forgetting factor value .98.

### 4.3.1 Testing With White Noise

Speech signals have been tested with added white noise at different SNR's. Again testing is done for both synthesized and natural male and female speakers. Figure 4.39 and 4.40 show the spectrogram and estimated formant frequencies for a synthesized male and female speaker saying "five woman played basketball" respectively. Figure 4.41 shows the spectrogram and estimated formant frequencies for a synthesized male speaker at AWGN 30dB. Figure 4.42 shows the spectrogram and estimated formant frequencies for a synthesized female speaker at AWGN 30dB. Figures 4.39-4.42 show that RLS

algorithm is able to estimate formant frequencies but the formant estimate is not smooth. There are long jumps in the signal tracking.

Algorithm has also been tested for natural male and female speakers. Fig 4.43 shows the spectrogram and formant frequencies for a natural male speaker saying “fifth yard contains big juicy peaches” at AWGN 25dB. Fig 4.44 shows the spectrogram and formant frequencies for a natural female speaker saying “a book of scholars” at AWGN 30dB.

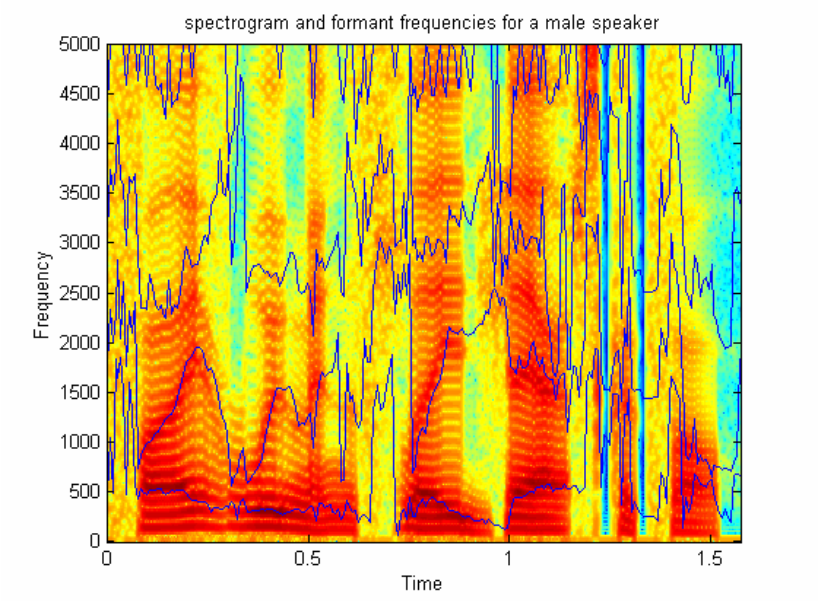



**Fig 4.39 Spectrogram and formant frequencies for a synthesized male**

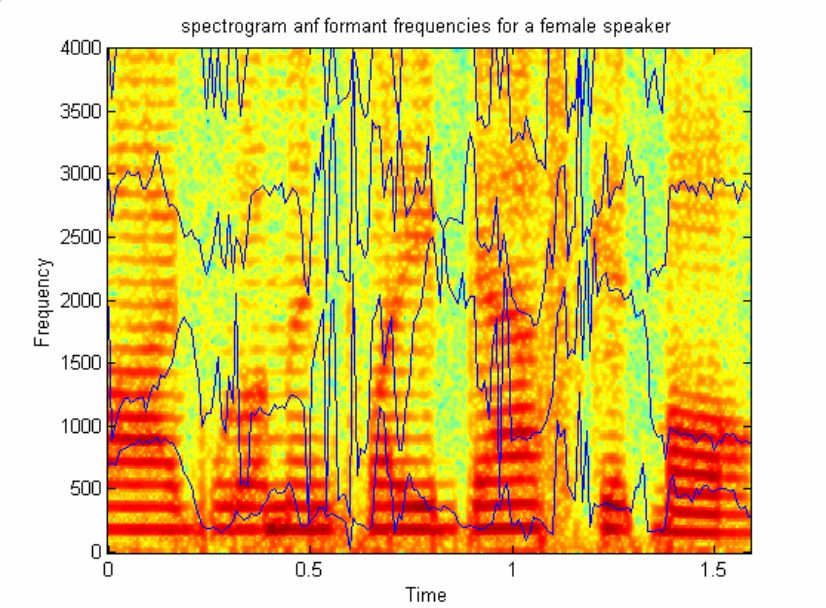



