DSP Compensation for Distortion in RF Filters

A Thesis Submitted
to the College of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in the Department of Electrical & Computer Engineering
University of Saskatchewan

by

Mehdi Alijan

Saskatoon, Saskatchewan, Canada

PERMISSION TO USE

In presenting this Thesis in partial fulfillment of the requirements for a graduate degree from the University of Saskatchewan, it is agreed that the Libraries of this University may make it freely available for inspection. Permission for copying of this Thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professors who supervised this Thesis work or, in their absence, by the Head of the Department of Electrical & Computer Engineering or the Dean of the College of Graduate Studies and Research at the University of Saskatchewan. Any copying, publication, or use of this thesis, or parts thereof, for financial gain without the written permission of the author is strictly prohibited. Proper recognition shall be given to the author and to the University of Saskatchewan in any scholarly use which may be made of any material in this thesis.

Request for permission to copy or to make any other use of material in this thesis in whole or in part should be addressed to:

Head of the Department of Electrical & Computer Engineering
57 Campus Drive
University of Saskatchewan
Saskatoon, Saskatchewan, Canada
S7N 5A9

ABSTRACT

There is a growing demand for the high quality TV programs such as High Definition TV (HDTV). The CATV network is often a suitable solution to address this demand using a CATV modem delivering high data rate digital signals in a cost effective manner, thereby, utilizing a complex digital modulation scheme is inevitable. Exploiting complex modulation schemes, entails a more sophisticated modulator and distribution system with much tighter tolerances. However, there are always distortions introduced to the modulated signal in the modulator degrading signal quality.

In this research, the effect of distortions introduced by the RF band pass filter in the modulator will be considered which cause degradations on the quality of the output Quadrature Amplitude Modulated (QAM) signal. Since the RF filter's amplitude/group delay distortions are not symmetrical in the frequency domain, once translated into the base band they have a complex effect on the QAM signal. Using Matlab, the degradation effects of these distortions on the QAM signal such as Bit Error Rate (BER) is investigated.

In order to compensate for the effects of the RF filter distortions, two different methods are proposed. In the first method, a complex base band compensation filter is placed after the pulse shaping filter (SRRC). The coefficients of this complex filter are determined using an optimization algorithm developed during this research. The second approach, uses a pre-equalizer in the form of a Feed Forward FIR structure placed before the pulse shaping filter (SRRC). The coefficients of this pre-equalizer are determined using the equalization algorithm employed in a test receiver, with its tap weights generating the inverse response of the RF filter. The compensation of RF filter distortions in base band, in turn, improves the QAM signal parameters such as Modulation Error Ratio (MER). Finally, the MER of the modulated QAM signal before and after the base band compensation is compared between the two methods, showing a significant enhancement in the RF modulator performance.

ACKNOWLEDGMENTS

I would like to extend my sincere gratitude and appreciation to my supervisor, Professor J. Eric Salt, for his thoughtful guidance and teaching, and generous support and encouragement during my pursuit of this Master of Science degree. His excellent comments had guided me throughout my research, and promoted my thinking. It has been a privilege and rewarding experience to work under Professor J. Eric Salt.

I would also like to extend my thanks to the management and staff of Department of Electrical Engineering and the College of Graduate Studies and Research, university of Saskatchewan for their support during this research.

Finally, I would like to extend my thanks to my friend and colleague Brian Berscheid for his generous help and comments on this thesis.

To my parents

Nayereh,

and

Ahmad.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

APF	All Pass Filter
BER	Bit Error Rate
CATV	Community Antenna TV
CSO	Composite Second Order
CTB	Composite Third order Beat
DAC	Digital to Analog Converter
DOCSIS	Data Over Cable System Interface Specification
DRFI	Downstream RF Interface
DSP	Digital Signal Processing
DTFT	Discrete-Time Fourier Transform
FIR/IIR	Finite/Infinite Impulse Response
FFF	Feed Forward FIR
FFT	Fast Fourier Transform
HDTV	High Definition TV
I/Q	In-phase/Quadrature
ISI	Inter Symbol Interference
LMS	Least Mean Square
LTI	Linear Time Invariant
MER	Modulation Error Rate
MF	Matched Filter
MSE	Mean Squared Error
PDF	Probability Density Function
PSD	Power Spectral Density
QAM	Quadrature Amplitude Modulation
QPSK	Quadrature Phase Shift Keying
RF	Radio Frequency
SNR	Signal-to-Noise Ratio
SRRC	Square Root Raised Cosine

1. Introduction

1.1 Background

CATV (Community Antenna Television) systems was created as an alternative for terrestrial TV broadcasting for areas where the transmitted signals are too weak for reception. One solution was to use coaxial cables in a CATV network to deliver reasonable quality signals to the customer home. At the beginning, the CATV antenna towers would receive analog TV channels off air, as would a TV set, and mapped them in the cable network spectrum. In north America, the bottom portion of the frequency band 50-550 MHz, is reserved for NTSC analog cable TV broadcast, as it is shown in figure 1.1. [1]

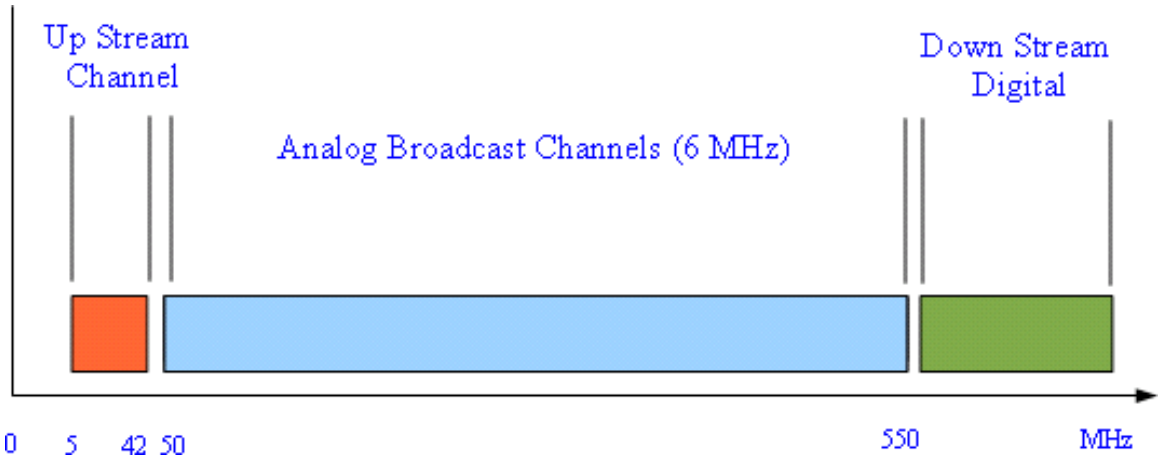

Figure 1.1 CATV Spectrum

The main reason the CATV service grew so quickly, is that the reach of over-the-air

television transmitter was very limited, and many communities had poor reception. For these communities the CATV network was a good solution. Eventually more channels were added to the network, which made CATV popular even in areas with good reception over-the-air not just in north America but all around the world.

The most outstanding feature of CATV plant is that it is broadband (0.5MHz-1.0GHz), and can carry many RF modulated signals. The plant consists of a network of coaxial cable that links a head end via the distribution network to the customer equipment (Refer to Figure 2.3 in page 12). The CATV plant can carry any information that can be modulated on a RF carrier.

The infrastructure of the CATV plant was designed to deliver analog TV signals to the end user. The topology and layout of its infrastructure was optimized for the network to have maximum cost efficiency in performing that goal. This lead to an architecture that is referred to as a *Tree* and *Branch* architecture [2]. The CATV infrastructure consists of various subsystems, the Head End, the Trunk cable, the distribution (or feeder) cable, the drop cable in house wiring, and the terminal equipment (Set Top Box).

1.2 Motivation

Cable TV service was meant to deliver high quality TV reception to the customer and has been successful in this sense. By 1999 almost 97 percent of U.S. Television households had cable television service available, and almost 66 million households subscribed to at least basic video service which is about 67 percent of U.S. TV households [3].

A growing demand for high quality reception for a large number of TV programs at reasonable cost has contributed to the evolution of new bandwidth efficient modulation schemes such as 64 and 256 digital Quadrature Amplitude Modulation (QAM) which have the capability to transfer high data rate digital signals over relatively small spectral bandwidth. This enables CATV operators to tightly pack a large number of

digital carriers and deliver a large number of TV programs at a lower cost.

Using bandwidth efficient modulation techniques comes at the price of higher signal to noise ratio requirement and more stringent amplitude/group delay specifications. In this regard, the pristine quality digital RF modulator must be used in the down stream path which is able to deliver large number of high quality signals such as HDTV at low cost. However, RF modulators usually use an RF filter which does not have an ideal amplitude/group delay response, degrading the quality of the QAM signal. The main motivation for this thesis was to compensate the RF filter distortions and improve the quality and performance of the RF modulator required to deliver high quality signals in the down stream path.

1.3 Problem statement

In a typical digital communication system the RF modulator is responsible to up-convert the base band signal into the RF frequency. At the output of the RF modulator in addition to the fundamental carrier frequency, second and third harmonics are also present which must be rejected. For this reason an RF filter will be used at the output of the RF modulator which ideally has a flat amplitude/group delay response in the pass band and a large attenuation in the stop band.

Different technologies can be used for the design of the bandpass RF filter. Surface Acoustic Wave (SAW) or ceramic bandpass filters are commonly used in industry. However, most manufacturers, due to proprietary concerns prefer to design their own bandpass RF filter using discrete passive components. However, regardless of the technology used to design the bandpass RF filter, there are always some amplitude/group delay distortions on the modulated QAM signal caused from the RF filter.

When an RF filter is designed using discrete components, a flat amplitude/group delay response in the pass band requires fine tuning of the filter which is a labor intensive and expensive task. In addition the RF filter components have certain tolerances which will create even more variation on the amplitude/group delay response

during the mass production. Figure 1.2 shows these distortions, also the position of the QAM carrier in the pass band of a typical RF filter. Since these distortions are not symmetrical with respect to the carrier frequency, after down conversion into the base band, will have a complex effect on the QAM signal and cross couple the in-phase and quadrature components which in turn causes symbol scattering and bit error rate degradation of the system.

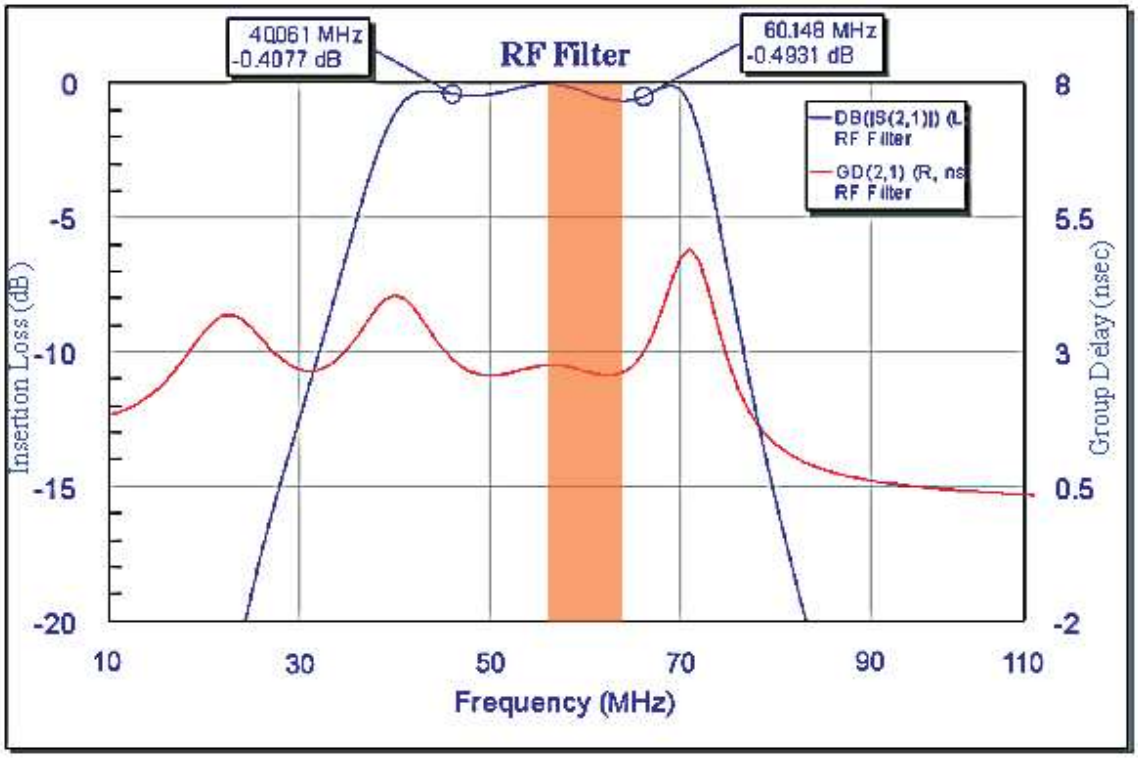

Figure 1.2 RF Filter distortion on the QAM signal

1.4 Research Objectives

The overall aim in this thesis, is to focus only on the amplitude and group delay distortion caused by the RF filter in the digital QAM modulator. The objective to reduce the distortion to the point where the system becomes DRFI[1] [4] compliant. These two type of distortions have significant effect on the quality of signal and CATV

[1]Downstream RF Interface Specification

system performance as a whole, and are therefore critical. The research plan that was followed is outlined below:

1. Set up a system level simulation and analysis, of a typical CATV digital communication system (using Matlab) to characterize the behavioral profile of the various types of amplitude and group delay distortions caused by RF filters. For the purpose of this simulation, a system performance figure of merit called *Bit Error Rate* was used in order to characterize the degradation effects of RF filter distortions.

2. Propose a feasible compensation technique using a base band Complex Digital Filter placed after the Square Root Raised Cosine (SRRC) filter in the modulator.

3. Propose a technique that calculates the optimum coefficients for the complex base band compensation filter.

4. Propose an alternative, in the form of a $Feed\ Forward$ structure and use channel equalization technique to calculate the tap weights for the filter.

5. Compare the simulation results of the optimization and equalization techniques.

1.5 Literature Review

The effect of amplitude and group delay distortion caused by the analog RF filter was studied in [5] for selective fading channels on digital radio. In the analysis, the amplitude distortions were modeled from a probabilistic point of view. The probability distribution of amplitude slopes were shown to characterize the amplitude distortion of the channel.

The effect of amplitude and delay slope of frequency selective fading channels on QPSK/8PSK modulation has been characterized in more detail by Douglas [6]. Specifically, Douglas characterized the sensitivity of 8PSK modulation by the slopes

of the amplitude distortions. Later, Mathiopoulos investigated the effect of amplitude and group delay distortions on the more complicated 1024-QAM signal [7]. Another special case study was accomplished by Ramadan [8] on the effects of group delay slope on the prediction of availability threshold of digital microwave system.

More specifically, in the case of 64-QAM and 512-QAM, performance degradation due to amplitude/group delay distortions was studied by Wu and Feher [9], Tricia [10], and Mathiopoulos [11], however, these studies, do not directly apply to the problem at hand in that they target different applications that have different system parameters. While the trends uncovered in the studies apply, they do not fully represent the sensitivity of the digital modulation schemes used in the CATV system. Furthermore, the studies are more limited in scope than the problem at hand, in that they only evaluated degradation effects of these imperfections, rather than proposing a solution to compensate for them. Therefore, a new investigation for the system under question, as well as proposing a solution to compensate these degradations, is necessary.

One approach to the problem is to derive a filter with an amplitude and group delay response that compensates for the distortions. A commonly used filter structure is a symmetric Finite Impulse Response (FIR) which has a linear phase response. Such structures can only compensate for amplitude distortions. The tap weights can be found using a *weighted Chebychev* algorithm. This yields a filter with equiripple error in the pass band. The development of this algorithm started with Herrmann [12] in 1970. Herrmann's work was followed by Hofstetter, Oppenheim and Siegel [13]. Then, a series of contributions were made in the 1970s by Parks, McClellan, Rabiner, and Herrmann [14–19]. The thrust of the work in the 1970s was to improve the convergence speed, efficiency, and other performance figures of the algorithm.

From this work, a computer algorithm rose to positions of prominence. This was a computer algorithm known as McClellan-Parks-Rabiner algorithm [20], which was published in 1979. This algorithm is considered better than the Remez Exchange Al-

gorithm [21] which was published in 1957. It is worth mentioning that improvements and contributions continue to be made [22, 23].

The approach taken in this research is to compensate for the amplitude distortions of the RF filter, with a linear phase digital filter. The coefficients for the filter are found with a variation of Newton's gradient method. Newton's method is described in [24] [14, 17, 19], and applies to the real filters so can not be directly applied to the design of a complex FIR low pass filter with complex coefficients.

To compensate the group delay distortion caused by the RF filter, a recursive all pass digital filter is used. The coefficients are found using the Quasi Newton Gradient Method, which is described in [24–29]. Again the method was adopted to find complex coefficients for the all pass filter.

A second approach, which is very different from the approach just described is also explored. In this second approach compensation is accomplished with a pre-equalizer feed forward structure. The coefficients for the pre-equalizer are determined using Least Mean Square (LMS) equalizer at the receiver, which is based on linear equalizer techniques which are widely used and extensively described by Haykin and Sayed [30] [31].

1.6 Thesis Organization

The thesis is organized into six chapters.

1. The first chapter provides an introduction to CATV system and to the problem of interest. It also outlines the two approaches that will be investigated to solve the problem.

2. Chapter two, provides information on the structure of a typical analog RF filter as well as its amplitude and group delay responses. A statistical analysis is performed on the RF filter to show the statistical behavior in terms of the

amplitude and group delay responses that result when component values are not precise. Three types of distortions are described and their degradation effects on the digital communication system performance are discussed. The compensation goal, is to sufficiently reduce the in-band amplitude and group delay distortions to be compliant with DOCSIS [32].

3. Chapter three focuses on the process of compensation of amplitude and group delay distortions of the RF filter, using a complex filter located at the base band portion of the digital modulator. An optimization algorithm that determines the coefficients of this complex base band filter is developed as well.

4. Chapter four parallels chapter three with a alternative compensation method, an *Adaptive Equalizer* is used to compensate for these distortions. An existing adaptive equalization algorithm is used. The compensation is accomplished with a pre-equalizer digital feed forward structure placed in the base band portion of the modulator. The tap weights for the pre-equalizer are obtained from an equalizer in a gold standard receiver. These tap weights, are exactly the same coefficients to be used for the pre-equalizer which compensates the RF filter distortions.

5. Chapter five centers on verification with simulation. Simulation is used to establish the methods to compensate the in band amplitude and group delay distortions sufficiently to comply to DOCSIS. The performance of these two methods are also compared. The performance measure used in the comparison is the *Modulation Error Ratio* (MER).

6. Chapter six contains the conclusions and discussion for future work.

2. Distortion in CATV networks

2.1 Introduction

In a CATV digital communication system, there are many sources of distortion. Virtually every component in the network introduces distortion. These include IQ modulators, RF filters, amplifiers and splitters. This section discusses the distortion caused by these components, starting from the IQ modulator and continuing along the downstream path.

2.2 Sources of Distortion

2.2.1 Digital IQ modulator

The input to the digital modulator section, is a digital data stream that has been mapped to the proper constellation points (typically 64 or 256 QAM) using an in-phase and quadrature phase carrier. There are two types of up conversion schemes in use today: the *super heterodyne* scheme and *direct up conversion* scheme. The method discussed here is the direct up conversion method of getting base band digital data into the QAM RF channel. Figure 2.1 illustrates a typical IQ modulator and its components. Ideally, the IQ modulator output has a Single Side Band (SSB) Suppressed carrier modulation format. Impairments in the modulator output occur when the I and Q local oscillators are not in perfect Quadrature. Another source of distortion is when the gains in the I and Q paths are not exactly equal. These two impairments are referred to as phase and gain imbalance, respectively. These will cause the modulator to create small side bands on the opposite side of the carrier (in single side band transmission). This spectral growth on the other side of the carrier is

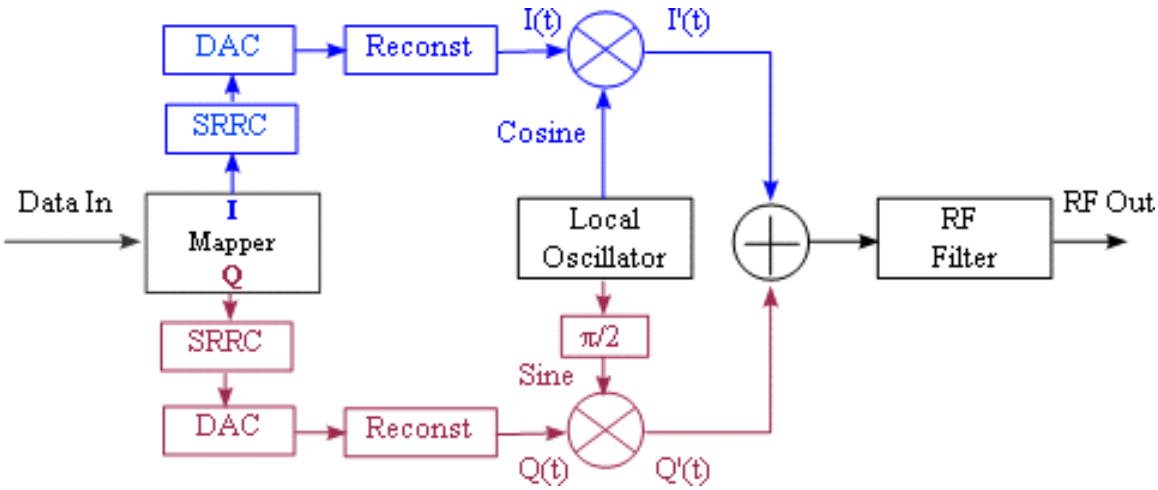

Figure 2.1 Block diagram of a Digital I/Q modulator (transmitter)

referred to as the *Image* of the original QAM carrier. Since the bandwidth allocation in CATV has the QAM carriers tightly packed, even small amounts of image (due to gain/phase imbalance) of one carrier degrade the signal to noise ratio of neighboring carriers.

Another source of impairment in an IQ modulator is caused by a DC offset in I and Q paths. The DC on I and Q paths will create carrier leakage at the output of the modulator. The gain/phase imbalance between I/Q will shift the modulated symbols from their ideal position in the constellation chart. Yet another source of distortion is differential delay, the delay difference between I and Q pathes causes a loss of *Orthogonality* between I and Q signals. One can take a closer look at the IQ modulation process in Figure 2.1. For 256 QAM modulation, I(t) and Q(t) signals would take on the values ± 1, ± 3, . . . ± 15. The distortions caused from the IQ modulator can be expressed in a mathematical form. For the time being, it is assumed all other CATV subsystems, such as *amplifiers*, cable, customer receiver functionalities such as *carrier recovery* and *symbol timing recovery* are perfect and do not impose any distortion. If the DC offsets for the I and Q channels, denoted as C_I and C_Q respectively, are added to the I and Q channels they become: $I + C_I$ and $Q + C_Q$. The amplitude imbalance of Q and I paths can be denoted by a coefficient

α. The mathematical form of the final RF signal would be:

$$R(t) = [(I(t) + C_I]cos(\Omega_{LO}t) + \alpha[Q(t) + C_Q]sin(\Omega_{LO}t + \varphi) \tag{2.1}$$

where φ indicates the phase imbalance. One can use trigonometric identities to simplify equation (2.1) to the form below:

$$R(t) = A(t)cos(\Omega t) - B(t)sin(\Omega t) \tag{2.2}$$

where A(t) and B(t) are:

$$A(t) = [I(t) + C_I] - \alpha[Q(t) + C_Q]sin(\varphi) \tag{2.3}$$

$$B(t) = \alpha[Q(t) + C_Q]cos(\varphi) \tag{2.4}$$

Equation (2.3) shows that for $\varphi \neq 0$ the I and Q channels are not *Orthogonal*.

2.2.2 Analog RF Filter

The output of the IQ modulator contains the fundamental carrier as well as second, third, fifth, and other harmonics of the fundamental carrier. Since CATV is a broad band system, it is obvious that the presence of these harmonics at modulator output is not desirable, because these harmonics will fall within higher frequency channels, causing interference. Further more, in order to maintain adequate *Broad Band Noise* (BBN) level at the digital IQ modulator output, it is necessary to use a band pass RF channel Filter at modulator output.

This filter is meant to have a flat *amplitude* and constant *group delay* response in the pass band, and a reasonably sharp transition band with adequate attenuation on the stop band. However, usually this is not the case and there are ripples in the amplitude and group delay responses. To get a better understanding of RF filter distortions, Figure 2.2 shows a simulated frequency response of a typical band pass RF filter (Chebyshev).

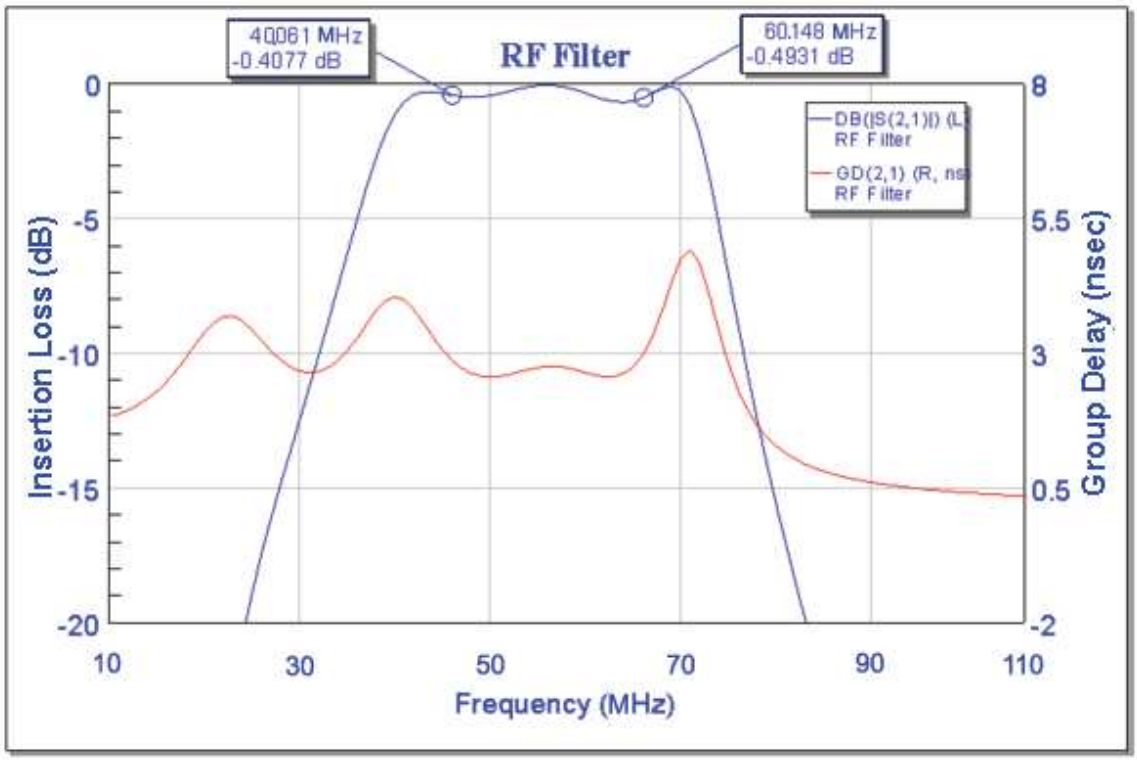

Figure 2.2 Typical RF band pass filter amplitude and group delay responses

Ideally a flat group delay response would be preferred, however, it is clear from Figure 2.2 the group delay is not constant across the band (right axis in nanosecond). In some regions, the slope of variations is constant, while close to the band edges, it follows a somewhat *parabolic* shape, and at some points it has a *sinusoid* shape. Amplitude response have similar variations across the band (left axis is the insertion loss in dB). This *group delay* and *amplitude* variations of the RF filter, are certainly a source of impairment on the IQ modulated signal.

2.2.3 Coaxial Cable

The transmission medium in CATV networks is a combination of optical fiber and coaxial cable. The distribution network, see Figure 2.3, is primarily co-axial cable. The trunk lines, which at one time were coaxial cable are now, for the most part, optical fiber.