**Fig 4.40 Spectrogram and formant frequencies for a synthesized female speaker**

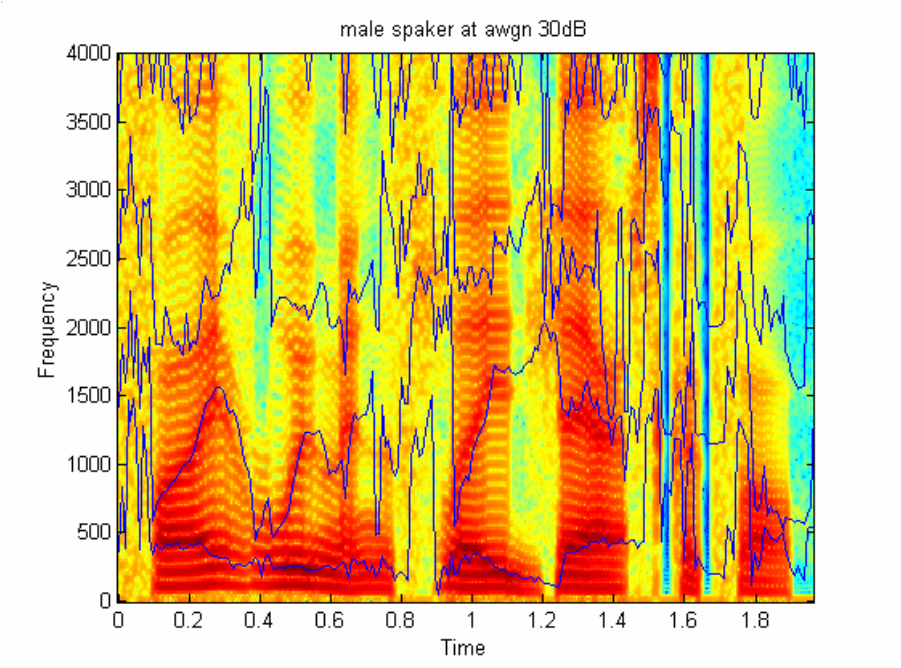



**Fig 4.41 Spectrogram and formant frequencies for a synthesized male speaker with AWGN 30dB**

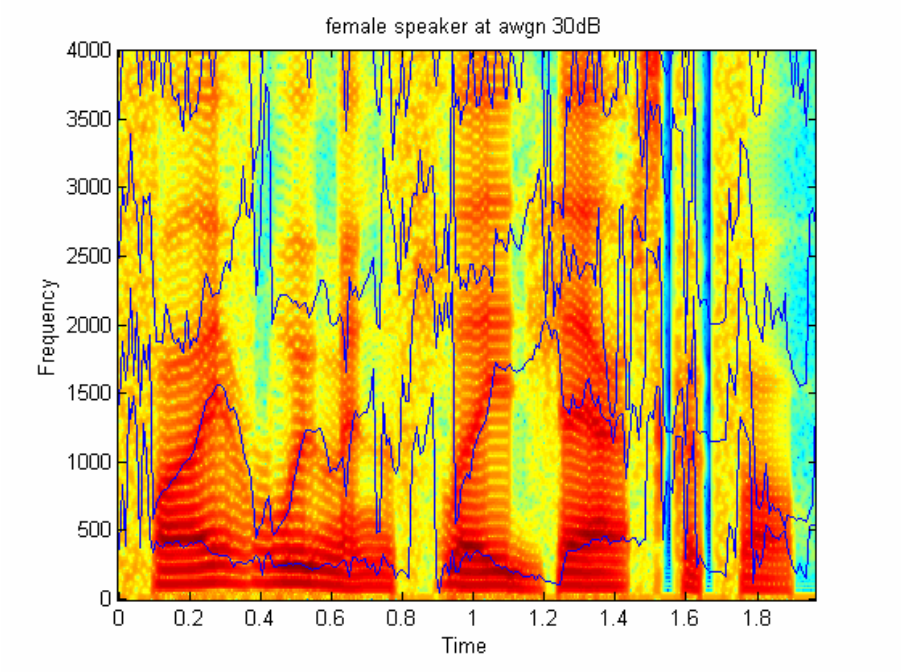



**Fig 4.42 Spectrogram and formant frequencies for a synthesized female speaker with AWGN 30dB**

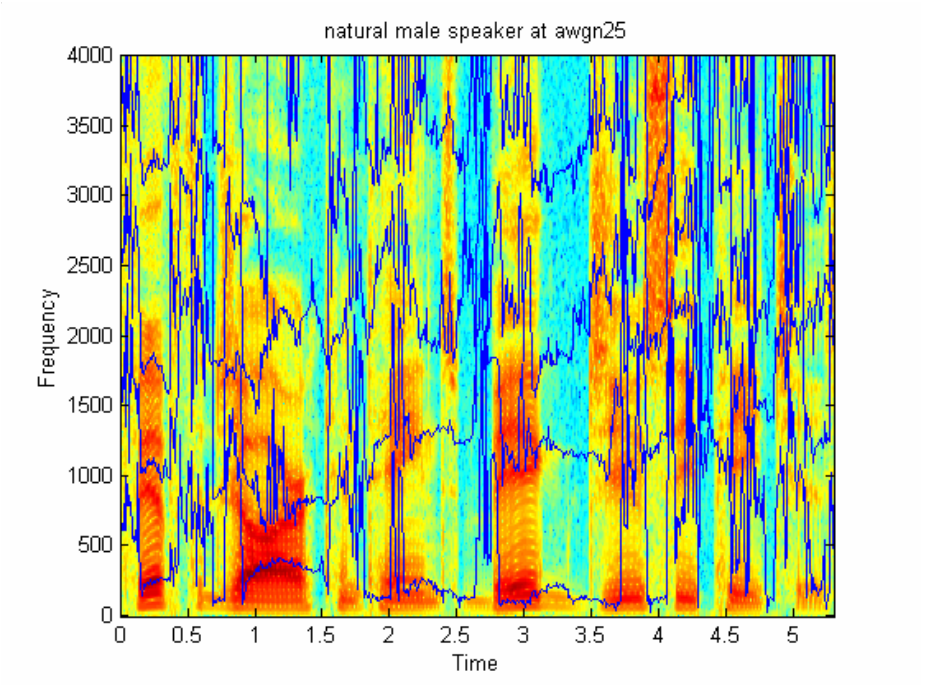



**Fig 4.43 Spectrogram and formant frequencies for a natural male speaker at AWGN 25dB**

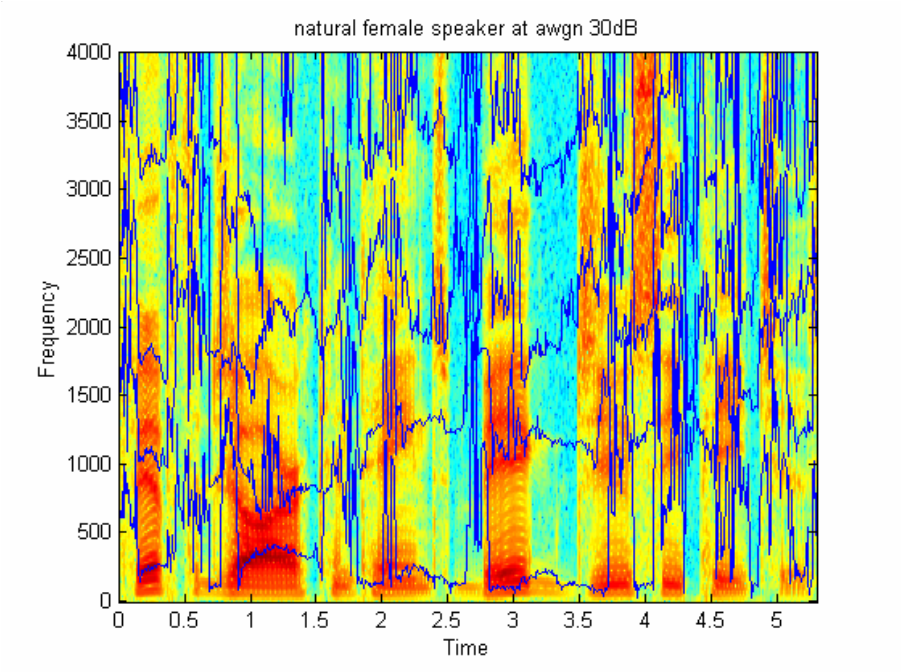



**Fig 4.44 Spectrogram and formant frequencies for a natural female speaker at AWGN 30dB**

### 4.3.2 Testing With Single Background Male Speaker

Similar to the robust formant tracking algorithm RLS algorithm has also been tested in the presence of single background male speaker. RLS algorithm is able to track formant frequencies but not accurately. Tracking with RLS algorithm is very noisy. Figure 4.45 shows the spectrogram and formant frequencies for a synthesized male speaker in the presence of single background male speaker at SNR 15dB. Figure 4.46 shows the spectrogram and formant frequencies for a female speaker in the presence of single background male speaker at SNR 15dB. From figures it can be easily said that this algorithm estimates formant frequencies accurately within the range shown in table 4.1.

Algorithm has also been tested for natural speakers. Figure 4.47 shows the spectrogram and formant frequencies for a natural male speaker in the presence of single background male speaker at SNR 25dB. Figure 4.48 shows the spectrogram and formant frequencies for a female speaker in the presence of single background male speaker at SNR 25dB.

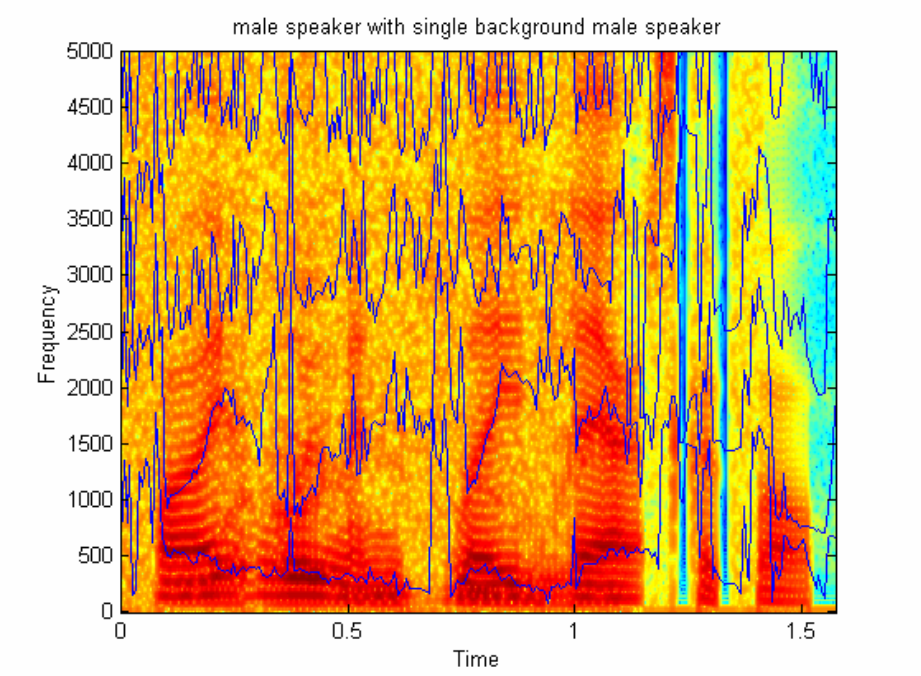



**Fig 4.45 Spectrogram and formant frequencies for a synthesized male speaker with single background male speaker at 15dB**

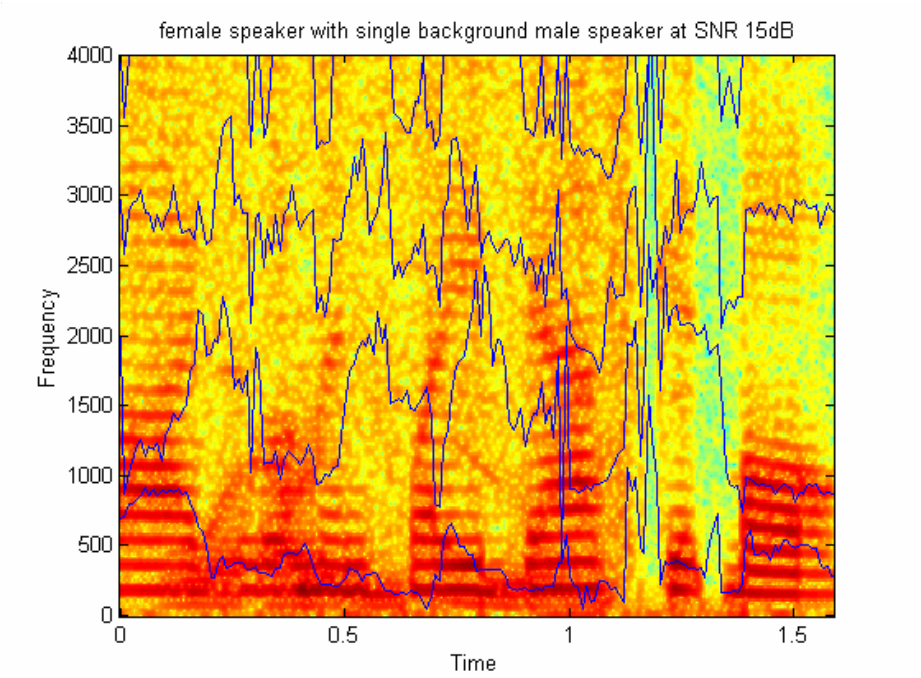



**Fig 4.46 Spectrogram and formant frequencies for a synthesized female speaker with single background male speaker at 15dB**