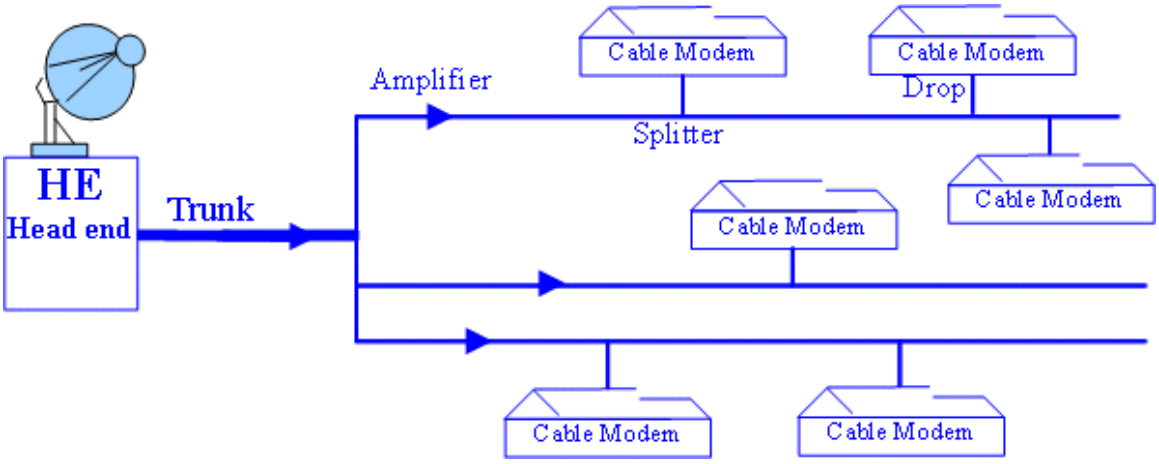

Figure 2.3 The general block diagram of the CATV system

The CATV network, which was designed to deliver analog television signals to the antenna terminal of subscribers TV sets, has evolved into a sophisticated multi-service network. In the evolution, some of the coaxial cable was replaced with linear optical fiber, however, the network remains largely the same. For instance the residential portion of the plant is largely coaxial cable and is still channelized into 6 MHz Frequency Division Multiplex (FDM) channels. While the channelization has not changed many of the carriers are digitally modulated signals. The digitally modulated signals are quite sensitive to distortion caused by non linearities. This requires all components in cable network be linear including amplifiers, passive components, and fiber optic links.

A coaxial cable has a center conductor surrounded by a concentric cross section dielectric, and by an outer conductor known as the shield. The RF signal sets up an electromagnetic field in the cable that has a configuration known as a *Transversal Electric* and *Magnetic* (TEM) field. The characteristic impedance of coaxial cable is related to the ratio of the diameter of the outer to the inner conductor and the dielectric constant of the insulator that separates them. The characteristic impedance is given by:

$$Z_o = \frac{138}{\sqrt{\epsilon_o}} \log(\frac{D}{d}) \tag{2.5}$$

where Z_o is the characteristics impedance in Ohms, D is the outer conductor diameter and d is the inner conductor diameter. Coaxial cable is a lossy medium which means the signal loses amplitude as it travels along the cable. the attenuation which is a function of the frequency, is caused by: radiation through the shield, resistive losses in the cable conductors, signal absorbtion in the dielectric, reflections at places where cables are joined, spliced, connected, or along the cable where the characteristic impedance is not uniform. The general equation for the residual loss of a coaxial cable is as follow:

$$\alpha = 4.344(\frac{R}{Z_o}) + 2.774F_p\sqrt{\epsilon}f \tag{2.6}$$

where, α = attenuation in (dB /100 ft), R = the effective Ohmic resistance of the cables, F_p = the power factor of the dielectric used, f = the frequency in MHz and ϵ = relative permittivity of the dielectric in the coaxial cable.

2.2.4 Amplifier

Amplifiers are used to compensate for the insertion loss and to replace the power tapped off and sent to a subscriber. Due to the thermal random noise inherent in the electronic components, amplifiers always introduce noise as a function of temperature and bandwidth in which the noise is measured. This *Broad Band Random Noise*, known as *Thermal* noise, has a power that depends on the temperature of the device and bandwidth given by: $\eta_p = KTB$, Where η is the noise power density in milli-Watts (-174 dBm/Hz), K is Boltzman constant (1.3807 x 10^{-23} $Joules/K$), T is the absolute temperate in degrees Kelvin, B is the bandwidth of the noise in Hz.

Since the signal power in CATV is expressed in terms of decibels referenced to 1 mV (dBmV) and the characteristics impedance of the cable is 75 Ohms, the thermal noise at room temperature (T = 300 K) equals to: $n_p(dBmV) = -125.1 + 10\log B$, where n_p is the noise power in dBmV and B is the bandwidth in Hz. To be more specific about effective noise power within a single CATV channel, although the defined bandwidth in a analog TV channel is 6 MHz, the receiver noise bandwidth is usually less. According to the FCC's rules [33] the effective bandwidth is about 4 MHz. The

thermal noise power at room temperature for bandwidths of 4 MHz and 6 MHz, are $n_p = -59.1\ dBmV$ and $n_p = -57.3\ dBmV$ respectively.

The amplifier itself, adds noise that can be considered as additive input noise. A figure of merit is often used to represent the noise behavior of an amplifier. This figure of merit is the ratio of the amplifier generated noise to the thermal noise of the input resistance. This ratio is known as the *noise figure* of the amplifier. Therefore an amplifier with a noise figure of F_A dB will have a total input noise power of n_A, such that

$$n_A = n_p + F_A \tag{2.7}$$

where n_A and n_p in dBmV and F_A is the noise figure in dB. In the same manner, the output noise power is the input noise power plus the gain of amplifier, where G is in dB and n_{out} is in dBmV.

$$n_{out} = n_p + F_A + G \tag{2.8}$$

One of the important attributes of amplifiers is linearity. However, the amplifiers usually have some degree of non linearity which causes intermodulation products. This distortion in CATV becomes significant due to the fact that there are many frequency multiplexed RF carriers passing through the amplifier at the same time, creating the mixing products in the same way that a communications mixer does. In general, a mixer is a non linear device (with quadratic input/output characteristic) that can be estimated using a quadratic second order polynomial, generating intermodulation products between every carrier pair. The frequency of these intermodulation products coincides with the addition and subtraction of the original carriers frequencies, and typically with lower amplitude. In the case of amplifiers, this non linear behavior is mostly caused by compression which is a saturation effect that occurs near the DC supply voltage of the amplifier. Compression can be formulated as a nonlinear input/output characteristic composed of second and third order polynomials. such compression creates second and third order distortion or intermodulation distortion to the amplified signal. The results of this distortion can be classified as even order distortion, odd order distortion, and cross modulation.

If an amplifier is perfectly linear, the output can be expressed as input multiplied by a constant, the constant is known as gain of amplifier. When the input / output relation (transfer function) is not linear, the output can be expressed as function of the input using this polynomial

$$e_o = Ae_i + Be_i^2 + Ce_i^3 + ...$$

where e_o is the output signal, e_i is the input signal, and the set of coefficients A, B, C, are the gains for various input signal power levels.

In this equation, the terms with even numbered powers, indicate the even number distortions, and terms with odd numbered powers indicate odd order distortions, the dominant types of distortion in CATV are second order and third order distortions. In solid state amplifiers, odd order distortions, especially third order distortions, are significant and create intermodulation products between two carriers.

Composite Triple Beats (CTB) is the spectrum produced between multiple carriers. These intermodulation products appears at three times the frequency of the original carriers. These products appears when the amplifier generates third order distortion and the amplitudes of all carriers are the same.

Another form of distortion will occur when multiple carriers with multiple modulated amplitudes pass through an amplifier with a third order nonlinearity. In this case, the intermodulation product is called Cross Modulation (XMOD). All these distortions, if too high, can create significant degradation to the quality of the signal, hence on the performance of the system. Excessive CTB and CTO can degrade a parameter in CATV known as Modulation Error Ratio (MER) which is reduced to the signal to noise ratio of the received signal.

2.2.5 Passive Coaxial Components

The CATV network has a branch and leaf structure in the downstream direction. The new branch and leaves are created by devices called splitters. In the up stream

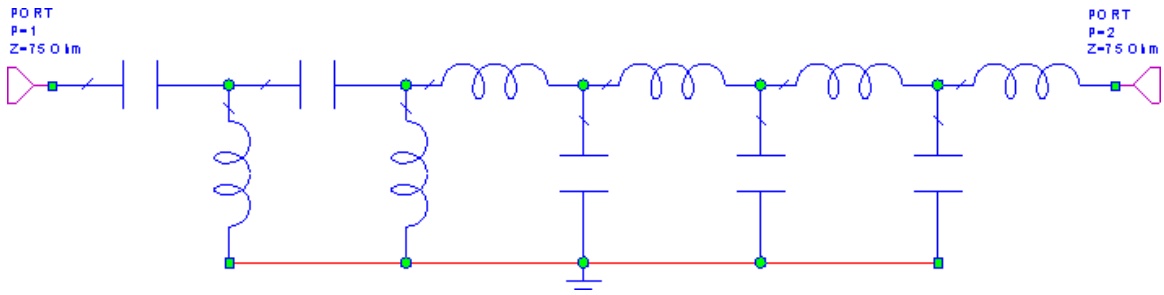

Figure 2.4 The Schematic of a typical Band Pass RF Filter

direction the branches and leaves are merged with devices called combiners. It is evident that thousands of signal splitters and signal combiners are needed to implement a CATV network.

Splitters are passive components that are used for splitting the signal into two or more signals with lower power, the most elementary splitter is a 1:2 splitter which splits the signal into two, each with equal power. Combiners are used to combine two or more signals into one signal in the up stream direction. Another commonly used device is directional coupler to divide the signal into outputs with un-even output powers which is necessary at many points in the network. The typical impairments caused by splitters and directional couplers are related to insertion loss. The insertion loss is controlled by using amplifiers along the path.

2.3 RF Filter Distortions

The focus of the research presented in this thesis is on the distortion caused by RF filters. An understanding of this distortion can be gained from the structure of the filter. The structure of a typical analog band pass RF filter will be used to explain amplitude and group delay responses. It will also be used to show the sensitivity of these responses to the tolerances of filter components.

The RF filter used in the modulator, is a *Band Pass Filter*. Such filters are explained in textbooks such as Pozar [34]. A typical schematic diagram of this filter is shown in Figure 2.4.

This filter consists of 11 passive components: six inductors and five capacitors. Generally, variable (i.e. tunable) components will be used and are tuned to get a flat response across the frequency band of interest. However, during mass production, the tuning process is very expensive so it is preferable to use fix components. The fixed components do not have precise values which translates into changes in the amplitude and group delay responses.

The component tolerances will cause both the amplitude and group delay responses to vary from the ideal response as shown in Figure 2.2. The concern is to what extent these two parameters vary, and the rate of change of the variation. It is important to know the shape and severity of these variations in order to develop an efficient compensation circuit.

One method to characterize the form and extent of variations is to use an RF simulation tool. Using this tool it is possible to setup a simulation to analyze the RF Filter schematic shown in Figure 2.4. Looking at this figure, one can setup a statistical simulation known as *Monte Carlo* simulation in which, the value of components will randomly change within ± 0.05 percent of the nominal value, which is a reasonable practical tolerance. In the Monte Carlo simulation setup, it is possible to define a uniform shape for the *probability density function* (*pdf*) of the component values.

In order to obtain enough statistical confidence in the simulation outcome, the simulation was performed for 5000 iterations. The result of the Monte Carlo simulation, is shown in the Figure 2.5.

Looking at the Monte Carlo simulation results in Figure 2.5, which is the result for a uniform distribution with 0.05 component tolerance, it is clear that both amplitude and group delay responses depart from the original response.

The maximum variations of attenuation in the pass band is about 2 dB with respect to the desired frequency response. Like wise, there is up to 2 nsec variations in group delay in the pass band. In terms of the shape of the variations, they are smooth and appear to be without high frequency content.

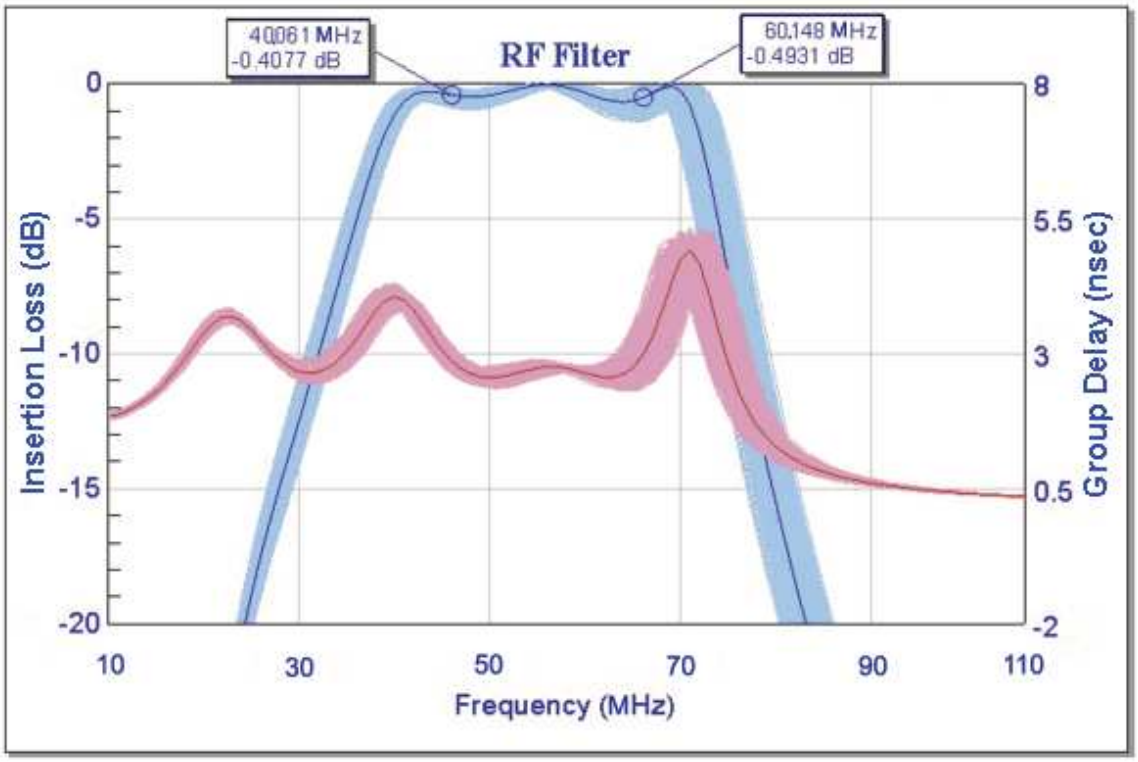

Figure 2.5 The Monte Carlo simulation result of RF Filter

One solution to get the desired response is to replace components that have poor tolerance, however, this solution is not practical for mass production. In addition there are limitations, to get a completely flat response even with exact component values due to resistive component losses, component value availability. For example, optimum tuning for the frequency band at near 950 MHz, produces an amplitude response that decreases with frequency due to the resistive losses in the components. One of the other common situations is non symmetrical ripples in amplitude response as the result of component value variations.

2.4 Effects of RF Filter on distortion of QAM signals

Depending on where the QAM signal is located within the RF Filter pass band frequency, the impairment effects vary in type and severity. The distorted frequency response can be broken into Hermitian symmetric and Hermitian antisymmetric com-

ponents with respect to carrier frequency, at baseband the Hermitian symmetric components give rise to real coefficients of the complex low pass filter, while the Hermitian antisymmetric give rise to complex coefficients. The antisymmetric components causes cross coupling between the I and Q channels.

To show this in a mathematical form, let $H(e^{\jmath\Omega})$ denote the transfer function of the low pass equivalent filter. If $H(e^{\jmath\Omega}) = H^*(e^{-\jmath\Omega})$, $H(e^{\jmath\Omega})$ has *conjugate or Hermitian symmetry*. This means that the magnitude of the transfer function has even symmetry and the phase of the transfer function has odd symmetry. This is true if and only if the impulse response is *real* [35].

If $H(e^{\jmath\Omega}) = -H^*(e^{-\jmath\Omega})$, $H(e^{\jmath\Omega})$ has *conjugate or Hermitian antisymmetry* or *odd symmetry* with its impulse response being purely imaginary. Moreover, any function $H(e^{\jmath\Omega})$ can be decomposed into its *odd* and *even* functions as $H(e^{\jmath\Omega}) = H_e(e^{\jmath\Omega}) + H_o(e^{\jmath\Omega})$, where

$$H_e(e^{\jmath\Omega}) = \frac{1}{2}[H(e^{\jmath\Omega}) + H^*(e^{-\jmath\Omega})] \tag{2.9}$$

$$H_o(e^{\jmath\Omega}) = \frac{1}{2}[H(e^{\jmath\Omega}) - H^*(e^{-\jmath\Omega})] \tag{2.10}$$

Obviously, $H_e(e^{\jmath\Omega})$ has *conjugate symmetry* and $H_o(e^{\jmath\Omega})$ has *conjugate antisymmetry*. Intuitively, the magnitude and group delay distortion caused by the low pass equivalent filter can be corrected with a low pass filter that has the inverse response.

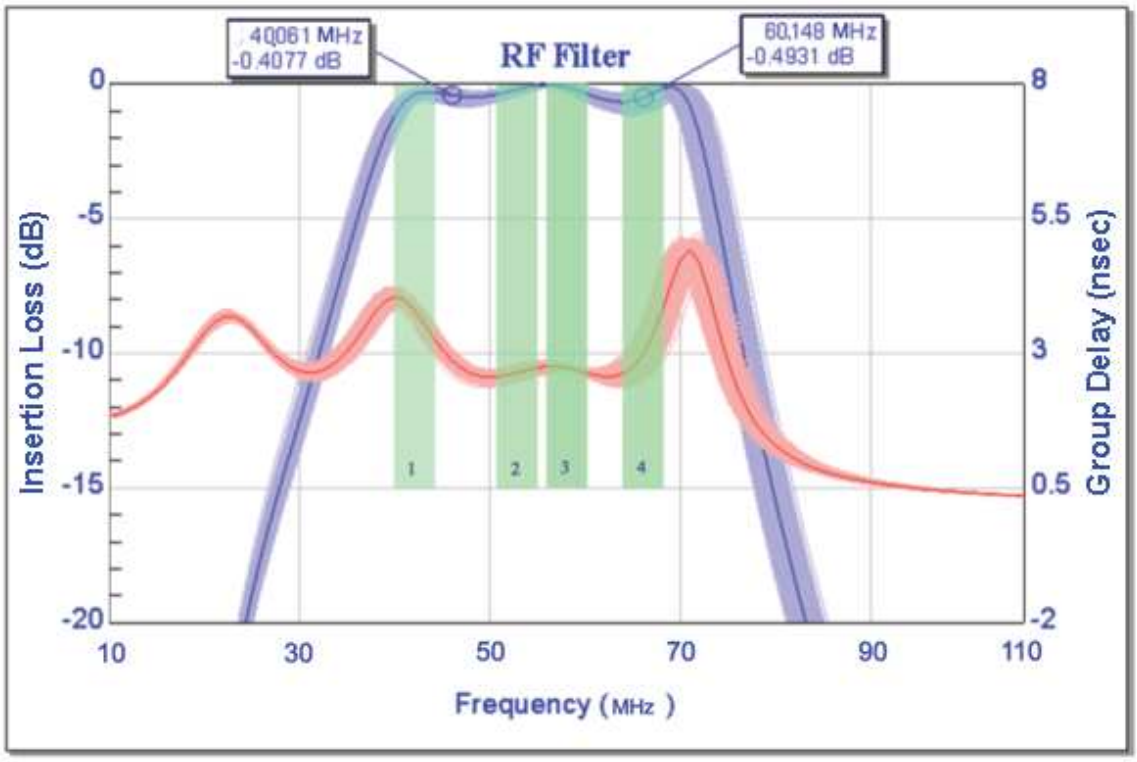

Figure 2.6 The position of QAM signal in the RF Filter response

2.4.1 Different types of distortions

Previous work on the effects of amplitude and group delay of the channel on the QAM signal [6] to [11], indicates that amplitude distortion across the bandwidth can be segmented into three categories: linear slope distortion, parabolic slope distortion and sinusoid slope distortion. This can be seen in Figure 2.6.

A QAM signal may span a bandwidth that is characterized with a single category or perhaps two or more categories. Figure 2.6 shows the response of an RF filter that is wide enough to pass several QAM signals. The four shaded areas show the location of four QAM signals. QAM carrier number one and four are located at the band edges. These signals experiences parabolic group delay response. QAM carriers number two and three experience a group delay response with linear slope.

As for amplitude response, carrier numbers two, three, and four experience linear

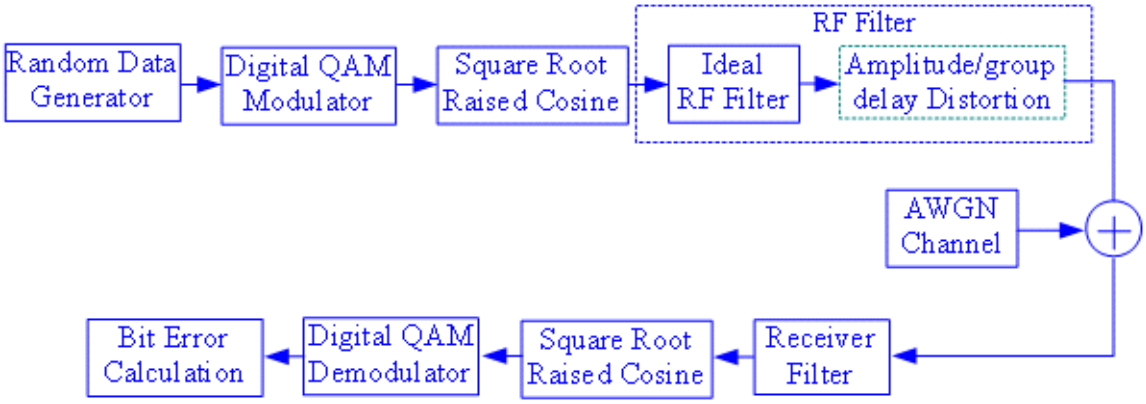

Figure 2.7 The Communication system used during the simulation

amplitude slope distortion and carrier number 1 experience parabolic amplitude distortion. The more rare category of sinusoid distortion is not illustrated in Figure 2.6.

2.4.2 Description of simulation parameters

A Matlab simulation was used to characterize the system performance degradation of the CATV system. The digital communication system used for the simulation is shown in the Figure 2.7. In this figure the RF filter is modeled as an ideal filter in cascade with one that has the equivalent distortion. The parameters used in the simulation are listed in Figure 2.8 on page 25. These parameters are listed as typical and quoted from [32] Annex-B 256 QAM modulation.

The simulation was setup entirely in the complex base band domain. Both SRRC filters at transmitter and receiver are ideal with $x/sin(x)$ compensation at the transmitter, hence the whole system satisfies the *Nyquist* first criterion. In the simulation, no carrier phase offset or timing offset is introduced, i.e, perfect timing and carrier are used. To generate the modulating signal source in Matlab, the integer random source (randint) function was used with enough data length (10^7) symbols to get adequate statistical confidence. For the modulation, the *qammod* function was used and for the pulse shaping, the *rcosine* function was used with a symbol rate of 5.360537 Msym/s, a roll off factor of 0.12 and an up sample factor of 16. Different amount of

amplitude/group delay distortions was generated and used as the input to the filter design tool in Matlab (FDATOOL) and the resulting filter coefficients were used in the simulation as the RF filter distortions. At the receiver, the function $rcosflt$ was used for the SRRC filter with identical parameters to the transmitter SRRC filter. The function *qamdemod* was used for demodulation in receiver after the match filter, and finally the bit error rate was measured using the *biterr* function.

To clarify the method by which the effects of the RF Filter distortions was generated, consider a band pass filter with equivalent base band transfer function $H(f)$ between the transmitter and receiver. The transfer function can be expressed in terms of magnitude and phase as follows:

$$H(f) = |H(f)|e^{j\theta(f)} \tag{2.11}$$

In Equation (2.11) the magnitude of the transfer function represents the amplitude frequency response $Amp(f) = |H(f)|$, the group delay is obtained from $Del(f) = -\frac{1}{2\pi}\frac{d\theta(f)}{df}$, also assuming the group delay of the filter is constant, e.g., $Del(f) = 1$. The frequency response of $H(f)$ may be shown as $Amp(f)$ as follows:

$$Amp(f) = \begin{cases} L_A f, & \text{for } \textit{linear} \text{ amplitude slope} \\ P_A f^2, & \text{for } \textit{parabolic} \text{ amplitude slope} \\ S_A sin(2\pi K f/2F_{BW}), & \text{for } \textit{sinusoid} \text{ amplitude slope} \end{cases} \tag{2.12}$$

where

$$F_{BW} \triangleq \frac{(1+\alpha)R_S}{2}$$

R_S is the symbol rate, and α is the roll off factor. In the case of 256-QAM Annex-B, $R_S = 5360537$ sym/s, therefore for roll off factor 0.12 the $F_{BW} \simeq 3$ MHz. Also K determines the number of sinusoidal cycles within the QAM signal bandwidth, which in this case equals four cycles (Although the sinusoidal distortion is not of concern

here, still the equation is presented).

Destructive effects of the filter distortions can be observed by using the BER curve described in the previous section (bit error rate indicates the ratio between the number of erroneous bits, to the total number of received bits). For instance, to characterize the amount of degradation of bit error curve for a 256-QAM modulation as a result of linear, parabolic, and sinusoid distortions, the worse case will be considered which is the maximum amplitude distortion within the filter bandwidth, introduced here as A_m, where

$$A_m = \begin{cases} L_A(2f_{BW}), & \text{for } \textit{linear} \text{ amplitude slope} \\ P_A(f_{BW})^2, & \text{for } \textit{parabolic} \text{ amplitude slope} \\ S_A, & \text{for } \textit{sinusoid} \text{ amplitude slope} \end{cases} \tag{2.13}$$

Starting with linear amplitude distortion, the value of $L_A = 0.1 - 0.5$, incrementing with 0.1 step size, produces five different BER curves which can be compared with the ideal 256 QAM BER curve to measure the amount of BER degradation as the result of the RF filter distortions (the ideal BER curve is the result of simulation with no distortion, which is fairly close to the theoretical 256 QAM BER). In order to get enough confidence on the BER simulation results, a large number of digitally modulated symbols were used during simulation to get consistent results for the low bit error probability ranges. This family of curves for the linear amplitude distortion is shown in Figure 2.9.

Looking at Figure 2.9, there are six BER curves, lower left curve shows the ideal BER curve for 256 QAM, the next five curves are the BER curves after applying the linear slope amplitude distortion for $L_A = 0.1$, 0.2, 0.3, 0.4, 0.5 respectively. A reference point on the BER curves can be chosen for a probability of error (P_e) equal to 10^{-3}, using this reference point, degradation of the bit to noise energy ratio(E_b/N_o) can be evaluated on these curves, and plotted versus the maximum amplitude slope (A_m) within the filter bandwidth.

This process can be repeated for parabolic, and sinusoid amplitude distortions.

Parameter	Specification					
	Annex A		Annex B		Annex C	
QAM Modulation	64 QAM	256 QAM	64 QAM	256 QAM	64 QAM	256 QAM
Symbol Rate	6952000 Sym/s	6952000 Sym/s	5056941 Sym/s	5360537 Sym/s	5274000 Sym/s	5274000 Sym/s
M/N	869/1280	869/1280	401/812	78/149		
Bits-Per_Symbol	6	8	6	8	6	8
RF BW	8MHz	8MHz	6MHz	6MHz	6MHz	6MHz
Information bit rate per-QAM channel (before FEC bits)	~38.4Mbps	~51.3Mbps	~26.9Mbps	~38.8Mbps	~29.162Mbps	~38.8Mbps
RF Transmission Bit-Rate per QAM channel (after FEC)	~41.7Mbps	~55.6Mbps	~30.3Mbps	~42.9Mbps	~31.6Mbps	~42.2Mbps

Figure 2.8 DOCSIS parameters used for BER degradation simulation

For instance, again for the parabolic amplitude distortion, the value of $P_A = 0.1 - 0.5$ is chosen, likewise, for sinusoid amplitude distortion the value of $S_A = 0.005 - 0.025$. For the sake of comparison, the resulting 256-QAM BER curves are plotted in the Figures 2.10 and 2.11.

Finally, comparison of results for all types of amplitude distortions are shown in Figure 2.12. Note that the horizontal axis indicates the amplitude slope in (dB/MHz) for the case of linear amplitude slopes, for the parabolic slope distortion this axis must be divided by four, also for the sinusoid slope, must be divided by ten.

It is obvious that for a given maximum amplitude distortion(A_m), sinusoid amplitude distortion provides the worst degradation effect on the QAM signal , followed in decreasing order of degradation by parabolic, and linear amplitude distortions which creates the least degrading effect.

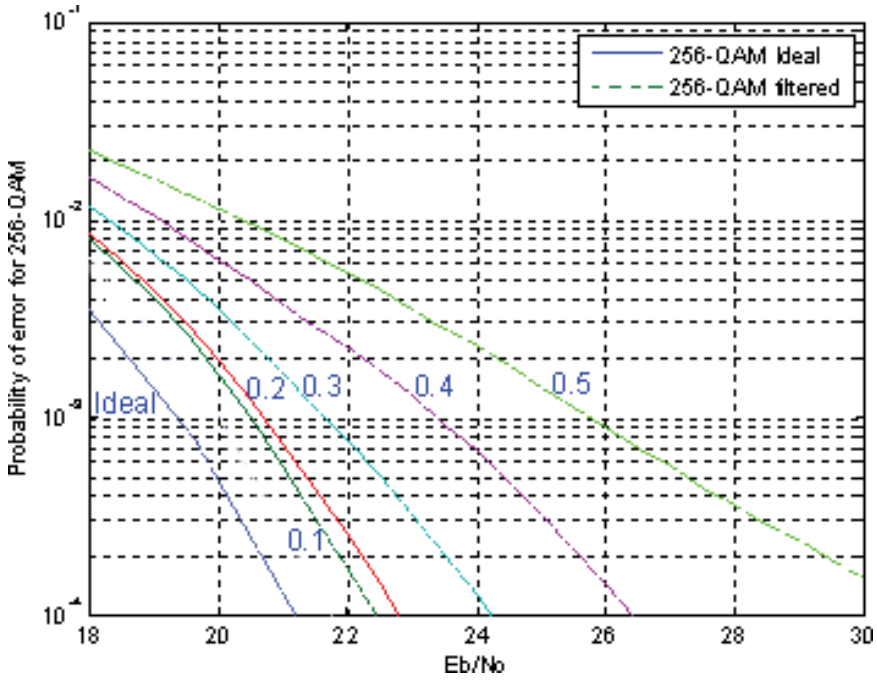

Figure 2.9 The BER degradation results due to linear amplitude distortion

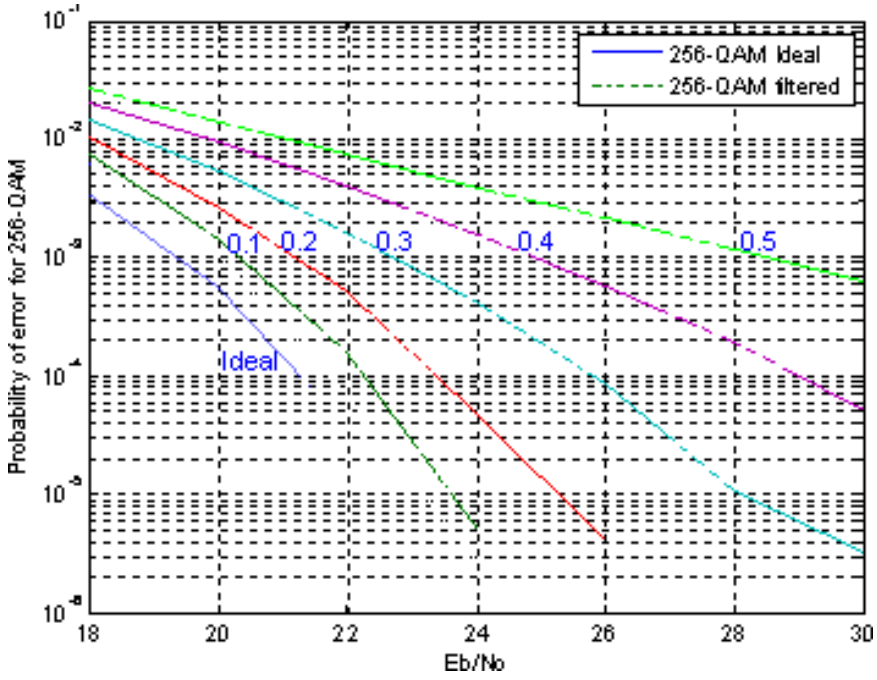

Figure 2.10 The BER degradation results due to parabolic amplitude distortion

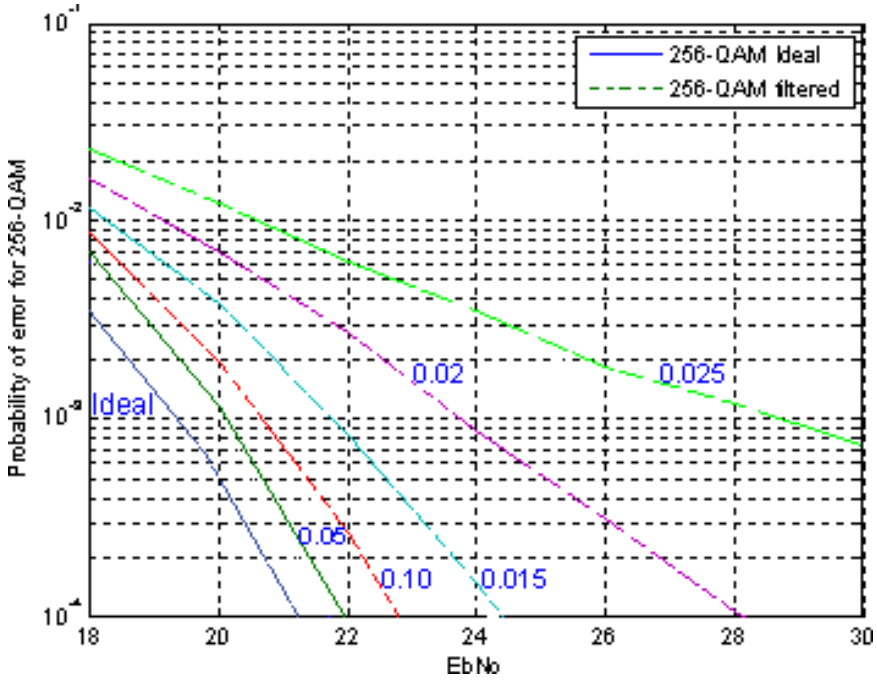

Figure 2.11 The BER degradation results for the sinusoid amplitude distortion

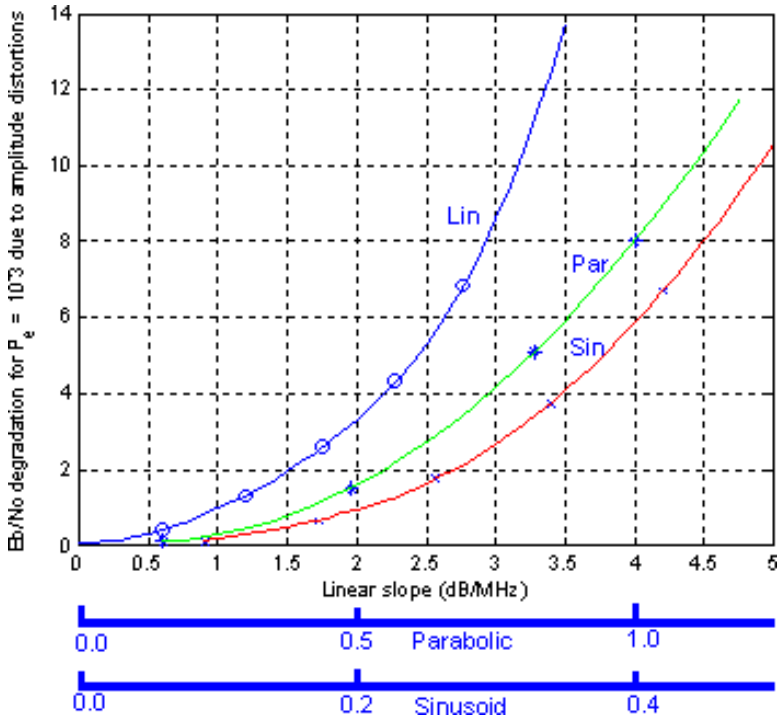

Figure 2.12 Eb/No degradation comparison graph between linear, parabolic, sinusoid slope amplitude distortions

Likewise, for the characterization of the BER degradation of the CATV system due to the group delay distortions, it is presumed the amplitude response is flat (i.e., $Amp(f) = 1$), a series of the group delay distortions are generated with linear, parabolic and sinusoid slopes. These distortions can be defined as follows:

$$Del(f) = \begin{cases} L_D.f, & \text{for } \textit{linear} \text{ group delay slope} \\ P_D.f^2, & \text{for } \textit{parabolic} \text{ group delay slope} \\ S_D.sin(2\pi Kf/2f_{BW}), & \text{for } \textit{sinusoid} \text{ group delay slope} \end{cases} \tag{2.14}$$

To show the simulation results, a maximum group delay τ_m will be defined in the filter bandwidth ($2f_{BW}$) as follows:

$$\tau_m \triangleq \begin{cases} L_D(2f_{BW})\, ns, & \text{for } \textit{linear} \text{ group delay slope} \\ P_D(f_{BW})^2\, ns, & \text{for } \textit{parabolic} \text{ group delay slope} \\ S_D\, ns, & \text{for } \textit{sinusoid} \text{ group delay slope} \end{cases} \tag{2.15}$$

Starting with the linear slope parameter $L_D = 0.15 - 0.75$ with 0.15 step size, five different BER curves were generated for a 256-QAM signal accompanied by the ideal curve for 256-QAM with no distortion. Similar setting will be used for the parabolic slope. The resulting BER curves are shown in the Figures 2.13 and 2.14.

In order to compare the results of BER degradation due to different group delay distortion slopes, the E_b/N_o degradation for $P_e = 10^{-3}$ versus the group delay slopes are graphed in the figure 2.15.

From the results shown in Figure 2.15, it is obvious that the degradation effects of the linear slope group delay distortion on the BER performance of the QAM signal is much higher than that of sinusoid and parabolic distortions respectively.

It is understood from the group delay comparison curves that only a few nanosecond linear group delay slope would cause significant BER degradation. In this simulation the DOCSIS parameter settings were used, SRRC roll off = 0.12, for a 6

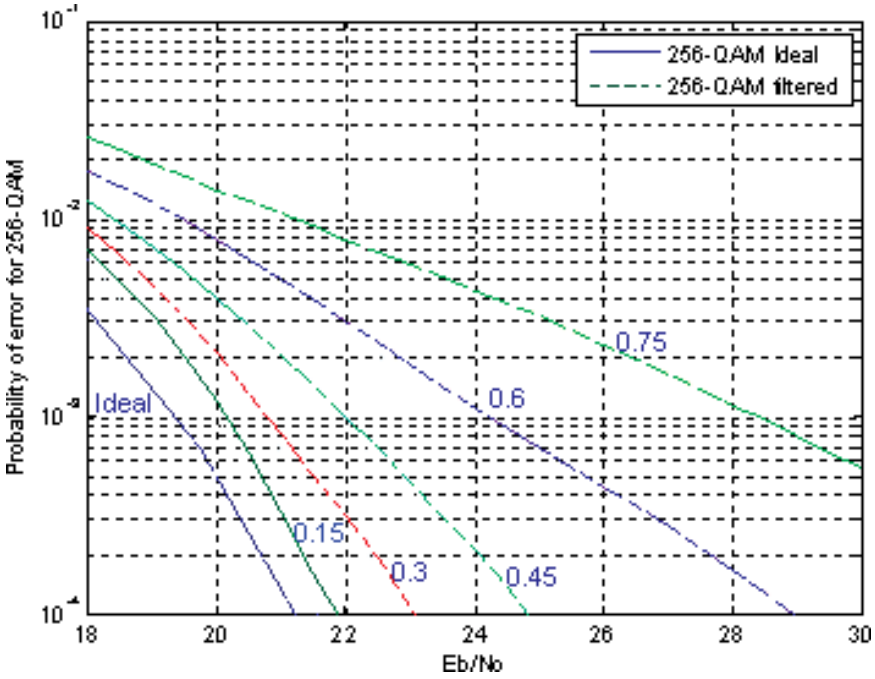

Figure 2.13 The BER degradation results for the linear group delay distortion

MHz wide QAM signal, up-sampling is 16, the sampling frequency will be about 85.7 Msym/s and sampling time is about 11.66 nanosecond. Therefore changes in order of a tenth of a sample period will have noticeable degrading effect on the BER.

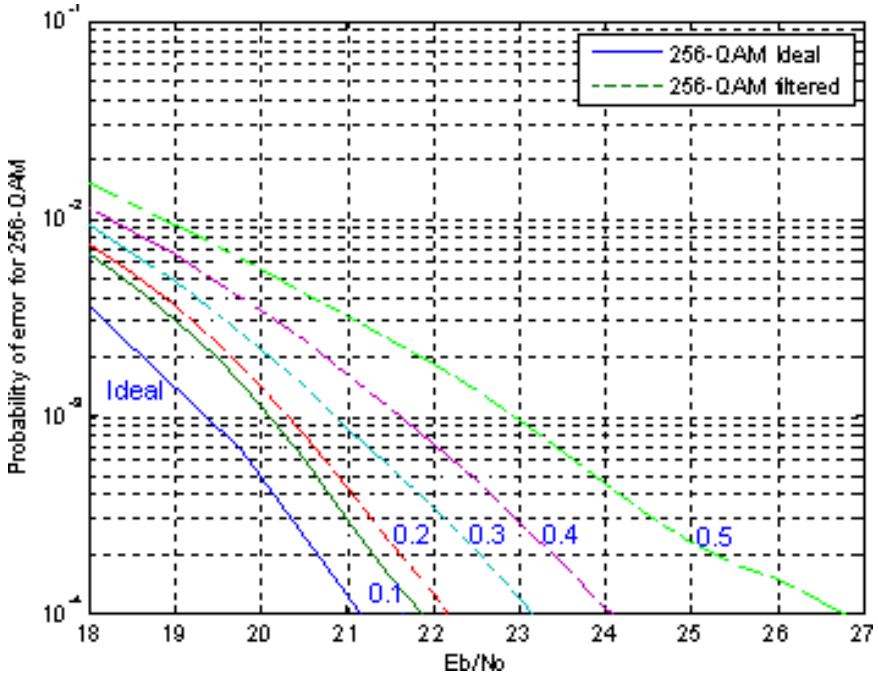

Figure 2.14 The BER degradation results for the parabolic group delay distortion

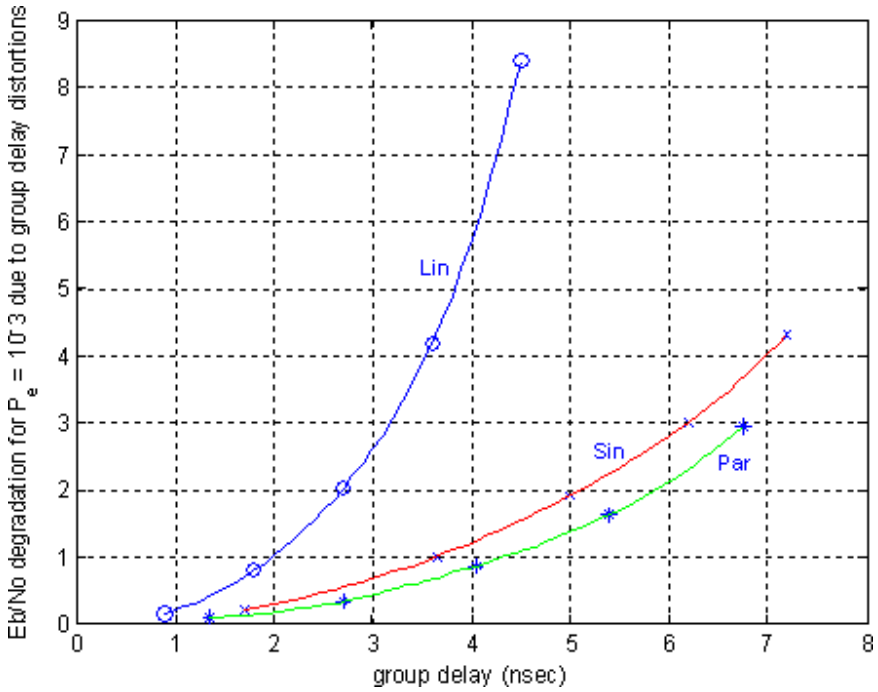

Figure 2.15 Eb/No degradation comparison graph between linear, sinusoid, parabolic slope group delay distortion

2.5 Summary

In this chapter various types of distortions in a CATV system were reviewed. These distortions impose degradation effects on the quality of the QAM signal. This thesis focuses on the distortions created in the RF filter in the head end block.

To evaluate the RF filter distortions in a statistical sense, a Monte Carlo analysis was performed using an RF simulation tool. The simulation showed the effects of the filter component tolerances, assuming a reactive lumped element low pass-high pass ladder circuit implementation. The results of this simulation indicated that the amplitude and group delay responses will follow the original response of the filter without any abrupt changes in the filter response as long as the filter is reasonably well tuned.

It is also understood that these distortions can be categorized into three major categories: linear, parabolic, and sinusoid distortions. From the Monte Carlo simulation result, considering the QAM signal bandwidth and using practical CATV system parameters, it is understood that it is more likely for a QAM signal to experience linear and parabolic distortions. For the QAM signal with a 6 or 8 MHz bandwidth, it is almost impossible to experience sinusoid distortion since the granularity of the variations of the RF filter response is much larger than an 8 MHz span. Nevertheless, degradation effects of sinusoid distortion were simulated.

One might be concerned as to which type of these distortions will cause the worse degradation effects on the QAM signal, or what is the sensitivity of the QAM signal to these distortions. To address this concern, a simulation carried out on the CATV system using Matlab, characterizing its BER degradation due to the RF filter distortions. For performance measurement, the BER performance parameter was used. A family of BER curves for the amplitude and group delay distortions were generated with three different slopes: linear, parabolic and sinusoid for each.

Using these curves and comparing the bit to noise energy ratio(E_b/N_o) degradation for a probability of error (P_e) of 10^{-3}, another curve was derived showing the

degradation of (E_b/N_o) versus amplitude and group delay slopes for three different distortion slopes.

Looking at the comparison graph for group delay distortion, it is understood that the linear group delay slope creates the worse degradation effect on the QAM signal, followed in order of decreasing degradation by sinusoid and the parabolic distortions.

For the amplitude distortions, on the other hand, the degradation of the BER performance to the sinusoid amplitude response is the worst. This suggests that to define a compensation method, sinusoidal amplitude distortion needs special attention. Note that for this simulation four cycles of the sinusoid within the QAM signal bandwidth were used, even though this case is very rare and less likely to happen in practice, but still this simulation was considered only for comparison with other types of distortions. This is mostly because the granularity of variations of the filter amplitude response is much larger than the QAM signal bandwidth. Therefore, in practice the sinusoid distortion does not happen; also, the parabolic distortion is less likely to happen within the pass band of the RF filter. The linear amplitude slope is the one that needs to be dealt with most of the time.

To compensate for the RF filter distortions, two methods will be proposed. In the first method, a complex digital low pass filter will be used after the SRRC pulse shaping filter in the base band. This filter will have a frequency response which is the inverse of the RF filter. The coefficients of this digital filter have to be determined in some way, which is explained in detail in Chapter three.

The second compensation method will use a pre-equalizer located before the SRRC filter in the base band. The coefficients of this FIR structure will be determined using an equalizer at the receiver. The equalizer tap weights will be used as the pre-equalizer coefficients. This method will be described in detail in Chapter four.

3. Complex Low Pass Filter Design

3.1 Introduction

The main focus of this chapter is to design a complex low pass filter to be placed after the SRRC pulse shaping filter in the base band portion of the IQ modulator. The amplitude/group delay response of this filter has to be the inverse of that of the low-pass complex equivalent of the RF filter. The compensation filter is a digital filter whose frequency response is determined by coefficients. Since these coefficients can be easily changed, the filter is easily adjusted or tuned.

A methodology to design a digital filter with arbitrary amplitude and group delay response is the subject of this chapter. In Appendix A the basic concept of filtering using the theory of *Linear Time Invariant* systems is explained, also the basic concept of digital low pass filtering will be described. Two types of digital filters will be described: *Finite Impulse Response* (FIR) and *Infinite Impulse Response* (IIR).

In order for compensating both amplitude and group delay responses, the digital filter can be designed using an *optimization* method which uses the inverse of RF filter amplitude and group delay responses in a *cost* function. This cost function is used to minimize the error between the digital LPF and RF filter amplitude/group delay responses. The optimization process needs two goals: one for the amplitude response, and the other for group delay response. Therefore two cost functions are needed. In the optimization algorithm it is possible to satisfy only one cost function at a time. Therefore, it is necessary that each of the amplitude and group delay responses be dealt with individually using separate optimization algorithms.

In order to compensate for the amplitude response, a digital filter has to be used that corrects the amplitude and has no effect on the phase or group delay. In other words, this filter will have an arbitrary amplitude response and flat group delay response. Likewise, the digital filter correcting the group delay, must have an arbitrary phase or group delay response and have flat amplitude response. The former can be realized using a FIR structure, and the latter using a specific IIR structure referred to as *All Pass Filter*.

The resulting two digital filters can be connected in cascade as is shown in Figure 3.1, compensating for both amplitude and group delay responses of the RF filter. In this figure, both FIR and IIR filters could have complex coefficients, so the overall cascaded filter. The structure and design of these two filters will be discussed in detail in sections 3.3 and 3.5 respectively.

In Appendix B, different optimization methods are described, however, all of the methods apply for digital filter design, but only to filters with real coefficients. Since a complex coefficient digital filter is required, a dedicated algorithm is developed to design complex digital FIR and IIR filter in Sections 3.3 and 3.5 respectively. Finally, this chapter is concluded in section 3.7 by a brief discussion on the implementation issues and the summary of the topics covered in this chapter.

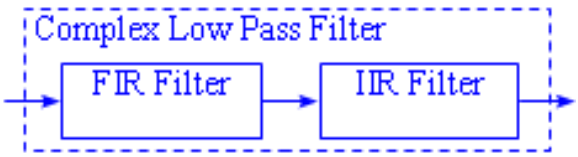

Figure 3.1 The Complex Low Pass Filter in base band, consisting of two separate filters, FIR for amplitude, and IIR for the group delay compensation

3.2 RF Filter Characterization

For the purpose of this thesis, it is assumed that the RF filter characterization is already performed and the RF filter amplitude and group delay responses are given. In practice, the RF filter could be characterized in a number of ways. One way is to excite the filter with a QAM signal and use the equalizer built into the a QAM signal Analyzer to determine amplitude and group delay responses. A second way is to excite the filter with a multi tone signal and use the FFT to compute the amplitude and group delay responses.

The second method seems a more viable option. In this method a set of harmonically related complex sinusoid tones will be generated. The tones can be generated inside the FPGA that is connected to the I and Q inputs of the modulator. The complex sinusoids will then pass through the RF filter. The total number of tones used must be high enough to span the bandwidth of the RF filter with small enough spacing that the frequency response of the RF filter can be obtained with linear interpolation. The SRRC pulse shaping filter has to be bypassed during this process. The desired target response is obtained by inverting the response of the RF filter in the pass band region. The procedures for designing the FIR and IIR all pass compensation filters will be discussed in details in section 3.3 and 3.5 respectively. Background material in the LTI systems and optimization algorithms have been included as appendices A and B.

3.3 Real FIR Filter Design Review

For a better understanding of the FIR real filter design method, it is preferred to refer to Appendix E for familiarization with the *Minmax* and the *Remez exchange* algorithms.

3.4 Finding the Coefficients for the Compensating FIR filter

3.4.1 Introduction

The coefficients for the complex FIR filter that compensate the amplitude distortion can be found using optimization methods [25–29]. The error function is formulated based on the difference between the desired and compensated amplitude response. A norm of the error function is minimized by changing the location of the zeros. As the value of the norm approaches zero, the resulting amplitude response approaches the desired amplitude.

3.4.2 Using Grid search Algorithm

The FIR compensating filter can be found using a grid search. Since the filter is complex, the zeros need not appear in conjugate pairs. however, the linear phase requirement forces any zeros with magnitudes other than one to appear with a zero equal to its reciprocal.

The following fundamental assumptions are made in searching for the coefficients of the compensating FIR filter: The width of the pass band is about 30 MHz and peak-peak amplitude variations across the band is less than 5dB.

The fundamental restriction for the compensating complex FIR filter is the linear phase property which means zeros must appear in pairs with their reciprocals. The search for coefficients starts with using only one complex zero (accompanying its reciprocal), by increasing the number of zeros, the amount of error between the desired response and the response of the filter under construction is controlled better.

Another reasonable approach is to use a predefined real filter with a symmetrical response to be used as default filter for the optimization process. This unique filter could be modified to a complex FIR filter during the optimization process, which is a reasonable assumption for compensating the RF filter amplitude response. Given the above specification for the default filter, the number of zeros in the pass band can be determined using some empirical methods (simulation). The starting point would be to use 6 well placed zeros and their 6 reciprocals. The angle of these zeros will be chosen so that 6 equally spaced zeros with equal magnitude are placed within the pass band of the desired filter, likewise, their reciprocal zeros with same angles and reciprocal magnitudes are placed outside of the unit circle in the z-plane. The simulation results indicated that a filter order of 16 with 12 zeros in the pass band yields a reasonable response with acceptable ripple in the pass band. A typical configuration of zeros is illustrated in Figure 3.2. In this figure $z_1, z_2, ..., z_6$ are considered the reciprocal pairs (in magnitude) of the $z_{17}, z_{18}, ..., z_{13}, z_{14}$ respectively.

The next step is to define an objective function to be used during the optimization process. A typical objective function could be the well known *mean square error*. It is well known that minimizing mean square error produces overshoots at the band edges which is referred to as Gibbs phenomenon. Nonetheless, this formulation yields satisfactory performance for this application. If the desired response is represented by $D(\omega)$ and symmetrical response by $P(\omega)$, the objective function would be

$$\Psi(\omega) = \lambda[D(\omega_i) - P(\omega_i)]^2 \quad i = 1, 2, ..., 6 \tag{3.1}$$

where λ is a positive real coefficient. The filter must have a linear phase response which forces the zeros that don't have a magnitude of one to be accompanied by their reciprocals. This means only one of the zeros in a reciprocal pair can be adjusted. A simple convention for this is to independently adjust the location of the zeros inside the unit circle and as part of the search then change the location of their reciprocals to keep them reciprocal.

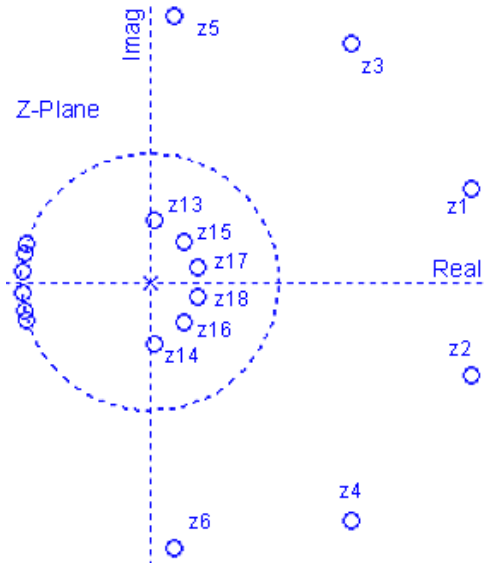

Figure 3.2 The typical zero layout in z-plane for the default lowpass filter

The objective function only spans the pass band of the filter, since only this region is contributing to the distortion effects. Furthermore, since the mean square error yields overshoots at the band edges, these spots are also excluded from the objective function frequency span to avoid overshoots. This would be a safe assumption, in that, in the CATV modulator, the QAM carrier frequency is not set to the regions close to band edges. The pass band is divided into six sub bands, this results in an objective function with six maximum peaks one in each of these sub bands in the pass band of the filter. The goal is to minimize these peak errors which means minimizing the difference between the default response and the desired response. For this purpose, a convergence factor Q is defined which is indicative of both the minimum and maximum peak errors in the objective function. This convergence factor, beside indicating the convergence of the optimization process, will force the objective function to have equal ripple in the error function, hence equiripple pass band.

$$Q = \frac{\max|\Psi_p(\omega_i)| - \min|\Psi_p(\omega_i)|}{\max|\Psi_p(\omega_i)|} \qquad i = 0, 1, ..., 6 \tag{3.2}$$

where $\Psi_p(\omega_i)$ represents the local peak of the objective function $\Psi(\omega_i)$ and Q factor represents the ratio of the difference of the maximum and minimum of these local

peaks to the max peak of the objective function, indicating the convergence of this objective function.