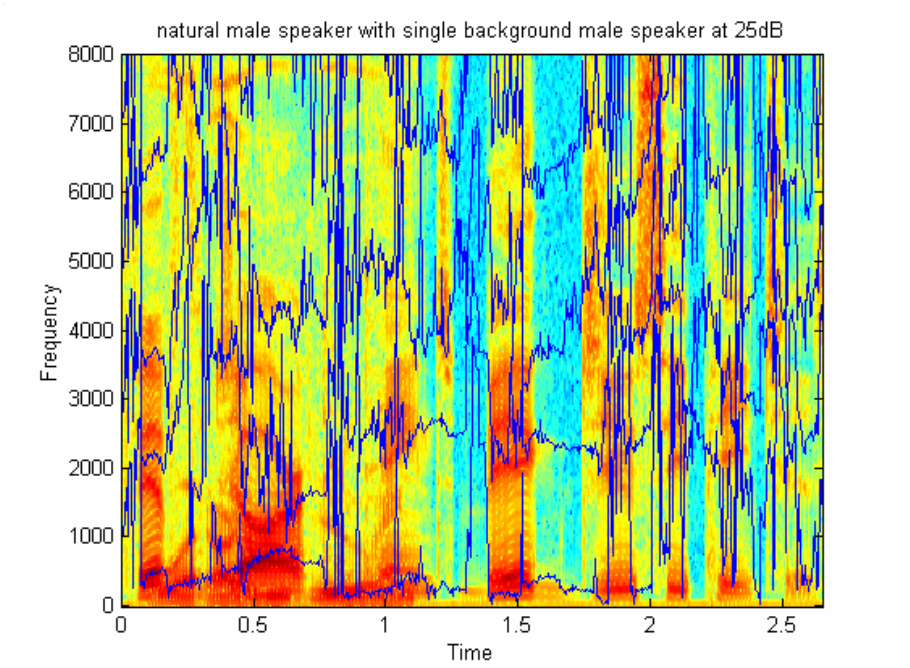



**Fig 4.47 Spectrogram and formant frequencies for a natural male speaker with single background male speaker at 25dB**

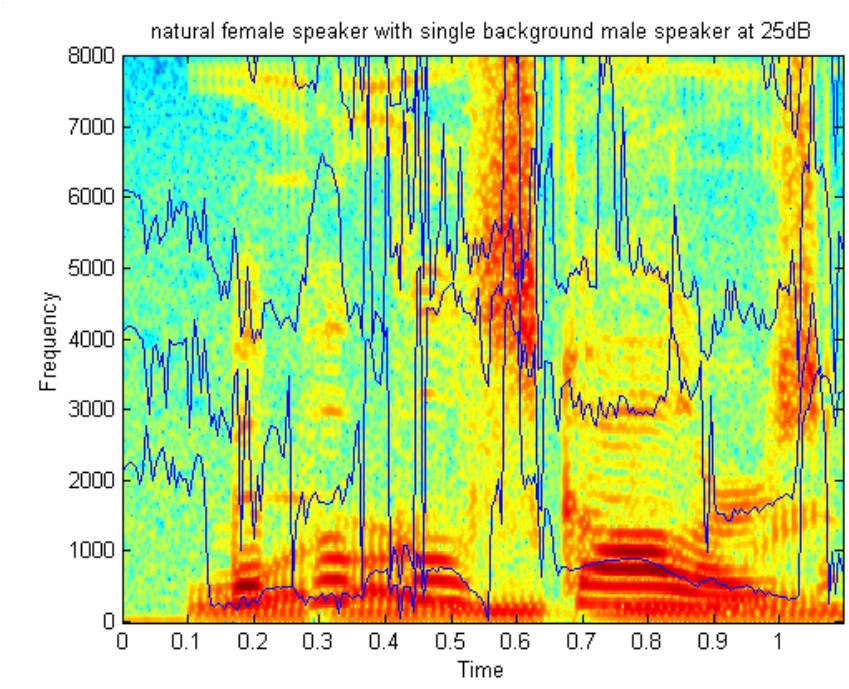


**Fig 4.48 Spectrogram and formant frequencies for a natural female speaker with single background male speaker at 25dB**

### 4.3.3 Testing with Single Background female Speaker

RLS algorithm has also been tested in the presence of single background female speaker. Figure 4.49 shows the spectrogram and formant frequencies for a synthesized male speaker in the presence of single background female speaker at SNR 5dB. Figure 4.50 shows the spectrogram and formant frequencies for a female speaker in the presence of single background female speaker at SNR 5dB. Figures 4.49 and 4.50 show that the formant tracking is very poor in the presence of single background female speaker.

Algorithm has also been tested for natural speakers. Figure 4.51 shows the spectrogram and formant frequencies for a natural female speaker in the presence of single background female speaker at SNR 20dB. Figure 4.52 shows the spectrogram and formant frequencies for a male speaker in the presence of single background female speaker at SNR 10dB.

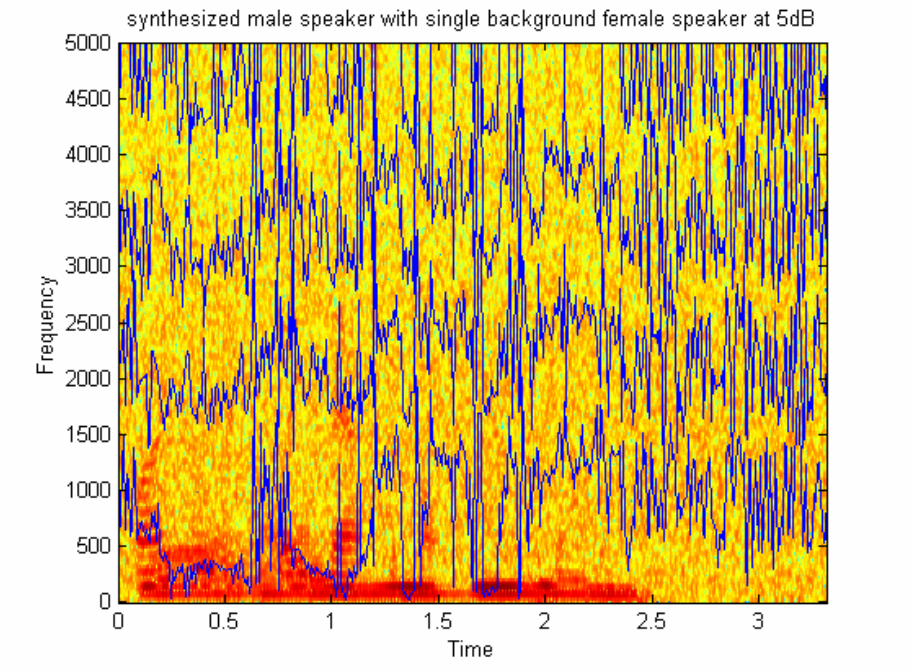


**Fig 4.49 Spectrogram and formant frequencies for a synthesized male speaker with single background female speaker at 5dB**

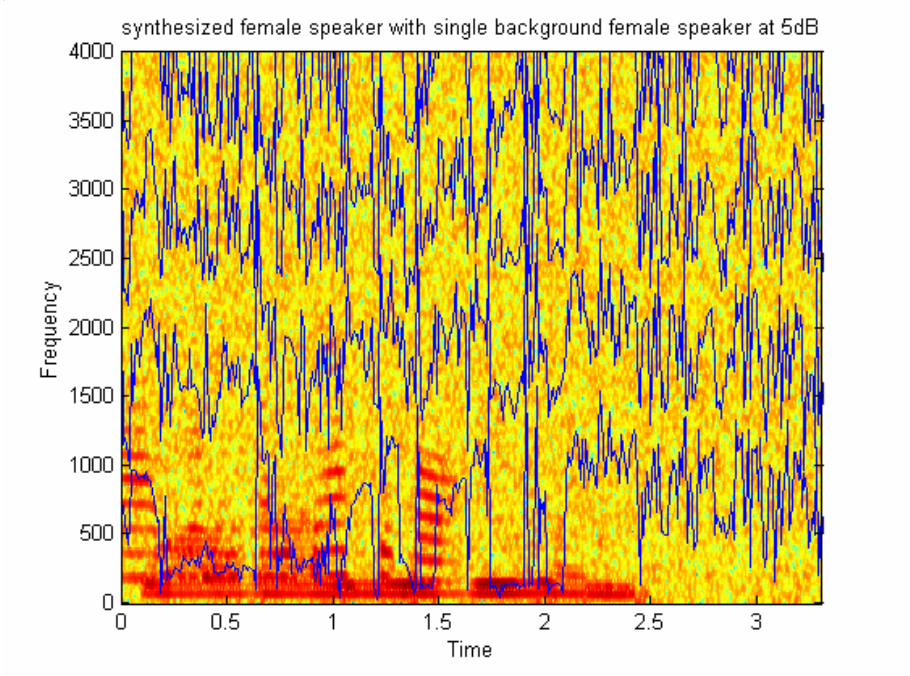



**Fig 4.50 Spectrogram and formant frequencies for a synthesized female speaker with single background female speaker at 5dB**