During the *univariate* grid optimization for the amplitude response, one zero at a time will be adjusted, so that the objective function yields the minimum error for that variable, then this process will be repeated for the other variables iteratively. This approach yields satisfactory results for the amplitude response within a moderate time interval. However, this method did not show similar performance and certainly not satisfactory performance when used for the optimization of the group delay response.

The main reason for this lack of performance for group delay is that group delay adjustment entails an accurate displacement of the zeros, and any slight change in the phase of the zeros will have a substantial change in the group delay of the corresponding sub band as well as the neighboring sub bands.

This method is iterative, as the result, it usually involves a large amount of computation which is another disadvantage, and the convergence time is significant for a reasonable grid size and tolerance.

3.4.3 Using Quasi Newton Algorithm

An alternative approach to the grid search, is some variant of the steepest descent method, known as the *Quasi* Newton method. For this method the formulation of the error function is slightly revised. Suppose the transfer function of the FIR filter is required to approach some specified amplitude response. Such a filter can be designed using two steps:

1. Formulate an objective function dependent on the difference between the actual and specified amplitude response.

2. Minimize the objective function with respect to the system function coefficients

Note that the second step (minimizing with respect to the coefficient independently) results in a real coefficient FIR filter, whereas, a complex FIR is needed. Therefore,

instead of using the filter coefficients in the objective function independently, the magnitude/phase of the zeros inside the unit circle will be changed independently without any restriction that zeros occur in conjugate pairs. However, the zeros outside the unit circle will be changed to maintain the reciprocal pairs. Then the resulting coefficients will be used in the objective function for the calculation of the error.

The amplitude response of the filter can be expressed as $M(\mathbf{a}, \omega) = |H(e^{\jmath\omega T})|$, where $\mathbf{a} = [a_0, a_1, ..., a_n]^T$, and ω is the frequency. Let $M_0(\omega)$ be the specified amplitude response. The difference between $M(\mathbf{a}, \omega)$ and $M_0(\omega)$ is the approximation error expressed as

$$e(\mathbf{a}, \omega) = M(\mathbf{a}, \omega) - M_0(\omega) \tag{3.3}$$

The error function is sampled at frequencies (defined by sub bands) $\omega_1, \omega_2, ..., \omega_k$ which results in the column vector

$$\mathbf{E}(\mathbf{a}) = [e_1(\mathbf{a})\ e_2(\mathbf{a})\ ...\ e_k(\mathbf{a})]^T \qquad for\ i = 1, 2, ..., k \tag{3.4}$$

where $e_i(a) = e(a, \omega_i)$

This approximation problem can be solved by finding the point $\mathbf{a} = \check{\mathbf{a}}$ such that

$$e_i(\check{\mathbf{a}}) \approx 0$$

For this equation to have a solution, a proper *objective* function must be formed which can satisfy a number of conditions. It should be a scalar quantity, and its minimization with respect to the point $\mathbf{a}$ should result in the minimization of all the elements of $\mathbf{E}(\mathbf{a})$. Another important requirement of this function is to be *differentiable*, an objective function satisfying all these requirements is what is known as the L_p norm of $\mathbf{E}(\mathbf{a})$ such that

$$L_p = ||\mathbf{E}(\mathbf{a})||_p = \left[\sum_{i=1}^{k} |e_i(\mathbf{a})|^p\right]^{1/p} \tag{3.5}$$

where p is integer. One special case of L_p norm is where $p = 2$. This is the popular *Euclidian* norm, the L_2 norm is expressed by

$$L_2 = ||\mathbf{E}(\mathbf{a})||_2 = \left[\sum_{i=1}^{k} |e_i(\mathbf{a})|^2\right]^{1/2} \tag{3.6}$$

In the case where $p = \infty$, one can define the maximum of the error function as:

$$\widehat{\mathbf{E}}(\mathbf{a}) = \max_{1\leq i\leq k} |e_i(\mathbf{a})| \neq 0$$

it follows

$$L_\infty = ||\mathbf{E}(\mathbf{a})||_\infty = \lim_{p\to\infty} \left\{\sum_{i=1}^{k} |e_i(\mathbf{a})|^p\right\}^{1/p}$$

$$= \widehat{\mathbf{E}}(\mathbf{a}) \lim_{p\to\infty} \left\{\sum_{i=1}^{k} \left[\frac{|e_i(\mathbf{a})|}{\widehat{\mathbf{E}}(\mathbf{a})}\right]^p\right\}^{1/p}$$

In the above equation each of the error elements are normalized, therefore the elements are equal to or less than unity, then

$$L_\infty = ||\mathbf{E}(\mathbf{a})||_\infty = \widehat{\mathbf{E}}(\mathbf{a})$$

meaning the infinity norm of the error function yields the maximum of the error function in the band of interest.

Now that the objective function is available, the required design will be obtained by solving the optimization problem

$$\min_{\mathbf{a}}\{||\mathbf{E}(\mathbf{a})||_\infty\} \tag{3.7}$$

The problem stated by Equation (3.7) can be solved using an un constrained optimization algorithm. Different classes of this algorithm have been developed including the *steepest descent* algorithm [36]. An important class of optimization algorithm that proved to be very efficient for the design of digital filters is the *quasi Newton* al-

gorithm. It is based on Newton's method to find the minimum point in the *quadratic convex*[1]function.

For more insight into the Newton's Algorithm, consider a function $f(\mathbf{a})$ of n variables, where $\mathbf{a} = [\mathbf{a}_1\, \mathbf{a}_2 \ldots \mathbf{a}_n]^T$ is a column vector. For small change $\mathbf{h} = [\mathbf{h}_1\, \mathbf{h}_2 \ldots \mathbf{h}_n]^T$, $f(\mathbf{a})$ can be approximated with a Taylor series about point $\mathbf{a}$. The error in a Taylor series is $o(||\mathbf{h}||_2^2)$ where $o(||\mathbf{h}||_2^2)$ is some function of $||\mathbf{h}||_2^2$ that has the property that it approaches zero faster than $||\mathbf{h}||_2^2$. This means:

$$f(\mathbf{a}+\mathbf{h}) = f(\mathbf{a}) + \sum_{i=1}^{n} \frac{\partial f(\mathbf{a})}{\partial \mathbf{a}_i}\mathbf{h}_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 f(\mathbf{a})}{\partial \mathbf{a}_i \partial \mathbf{a}_j}\mathbf{h}_i\mathbf{h}_j + o(||\mathbf{h}||_2^2) \tag{3.8}$$

In Equation (3.8), if the reminder $o(||\mathbf{h}||_2^2)$ is negligible a stationary point exists in the vicinity of point $\mathbf{a}$. That stationary point can be found by setting the derivative of $f(\mathbf{a}+\mathbf{h})$ with respect to $\mathbf{h}_k$ for $k = 1, 2, ..., n$, and solving for $\mathbf{h}$. From Equation (3.8) it follows

$$\mathbf{g} = -\mathbf{H}\mathbf{h} \tag{3.9}$$

where

$$\mathbf{g} = \nabla f(\mathbf{a}) = \left[\frac{\partial f(\mathbf{a})}{\partial a_1}\ \frac{\partial f(\mathbf{a})}{\partial a_2} \ldots \frac{\partial f(\mathbf{a})}{\partial a_n}\right]^T$$

where $\mathbf{g}$ and $\mathbf{H}$ are the *gradient* vector and Hessian matrix of $f(\mathbf{a})$, respectively. Therefore, the value of $\mathbf{h}$ that provides the stationary point of $f(\mathbf{a})$ is obtained by

$$\mathbf{h} = -\mathbf{H}^{-1}\mathbf{g} \tag{3.10}$$

Equation (3.10) provides a solution, if and only if the following two conditions hold:

1. The error $o(||\mathbf{h}||_2^2)$ in Equation (3.8) is negligible.

2. The Hessian is non singular.

[1]A two variable convex function is one that represents a surface whose shape resembles a punch bowl

Furthermore, if $f(\mathbf{a})$ is a *quadratic* function, its second partial derivatives are constants, thus $\mathbf{H}$ is a constant symmetric matrix, and its third and higher derivatives are zero. Therefore condition (1) holds. If $f(\mathbf{a})$ has a stationary point and there exists the sufficiency condition for a minimum in the vicinity of the stationary point, then the Hessian matrix is *positive definite*, hence *non singular*. Under these circumstances for an arbitrary point $\mathbf{a}$ in the n-dimensional Euclidian space, the minimum point can be found at $\check{\mathbf{a}} = \mathbf{a} + \mathbf{h}$ using Equation (3.10).

If $f(\mathbf{a})$ is a general nonquadratic convex function that has a minimum point $\check{\mathbf{a}}$, then for $f(\mathbf{a})$ in the vicinity of $\check{\mathbf{a}}$, i.e. $||\mathbf{a} - \check{\mathbf{a}}|| < \epsilon$, the reminder $o(||\mathbf{h}||_2^2)$ in Equation (3.8) becomes negligible and the second partial derivatives of $f(\mathbf{a})$ become approximately constant. To this end, the function $f(\mathbf{a})$ acts as if were a quadratic function and conditions (1) and (2) are satisfied again. Thus, for any point $\check{\mathbf{a}}$ such that $||\mathbf{a} - \check{\mathbf{a}}|| < \epsilon$, Equation (3.10) yields an accurate estimate of the minimum point.

Furthermore, if during the minimization of a general function $f(\mathbf{a})$, an arbitrary point $\mathbf{a}$ in an n-dimensional Euclidian space is considered, conditions (1) and/or (2) maybe violated in which case Equation (3.10) will not yield the solution. If condition (2) is violated, Equation (3.10) either has an infinite number of solutions or has no solution at all. In this case, these problems can be overcome by exploiting an iterative procedure in which the value of the function is progressively reduced by applying a series of corrections to $\mathbf{a}$ until a point in the vicinity of the solution is obtained. When the reminder $o(||\mathbf{h}||_2^2)$ in equation (3.8) becomes negligible, an accurate estimate of the solution can be obtained by using Equation (3.10). A suitable strategy to achieve this goal is based on the fundamental property that if $\mathbf{H}$ is positive definite, then $\mathbf{H}^{-1}$ is also positive definite.

Furthermore, in this case, it can be shown using the Taylor series that the direction pointed to by vector $-\mathbf{H}^{-1}\mathbf{g}$ of Equation (3.10), which is known as the *Newton direction*, is a *descent* direction of $f(\mathbf{a})$. As a result, if at some initial point $\mathbf{a}$, $\mathbf{H}$ is positive definite, a reduction can be achieved in $f(\mathbf{a})$ by simply applying the correction of the form $\mathbf{h} = \alpha\mathbf{d}$ to $\mathbf{a}$, where α is a positive factor and $\mathbf{d} = -\mathbf{H}^{-1}\mathbf{g}$.

On the other hand, if $\mathbf{H}$ is not positive definite it can be forced to become positive definite by using some algebraic manipulation (for example, it can be changed to the unity matrix) and again, $f(\mathbf{a})$ can be reduced. In either case, the largest possible reduction in $f(\mathbf{a})$ with respect to direction $\mathbf{d}$ can be achieved by choosing variable α such that $f(\mathbf{a} + \alpha\mathbf{d})$ is minimized. This can be performed using one of many available one-directional minimization algorithms known as *line* search algorithms [36]. Repeating these steps a number of times will yield a value of $\mathbf{a}$ in the neighborhood of the solution and eventually the solution itself. An algorithm based on these steps known as *Newton's algorithm*, is briefly described below.

Algorithm 1: Basic Newton algorithm

1. Input $\mathbf{a}_0$ and ϵ, set $k = 0$.
2. Calculate the gradient $\mathbf{g}_k$ and Hessian $\mathbf{H}_k$, if $\mathbf{H}_k$ is not positive definite, force it to become positive definite.
3. Calculate $\mathbf{H}_k^{-1}$ and $\mathbf{d}_k = -\mathbf{H}_k^{-1}\mathbf{g}$.
4. Find α_k, such that minimizes $f(\mathbf{a}_k + \alpha\mathbf{d}_k)$, exploiting a line search.
5. Set $\mathbf{a}_{k+1} = \mathbf{a}_k + \mathbf{h}_k$, where $\mathbf{h}_k = \alpha_k\mathbf{d}_k$, and compute $f_{k+1} = f(\mathbf{a}_{k+1})$.
6. If $||\alpha_k\mathbf{d}_k||_2 < \epsilon$, then output $\breve{\mathbf{a}} = \mathbf{a}_{k+1}$, $f(\breve{\mathbf{a}}) = f_{k+1}$, and stop. Otherwise, set $k = k + 1$ and repeat from step 2.

Finally the algorithm is terminated if the $\mathbf{L}_2$ norm of $\alpha_k\mathbf{d}_k$, for example, the magnitude of the change in $\mathbf{a}$, is less than ε, this parameter is known as *termination tolerance*, a small positive constant whose value is determined by the application under question.

So far, it was assumed that the optimization problem has only one global minimum. However in practice, this may not be true. An optimization problem may have more than one local minimum, therefore a well-defined minimum may not exist. It is

reasonable to abandon the idea of having the best solution available, and one should limit expectation to a solution that satisfies a number of the required specifications.

Quasi Newton Algorithm

The algorithm described in the previous section has three disadvantages. First, to obtain the gradient and Hessian, both the first and second partial derivatives of $f(\mathbf{a})$ must be calculated in each iteration, respectively. Second, in each iteration the Hessian must be checked for positive definiteness , and if it found to be nonpositive definite, force it to become positive definite. Third, in each iteration a matrix inversion is required.

In contrast, the quasi-Newton algorithm only needs to compute the first derivative, and it is also not necessary to invert or manipulate the Hessian. Consequently, for general optimization problems other than convex quadratic optimization problems, the quasi-Newton algorithm is much more efficient, and it is preferred.

The quasi-Newton algorithm, like the Newton algorithm, was originally developed for a convex quadratic problem, and then extended to the general problem. The basic principle is based on an approximation to the $n \times n$ inverse Hessian matrix. The approximation matrix denoted by $\mathbf{S}$ is constructed using available data so that $\mathbf{S} \simeq \mathbf{H}^{-1}$. Furthermore, as the number of iterations increases, $\mathbf{S}$ becomes a more accurate representation of $\mathbf{H}^{-1}$. For convex quadratic objective functions, the $n \times n$ matrix $\mathbf{H}^{-1}$ becomes identical to $\mathbf{H}$ in $n + 1$ iterations, where n is the number of variables.

Basic quasi Newton Algorithm

If the gradient of $f(\mathbf{a})$ at point $\mathbf{a}_k$ and $\mathbf{a}_{k+1}$ is $\mathbf{g}_k$ and $\mathbf{g}_{k+1}$, respectively, and

$$\mathbf{a}_{k+1} = \mathbf{a}_k + \mathbf{h}_k$$

then the elements of $\mathbf{g}_{k+1}$ computed by the Taylor series are

$$g_{(k+1)m} = g_{km} + \sum_{i=1}^{n} \frac{\partial g_{km}}{\partial \mathbf{a}_{ki}} \mathbf{h}_{ki} + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2 g_{km}}{\partial \mathbf{a}_{ki} \partial \mathbf{a}_{kj}} \mathbf{h}_{ki} \mathbf{h}_{kj} + o(||\mathbf{h}||_2^2) \tag{3.11}$$

For $m = 1, 2, ...n$. If $f(\mathbf{a})$ is quadratic, the third and higher orders derivatives of $f(\mathbf{a})$ are zero; likewise, the second and higher derivatives of g_{km} will vanish. Therefore

$$g_{(k+1)m} = g_{km} + \sum_{i=1}^{n} \frac{\partial g_{km}}{\partial \mathbf{a}_{ki}} \mathbf{h}_{ki} \tag{3.12}$$

and since

$$g_{km} = \frac{\partial f_k}{\partial \mathbf{a}_{km}}$$

it follows

$$g_{(k+1)m} = g_{km} + \sum_{i=1}^{n} \frac{\partial^2 f_k}{\partial \mathbf{a}_{ki} \partial \mathbf{a}_{km}} \mathbf{h}_{ki} \tag{3.13}$$

for $m = 1, 2, ..., n$. Therefore, g_{k+1} is given by

$$g_{k+1} = g_k + \mathbf{H}\mathbf{h}_k \tag{3.14}$$

where $\mathbf{H}$ represents the Hessian matrix of $f(\mathbf{a})$. One can also write

$$\gamma_k = \mathbf{H}\mathbf{h}_k \tag{3.15}$$

where

$$\mathbf{h}_k = \mathbf{a}_{k+1} - \mathbf{a}_k$$

and

$$\gamma_k = \mathbf{g}_{k+1} - \mathbf{g}_k$$

This analysis indicates that if the gradient of $f(\mathbf{a})$ is known at two points $\mathbf{a}_k$ and $\mathbf{a}_{k+1}$, it is possible to deduce a relation that gives a certain amount of information about H, for instance, Equation (3.15). Because $\mathbf{H}$ is real symmetric matrix with $n \times (n+1)/2$ unknowns and Equation (3.15) provides only n equations, $\mathbf{H}$ can not

be uniquely determined using Equation (3.15). One can overcome this situation by evaluating the gradient sequentially at $n+1$ points, i.e. at $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_n$, such that the changes in $\mathbf{a}$ will form a set of linearly independent vectors

$$\begin{aligned} \mathbf{h}_0 &= \mathbf{a}_1 - \mathbf{a}_0 \\ \mathbf{h}_1 &= \mathbf{a}_2 - \mathbf{a}_1 \\ \vdots &\quad \vdots \\ \mathbf{h}_{n-1} &= \mathbf{a}_n - \mathbf{a}_{n-1} \end{aligned}$$

With these conditions, Equation (3.15) yields

$$[\gamma_0 \ \gamma_1 \ ... \ \gamma_{n-1}] = \mathbf{H}[\mathbf{h}_0 \ \mathbf{h}_1 \ ... \ \mathbf{h}_{n-1}]$$

Consequently, $\mathbf{H}$ can be uniquely determined as

$$\mathbf{H} = [\gamma_0 \ \gamma_1 \ ... \ \gamma_{n-1}][\mathbf{h}_0 \ \mathbf{h}_1 \ ... \ \mathbf{h}_{n-1}]^{-1} \tag{3.16}$$

The above principles gives rise to the alternative Newton algorithm

Algorithm 2: Alternative Newton algorithm

1. Input $\mathbf{a}_{k0}$ and ϵ, input a set of n linearly independent vectors $\mathbf{h}_0, \mathbf{h}_1, ..., \mathbf{h}_{n-1}$, and set $k = 0$.
2. Calculate the gradient $\mathbf{g}_{k0}$.
3. For $i = 0$ to $n - 1$ perform:
   - Set $\mathbf{a}_{k(i+1)} = \mathbf{a}_{ki} + \mathbf{h}_i$
   - Compute $\mathbf{g}_{k(i+1)}$.
   - Set $\gamma_{ki} = \mathbf{g}_{k(i+1)} - \mathbf{g}_{ki}$.

4. Compute $\mathbf{H}_k$, using equation 3.16. If $\mathbf{H}_k$ is not positive definite, force it to become positive definite.

5. Determine $\mathbf{S}_k = \mathbf{H}_k^{-1}$.

6. Set $\mathbf{d}_k = -\mathbf{S}_k\mathbf{g}_{k0}$ and find α_k, the value of α that minimizes $f(\mathbf{a}_{k0}+\alpha\mathbf{d}_k)$, using a line search.

7. Set $\mathbf{a}_{(k+1)0} = \mathbf{a}_{k0} + \alpha_k\mathbf{d}_k$ and compute $f_{(k+1)0} = f(\mathbf{a}_{(k+1)0})$.

8. If $||\alpha_k\mathbf{d}_k||_2 < \epsilon$, then output $\check{\mathbf{a}} = \mathbf{a}_{(k+1)0}$, $f(\check{\mathbf{a}}) = f_{(k+1)0}$, and stop.

   Otherwise, set $k = k + 1$ and repeat from step 2.

In this algorithm, the parameter $\mathbf{a}_{k0}$ denotes the initial point and ϵ denotes the tolerance. $\mathbf{a}_{k(i+1)}$ denotes the point $\mathbf{a}_{i+1}$ at k^{th} iteration. $\mathbf{g}_{ki}$ denotes the gradient of the $f(\mathbf{a}_{ki})$ at k^{th} iteration. γ_{ki} also represent the difference between the Hessian matrix $\mathbf{H}$ and pseudo Hessian matrix $\mathbf{S}$ which in each iteration become closer to the Hessian matrix, therefore γ_{ki} tends toward zero as the number of iterations increases. $\mathbf{d}_k$ represents the dierection of descent in each iteration

It is understood that the above algorithm does not compute $\mathbf{H}^{-1}$ using the second derivatives, but rather uses the information concealed in the computed data. However, like the Newton algorithm, it is necessary, for a general nonquadratic problem, to check, manipulate, and invert the Hessian in every iteration. In addition, the algorithm requires a set of linearly independent vectors i.e., $\mathbf{h}_0, \mathbf{h}_1, ..., \mathbf{h}_{n-1}$, hence the algorithm is of little practical use.

One can overcome this problem by generating $\mathbf{H}^{-1}$ from the data using a set of linearly independent vectors $\mathbf{h}_0, \mathbf{h}_1, ..., \mathbf{h}_{n-1}$ that are themselves generated from available data. This can be performed by generating the vectors

$$\mathbf{h}_k = -\mathbf{S}\mathbf{g}_k \tag{3.17}$$

$$\mathbf{a}_{k+1} = \mathbf{a}_k + \mathbf{h}_k \tag{3.18}$$

and

$$\gamma_k = \mathbf{g}_{k+1} - \mathbf{g}_k$$

By making an additive correction to $\mathbf{S}_k$ of the form

$$\mathbf{S}_{k+1} = \mathbf{S}_k + \mathbf{C}_k \tag{3.19}$$

for $k = 1, 2, ..., n$. If a correction matrix $\mathbf{C}_k$ is found such that conditions

$$\mathbf{S}_{k+1}\gamma_i = \mathbf{h}_i \qquad for\ \ 0 \leq i \leq k \tag{3.20}$$

hold, and the vectors $\mathbf{h}_0, \mathbf{h}_1, ..., \mathbf{h}_{n-1}$ and $\gamma_0, \gamma_1, ..., \gamma_{n-1}$ generated during this process are linearly independent, then for the case $k = n - 1$ one can write

$$S_n[\gamma_0\ \gamma_1\ ...\ \gamma_{n-1}] = [\mathbf{h}_0\ \mathbf{h}_1\ ...\ \mathbf{h}_{n-1}]$$

or

$$S_n = [\mathbf{h}_0\ \mathbf{h}_1\ ...\ \mathbf{h}_{n-1}][\gamma_0\ \gamma_1\ ...\ \gamma_{n-1}]^{-1} \tag{3.21}$$

One can conclude from Equations (3.16) and (3.21) that

$$S_n = \mathbf{H}^{-1} \tag{3.22}$$

for $k = n$, Equations (3.17) and (3.22) yield the Newton direction, which is the steepest descent direction

$$\mathbf{h}_n = -\mathbf{H}^{-1}\mathbf{g}_n \tag{3.23}$$

Therefore, subject to conditions (i) and (ii) described earlier, one can obtain the solution for a convex quadratic problem from Equation (3.18) and Equation (3.23) as

$$\check{\mathbf{a}} = \mathbf{a}_{n+1} = \mathbf{a}_n - \mathbf{H}^{-1}\mathbf{g}_n$$

The above principle gives rise to the basic quasi Newton algorithm described next.

Algorithm 3: Basic Quasi Newton algorithm

1. Input $\mathbf{a}_0$ and ϵ, set $\mathbf{S}_0 = \mathbf{I}_0$, and $k = 0$. Compute $\mathbf{g}_0$
2. $\mathbf{d}_k = -\mathbf{S}_k\mathbf{g}_k$, and find α_k, the value of α that minimizes $f(\mathbf{a}_k + \alpha\mathbf{d}_k)$, using a line search.
3. Set $\mathbf{h}_k = \alpha_k\mathbf{d}_k$, and $\mathbf{a}_{k+1} = \mathbf{a}_k + \mathbf{h}_k$, and compute $f_(k+1) = f(\mathbf{a}_{k+1})$.
4. If $||\mathbf{h}_k||_2 < \epsilon$, then output $\breve{\mathbf{a}} = \mathbf{a}_{k+1}$, $f(\breve{\mathbf{a}}) = f_{k+1}$ and stop.
5. Compute $\mathbf{g}_{k+1}$, and set $\gamma_k = \mathbf{g}_{k+1} - \mathbf{g}_k$.
6. Compute $\mathbf{S}_{k+1} = \mathbf{S}_k + \mathbf{C}_k$.
7. Check $\mathbf{S}_{k+1}$ for positive definiteness, and if it found nonpositive definite, force it to become positive definite.
8. Set $k = k + 1$ and go to step 2.

In this algorithm, the initial value of pseudo Hessian matrix $\mathbf{S}$ is set equal to the identity matrix $\mathbf{I}$. The vector $\mathbf{d}$ denotes the direction of descent in each iteration. The parameter α in each iteration minimizes the function $f(\mathbf{a}_k + \alpha\mathbf{d}_k)$. The parameter $\mathbf{h}$ represents the direction of descent at a local minimum point in each iteration. $\mathbf{C}_k$ denotes the correction matrix in each iteration that builds the pseudo Hessian matrix for the next iteration.

In algorithm 3, the set of linearly independent vectors $h_0, h_1, ..., h_{n-1}$, are not used as an input. In addition, the inversion of $\mathbf{H}_k$ is avoided. Instead, an approximation is constructed by additive operations to $\mathbf{S}_k$. However, matrices $\mathbf{S}_1, \mathbf{S}_2, ...$ have to be checked for positive definiteness and if they are not, modified to be so. This may be done through *diagonalization* of $\mathbf{S}_{k+1}$ and replacing any nonpositive diagonal *eigenvalues* by corresponding positive ones. Although at first glance this seems to be a radical change, it does not change usefulness. The modification serves its purpose which is to force the iterative optimization to converge.

Moreover, in step 2 of algorithm 3, the vector $-\mathbf{S}_k\mathbf{g}_k$ is denoted as $\mathbf{d}_k$, instead of $\mathbf{h}_k$ in Equation (3.17), and $f(\mathbf{a}_i + \alpha\mathbf{d}_k)$ is minimized during a line search process, with respect to α. This was done to make the algorithm applicable for both quadratic and general nonquadratic problems, since $-\mathbf{S}_k\mathbf{g}_k$ may not be the Newton direction. Matrix $\mathbf{S}_k$ has to be positive definite in each iteration k to make sure that vector $\mathbf{d}_k$ is pointing toward the descent direction.

To obtain a descent direction in step 1 of the first iteration, $\mathbf{S}_0$ is assumed to be the $n \times n$ unity matrix. In step 5, the calculation of vector γ_k is necessary for the computation of correction matrix $\mathbf{C}_k$ (in step 6), this calculation is explained in more details in the following subsection.

Generating matrix $\mathbf{S}_{k+1}$ in each iteration

In Equation (3.19), the update formula for matrix $\mathbf{S}_{k+1}$ has to satisfy tight requirements to be useful in algorithm 3. For a convex quadratic problem, Equation (3.20) has to be satisfied and vectors $\mathbf{h}_0, \mathbf{h}_1, ..., \mathbf{h}_{n-1}$ and $\gamma_0, \gamma_1, ..., \gamma_{n-1}$ must be linearly independent. Several distinct formulas have been derived in literature to address the updating of this type of formula. Among the early works, a so-called *rank-one* formula was proposed, in which the correction matrix $\mathbf{C}_k$ is of rank-one. In recent years, rank-two formulas were developed by Davidon-Fletcher-Powell (DFP) and Broyden-Fletcher-Goldfarb-Shanno (BFGS) [36] [37] [38] [39] [40]. These formulas have the important property that a positive definite matrix $\mathbf{S}_k$ yields a positive definite $\mathbf{S}_{k+1}$ for both the convex quadratic problems and general nonquadratic problem. This is contingent on the line search in step 2 of the algorithm be exact [37]. Even if an inexact line search performed, this property may still hold, contingent to forcing a scalar quantity inherent in the computation of $\mathbf{C}_k$ to remain positive.

It is understood that maintaining a positive definite sequence $\mathbf{S}_k, \mathbf{S}_{k+1}, ...$ is very useful in algorithm 3, since checking and manipulation of $\mathbf{S}_{k+1}$ in step 7 of the algorithm is unnecessary, hence a significant computational load is avoided, which is why

this algorithm is very desirable. The DFP and BFGS updating formulas are given by

$$\mathbf{S}_{k+1} = \mathbf{S}_k + \frac{\mathbf{h}_k \mathbf{h}_k^T}{\gamma_k^T \mathbf{h}_k} - \frac{\mathbf{S}_k \gamma_k \gamma_k^T \mathbf{S}_k}{\gamma_k^T \mathbf{S}_k \gamma_k} \tag{3.24}$$

and

$$\mathbf{S}_{k+1} = \mathbf{S}_k + \left(1 + \frac{\gamma_k^T \mathbf{S}_k \gamma_k}{\gamma_k^T \mathbf{h}_k}\right) \frac{\mathbf{h}_k \mathbf{h}_k^T}{\gamma_k^T \mathbf{h}_k} - \frac{\mathbf{h}_k \gamma_k^T \mathbf{S}_k + \mathbf{S}_k \gamma_k \mathbf{h}_k^T}{\gamma_k^T \mathbf{h}_k} \tag{3.25}$$

The condition for positive definiteness of $\mathbf{S}_{k+1}$ in both formulas is

$$\mathbf{h}_k^T \gamma_k = \mathbf{h}_k^T \mathbf{g}_{k+1} - \mathbf{h}_k^T \mathbf{g}_k > 0 \tag{3.26}$$

One can utilize the principles stated in algorithm 3 to design the complex FIR filter. The formulation of the objective error function can be based on the complex low pass amplitude response and the specified desired response (as the result of RF Filer characterization). For more clarification, a hypothetical band pass non-symmetrical amplitude response (result of RF filter characterization) is shown in the Figure 3.3, with its complex lowpass equivalent. Note that the RF filter is a real filter but its response is non-symmetrical with respect to the center of its pass band. The resulting low pass equivalent filter has complex coefficients and a frequency response that is not conjugate symmetric. The problem is to design a low pass complex coefficient filter with the inverse of this amplitude response using the Quasi-Newton optimization method. The target inverse amplitude response for the complex lowpass filter is plotted as shown in Figure 3.4. The algorithm finds the coefficients that forces the response of the compensating filter to that of the target response in only a portion of the pass band. This region is shown in the Figure 3.4, bounded within the interval [A,B]. For more clarity, four QAM carriers are also shown in this figure to make a comparison with the band of interest used during the optimization.

The basic optimization procedure is to adjust the default filter response by changing the location of its zeros in an iterative manner, such that the amplitude response become close to that of the target response. Experimentation with Matlab simulations, indicated that the best convergence speed and least error is achieved by starting

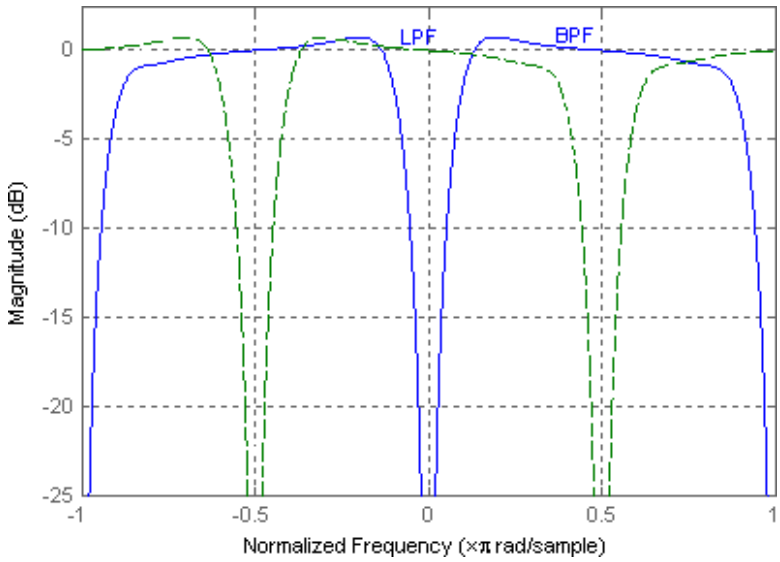

Figure 3.3 Amplitude response of band pass (solid line) and equivalent complex low pass (dotted line)

with a symmetrical amplitude response. The initial response uses conjugate zeros and their reciprocals within the pass band. The independent variables were chosen to be the zeros which determine the pass band response and located inside the unit circle (in conjunction with their reciprocals). This can be used as a constraint on the optimization routine of the FIR section, using the zeros in the pass band as independent variables and their reciprocals which will be computed correspondingly. The magnitude and phase of zeros will be used as independent variables in each round of optimization in a successive manner in order to find the best solution. Another constraint was to define a limit for the minimum angle between adjacent zeros to be greater than 0.01 radian/sample to avoid zero overlapping.

As mentioned earlier, the general objective function for the optimization problem is being formulated based on the coefficients of the real FIR filter. By iteratively changing the coefficients, the amplitude response of the compensating FIR filter will approach the target amplitude response. That is to say, the response of the compensating filter converges to the target response and the error function approaches zero. However, in the case of a complex FIR filer with linear phase property, this

method can not be used, since the independent adjustment of the coefficients will not necessarily guaranty the linear phase attribute of the resulting filter.

In this particular design, in order to maintain the linear phase property, priority will be given to the configuration of zeros and how position in the z-plane, particularly for those located in the pass band and their reciprocal pairs. After setting the zeros based on the quasi Newton algorithm in each iteration, the filter coefficients will be computed according to the optimized zeros. This strategy has the pass band zeros inside the unit circle treated as independent variables (and paired with their reciprocals).

The zeros for a typical complex coefficient filter is shown in Figures 3.5. This filter has 16 zeros in total, with 8 of them in the pass band. It is obvious the pass band amplitude/phase response is mostly affected by zeros located in this region, likewise, these zeros determine the response of the objective function. In this example four zeros are used as independent variables and the other four zeros in the pass band paired with them reciprocally as shown in Figure 3.5, to maintain the linear phase property.

In general, the total number of zeros can be determined during the optimization procedure. One can start with four zeros in the pass band and try to increase /decrease the number of zeros in groups of two, until an optimized response is achieved. It was found, however, for this application and given bandwidth, that the total number of six zeros in the pass band yields the best results.

To be more specific, in the context of the quasi-Newton algorithm, referring to Figure 3.2, six zeros $z_{13}, z_{14}, z_{15}, z_{16}, z_{17}, z_{18}$ will be used as the independent variable vector $[\mathbf{a}_1\ \mathbf{a}_2\ ...\ \mathbf{a}_6]$ in the quasi-Newton algorithm. The differential increase for these variables is the vector $[\mathbf{h}_1\ \mathbf{h}_2\ ...\ \mathbf{h}_6]$. This vector is used for the calculation of the gradient vector $\mathbf{g}_k$ and the direction of pseudo steepest descent vector $-\mathbf{S}_k\mathbf{g}_k$. The step size, which controls the amount by which $\mathbf{h}_k$ can change, controls the amount by which the magnitude/phase of zeros will change from one iteration to the next. The

step size decreases from one iteration to the next as the optimization progresses.

In each iteration, a line search is performed to minimize the function $f(\mathbf{a}_i + \alpha \mathbf{d}_k)$, the result of this line search is the parameter α which will be used for the calculation of the next point and its gradient, followed by the calculation of the correction vector. This process will be repeated iteratively until the error function become less than the tolerance. Algorithm 3 was found to converge very quickly, usually within 8-14 iterations. In each iteration the values for all variables (in this case six) are updated, and pertinent coefficients are calculated. Following that, the amplitude response is computed, and then the error function.

It was found during the optimization that the euclidian norm L_2 yields good results in terms of the convergence speed and performance accuracy. Higher order norms were tried, but did not yield better result. In order to avoid the overshoot effects due to the L_2 norm, regarding the objective function calculations, only the pass band interval [A,B] excluding the band edges was used. Another constraint was applied to prevent superimposing two zeros on each other.

It is worth mentioning that in the CATV system context, the whole system needs to meet tight specifications. Among them, each modulator must comply with a parameter known as *latency*, which is partly related to the response time of each individual signal source in the system supplying the QAM signals. The compensating complex lowpass filter will certainly add latency which is about 2 or 3 symbols.

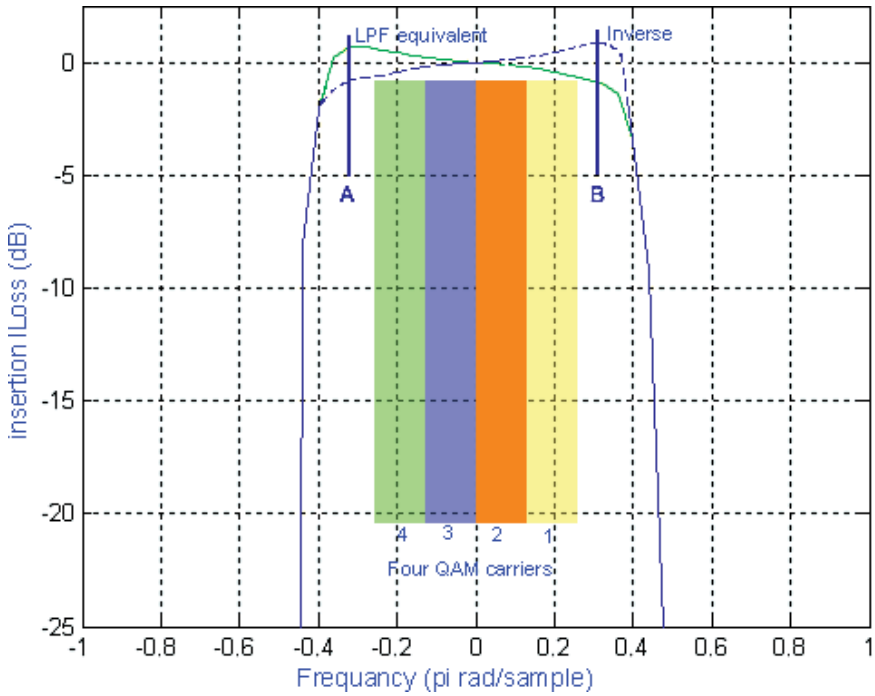

Figure 3.4 Amplitude response of the equivalent complex lowpass (solid line), its inverse (dotted line) in the pass band, and the four QAM carriers.

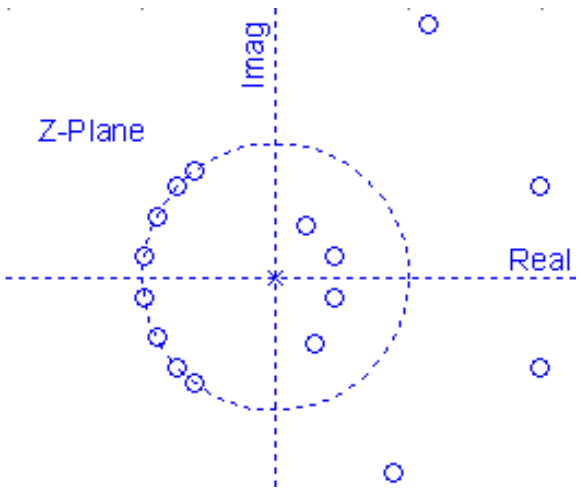

Figure 3.5 Zero configuration for a typical complex FIR lowpass filter

3.5 IIR Filter Design

The well known *minmax* algorithm, which is discussed in Appendix E, can also be used to determine the coefficients for an IIR filter. The details of this algorithm and its various augmentations to make it faster and less computationally intensive is extensively discussed in the literature [22] [23] [28] [29]. Therefore a detailed description will not be presented here.

The major difference between the FIR and IIR filters is that FIR filters are unconditionally stable while IIR filters are not. The poles of the IIR filters must be inside the unit circle for the filter to be stable. The minmax algorithm yields poles that are inside the unit circle, hence the stability of the optimized filter is not in question. One can refer to *Mitra* [41] for a fairly extensive treatment of this topic.

3.5.1 All Pass Filter Review

It was mentioned that for compensation of the RF filter amplitude distortion and group delay, two separate filters can be used in the base band. An FIR filter with constant group delay and arbitrary amplitude response to compensate the amplitude response distortions, and an IIR filter with constant amplitude response and arbitrary group delay response to compensate for the group delay distortion. A filter that has a flat amplitude response is called an *allpass filter*. The allpass filter belongs to the class of recursive filters, which means it has both poles and zeros in the system function. An allpass filter has a system function with *unity* magnitude for all frequencies. The system function of an all pass filter has the form [41]

$$A_M(z) = \pm \frac{q_M + q_{M-1}z^{-1} + \ ... \ + q_1 z^{-M+1} + z^{-M}}{1 + q_1 z^{-1} + \ ... \ q_{M-1}z^{-M+1} + q_M z^{-M}} \tag{3.27}$$

If the denominator of $A_M(z)$ is denoted by $Q_M(z)$, then the system function can be written as

$$A_M(z) = \pm \frac{z^{-M} Q_M(z^{-1})}{Q_M(z)} \tag{3.28}$$

It is understood from Equation (3.28) if the transfer function has a pole at $z = re^{\jmath\phi}$, then it also has zero at $z = (1/r)e^{-\jmath\phi}$. Moreover, it is clear from Equation (3.27), the coefficients of the numerator are the *reflected* coefficients of the denominator polynomial i.e. the numerator coefficient for z^k is equivalent to the denominator coefficient for z^{M-k}. The reflected coefficients of a degree-M polynomial can be shown to satisfy $Q_M(z) = z^{-M}Q_M(z^{-1})$. Equation (3.28) implies the poles and zeros in a real coefficient all pass filter appear in reciprocal form. Thus, Equation (3.28) can be written as

$$A_M(z^{-1}) = \pm\frac{z^M Q_M(z)}{Q_M(z^{-1})} \tag{3.29}$$

Therefore

$$A_M(z)A_M(z^{-1}) = \frac{z^{-M}Q_M(z^{-1})}{Q_M(z)}\frac{z^M Q_M(z)}{Q_M(z^{-1})} = 1. \tag{3.30}$$

the frequency response is therefore

$$|A_M(e^{\jmath\omega})|^2 = A_M(z)A_M(z^{-1})|_{z=e^{\jmath\omega}} = 1 \tag{3.31}$$

For recursive filters to be stable all poles must be inside the unit circle. Group delay which is denoted by $\tau(\omega)$, is denoted by

$$\tau(\omega) = -\frac{d}{d\omega}[\theta(e^{\jmath\omega})], \tag{3.32}$$

where $\theta(\omega) = \arg\{A(e^{\jmath\omega})\}$. A property of an allpass filter, which must have all its poles inside the unit circle and all its zeros outside, is the unwrapped phase function $\theta(\omega)$ *monotonically* decreases with ω for $0 \leq \omega \leq \pi$. Furthermore, this implies $\theta(\pi) - \theta(0) = M\pi$, this make $\tau(\omega)$ positive everywhere between $\omega = 0$ to $\omega = \pi$. For an all pass filter the delay for a sinusoid with frequency ω_o is

$$Delay(\omega) = -\frac{\theta(\omega_o)}{\omega_o} = \frac{1}{\omega_o}\int_0^{\omega_o}\tau(\omega)d\omega \tag{3.33}$$

This means that the delay at $\omega_o = \pi$ of an order M allpass filter is equal to M samples. For the case of the complex allpass filter, the above relation does not hold.

The complex allpass function must have its zeros located at the reciprocal of the poles, but neither the zeros or the poles need a conjugate. This translates into $\theta(\omega)$ decreasing with ω and $-\pi \leq \theta(\pi) - \theta(0) \leq 0$

3.6 Complex IIR Design using Quasi Newton

The purpose of the complex recursive filter structure is to compensate for the group delay distortion caused by the RF filter. The objective group delay response is obtained through the RF filter characterization process, in the same way as the amplitude of the RF filter is measured. It is assumed that the target group delay response of the compensation filter is calculated from the measured response, as was done for the amplitude distortion. The basic concepts presented in sections 3.4.3 and 3.5.1 can be exploited to find the coefficients for the complex recursive group delay compensation filter. The same quasi-Newton algorithm may be used. The important point here is that in the case of the coefficients for the FIR filter, there was no concern regarding the stability of the optimized filter. In the case of allpass filter, which has poles, it is vital to make sure the filter is stable; this entails performing optimization under a constraint. The constraint is that the magnitude of the poles must be less than unity. A practical value, due to round off error and with a safe margin, that could be used is $r \leq 0.99$. This will guaranty that during the optimization process the poles are always inside the unit circle.

A good strategy for the coefficient finding algorithm is to start from a pre-defined all pass filter with a flat group delay response. Starting with four poles located inside the unit circle (and four reciprocal zeros outside the unit circle) yields a satisfactory compensation group delay response in 10-15 iterations. The magnitude and the phase of the poles change independently, whereas the zeros change as reciprocals of the poles. For more clarification on the subject, the pole/zero configuration and the group delay response of a typical allpass filter is shown in Figures 3.6, and 3.7 respectively. The arbitrary group delay of the compensating all pass filter make it possible to compensate for most of the group delay distortions. The starting point for the optimization process is to define the objective function as the difference between the predefined group delay with flat response and the measured group delay response found during the RF filter characterization.

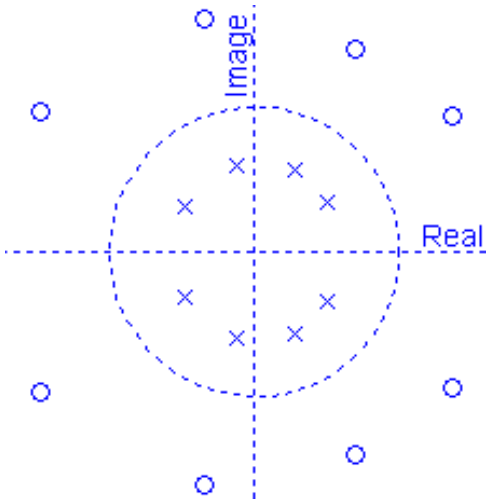

Figure 3.6 Pole/Zero configuration for a complex allpass filter

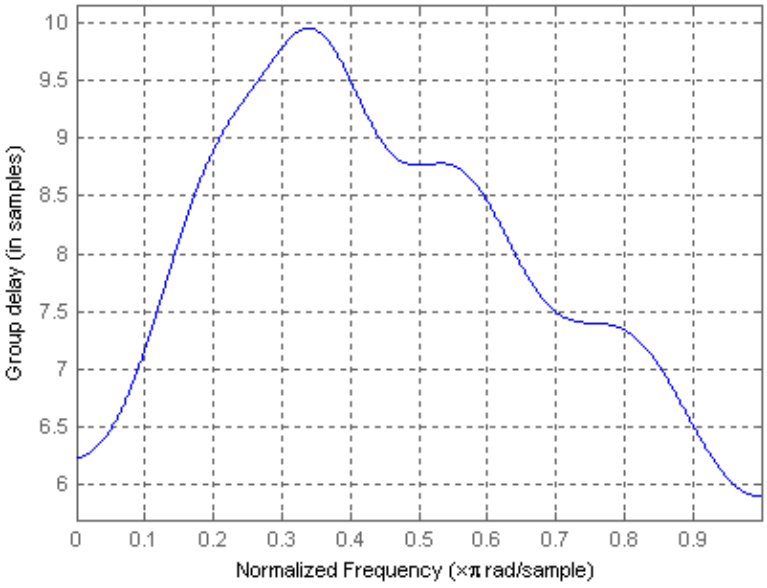

Figure 3.7 The group delay response for the complex allpass filter with poles and zeros as shown in Figures 3.6

For the allpass group delay compensating filter, only the poles were considered independent variables in optimization. The zeros were paired with poles as their reciprocals. Using the quasi Newton algorithm explained on page 50, the variable vector $\mathbf{a}_i$ is defined by the vector of poles of the allpass filter as $\mathbf{a}_i = [p_1\, p_2 \ldots p_i]$. Also the vector used to make the differential increase to the variable $\mathbf{a}_i$ is the vector $\mathbf{h} = [\mathbf{h}_1\, \mathbf{h}_2 \ldots \mathbf{h}_i]$, applying the constraint on the magnitude of the poles $p_i = r_i e^{\jmath\phi_i}$ such that $r_i \leq 0.99$ for stability concerns.

The group delay response of the allpass filter is calculated and compared with the target group delay, using the L_2 norm to measure the difference. The difference is minimized by finding the gradient $\mathbf{g}_k$ and the descent vector $-\mathbf{S}_k\mathbf{g}_k$, and moving in that direction. Also, the matrix $\mathbf{S}$ and its positive definiteness is checked in each iteration by calculating the *eigenvalues* of this matrix, and is forced to be positive definite if needed. As explained earlier, $\mathbf{S}$ must be positive definite for $-\mathbf{Sg}$ to point toward the descent direction in each iteration. To find the amount by which the steepest descent vector needs to be scaled for the optimum step size on each iteration, a line search is performed by minimizing the function $f(\mathbf{a}_i + \alpha\mathbf{d}_k)$ versus α, this yields a reasonable step size for the next iteration. This process is repeated in an iterative manner until the minimum of the objective function is found which makes the error function as small as possible. The convergence speed depends on the step size chosen for the differential increment $\mathbf{h}$, the resolution of the line search to find the parameter α, the allpass filter order, the required accuracy, and the target response. If these parameters are chosen with discretion, the optimization converges within 10-20 iterations. During the optimization process it was observed that the group delay of the allpass filter converges to the target group delay in the band of interest satisfactorily with small error. However, the norm square of the error is not as low as what was obtained from the amplitude optimization. Increasing the filter order reducing the step size of parameters $\mathbf{h}$ and α, did not noticeably reduce the mean acquired error. This is due to the structure of the allpass filter with magnitude of the poles less than 0.99. Although the allpass filter does allow it to have arbitrary

group delay, the phase is monotonically decreasing with ω and the group delay is always positive. Another interpretation is that in the case of amplitude optimization it is possible to define the amplitude response in terms of a set of linear combination of orthogonal functions (cosine function), whereas for the group delay this is not possible.

3.7 Implementation

3.7.1 Introduction

After finding the coefficients for the polynomials that determine the nonrecursive and recursive compensation digital filters, a structure for these filters must be selected. Some of the available structures are: Direct Form I, Direct Form II, cascade of second order sections and polyphase. Each structure has advantages and disadvantages. Each also has different sensitivity to finite register length effect. In this thesis however, the advantages and disadvantages are not going to be treated comprehensively, rather only a short discussion is included.

3.7.2 FIR implementation

The FIR filter structure can be categorized as:

1. Direct form
2. Cascade second-order-section form
3. Polyphase realization

Direct form

A well known structure to implement FIR filters is the *direct* form, in which, the multiplier coefficients and the coefficients of the system function polynomial are exactly the same. A filter with the direct form structure is also referred to as *tapped* delay line or a *transversal* filter. The transpose of this form is known as direct form

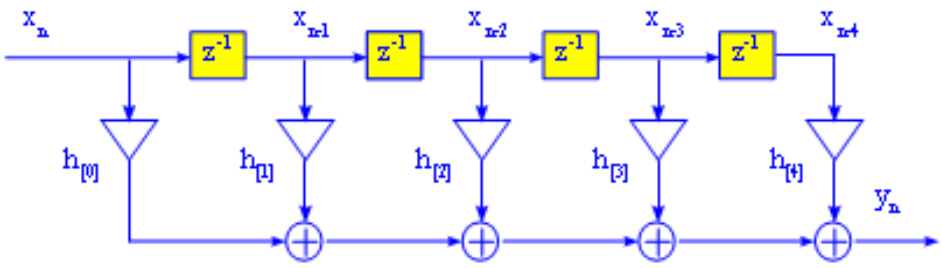

Figure 3.8 FIR Direct form I structure

II structure. Both direct forms are canonic in terms of delays, meaning the number of delay lines is optimized and is minimum. A direct form I, FIR is shown in Figure 3.8.

Cascade form

A higher order FIR filter transfer function can be realized as cascaded second-order FIR sections. In this regard, the system function $H(z)$ can be factorized and written in the form

$$H(z) = h[0] \prod_{k=1}^{k} (1 + \beta_{1k} z^{-1} + \beta_{2k} z^{-2}), \tag{3.34}$$

where, $k = (M-1)/2$ if M is odd, and $k = M/2$ if M is even, also $\beta_{2k} \neq 0$. Each second order section can be implemented using direct form or transposed direct form. The virtue of exploiting the second-order implementation is that the truncation errors will not propagate through the whole FIR structure. Figure 3.9 illustrates the cascade structure.

Polyphase realization

An interesting form of FIR realization is based on *polyphase* decomposition of its system function, which produces a parallel structure [42]. The system function can be expressed as a sum of two terms: one term containing the even index coefficients

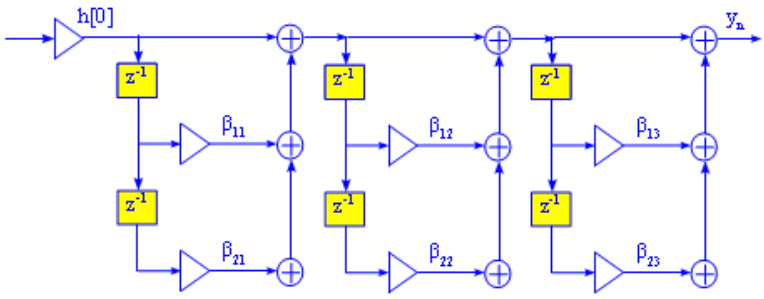

Figure 3.9 FIR cascade structure for sixth-order filter

and the other the odd index coefficients, as shown in Figure 3.10. That is to say:

$$H(z) = E_0(z^2) + z^{-1}E_1(z^2) \tag{3.35}$$

where

$$\begin{aligned} E_0(z) &= h_0[0] + h_0[2]z^{-1} + h_0[4]z^{-2} + h_0[6]z^{-3} + h_0[8]z^{-4}, \\ E_1(z) &= h_1[1] + h_1[3]z^{-1} + h_1[5]z^{-2} + h_1[7]z^{-3}, \end{aligned} \tag{3.36}$$

and $h[n]$ is the impulse response of the filter.

Equation (3.35) is generally referred to as a *polyphase decomposition* of the system function. The sub filters $E_0(z)$ and $E_1(z)$ can be realized using any FIR implementation technique discussed earlier. The polyphase realizations are generally used in *multirate* digital signal processing applications where the output is immediately down sampled [42]. For some applications this structure is computationally efficient, but for this application there is no computational advantage.

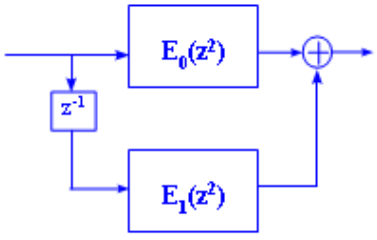

Figure 3.10 FIR Poly Phase realization of an FIR transfer function

Finite word length effects

Another possible source of error that needs special attention is the effect of *finite word length*, which encompasses the effects of truncating the output of multipliers and adders as well as the effect of representing filter coefficients with a finite length binary numbers. The coefficients of the compensating complex FIR filter, are assumed to have infinite word length. However, in practice the coefficients are represented by finite words, the length of which is dictated by the available hardware. The coefficients have to be quantized to a certain resolution which introduces error. This error can be treated as additive error. This quantization error generated due to the truncation of the coefficients, can be modeled as follows:

$$\hat{H}(z) = \sum_{n=0}^{M-1} \hat{h}[n]z^{-n} = \sum_{n=0}^{M-1} (h[n] + e[n])z^{-n}, \tag{3.37}$$

where $\hat{H}(z)$ is the system function of the filter that is implemented. This can be rewritten as:

$$\hat{H}(z) = H(z) + E(z), \tag{3.38}$$

where

$$E(z) = \sum_{n=0}^{M-1} e[n]z^{-n}. \tag{3.39}$$

The error $e[n]$ depends on the type of FIR implementation. For the direct form I FIR structure, the error can be propagated through the filter. However, for the cascade of second order sections structure the error will not propagate and this structure produces a small error if not the minimum $e[n]$.