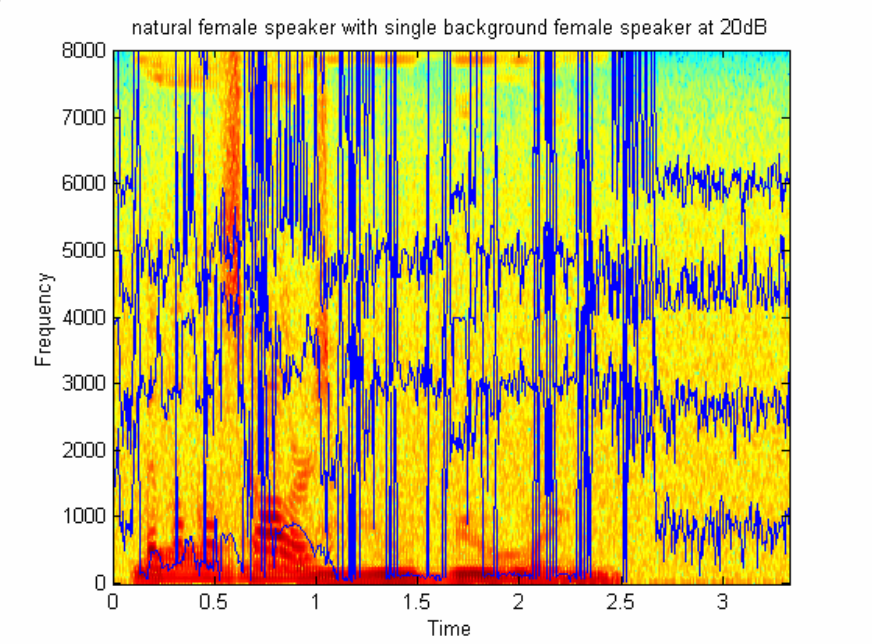



**Fig 4.51 Spectrogram and formant frequencies for a natural female speaker with single background female speaker at 20dB**

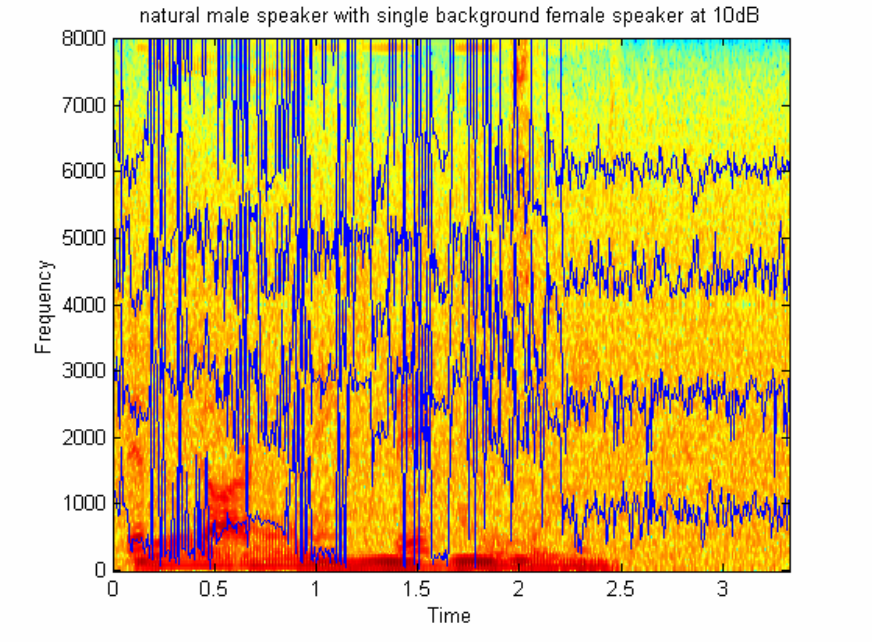



**Fig 4.52 Spectrogram and formant frequencies for a natural male speaker with single background female speaker at 10dB**

### 4.3.4 Testing In The Presence Of Echo

RLS algorithm has also been tested in the presence of echo in the speech signal. Figure 4.53 shows the spectrogram and formant frequencies for a synthesized male speaker in the presence of echo. Figure 4.54 shows the spectrogram and formant frequencies for a synthesized female speaker in the presence of echo.