3.7.3 IIR implementation

The basic IIR filter system function is represented by a rational function in the z-domain, as given by

$$H(z) = \frac{Y(z)}{X(z)} = \frac{p_0 + p_1 z^{-1} + p_2 z^{-2} + ... + p_M z^{-M}}{d_0 + d_1 z^{-1} + d_2 z^{-2} + ... + d_N z^{-N}} \tag{3.40}$$

In the time domain, for a causal IIR filter, one can use the difference equation of the form,

$$y[n] = -\sum_{k=1}^{N} \frac{d_k}{d_0} y[n-k] + \sum_{k=0}^{M} \frac{p_k}{d_0} x[n-k] \tag{3.41}$$

From Equation (3.41) it is obvious that to compute the n^{th} output sample, the knowledge of several past samples of the output sequence is needed. This means that the realization of the IIR filter needs feedback from its output.

In general, the realization of an IIR filter with order N requires $2N + 1$ unique coefficients, this means, it requires $2N + 1$ multipliers and $2N$ adders. As with FIR filters, IIR filters have a direct form structure in which the multiplier coefficients are exactly the same as the system function polynomial coefficients.

A possible IIR realization of this form can be obtained by decomposing $H(z)$ into $H(z) = H_1(z) * H_2(z)$ where

$$\begin{aligned} H_1(z) &= \tfrac{Y(z)}{1} = P(z) = p_0 + p_1 z^{-1} + \ ... \, p_M z^{-M} \\ H_2(z) &= \tfrac{1}{X(z)} = \tfrac{1}{D(z)} = \tfrac{1}{d_0 + d_1 z^{-1} + ... \, d_N z^{-N}} \end{aligned} \tag{3.42}$$

It can be seen that the $H_1(z)$ is an FIR filter which can be realized using FIR direct form technique as before. The $H_2(z)$, in the case of $N = 3$ can be looked at in the time domain as

$$y[n] = w[n] - d_1 y[n-1] - d_2 y[n-2] - d_3 y[n-3], \tag{3.43}$$

Two structures can be connected in cascade and the resulting structure known as *direct form I*, consists of an FIR form and an IIR form which is not canonic, since it uses 2N delays to implement an N order IIR filter. After some manipulation of the transfer function, the canonic form shown in Figure 3.11 can be derived. Different types of IIR implementation techniques are as follows

1. Direct form

2. Cascade of second-order-sections

3. Parallel Realization

One can refer to text books such as *Mitra* [41] for a comprehensive description of various types of implementations for the IIR filter.

Stability concern

For a recursive digital filter, the main concern is how the stability of the filter is affected by the truncation of the coefficients. The truncation error can shift poles from inside to outside the unit circle, making the IIR filter unstable. The poles must be far enough from the unit circle so that practical truncation resolution will not move them outside the unit circle.

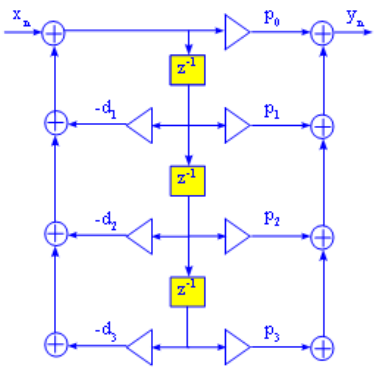

Figure 3.11 IIR Direct form II structure

3.8 Summary

In this section, in order to explain how the coefficients for the complex lowpass baseband filter are found, a brief mathematical representation of Discrete-Time LTI systems in the time domain and the frequency domain was reviewed. During this process, the system function in the z-domain for nonrecursive and recursive digital filters was introduced. The concept of filtering and FIR and IIR structures was described. As well, the concepts of linear phase and constant amplitude response were discussed.

Since the aim of this chapter was to find the complex compensation filters, which was done with an iterative computer search, a short review on the subject of optimization was presented. This was followed by a brief review on the application of some algorithm such as weighted Chebyshev, minmax algorithm, the Remez exchange algorithm, also their mathematical formulations were studied. After the background was established, the well known Quasi-Newton algorithm was described, this algorithm proved to be well behaved, yielding a satisfactory accuracy for the amplitude response of the amplitude compensation FIR filter.

The allpass structure was used for the group delay compensation filter. The coefficients were found using the same quasi Newton algorithm by adjusting the magnitude and phase of the poles.

Finally, a brief description on the implementation techniques for FIR and IIR filter structures and some issue related to the hardware restrictions was presented. These issues included truncation error and techniques to alleviate artifacts of their errors.

4. Complex pre Equalizer Design

4.1 Introduction

In this chapter, the RF filter distortion is compensated using a filter located prior to the pulse shaping filter that operates on the data prior to up sampling. Such a filter is referred to as a *pre-equalizer*. The algorithm that determines the coefficients of the pre-equalizer does not need a target response so does not require characterization of the RF filter. A brief review of the equalization algorithms will be presented here, then its application for finding the coefficients for a complex baseband compensator will be discussed. For a more detailed description of the equalizer theory, refer to Appendix F.

4.2 Statement of the Problem

The RF filter distortions in the IQ modulator of a CATV system will create *Inter Symbol Interference*, and cause some impairments on the QAM signal attributes such as signal to noise ratio, modulation error ratio (MER), and the received bit error rate. These errors need to be dealt with by employing some compensation techniques.

In this chapter, the equalization method will be used to obtain a complex base band compensator. The equalizer used to obtain the coefficients of the pre-equalizer will be placed in the receiver after the matched filter. The equalizer adaptively characterizes the channel response, employing the inherent statistical metrics embedded in the received signal. Once the equalization is completed, the equalizer's *tap weights*, represents approximately the inverse of the RF filter response. These tap weights

are used as the coefficients of the complex base band compensator which has a *feed forward* structure. This will force the ISI of the whole system to zero, hence a flat channel response with no distortion.

To be more precise, this process has two different modes. In characterization mode, the compensator acts transparently and has no effect on the system, and the equalizer characterizes the RF filter frequency response. In the second mode, once the characterization is performed, the tap weights generated as a result of the equalization, will be used as the coefficients of the complex pre-equalizer in the base band that compensates for the distortions of the RF filter.

There are two types of equalizers: a *symbol spaced* equalizer and a *fractional spaced* equalizer. A symbol spaced equalizer operates at the symbol rate and uses the decision variable as input. The taps are updated on each symbol time T_{sym}. A fractional spaced equalizer uses a higher rate input. The equalizer input is the output of the match filter, which operates at a rate of at least twice the symbol rate. The taps are updated at a fraction of symbol time, e.g. at half the symbol time, $T/2$. In symbol spaced equalization, the equalizer input is after the down sampling of the matched filter (SRRC) in the receiver, while in the fractional spaced equalizer, the equalizer is immediately after the matched filter in the receiver, and before the down sampling.

The fractional space equalizer has an advantage over the symbol spaced equalizer in that it is less sensitive to timing error, the disadvantage is that it can not be used to characterize the amplitude response in the region between the adjacent carriers. However, For the application in hand, the equalizer used to find the tap weights is in a receiver with no timing error, so the symbol spaced equalizer is used. The block diagram of a CATV modulator that shows the locations of the symbol spaced equalizer in the receiver and the symbol spaced pre-equalizer in the transmitter is shown in Figure 4.1. The tap weights for the pre-equalizer are determined without the pre-equalizer in the circuit. The block labeled "pre-equalizer" in Figure 4.1 shows the location in which the pre-equalizer resides after the tap weights are determined. The

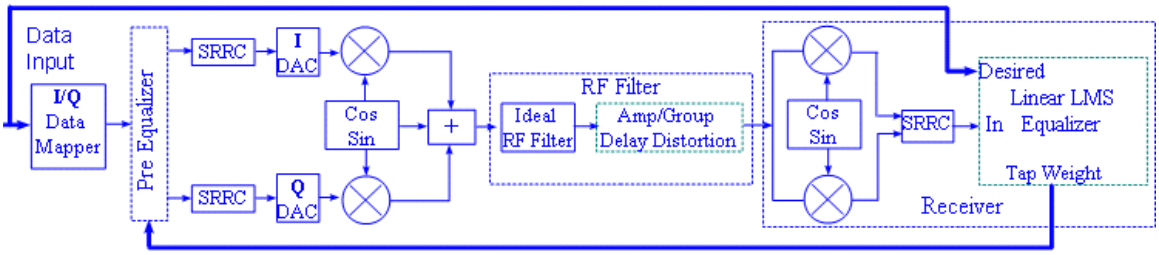

Figure 4.1 The modulator with pre-equalizer placed before the SRRC pulse shaping filter, followed by the test receiver containing equalizer block after the match filter.

tap weights of the pre-equalizer are determined in the receiver of the system shown in Figure 4.1. The receiver first demodulates the I and Q signals and then filters them with a match filter which has an inherent down sampler. The down sampled signal is then used as the input to the equalizer. The equalizer also receives the original base band I and Q signals from the modulator on another input port called "desired input". Using the statistical properties of the I and Q signals, also comparing with the desired signal, the equalizer is able to characterize the channel frequency response, which in turn will be used to generate the channel tap coefficients or *tap weights*. These tap weights, if used as the coefficients of the complex base band equalizer in the modulator, can compensate for the distortions caused by the RF filter.

4.3 Equalization Theory

4.3.1 Introduction

The aim of this section is to give some background on equalization theory. To this end, the topics relevant to this research are covered. A more detailed description is presented in Appendix F. More sophisticated treatment can be found in the literature: Haykin [30], Farhang [43], Sayed [31], Proakis [44].

4.3.2 Optimum Filtering

Generally, during an equalization process in the receiver, a reference data sequence is used which is referred to as the *desired* sequence, trying to equalize the received data sequence such that it becomes very similar to the desired sequence. The equalized sequence is the result of the received sequence passing through an FIR structure with certain coefficients. *Optimum* filtering refers to a process that finds the optimum coefficients for this FIR structure (referred to as tap weights) in the equalizer that minimize the error between the equalized sequence and the desired sequence. This error function is also referred to as the *mean square error* which is the mean of the square of the magnitude of the error between the desired and equalized sequences. This function is represented by

$$\epsilon^2 = d(n) - w(n)^* x(n)$$

where d(n) is the desired sequence, x(n) is the input sequence and w(n) is the coefficients of the FIR filter. The filter coefficients that yield the minimum mean square error, is referred to as *Optimum* or *Winner* filter. A block diagram of this filter is shown in Figure 4.2.

The coefficients of this filter can be derived using the statistical properties of the input and desired sequence, such as the autocorrelation of the input sequence denoted by R, and also the cross correlation between the input and desired sequences, denoted by P, then the optimum FIR filter weights can be computed using

$$w_{opt} = R^{-1}P$$

where w_{opt} is the optimum filter weight vector, assuming the input/desired sequence is a WSS stationary process [1] [45].

[1] An stochastic process is said to be stationary if its statistics do not depend on the time origin. In addition, if its autocorrelation function only depends on the time difference (τ), it is Wide Sense Stationary(WSS) as well.

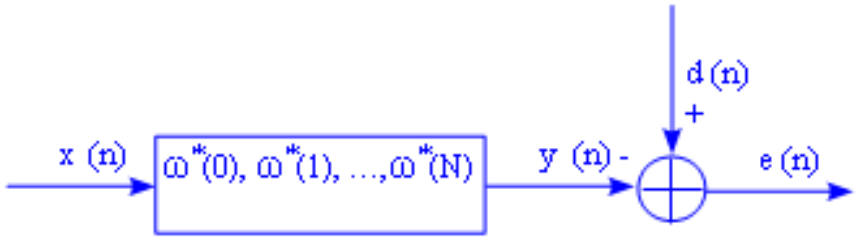

Figure 4.2 The block diagram for the optimum linear filter design

4.3.3 Adaptive Filtering

In the optimum filtering problem, *a priori* knowledge of the statistics (R and P) of the input sequence is needed, however, in practice this information is not available. To resolve this issue, an adaptive algorithm can be used which starts from an initial point (with small amount of statistical information), and iterates using the information embedded inside the input sequence to compute the next iteration coefficients. In this way the algorithm converges toward the final coefficients that yield the lowest mean square error and is very close to the optimum solution, contingent on the input sequence being a WSS stationary process. The important parameter for this equalizer is the convergence rate, which determines how fast the algorithm converges.

4.3.4 Linear LMS Equalizer

This equalizer uses a recursive adaptive algorithm in which the necessary statistical information is obtained iteratively. the error function $\epsilon^2 = d(n) - w(n)^* x(n)$, where n represents the iteration number, is a convex function that has a local minimum. This minimum point can be obtained using an algorithm referred to as *steepest descent* [43]. The general formulation for this algorithm is:

$$w(i+1) = w(i) + \mu(P - Rw(i))$$

where i indicates the iteration number, $R = x(n)x(n)^H$ and $P = x(n)d(n)^H$ in each iteration, $w(i)$ represents tap weights in each iteration and μ is the convergence rate. Although, R and P are not necessarily an accurate representation of the au-

tocorrelation/ cross correlation function, but if computed and averaged over a large number of iterations, they become very close to these functions, and the tap weights converge asymptotically to the optimum minimum of the convex error function. Substituting for R, P and n for i, it yields

$$w(n+1) = w(n) + \mu x(n) e^*(n)$$

where μ is the convergence rate and determines how fast the algorithm converges.

Design consideration

It was mentioned earlier that the intention is to design complex base band compensation filters exploiting the features of the equalization technique. The very fundamental question is: What is the appropriate equalizer for this application? To address this question, the attributes of this application that must be taken into account are listed as:

1. The characterization process will be performed in the lab, and the transmission medium is coaxial cable, hence the effect of noise is negligible and the channel is not time varying. Therefore, the test conditions are almost ideal in terms of the channel additive noise or fading effects, etc.

2. In this application, the original I and Q signals will be provided for the equalizer to be used as the desired signal during the equalization process.

3. The whole process needs to be software controllable.

4. The characterization of the amplitude and group delay response requires a certain amount of accuracy dictated by DOCSIS spec, hence the equalization algorithm must meet the minimum of these requirements.

5. The characterization of the RF Filter should be done within a reasonable time period, therefore the equalization convergence time must meet this requirement.

In the above list, the first and second items alleviate the burden of dealing with noise and channel variations for the candidate equalizer in that the channel is very well behaved and does not indicate any unexpected variations. Also, the effect of additive noise are to be excluded from the process. Moreover, the original symbols are provided to the equalizer to be used as the desired signal, hence there is no concern about the *symbol timing* errors or timing *jitter* and *carrier recovery* or *synchronization* errors, which are common concerns during the demodulation process. The third item is also assumed to be a given by default and the whole process is to be performed using automatic test software.

Therefore, the main parameters contributing to the selection of the proper algorithm seems to be items four and five. The accuracy needed for the amplitude and group delay responses are specified by DOCSIS [32] standard for gain variation and MER, i.e., the gain variation between any two adjacent carriers must be less than 0.2 dB, although this does not directly impose any requirement for the amplitude variations within the QAM carrier, but this implies that the accuracy of compensation needs to be comparable to this limit. On the other hand, the amplitude/group delay variation within the QAM signal bandwidth will have degradation effects on the quality of the QAM signal itself, and there are performance metrics such as MER that can be used to evaluate the accuracy of the equalizer in terms of amplitude or group delay accuracy as was indicated in chapter two.

At the first glance, the linear LMS algorithm using the descent gradient method appears to be a viable candidate as it converges reasonably fast, and also has reasonable accuracy. however, there are two main concerns about this algorithm as

1. Finite FIR length
2. Noise enhancement effect

The FIR length concern arises from the fact that, the effect of the channel (in this case an RF filter) can be considered as an FIR structure with its transfer function

denoted as $H(z)$. In the equalizer the effect of the channel will be compensated, which implies that in the equalizer almost the inverse of the channel transfer function will be produced, therefore the equalizer generates a transfer function of $\frac{1}{H(z)}$. This suggests that ideally an IIR structure should be used in the equalizer, or an FIR structure with infinite length, for a perfect compensation. This is not, however, realizable. A finite FIR will be implemented and this imposes some error on the accuracy of the resulting compensator. Furthermore, in the linear LMS algorithm method, since the estimated channel is in the form of $\frac{1}{H(z)}$, any big notch in the channel will give rise to a big amplitude peak in the compensator filter. The channel nulling on the amplitude response occurs once the transfer function of the channel has a zero near the unit circle. In case of the presence of additive noise on the channel, this noise will be amplified or enhanced considerably, degrading the signal to noise ratio. This concern, however, may not be important in this particular application since the transmission medium is the coaxial cable and the effect of noise is almost negligible. In addition, as it was mentioned in Chapter two, the variations of the RF filter amplitude response are controlled by the component tolerances, thus such a big null in amplitude response is not likely to occur.

The other concern is related to the accuracy of frequencies close to the Nyquist range, this concern arises when a *symbol* spaced equalizer is used. Generally the symbol spaced equalizers are prone to the timing jitter errors which causes the frequencies close to Nyquist range to not exhibit an ideal accuracy.

Looking at the Figure 4.3 the frequency response of the matched filter (SRRC) for a typical communication system is plotted indicating the frequency $R_s/2$ as half the symbol rate. In the presence of timing error, the resulting frequency response of the matched filer will be deflected from the ideal response. To investigate the effect of timing error on the performance of the match filter, a simulation was performed and the results for different timing errors as a portion of symbol time of $0.0625, 0.125, 0.250\, T_{sym}$ are shown in Figure 4.4. As a result, the amplitude response degradation at near Nyquist rate are almost 0.25, 1.5, and 5 dB respectively. The al-

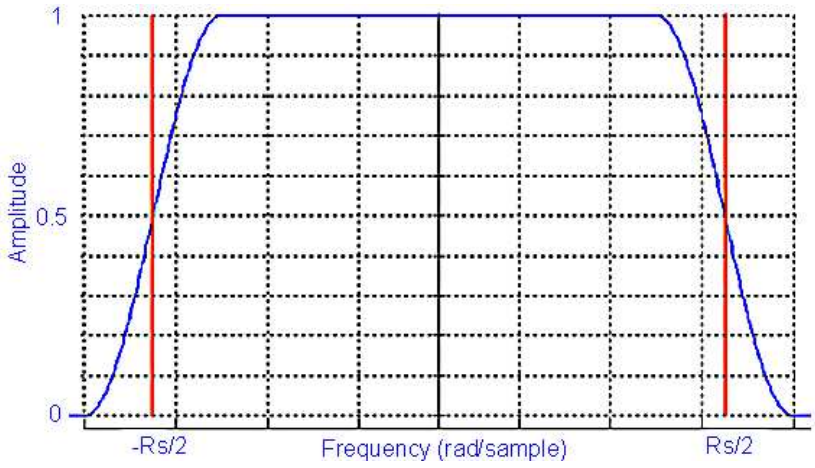

Figure 4.3 The matched filter frequency response including the Nyquist range

ternative viable option would be to use a *fractionally* spaced equalizer. In this method the equalization process occurs in a fraction of symbol time i.e., half or quarter of symbol time, and the update rate of equalizer parameters is at frequencies higher than the Nyquist rate. The fraction-based equalizer precedes the match filter and is not sensitive to the timing error like the symbol based equalizer. This entails putting the compensator after the match filter in the transmitter which implies running the compensator filter at higher speed, meaning the whole operation needs to run much faster than its baseband counterpart.

A reasonable solution is to use a symbol based equalizer that works in *training mode*, this equalizer uses some known symbols (pilot) to characterize the transmission channel (RF filter) and adjust the symbol timing, after the correct timing is acquired the timing error vanishes, yielding adequate accuracy in the Nyquist frequency region. As stated before, due to the ideal test condition the symbol timing jitter is not a concern in this application. For the purpose of this thesis, a symbol spaced equalizer with a LMS algorithm was chosen, and the simulation results prove that this technique performs satisfactorily.

In terms of the FIR filter length used in equalizer, ideally the FIR length should be infinite to represent an IIR filter with a transfer function $\frac{1}{H(z)}$, but it can be

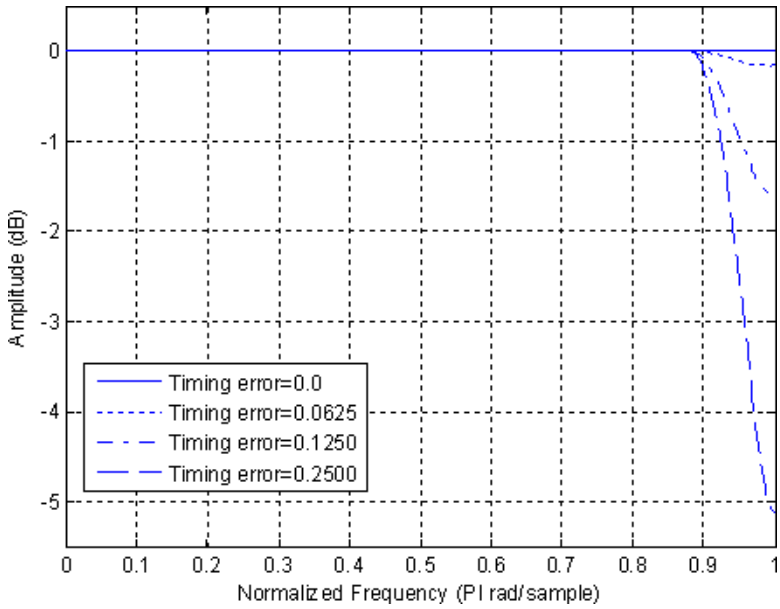

Figure 4.4 The matched filter frequency response, and its sensitivity to timing inaccuracy close to the Nyquist range

estimated with a practical finite length that yields the required accuracy. The FIR length depends on the extent to which the amplitude of the RF filter varies, and also the required accuracy. To investigate this further, a simulation was set up in Matlab and an RF filter with sharp transitions of about 25dB was characterized using an LMS equalizer in training mode, this range of variations is much more than what is needed in reality but it can prove the ability of this type of equalizer in compensating large distortions.

The simulation was performed in Matlab using the actual parameters that will be used in a typical QAM signal i.e., the symbol rate $R_s = 5360537$ sym/s, roll off factor $= 0.12$. The simulation results indicated that an FIR structure with a sufficient number of taps, in this case 69, achieves adequate accuracy for the whole band, including frequencies in the vicinity of the Nyquist frequency, with a step size of $\mu = 0.001$.

Looking at Figure 4.5, the solid line represents the original RF filter amplitude response, and the dotted line shows the equalized inverse amplitude. It is obvious

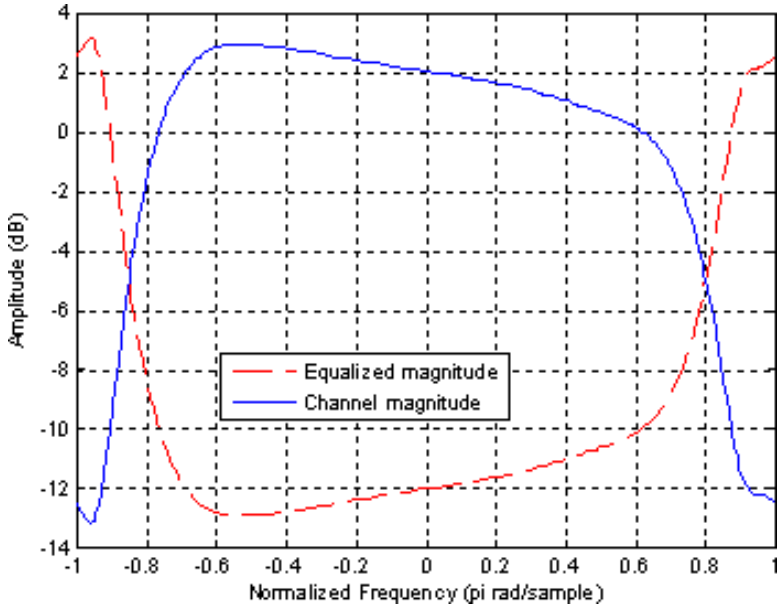

Figure 4.5 The channel amplitude frequency response and its equalized inverse

that the inverse amplitude has an acceptable accuracy. Also, looking at Figure 4.6, the solid line shows the RF filter group delay variations within the QAM signal bandwidth and the dotted line shows the equalized group delay showing the inverse of the RF filter group delay distortion with a very good accuracy. Figure 4.7, the error function versus time (symbols), indicates that the error approaches zero within 5 msec (for $R_s = 5360537$ sym/s), which is a reasonably short time interval. The un-equalized constellation is shown in Figure 4.8, and the equalized constellation is shown in Figure 4.9. It should be noted that in reality the amplitude variations within a QAM signal are much less than what was used in this simulation in the order of 2-5 dB without a null in the band. In reality the number of taps needed for the equalizer is much less than 69, and typically about 10-30 taps is needed, also the convergence time is much less.

Given the above simulation results on the requirement of this thesis, the LMS algorithm appears to perform satisfactorily and this algorithm was chosen for the equalization process.

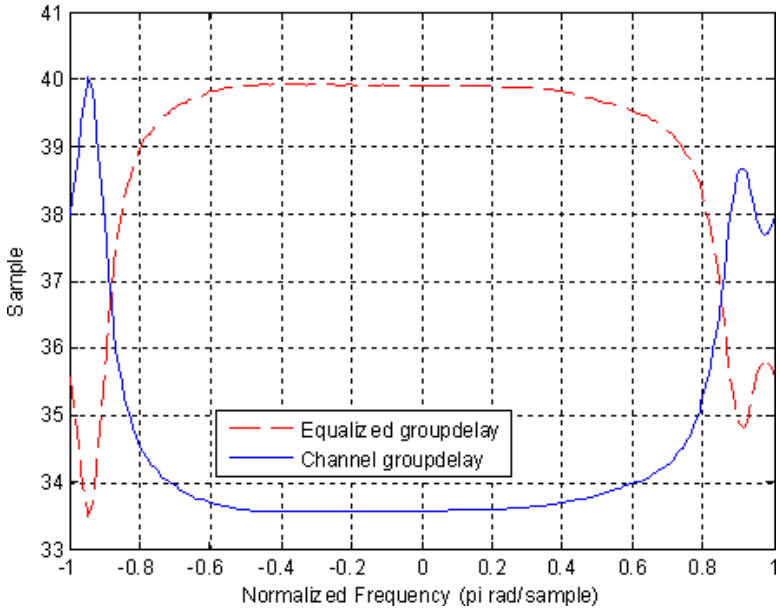

Figure 4.6 The channel group delay response and its equalized inverse

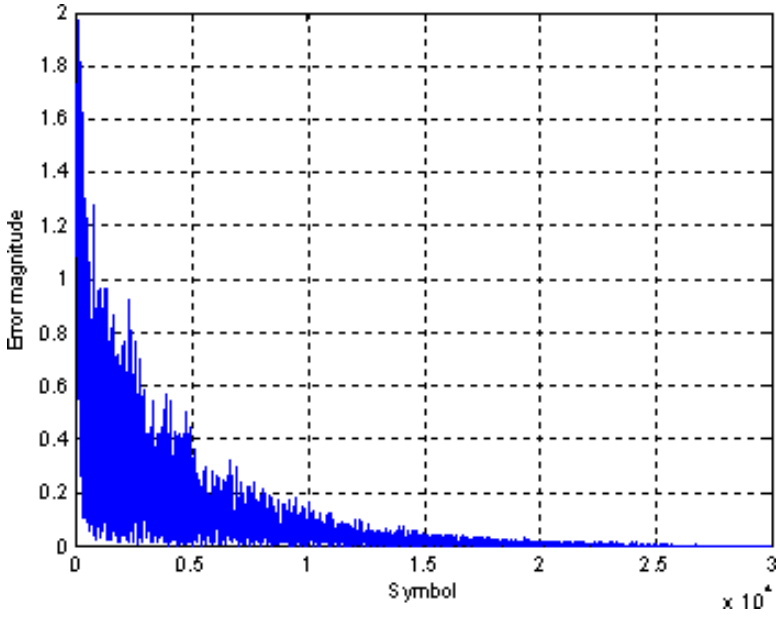

Figure 4.7 The error signal versus time (symbol)

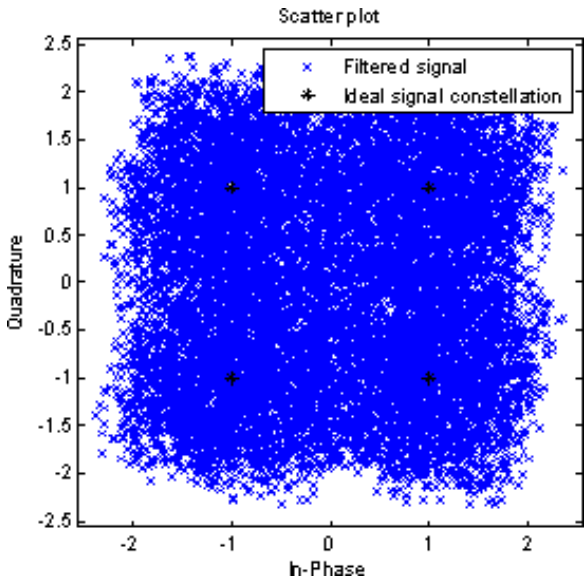

Figure 4.8 The QAM signal constellation before equalization

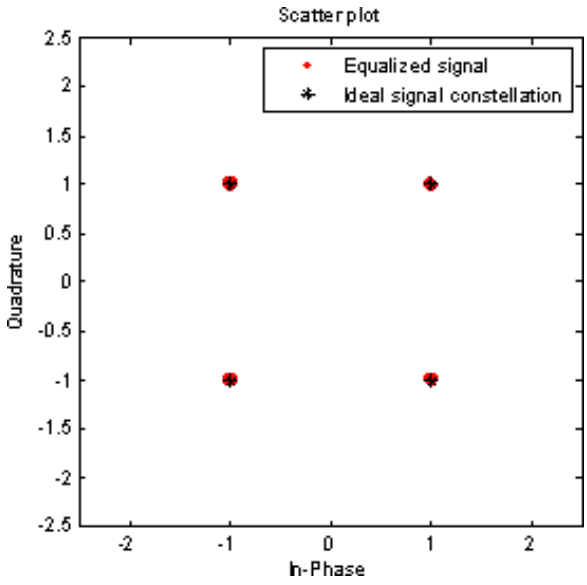

Figure 4.9 The QAM signal constellation after equalization

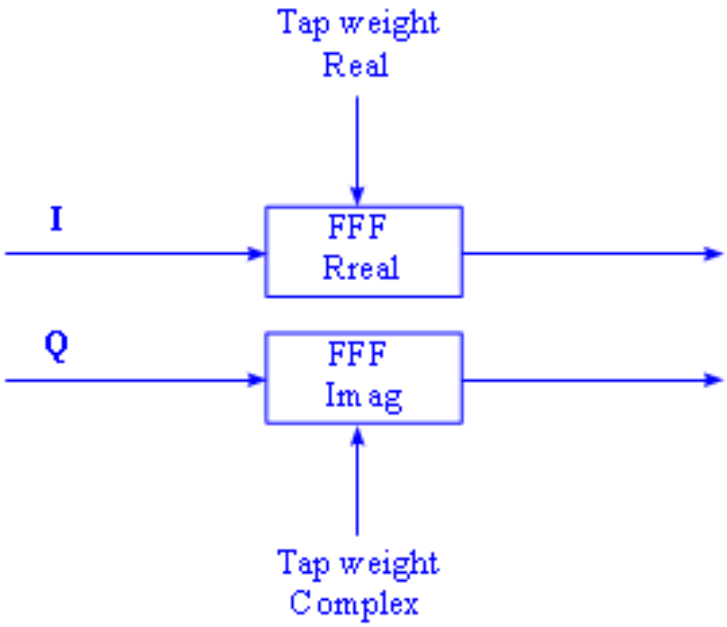

Figure 4.10 The Complex Feed Forward structure placed before the SRRC pulse shaping filter in the modulator

4.4 Implementation

The implementation of a symbol spaced equalizer involves using a pre-equalizer in the form of a feed forward structure placed before the match filter in the transmitter. The coefficients of this structure are determined during the equalization process as the tap weights.

Since the distortion effects on the QAM signal are complex in nature, the resulting inverse response is also complex, implying that the feed forward structure has to be in complex form comprised of two real and imaginary sections. The real and imaginary parts of the tap weights will be used for each part individually. Figure 4.10 shows this structure in more detail.

4.5 Summary

In this chapter, an alternative method was investigated for the purpose of compensation of the RF filter distortion. After some equalizer theory review, some key parameters of the equalizer were studied, an efficient algorithm in terms of convergence and accuracy was evaluated, and its attributes such as convergence and mean square error were characterized.

A particular type, so called symbol-spaced algorithm was studied. It has some disadvantages such as sensitivity to the timing error, but one can overcome this error by using a training mode equalizer to acquire accurate symbol timing.

In order to evaluate the performance of LMS equalizer for compensating large channel distortions, a simulation was setup in Matlab. The simulation results indicate that this equalizer is able to converge in a reasonably short time with adequate accuracy. This type of equalizer was chosen, and in the next chapter its performance will be compared with the first compensation method by using the complex low pass filter to improve the MER at the output of the IQ modulator.

5. Simulation Results

5.1 Introduction

The focus of this chapter is to determine the effectiveness of correcting distortion of an RF filter using a low pass baseband filter. The results for the complex low pass filter designed by optimization and equalization methods are compared. In addition the sensitivity of optimization with respect to key parameters is discussed. The performance measures of *convergence* and *mean square error* are both used. Finally, the performance of the compensating filter is evaluated by simulation in terms of the modulation error ratio (MER)[1] in a typical CATV setting.

The test bed for the simulation is a CATV digital communication system using the parameters dictated by DOCSIS. These parameters are: 16 QAM modulation with a symbol rate of 5360537 sym/s, and roll off factor of 0.12. The simulation setup is shown in Figure 5.1. The output of the IQ mapper is up sampled by 16 (this block is not shown in Figure 5.1). Referring to this Figure, it should be noted that during the compensation through the complex base band filter, the feed forward structure is bypassed. Also during the MER calculation, the MER block is switched to the output of the match filter (SRRC). During equalization the MER block is switched to the equalized output of the equalizer block. Finally, the performance of these two methods in tandem will be compared.

[1]For the definition of MER refer to the appendix D

5.2 Optimization results

5.2.1 Amplitude response

There is no universal method to determine the point in time that an algorithm converges. In each iteration the descent gradient vector $-\mathbf{S}_k\mathbf{g}_k$, which is pointing in the descent direction, needs to be calculated. Also, the parameter α_k is calculated through a univariate grid search. The quantity $-\alpha_k\mathbf{S}_k\mathbf{g}_k$ is the optimum step in the descent direction. As the optimization proceeds in each iteration the magnitude of $-\alpha_k\mathbf{S}_k\mathbf{g}_k$ becomes smaller and smaller as the point of the minimum is approached. There is a correspondence between the magnitude of $-\alpha_k\mathbf{S}_k\mathbf{g}_k$ and the closeness to the minimum point. Therefore, this parameter is used to indicate the point in time that the algorithm converges. Once this parameter is smaller than a required tolerance (tol), the optimization will be terminated. Moreover, the norm square of error function (**MSE**), as the result of difference between the specified response and optimized response, can be used as alternate convergence metric.

To examine the capability of the optimization algorithm in terms of optimizing different amplitude and group delay responses, one can refer to the Monte Carlo simulation results in chapter two, the range of variation of amplitude /group delay responses is limited to ±2dB. In order to show the optimization performance, one approach would be to show the evolution of the amplitude response in each iteration.

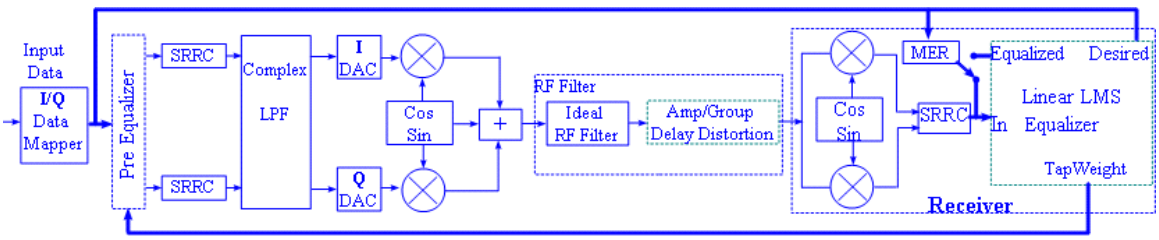

Figure 5.1 The simulation setup, FFF pre equalizer placed before SRRC filter, and complex LPF placed after the SRRC filter in the modulator, followed by the test receiver containing equalizer and MER measurement block

However, this is not practical, therefore only the results of some selected iterations are presented in Figure 5.2. It is obvious that the amplitude response is approaching the target response iteratively. For this particular problem it takes some 17 iterations to complete this optimization with ($| - \alpha_k \mathbf{S}_k \mathbf{g}_k | < 10^{-5}$). To better understand the convergence of the optimization process, the variations of the pseudo optimum descent vector pointing in the descent direction in each iteration is indicated in Figure 5.3. Also the zero plot of the optimized filter is shown in Figure 5.4.

5.2.2 MER performance results

The improvement in MER is demonstrated with and without the compensation filter present. Referring to the BER simulation results performed in Chapter 2, the BER performance is more susceptible to parabolic and linear amplitude distortions. In this simulation, different amplitude slopes were examined and the resulting MER degradations for a CATV 64 QAM signal with a symbol rate of 530537 sym/s and roll off factor of 0.12 are measured. The results are shown in Figure 5.5

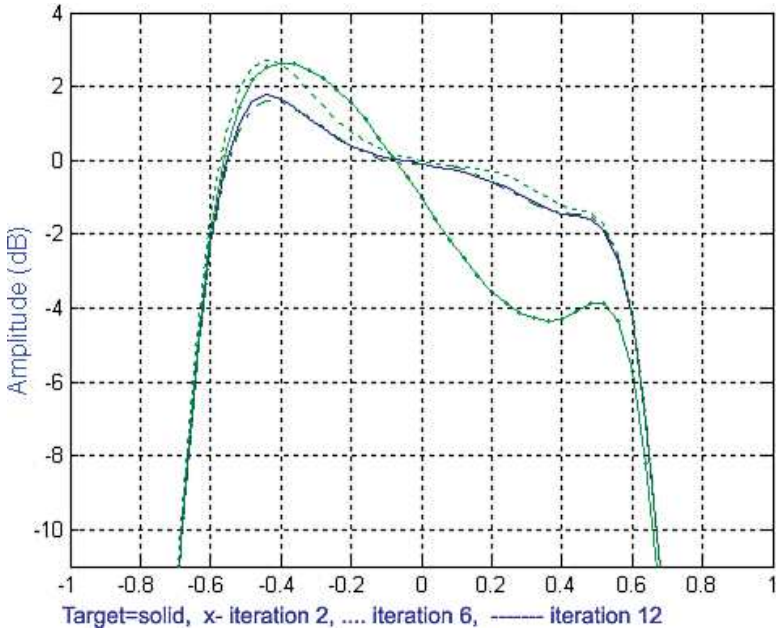

Figure 5.2 The evolution of optimized amplitude response in the 2nd, 6th, and 12th iterations toward the desired response.

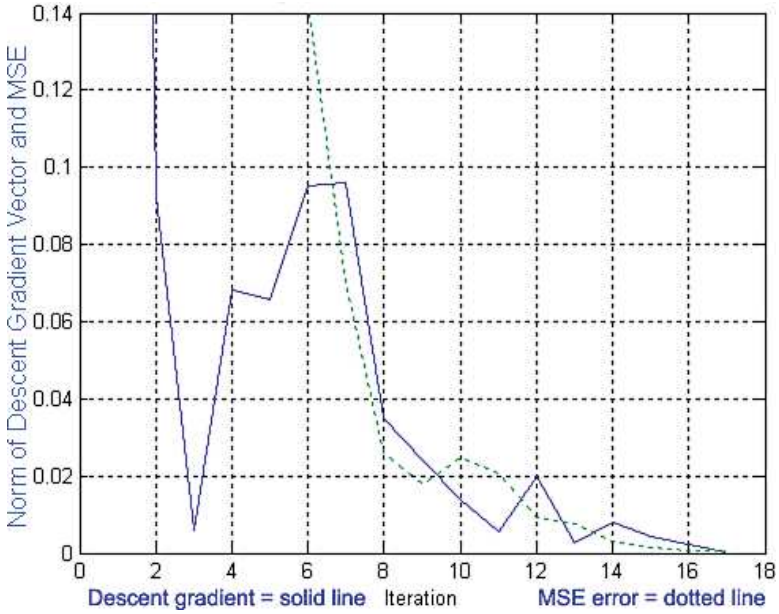

Figure 5.3 Magnitude of descent gradient vector $-\alpha_k \mathbf{S}_k \mathbf{g}_k$(solid line), and MSE in each iteration(dotted line)

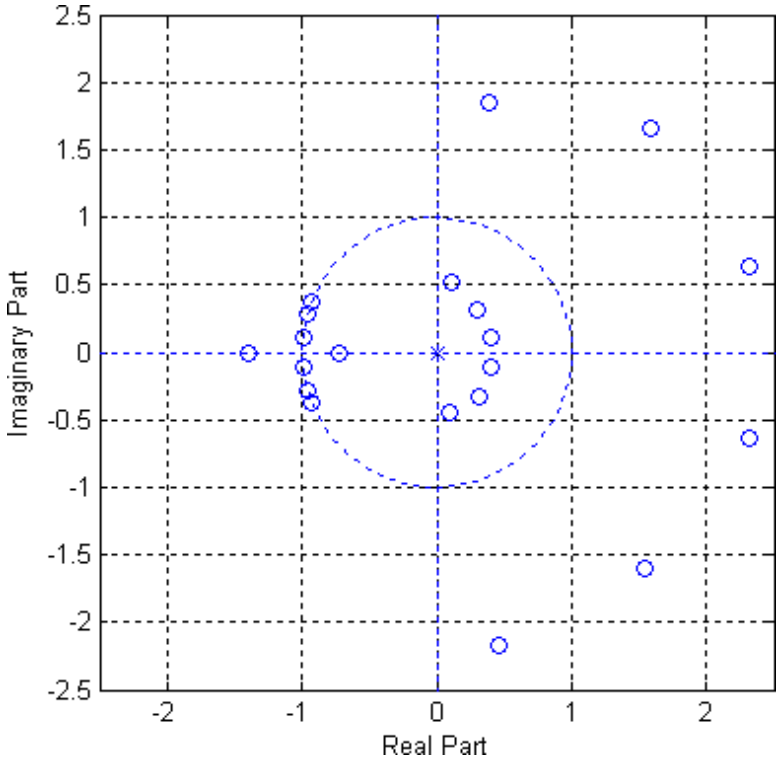

Figure 5.4 The zero plot of the compensation filter

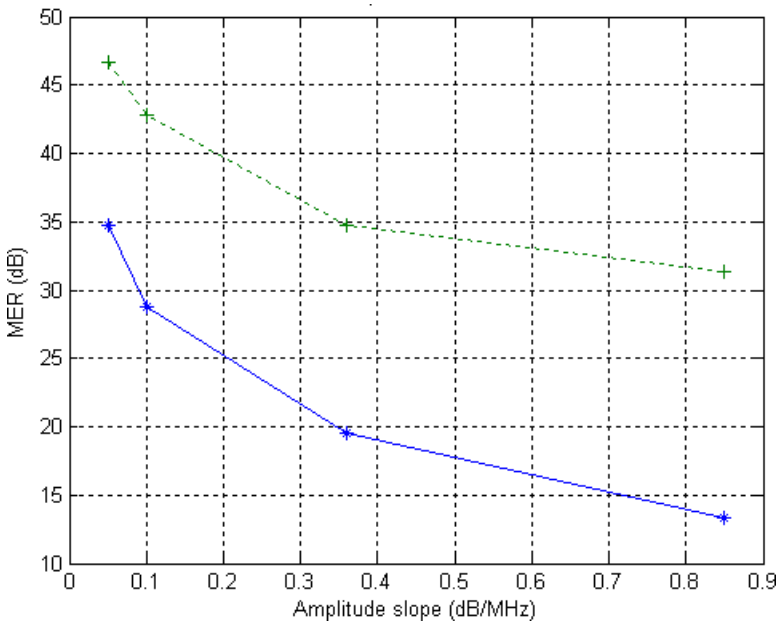

Figure 5.5 MER without the compensation filter (solid line) and MER with the compensation filter (dotted line) vs. amplitude slopes in the RF filter.

5.2.3 Group delay response

The quasi-Newton algorithm is used to find the coefficients of the allpass filter that compensates for the group delay distortion. The initial filter is a real allpass filter with IIR structure. Since the poles and zeros are reciprocally paired to yield a flat amplitude response, the poles were chosen as independent variables of objective function and the zeros were considered as reciprocals of the poles. This objective function is iteratively optimized so that the resulting group delay response approaches the target response. This implies that an objective function will be formed which is defined in the form of norm square of the difference between the specified response and the response of the filter under optimization. This particular optimization is a constrained optimization with the constraint being the magnitude of poles, $\rho_{pole} < 0.99$ to ensure stability. Another constraint is that the angle between two poles must be greater than 0.01 radians/sample to prevent overlapping the poles. The group delay optimization is accomplished only in the band of interest (pass band) not the whole Nyquist band. From the optimization results, it is understood that optimizing the group delay is not as efficient as the amplitude optimization. This is partly due to the fact that the phase response in an allpass filter is monotonically decreasing with frequency, and by changing the group delay response for one spot, the group delay response of other regions will also change and is very sensitive to the pole/zero variation. The group delay response is not well behaved and usually the gradient of the error function with respect to the pole/zero variations yields very large values.

For the group delay optimization, like amplitude optimization, the descent gradient vector pointing to the descent direction is represented by $-\alpha_k \mathbf{S}_k \mathbf{g}_k$, when $-\alpha_k$ is being computed iteratively via a univariate grid search. Once the norm square of the descent gradient vector is less than a specified tolerance ($|-\alpha_k \mathbf{S}_k \mathbf{g}_k| < 10^{-3}$), the optimization will be terminated. For more clarification, the optimization result for one group delay response of a typical RF filter is shown in Figure 5.6, followed by the mean square error in Figure 5.7, and the norm square of the descent gradient vector in each iteration in Figure 5.8.

5.2.4 MER performance results

Since the optimization results indicated a moderate performance for the group delay optimization process, one would intuitively expect that the MER improvement as a result of this optimization is reasonably well. To further investigate this intuitive observation, the same simulation setup was used and the RF filter with various group delay slopes was plugged into the CATV system. The linear slope group delay distortion has the most degradation effect on the BER. Thus this type of distortion which outweighs all other distortions, was used during the simulation and performance test.

In Figure 5.9 this linear group delay and the optimized response is shown, it is understood that the compensated response is very close to the desired response. Various linear group delays with increasing slopes were generated for this simulation. The results agree with the preliminary intuitive observation, indicating that the group delay optimization has a modest performance for the small amount of group delay distortions.

The results of MER simulation for a CATV 64 QAM signal with roll off = 0.12 and symbol rate = 5360537 sym/s are shown in Figure 5.10. Looking at Figure 5.10, it is obvious that the amount of MER improvement for a group delay slopes of 0.2 and 1.8 nsec/MHz is 2dB and 14 dB respectively. In other words, the precision of group delay optimization is about to 0.2 nsec/MHz for the symbol rate of 5360537 sym/s. One could increase this resolution by *increasing* the up-sampling rate.

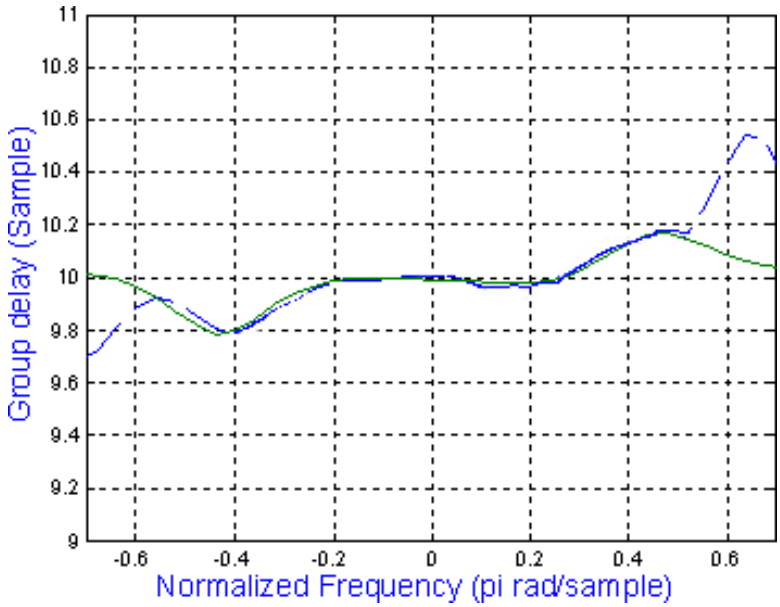

Figure 5.6 Group delay response of optimizations process, desired (solid line), and optimized(dashed line)

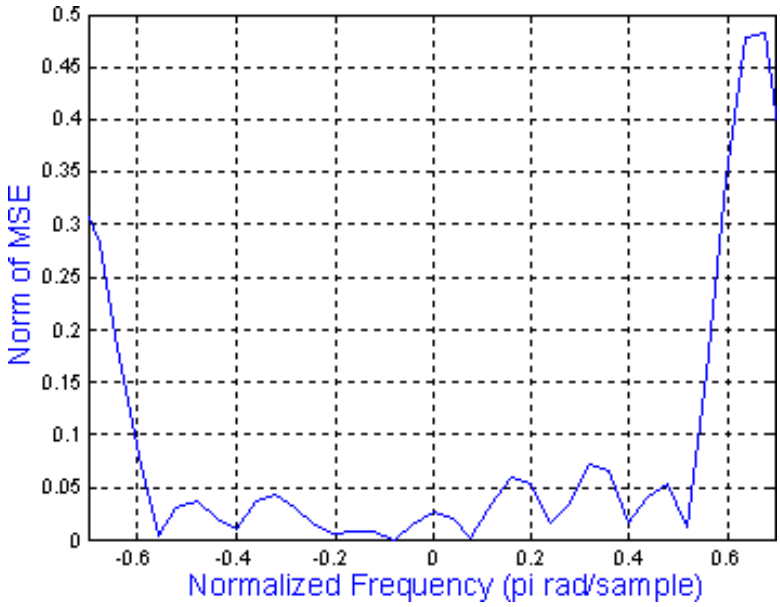

Figure 5.7 The mean square error between desired and optimized response

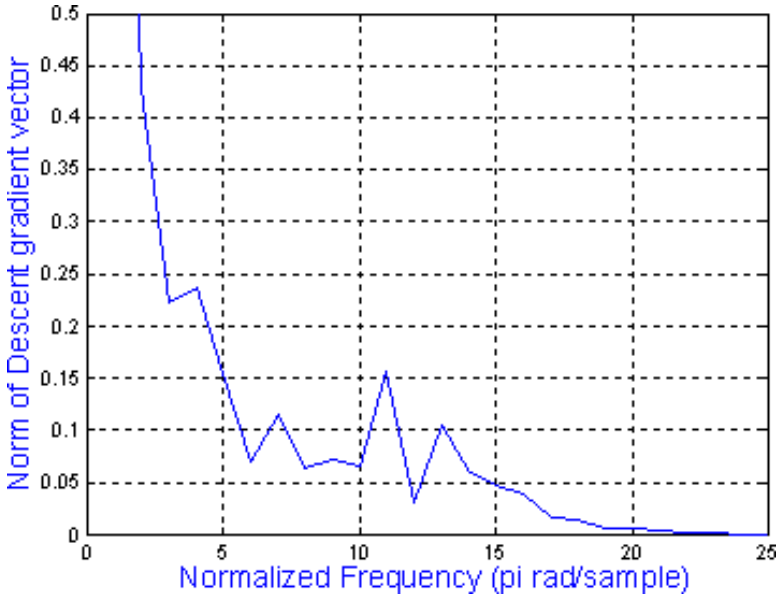

Figure 5.8 The norm square of descent gradient vector $|-\alpha_k \mathbf{S}_k \mathbf{g}_k|$

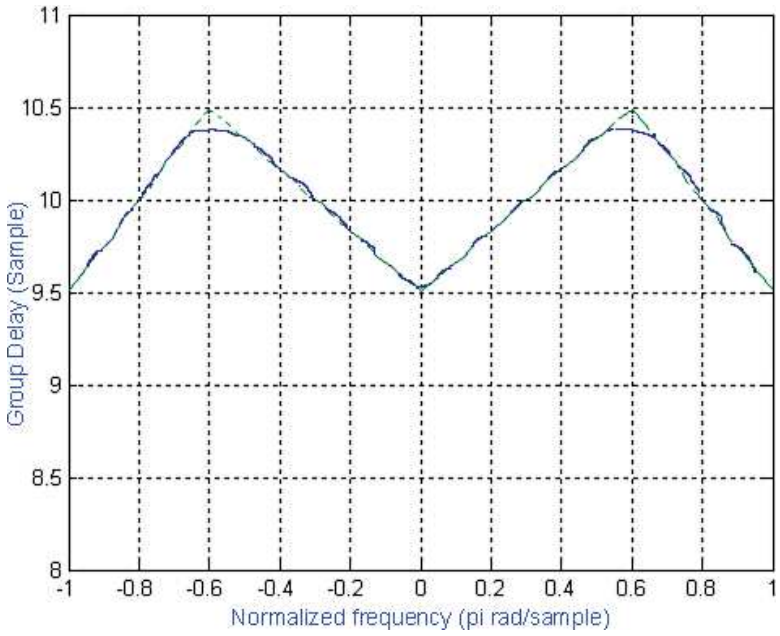

Figure 5.9 The desired linear slope group delay (dashed line) and compensated response (solid line)

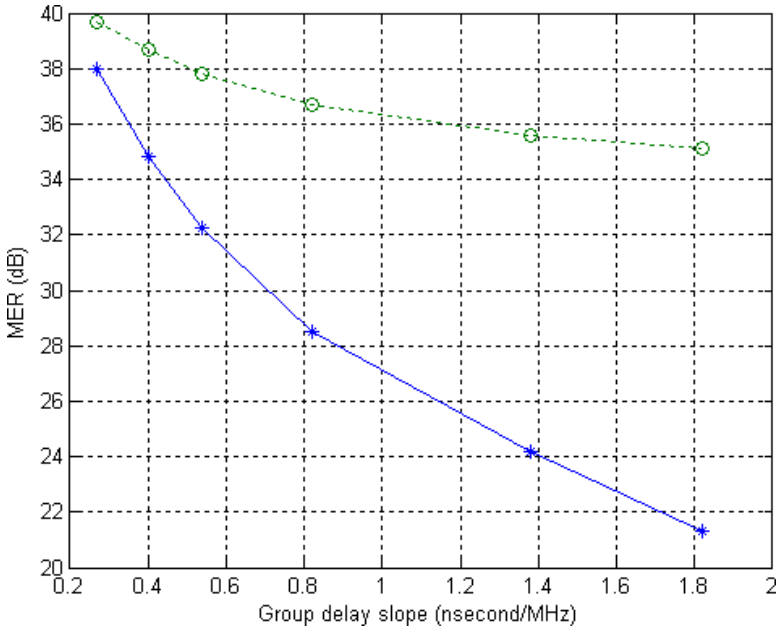

Figure 5.10 The MER without group delay compensation filter (solid line) and with compensation filter (dotted line) versus the slope of the group delay distortions.

5.3 Equalization results

The performance of the equalization-based compensation is investigated using a simulation setup similar to the one used for the optimization method. To perform an exhaustive test on the performance of the equalizer, the worst case group delay type, linear slope group delay, was generated with different slopes (nsec/MHz). In order to obtain satisfactory results, the LMS equalizer parameters need to be optimized. While running in training mode, it was empirically found that 19-30 taps will work reasonably well for a range of group delay slopes from 0.27 to 1.9 nsec/MHz, provided the reference tap is in the range of tap number 9 to 21. The step size, μ, was chosen to be 0.001 for all simulations.

The results with and without a pre-equalizer are shown in Figure 5.12. From this Figure it is obvious that the equalizer has very good performance even at very low group delay slopes, where the improvement in the MER is approximately 17 dB. This method will add to the latency of the system, the typical latency for the equalization method is about 15-21 symbol time.