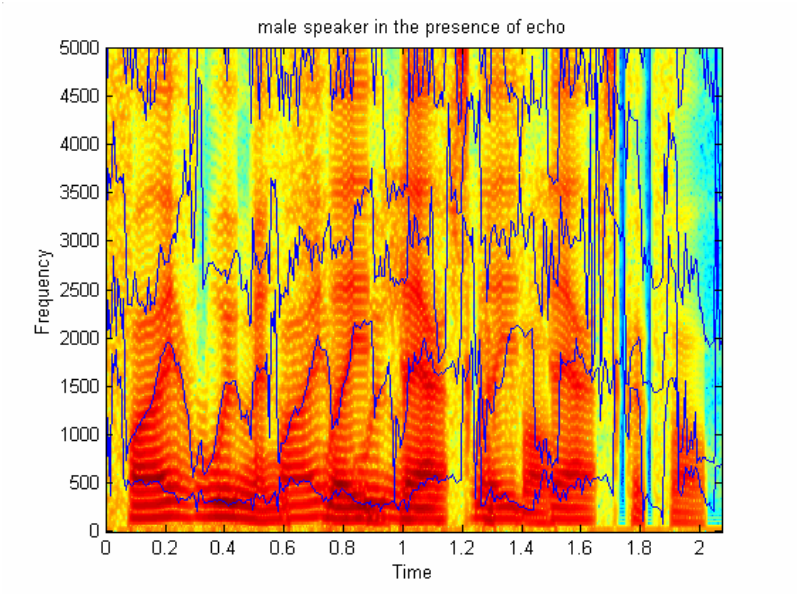



**Fig 4.53 Spectrogram and formant frequencies for a synthesized male speaker in the presence of echo**

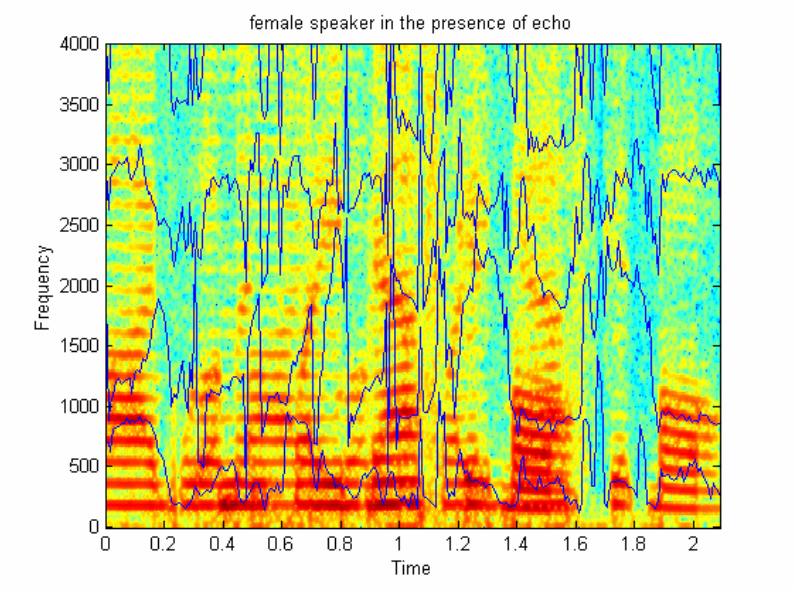


**Fig 4.54 Spectrogram and formant frequencies for a synthesized female speaker in the presence of echo**

### 4.3.5 Testing In the Fading Environment

RLS algorithm has also been tested with the speech signal in the fading environment. Test was done for both male and female speakers with speech fading 'in' and fading 'out'. Figure 4.55 shows the spectrogram and formant frequencies for a synthesized male speaker with speech signal fading 'in'. Figure 4.56 shows the spectrogram and formant frequencies for a synthesized male speaker with speech signal fading 'out'.

Figure 4.57 shows the spectrogram and formant frequencies for a synthesized female speaker with speech signal fading 'in'. Figure 4.58 shows the spectrogram and formant frequencies for a synthesized female speaker with speech signal fading 'out'. Figures 4.55-4.58 show that algorithm is able to track frequencies but not accurately. From visual inspection it can be seen that for fading environment also formant frequency values are within the standard formant frequency range shown in table 4.1.

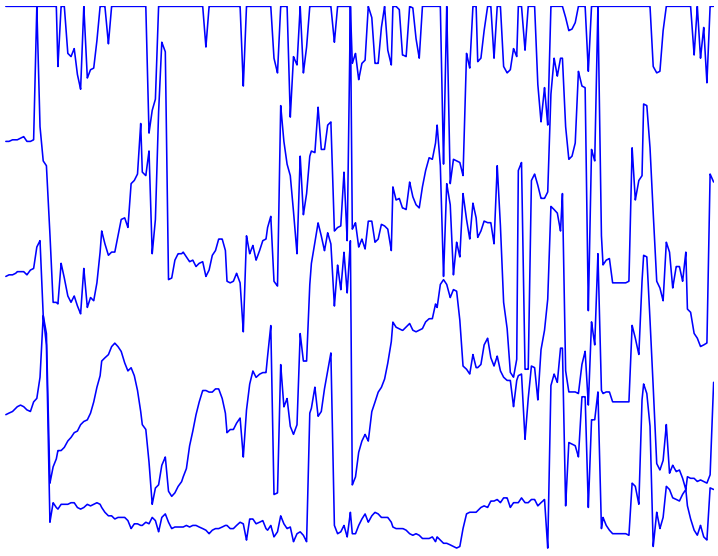


**Fig 4.55 Spectrogram and formant frequencies for a synthesized male speaker with speech signal fading 'in'.**

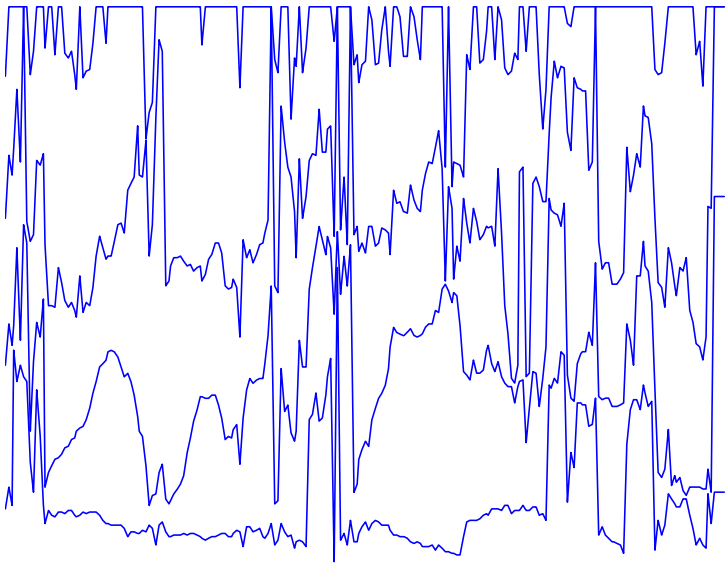


**Fig 4.56 Spectrogram and formant frequencies for a synthesized male speaker with speech signal fading ‘out’**

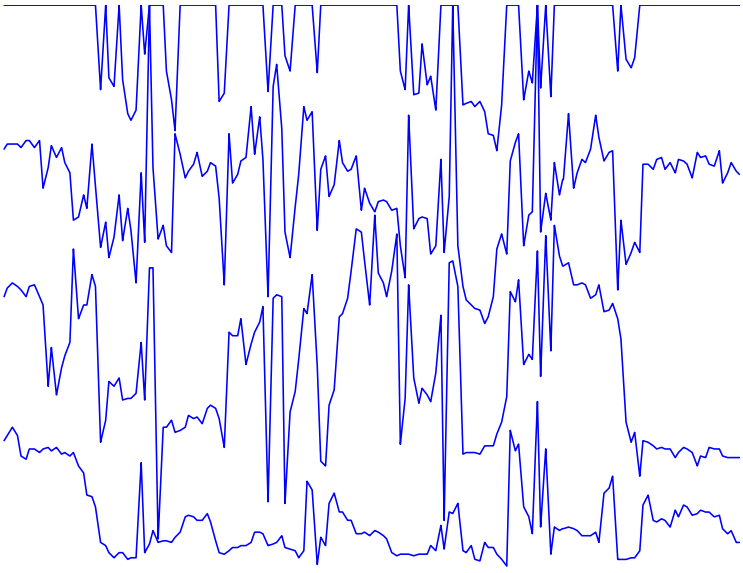


**Fig 4.57 Spectrogram and formant frequencies for a synthesized female speaker with speech signal fading ‘in’**

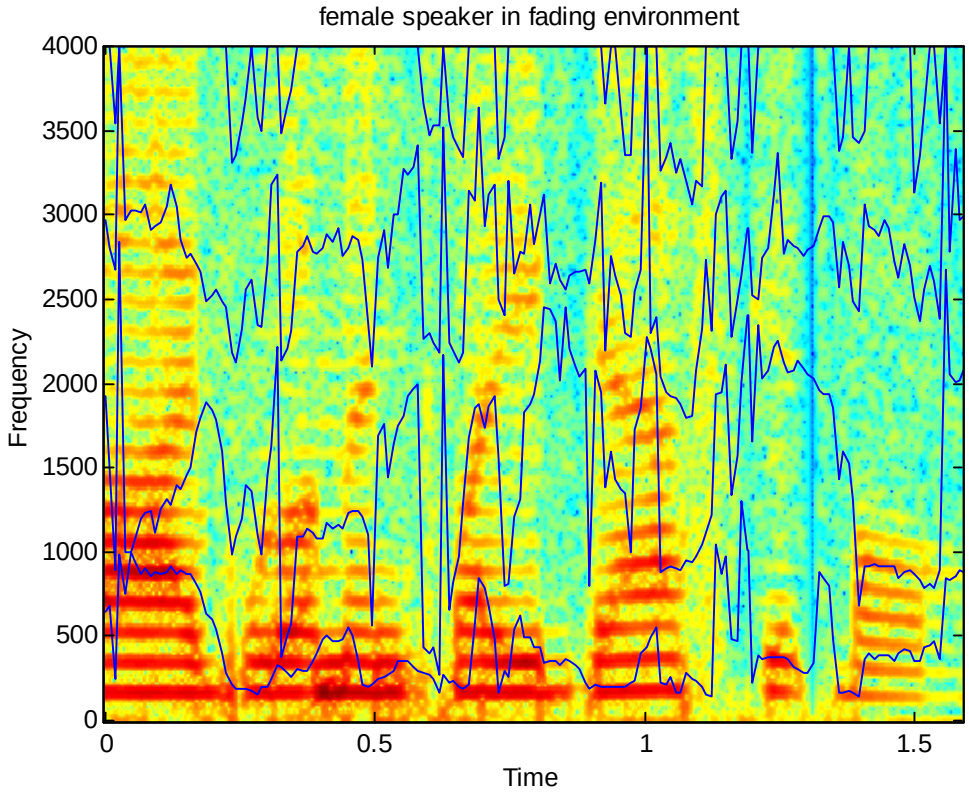



**Fig 4.58 Spectrogram and formant frequencies for a synthesized female speaker with speech signal fading 'out'**

In this way both algorithms have been tested in different environmental conditions. Testing can also be done in some other conditions like with reverberant speech, in the presence of traffic noise, other atmospheric noises. Results from both algorithms show that the formants can be estimated and tracked with both algorithms. But formant estimation and tracking is accurate only with robust formant tracking algorithm. RLS algorithm cannot track frequencies as accurately as with robust formant tracking algorithm.

**Chapter 5**

# CONCLUSIONS AND FUTURE SCOPE

## 5.1 Conclusions

Two different formant tracking algorithms have been discussed in the present work. These algorithms are: Robust formant tracking algorithm and formant tracking with RLS algorithm. Quantitive analysis of both formant tracking algorithms have shown that it provides accurate formant frequency estimates for both male and female speakers for a wide range of SNR's in real-time noise conditions such as AWGN, a single competing background speaker (male and female), multiple background speakers, in the presence of echo, with fading speech. Robust formant tracking algorithm provides mostly smooth formant frequency estimates than RLS algorithm. The robust formant tracking algorithm recovers quickly after erroneous estimates to go back to tracking the actual formant frequencies in the speech signal, which is not the case with RLS algorithm. Because of this reason RLS algorithm shows noisy tracking. Information about the gender is not available with RLS algorithm. But the computation complexity of RLS algorithm is less as compared to robust formant tracking algorithm. There have been some problems identified with the robust formant tracker. The algorithm occasionally gives 'choppy' and oscillating formant frequency estimates. This is an undesirable result because the actual formant frequencies of speech normally vary slowly with time and have smooth transitions. This problem is only encountered when the SNR is very low and occurs due to the algorithm tracking the excess energy added outside the formant frequency regions from the background noise source. However, the overall performance of the robust formant tracking algorithm is still much better than that of formant tracking with RLS algorithm.

The algorithms discussed in the present work are geared primarily towards use for CEFS amplification. It was identified earlier that in order to apply CEFS to continuous speech the second formant frequency has to be estimated accurately and in real-time. Furthermore, the estimated formant frequencies have to be smooth and the algorithm has to be able to identify formant transitions accurately so that the proper frequency

dependent amplification is applied to the speech signal. Testings on the algorithms have shown that the formant frequency estimates are smooth and the formant frequency transitions are tracked accurately. Therefore, the formant tracking algorithms presented here can be used to implement CEFS amplification. With RLS algorithm, the formant frequency estimates are not very smooth and also have large jumps. Because of this reason robust formant tracking algorithm is a better choice for CEFS amplification.

## 5.2 Future Scope

The oscillating formant frequency problem may be solved in future updates to the formant tracking algorithms by either smoothing the formant frequency estimates or by incorporating additional logical limitations to prevent abnormal jumps in the formant estimates.

Another future improvement may be to modify the formant pre-filters to have variable bandwidths that are dependent on the magnitudes of the poles estimated by the linear prediction coefficients. This may further improve the formant estimates during rapid formant transitions at high SNR’s, but the performance at low SNR’s would likely remain unchanged.

# REFERENCES


[1] P.F. Assmann, “The role of formant transitions in the perception of concurrent vowels,” *J. Acoustic. Soc. Am.,* vol. 97, no. 1, pp. 575–584, Jan. 1995.

[2] M. B. Sachs, I. C. Bruce, R. L. Miller, and E. D. Young, “Biological basis of hearing-aid design,” *Ann. Biomed. Eng.,* vol. 30, no. 2, pp. 157–168, Feb. 2002.

[3] J. R. Schilling, R. L. Miller, M. B. Sachs, and E. D. Young, “Frequency shaped amplification changes the neural representation of speech with noise-induced hearing loss,” *Hear. Res*., vol. 117, pp. 57–70, 1998.

[4] R. L. Miller, B. M. Calhoun, and E. D. Young, “Contrast enhancement improves the representation of */E/*-like vowels in the hearing-impaired auditory nerve,” J. *Acoust. Soc. Am*., vol. 106, no. 5, pp. 2693–2708, 1999.

[5] I. C. Bruce, “Physiological assessment of contrast-enhancing frequency shaping and multiband compression in hearing aids,” *Physiol. Meas*., vol. 25, no. 4, pp. 945–956, Aug 2004.

[6] R. W. Schafer and L. R. Rabiner, “System for automatic formant analysis of voiced speech,” *J. Acoust. Soc. Am*., vol. 47, no. 2, pp. 634–648, Feb. 1970.

[7] B. S. Atal and S. L. Hanauer, “Speech analysis and synthesis by linear prediction of the speech wave,” *J. Acoust. Soc. Am*., vol. 50, no. 2, pp. 637–655, Aug. 1971.

[8] K. Mustafa and I. C. Bruce, “Robust formant tracking for continuous speech with speaker variability,” in Proceedings of the *Seventh International Symposium on Signal Processing and Its Applications (ISSPA), Vol. 2*. Piscataway, NJ: IEEE, 2004, pp. 623–624

[9] S. Haykin, "*Adaptive Filter Theory*". Prentice-Hall. 1996. 3rd Edition.

[10] 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.

[11] J. L. Flanagan, “Automatic extraction of formant frequencies from continuous speech,” *J. Acoust. Soc. Am*., vol. 28, pp. 110–118, 1956.

[12] S. McCandless, “An algorithm for automatic formant extraction using linear prediction spectra,” *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. ASSP 22, pp. 135–141, 1974.

[13] S. W. Metz, J. A. Heinen, R. J. Niederjohn, and T. V. Sreenivas, “Auditory modeling applied to formant tracking of noise-corrupted speech,” *in Proc. Intl. Conf. Industrial Electronics, Control and Instrumentation*, vol. 3, 1991, pp. 2120 2124

[14] M. M. Sondhi, “New methods of pitch extraction,” *IEEE Trans. Audio and Electroacoustics*, vol. AU-16, no. 2, pp. 262–266, 1968.

[15] B. Widrow, & S. Stearns, "*Adaptive Signal Processing*". Prentice-Hall. 1985.

[16] Hodgkiss, W. & Presley, J. "Adaptive Tracking of Multiple Sinusoids Whose Power Levels are Widely Separated". IEEE *Trans. on Circuits and Systems*. Vol. CAS-28, n. 6, June, 1981.

[17] L. R. Rabiner and R. W. Schafer, “*Digital Processing of Speech Signals”* Englewood Cliffs, NJ: Prentice Hall, 2004.

[18] L.R.Rabiner and B.H Juang. (1993) “*Fundamentals of Speech Recognition*” Prentice-Hall, Englewood Cliffs, NJ.

[19] G. E. Peterson and H. L. Barney, “Control methods used in a study of the vowels,” *J. Acoustic. Soc. Am.*, vol. 24, no. 2, pp. 175–184, Mar. 1952.

[20] B. S.Atal and L. R Rabiner, “A pattern recognition approach to voiced-unvoiced silence classification with applications to speech recognition,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 201-212, June 1976.

[21] A. M. Liberman, P. C. Delattre, F. S. Cooper, and L. J. Gerstman, “The role of consonant-vowel transitions in the perception of the stop and nasal consonants,” *Psychological Monographs*, vol. 68, pp. 1–13, 1954.

[22] L. C. Pols, L. J. van der Kamp, and R. Plomp, “Perceptual and physical space of vowel sounds,” *J. Acoust. Soc. Am.*, vol. 46, no. 2, pp. 458–467, Aug. 1969.

[23] R.W., Schafer, L.R Rabiner, “System for automatic formant analysis of voiced speech,” *Journal of the Acoustical Society of America* vol.47, no.2, 1970, pp 634 650.

[24] L.R Rabiner,. and R.W Schafer,. “On the behavior of minimax FIR digital Hilbert transformers,” *The Bell System Technical Journal* Vol. 53, No. 2, 1974.

[25] J. Makhoul, “Linear prediction: A tutorial review,” *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[26] I. C. Bruce, N. V. Karkhanis, E. D. Young, and M. B. Sachs, “Robust formant tracking in noise,” *in Proc. ICASSP 2002, Vol. I. Piscataway, NJ: IEEE, 2002*, pp. 281–284.

[27] J.M. Cioffi and T. Kailath, “Fast recurisve-least-square, transversal filters for adaptive filtering,” *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-32, no. 2, pp. 304-334, 1984.

[28] O. Macchi and N. Bershad, “Adaptive recovery of a chirped sinusoid in noise, Part 1: Performance of the RLS algorithm,” *IEEE Trans. Signal Processing*, ASSP 39, pp. 583-594, Mar. 1991.

[29] B. Widrow et al., “Adaptive noise cancelling: Principles and applications,” *Proc. IEEE*, vol. 63, pp. 1692- 1716, Dec. 1975.

[30] T. R. Fortescue, L. S. Kershenbaum, and B. E. Ydstie, “Implementation of selftuning regulators with variable forgetting factors,” *Automatica*, vol. 17, pp. 831- 835, 1981.

[31] R. Kulhavy, “Restricted exponential forgetting in real-time identification,” *Automatica*, vol. 23, no. 5, pp. 589-600, 1987.

[32] D. J. Park, B. E. Jun and J. H. Kim, “Fast tracking RLS algorithm using novel variable forgetting factor with unity zone,” *Electronic Letters*, vol. 27, no. 23, pp. 2150-2151, November 1991.

.

# LIST OF PUBLICATIONS

[1] **"Speech Coding Using LPC – A Study"** at **National Conference on Electronic Circuits and Communication Systems- 2004,** held on 23rd-24th September, 2004 at Thapar Institute of Engineering And Technology, Patiala.

[2] **"Various DTMF Detection"** at **8th Punjab Science Congress 2005,** held on February 7-9, 2005 at Punjabi University, Patiala.

[3] "**Algorithms for Tracking Formant Frequencies of A Continuous Speech with Speaker Variability"** communicated to **4th International Conference on Natural Language Processing (ICON-2005)** will be held in IIT Kanpur, India during December 18-20, 2005.

[4] **"Formant tracking using RLS and LMS algorithms"** communicated to **Institute Of Electronics And Telecommunication Engineers (IETE) Journals.**