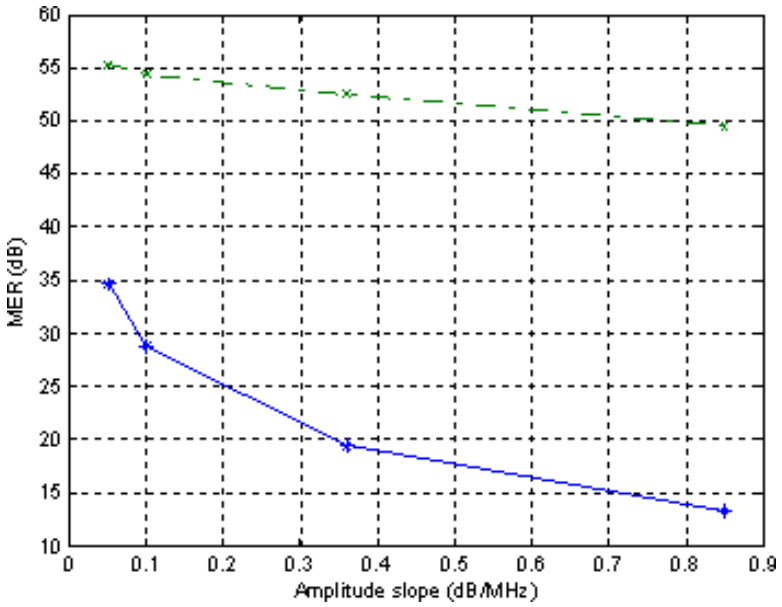

Figure 5.11 The MER results for the distorted amplitude (solid line)and equalized channel(dotted line)for different linear slopes

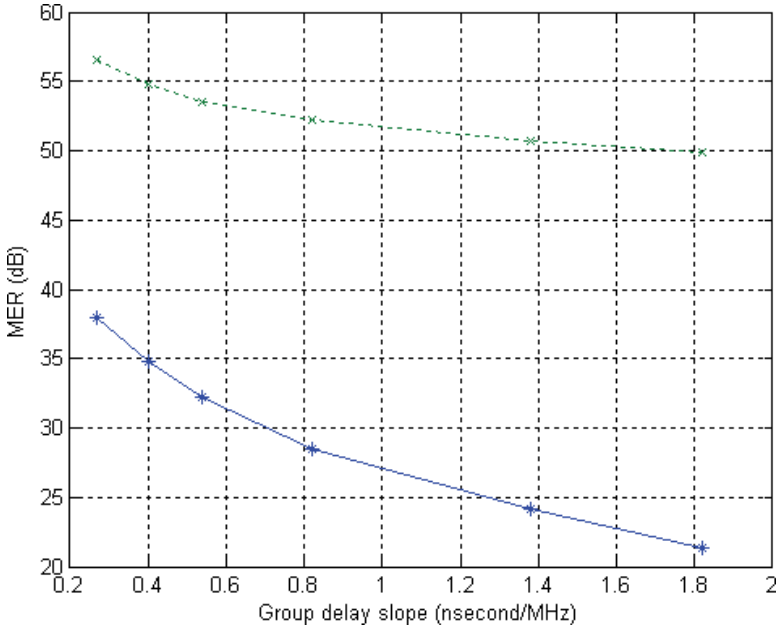

Figure 5.12 The MER results without pre-equalizer (solid line) and with pre-equalizer (dotted line) versus the slope of the group delay distoriton

5.4 Performance Comparison

Finally in this section the performance of the complex lowpass compensation filter is compared with the performance of the pre-equalizer in terms of their improvement on the MER. The MER performance comparisons for compensating the amplitude response distortions are presented in Figure 5.13. At 0.05 dB/MHz the complex lowpass filter improves the MER by 12 dB, while the pre-equalizer improves the MER by 20 dB. Clearly the pre-equalizer is better by at least 8dB.

The MER performance of both compensation methods with respect to group delay is shown in Figure 5.14. The pre-equalizer offers 17 dB better improvement for group delay distortions with slopes as low as 0.2 nsec/MHz. Clearly the pre-equalizer is a better method by at least 15 dB for group delay distortions with slopes from 0.22 nsec/MHz to 1.9 nsec/MHz. However, in terms of the latency, the optimization method adds less latency to the system (about 2 symbol time), whereas the equalization method adds some 15-21 symbol time. In the applications that the latency is very important the optimization method is preferred. Furthermore, the optimization method compensates for multiple carriers at a time, whereas the equalization method compensates only for single carrier at a time.

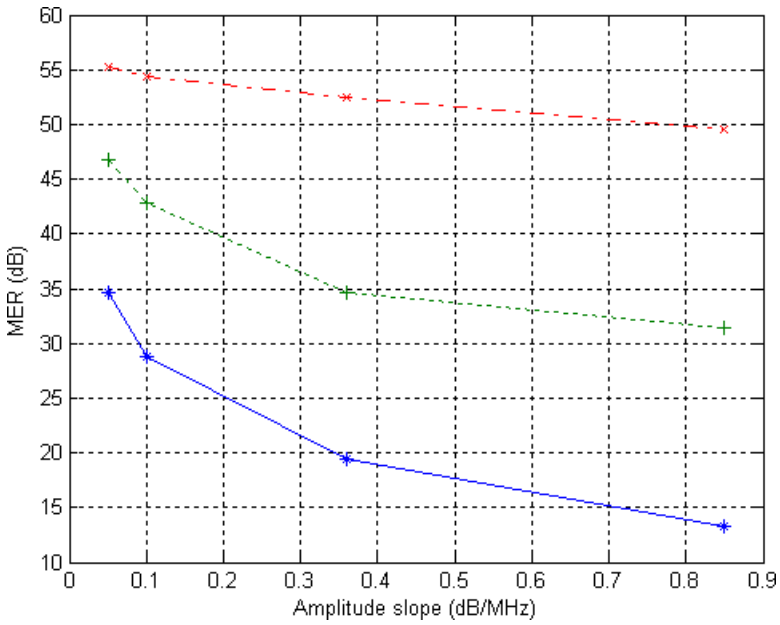

Figure 5.13 The MER without any compensation (solid line), with the complex lowpass filter compensation (dotted line+) and with pre-equalization (dash -dot-$*$) versus the amplitude distortions of the RF filter

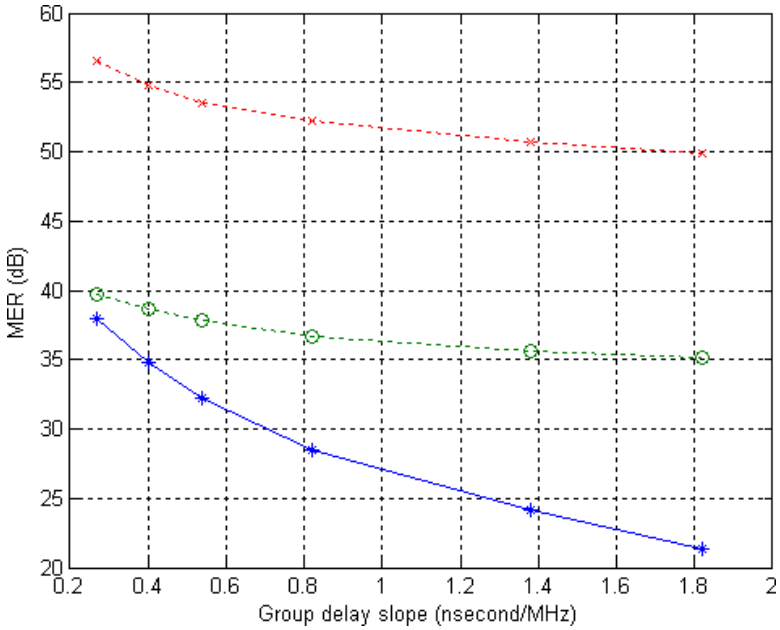

Figure 5.14 The MER without any compensation (solid line), with the complex lowpass compensation filter (dotted line o) and with pre-equalization (dot-$*$) versus the slope of the group delay distortions in the RF filter

6. Conclusion

6.1 Summary

In a CATV system, the RF Filter amplitude and group delay distortions will cause a reduction in the MER. In chapter two it was shown that these distortions have a complex degradation effect on the QAM signal when translated into base band frequency. Also the degree and shape of the distortion, which were classified as linear, parabolic and sinusoid, affect the MER which directly effects the BER performance.

The distortion is corrected in two different ways. One is with a complex coefficient filter whose amplitude and group delay responses are the inverse of the RF filter response. The other method is to use a symbol spaced pre-equalizer that operates on the data. In the former method, the coefficients for the filter that compensates for the amplitude/group delay were found iteratively using the mean squared error between the compensated and target responses. The total compensating filter was broken into two filters in cascade: One compensating for amplitude and the other for group delay. For compensating amplitude distortion, an FIR structure was employed for the lowpass filter. For compensating the group delay distortion an IIR structure in the form of an allpass filter was employed.

The Quasi-Newton optimization algorithm [24] was used to determine the complex coefficients of these two structures. The very important feature of the group delay compensation filter is the reciprocal relationship between poles and zeros. The amplitude compensation filter has a similar relationship between the zeros inside the unit circle and the zeros outside the unit circle. In each iteration, the steepest descent direction vector with optimum magnitude, $-\alpha \mathbf{S}_k \mathbf{g}_k$, is obtained to compute the

coordinates of the next point. During this optimization, the optimum range of step sizes for the gradient direction vector as well as the line search required to compute the optimum magnitude of direction vector α_k was obtained that yielded the best results. The resulting FIR filter required 32 zeros for acceptable performance. A similar strategy was employed to design the allpass IIR filter with constant amplitude and arbitrary response, but here the poles were used as independent variables and zeros were chosen as their reciprocals. To ensure stability, the magnitudes of the poles were constrained to be less than 0.99. For the allpass filter, an order of 6 to 20 was required to get acceptable performance. These two structures, were connected in cascade and placed after the matched filter in the modulator.

The pre-equalizer method used a complex coefficients feed forward structure (FIR) placed prior the pulse shaping filter in the modulator. A well suited equalization algorithm with satisfactory performance in terms of convergence rate and accuracy, was found to be the linear LMS algorithm. This algorithm when used with sufficient number of taps and small convergence step size, yields satisfactory results for the range of variations that are typical and shown through the RF simulation results in chapter two. The equalizer was trained in a test receiver. Once the tap weights were acquired, they were used as the complex coefficients of the pre-equalizer prior to the pulse shaping filter in the modulator.

The simulation results presented in Chapter 5 suggest that the quasi-Newton algorithm performs reasonably well for the amplitude compensation. Examining this functionality, different amplitude responses with variations up to ± 2.5 dB were tested, and the developed algorithm was able to make the compensation filter response converge to the target response, yielding a very good MSE error of within 0.02 dB.

For the case of group delay optimization it was shown that the group delay compensation has modest performance for the subtle group delay distortions within 0.05 sample, this resolution translates to a group delay of 0.58 nanosecond for a symbol rate of 5360537 sym/s.

6.2 Results

The performance was verified for amplitude distortion that was proportional to frequency, the slope of the distortion was varied from 0.05 to 0.85 dB/MHz. The results indicated that uncompensated MER is between 34.7 dB and 13.3 dB for this range of amplitude slope. Once the FIR amplitude compensator is used, the MER is between 46.7dB and 31.4 dB, the improvement is between 12 and 18.1dB. This means that the amplitude optimization method, improves the MER of distorted QAM signal by at least 12 dB.

In the case of group delay optimization, the optimization result was verified by simulating the effect of different group delay slopes on the QAM signal. The slope ranged from 0.27 to 1.8 nsec/MHz. The uncompensated MER ranged from 38 dB to 21.3dB. Applying the IIR group delay compensator, the MER ranges from 39.7 dB to 35.1 dB, which is an improvement of 1.8 dB to 13.8 dB. This simulation indicates that the group delay optimization algorithm offers smaller improvement on MER at small group delay distortions. The MER improvement diminishes for subtle distortions as predicted from the optimization results. This is mostly due to the fact that allpass filter has a monotonically decreasing phase response and is not able to compensate for certain shapes of group delay.

The equalization method, shows much better results under similar conditions. In terms of amplitude slopes for the same range of variations, the MER with pre-equalizer is 55.2 dB to 49.5 dB, which shows 20.5 to 36.2 dB improvement with respect to the un-equalized QAM signal. This amount of improvement outperforms the optimization method by 8.5 dB to 18.1 dB for different slopes. The superiority of the pre-equalizer is more obvious from the group delay compensation stand point. When the pre-equalizer is used for various group delay slopes from 0.27 to 1.8 nsec/MHz, the pre-equalized MER ranges from 56.5 dB to 49.9 dB showing an improvement of 18.5 dB to 28.6 dB. This is better than the optimization performance by at least 16.8 dB to 14.8 dB. This suggests that the equalization method has very good precision and can improve

MER substantially, even for small group delay slopes.

Given the above simulation results, one can conclude that the equalization method outperforms the lowpass filter method in both the amplitude and group delay compensations by a least 8.5 dB and 16.8 dB respectively. However, from the latency point of view, the optimization adds less latency to the system than the equalization method, this can be very important is some applications. In addition, optimization method can compensate for multiple carriers at a time, as oppose to the equalization method which can only compensate for a single carrier at a time.

6.3 Future work

The important aspects of the DSP compensation was discussed in this thesis. However, some topics such as finite word length for a practical implementation of FIR/IIR structure was left for the future work. Here is a list of topics that will be left for work to be done in the future:

1. Comprehensive mathematical analysis on the effect of distortions on the QAM signal.
2. Implementation and filter structure effects such as overflow and finite length word.
3. Detailed filter characterization process.
4. The effect of compensation filters on the latency of the modulator in the whole CATV system.

REFERENCES

[1] A. Azzam, *High Speed Cable Modems*. McGraw Hill, 1997.

[2] W. Ciciora, *Cable Television in United States, An overview*. Loiuse Ville CO, Cable Labs, 1995.

[3] W. C. et al, *Modern Cable Television Technology*. Elsevier, CA, Morgan Kaufmann Publishers, 2004.

[4] C. T. Labs, *Data Over Cable Service Interface Specifications DOCSIS 3.0 Physical Layer Specification Down Stream RF Interface Specification*. CM SP DRFI I07 081209, 2008.

[5] R. N. P. Carl W. Anderson, Stephen G. Barber, "The effect of selective fading on digital radio," *IEEE Transaction on Communicatons*, vol. COM-27, pp. 1870–1875, Dec 1979.

[6] K. F. Douglas H. Morais, AKE Sewerinson, "The effects of the amplitude and delay slope components of frequency selective fading on qpsk and 8psk systems," *IEEE Transaction on Communicatons*, vol. COM-27, pp. 1849–1853, Dec 1979.

[7] P. Mathiopoulos, H. Ohnishi, and K. Feher, "Study of 1024-qam system performance in the presence of filtering imperfections," *IEE Proceeding Communications*, vol. 136, pp. 175–179, April 1989.

[8] M. Ramadan, "Availability prediction of 8psk digital microwave syatems during multipath propagation," *IEEE Transaction on Communicatons*, vol. COM-27, pp. 1862–1869, Dec 1979.

[9] K.-T. Wu and K. Feher, "256-qam modem performance in distorted channels," *IEEE Transaction on Communicatons*, vol. COM-33, pp. 487–491, May 1985.

[10] K. F. Tricia Hill, "A performance study of nla 64-state qam," *IEEE Transaction on Communicatons*, vol. COM-31, pp. 821–826, June 1983.

[11] P. Mathiopoulos, , and K. Feher, "Performance evaluation of a 512-qam system in distorted channels," *IEE Proceeding Communications*, vol. 133, pp. 199–204, April 1986.

[12] O. Herrmann, "Design of non recursive digital filters with linear phase," *Electron Letter*, vol. 6, pp. 182–184, May 1970.

[13] A. o. E. Hofstetter and J. siegel, "A new thechnique for the design of non recursive digital filters," *5th Annual Princton Conf. Information sciences and Systems*, vol. 1, pp. 64–72, March 1971.

[14] T. W. Parks and J. H. McClellan, "Chebychev approximation for non recursive digital filters," *IEEE Transaction on Cicuit Theory*, vol. 19, pp. 189–194, March 1972.

[15] T. W. Parks and J. H. McClellan, "A program for the design of linear phase finite impulse response digital filters," *IEEE Transaction on Audio electroacoust*, vol. 20, pp. 195–199, Aug 1972.

[16] L. R. Rabiner and O. Herrmann, "on the design of optimum fir low pass filters wit even impulse response duration," *IEEE Transaction on Audio electroacoust.*, vol. 21, pp. 329–336, Aug 1973.

[17] J. H. McClellan and T. W. Parks, "A unified approach to the design of opimum fir linear phase digital filters," *IEEE Transaction on Circuit Theory*, vol. 20, pp. 697–701, Nov 1973.

[18] T. W. P. J. H. McClellan and L. R. Rabiner, "A computer program for designing opimum fir linear phase digital filters," *IEEE Transaction on Audio electroacoust.*, vol. 21, pp. 506–526, Dec 1973.

[19] J. H. M. L. R. Rabiner and T. W. Parks, "Fir digital filter design techniues using weighted chebychev approximation," *IEEE Proc.*, vol. 63, pp. 595–610, Apr 1975.

[20] T. W. P. J. H. McClellan and L. R. Rabiner, "Fir linear phase filter design program," *IEEE Press*, vol. programs for digital signal processing, pp. 5.1 1–5.1 13, Apr 1979.

[21] E. Y. Remez, "General computational method for tchebycheff approximation," *Atomic Energy Commission Translation*, vol. COM-27, pp. 1–85, Dec 1957.

[22] A. Antoniou, "Accelerated procedure for the design of eqiripple non recursive digital filters," *IEE Proc.*, vol. 129, pp. 1–10, Feb 1982.

[23] A. Antoniou, "New improved method for the design of weighteed chebychev non recursive digital filters," *IEEE Trans. Circuit and Systems*, vol. 30, pp. 740–750, Oct 1983.

[24] A. Antoniou, *Digital Signal Processing Signals Systems and Filters.* McGraw-Hill, 2006.

[25] K. steiglitz, "Computer aided design of recursive digital filters," *IEEE Trans. Audio electroacoust.*, vol. 18, pp. 123–129, June 1970.

[26] A. G. Deczky, "Synthesis of recursive digital filters using the minimump-error criterion," *IEEE Trans. Audio electroacoust.*, vol. 20, pp. 257–263, Oct 1972.

[27] J. W. Bandler and B. L. Bardakjian, "Least pth optimization of recursive digital filters," *IEEE Trans. Audio electroacoust.*, vol. 21, pp. 460–470, Oct 1973.

[28] C. charalambous, "Minmax design of recursive digital filters," *Computer Aided Design*, vol. 6, pp. 73–81, Apr 1974.

[29] C. charalambous, "Minmax optimization of recursive digital filters using recent minmax results," *IEEE Transaction Acoustics speech Signal Processing*, vol. 23, pp. 333–345, Aug 1975.

[30] S. Haykin, *Adaptive Filter Theory.* Pearson education Inc, 2008.

[31] A. H. Sayed, *Adaptive Filters.* John Wiley, 2008.

[32] C. T. Labs, *Data Over Cable Service Interface Specifications DOCSIS 3.0 Physical Layer Specification.* CM SP PHY v3.0 I03 070223, 2007.

[33] U. Government, *Code of Federal Regulations, C.F.R. 47 76.5 (w)DOCSIS 3.0 Physical Layer Specification.* Government Printing Office, Washington, DC., 1999.

[34] D. M. Pozar, *Microwave Engineering.* John Wiley, Third Edition.

[35] A. V. O. R. W. Schafer, *Discrete Time Signal Processing.* Prentice Hall, 1989.

[36] D. M. Himmelblau, *Applied NonLinear Programming.* McGraw-Hill, 1972.

[37] R. Fletcher, *Practical Methods for optimization, Unconstrained Optimization 2nd edition.* New York Wiley, 1990.

[38] D. G. Luenberer, *Linear and Nonlinear Programming 2nd editon.* addison-Wesley, 1984.

[39] P. E. G. W. Murray, *Practical Optimization.* New York Academic, 1981.

[40] B. D. Bunday, *Basic Optimization Methods.* London edward Arnold, 1984.

[41] S. K. Mitra, *Digital Signal Perocessing a Computer Based Approach.* The McGraw-Hill companies Inc, 1998.

[42] Vaidyanathan, *Multirate Systems and Filter Banks.* Pearson Education Inc, 1993.

[43] F. Boroujeny, *Adaptive Filters Theory and Applications.* John Wiley, 1999.

[44] J. G. Proakis, *Digital Communications.* McGraw-Hill Forth edition, 2007.

[45] A. Papoulis, *Probability Random Variables and Stochastic Processes.* McGraw-Hill Fourth edition, 2002.

[46] M. Rice, *Digital Communication, A Discrete Time Approach.* Printice Hall, United States ed edition, 2008.

[47] L.R.Rabiner and B. Gold, *Theory and application of Digital Signal Perocessing.* The Prentice Hall Englewood Cliffs NJ, 1975.

[48] C. McMillan, *An introduction to design and application of optimial decision machines.* John Wiley, 1974.

[49] N. Chowdhury, *Mathematical Tecchniques in Engineering.* E840 Cours material U of S, 2008.

[50] G. Strang, *Linear Algebra and its Applications.* Thomson Brooks/Cole, 2006.

[51] E. W. Cheney, *Introduction to approximation theory.* McGraw-Hill New York, 1996.

A. LTI Discreet-Time Systems Review

To get a better understanding of how complex low pass filtering works, it is helpful to refer to the basic theory of the LTI systems, where the very basic concept of filtering comes from. One can refer to text books such as Oppenheim, [35] Rice [46] for a detailed description of *Linear Time Invariant* systems.

The input-output relation of an LTI discrete-time system with an impulse response $h[n]$ is defined by the convolution sum of equation A.1 represented in the mathematical form of:

$$y[n] = \sum_{k=-\infty}^{k=\infty} h[k]x[n-k] \tag{A.1}$$

Where, $y[n]$ and $x[n]$ are the output and the input sequences respectively. The input sequence $x[n]$, is in the form of complex exponential as

$$x[n] = e^{\jmath\omega n}, \qquad -\infty < n < \infty \tag{A.2}$$

Then from equation A.1 the output is

$$y[n] = \sum_{k=-\infty}^{k=\infty} h[k]e^{\jmath\omega(n-k)} = \left(\sum_{k=-\infty}^{k=\infty} h[k]e^{-\jmath\omega k} \right) e^{\jmath\omega n} \tag{A.3}$$

The equation A.3 can be rewritten as

$$y[n] = H(e^{\jmath\omega})e^{\jmath\omega n}, \tag{A.4}$$

Where the system transfer function is

$$H(e^{\jmath\omega}) = \sum_{k=-\infty}^{k=\infty} h[n]e^{-\jmath\omega n} \tag{A.5}$$

Here the quantity $H(e^{\jmath\omega})$ is known as *frequency response* of the LTI discrete time system providing the frequency domain behavior of the system. from equation A.5 it is understood that to be more accurate, the $H(e^{\jmath\omega})$ is representative of the discrete-time Fourier Transform (DTFT) of the system impulse response $h[n]$.

From the equation A.3 it is obvious that with a complex sinusoidal input sequence $x[n]$ with angular frequency ω, the output is also a sinusoidal complex sequence with the same angular frequency, but is weighted with a complex amplitude $H(e^{\jmath\omega})$ that is a function of input frequency ω and the system's impulse response coefficients $h[n]$. It can be shown that the $H(e^{\jmath\omega})$ completely characterizes the LTI discrete-time system in the frequency domain.

In general, the discrete-time Fourier transfer function $H(e^{\jmath\omega})$ is a complex function of ω and period of 2π can be represented based on its real and imaginary parts

$$H(e^{\jmath\omega}) = H_{re}(e^{\jmath\omega}) + \jmath H_{im}(e^{\jmath\omega}) = |H(e^{\jmath\omega})|e^{\jmath\theta(\omega)}$$

Where $H_{re}(e^{\jmath\omega})$ and $H_{im}(e^{\jmath\omega})$ are the real and imaginary parts of $H(e^{\jmath\omega})$ respectively, and

$$\theta(\omega) = arg\{H(e^{\jmath\omega})\}$$

In the definition declared above, the quantity $|H(e^{\jmath\omega})|$ is called *magnitude* response and the quantity $\theta(\omega)$ is known as *Phase* response of the LTI discrete-time system.

It should be noted that the magnitude and phase functions are usually real function of ω, whereas, the frequency response is complex function of ω. As it discussed before, with a real impulse response coefficients $h[n]$, the magnitude is an *even* function of ω, this implies that $|H(e^{\jmath\omega})| = |H(e^{-\jmath\omega})|$, and the phase function is an odd function of ω,i.e., $\theta(\omega) = -\theta(-\omega)$. In the same manner $H_{re}(e^{\jmath\omega})$ is even, and $H_{im}(e^{\jmath\omega})$

is odd function.

A.1 LTI Discrete-Time system in Frequency-Domain

The input-output relation of LTI discrete-time system in the frequency domain can be represented as

$$Y(e^{\jmath\omega}) = \sum_{n=-\infty}^{n=\infty} y[n]e^{-\jmath\omega n} = \sum_{n=-\infty}^{n=\infty} \left(\sum_{k=-\infty}^{k=\infty} h[k]x[n-k] \right) e^{-\jmath\omega n},$$

After interchanging the summation signs and rearranging it follows

$$Y(e^{\jmath\omega}) = \sum_{k=-\infty}^{k=\infty} h[k] \left(\sum_{n=-\infty}^{n=\infty} x[n-k] \right) e^{-\jmath\omega n}$$

It can be further simplified to

$$Y(e^{\jmath\omega}) = \sum_{k=-\infty}^{k=\infty} h[k] \left(\sum_{l=-\infty}^{l=\infty} x[l]e^{-\jmath\omega l} \right) e^{-\jmath\omega k}$$

It is obvious that the quantity inside the parentheses is $X(e^{\jmath\omega})$ which is the DTFT of the input sequence $x[n]$, a final rearrangement yields

$$Y(e^{\jmath\omega}) = \left(\sum_{k=-\infty}^{k=\infty} h[k]e^{-\jmath\omega k} \right) X(e^{\jmath\omega}) = H(e^{\jmath\omega})X(e^{\jmath\omega}) \tag{A.6}$$

In equation A.6 $H(e^{\jmath\omega})$ is the frequency response of the LTI discrete-time system as defined earlier in the equation A.5, therefore equation A.6 relates the input-output of the LTI system in the frequency domain. Further more from this equation one can obtain the frequency response of the LTI discrete-time system as a ratio of the output to the input in the frequency domain.

$$H(e^{\jmath\omega}) = \frac{Y(e^{\jmath\omega})}{X(e^{\jmath\omega})} \tag{A.7}$$

Generalization of the frequency response function $H(e^{\jmath\omega})$ gives rise to the concept of transfer function which is more helpful for the realization of the digital filter. On the other hand, considering the z-transform of the impulse response of an LTI system, known as transfer function, is a polynomial in z^{-1}, for a system with real impulse response, this is a polynomial with real coefficients. Further more, usually the LTI digital filter can be represented using linear difference equation with constant and real coefficients. This will yield a real rational function of variable z^{-1} which is the ratio of two polynomials in z^{-1} with real coefficients that is easier to handle for synthesis.

Therefore, using the development of input-output relation of LTI system, a different form of representation for the transfer function of the system can be derived. One can describe this input-output relationship using the z-transform convolution properties, for the input-output relation $y[n] = x[n] \circledast h[n]$, gives $Y(z) = H(z)X(z)$. This in turn yields

$$H(z) = \frac{Y(z)}{X(z)}. \tag{A.8}$$

The quantity $H(z)$ is z-transform of the impulse response sequence of the system $h[n]$ and is referred to as transfer function. In a particular condition, if the *Region* Of *Convergence* (ROC) of $H(z)$ includes the unit circle, it can represent the frequency response of $H(e^{\jmath\omega})$, in other words, the discrete-time Fourier transform of the impulse response of LTI digital filter, by $H(e^{\jmath\omega}) = H(z)|_{z=e^{\jmath\omega}}$. For a real transfer function $H(z)$ the following expression is true

$$|H(e^{\jmath\omega})|^2 = H(e^{\jmath\omega})H^*(e^{\jmath\omega}) = H(e^{\jmath\omega})H(e^{-\jmath\omega}) = H(z)H(z^{-1})|_{z=e^{\jmath\omega}}$$

The *inverse* z-transform of the transfer function $H(z)$ yields the impulse response $h[n]$. As an alternative, once $H(z)$ is written in the form of the ratio of two polynomials, the partial fraction expansion method can be used to derive $h[n]$.

A.2 Filtering

The discrete-time LTI system can be used as a filter, that is, to pass certain frequency components in an input sequence, while attenuating the other frequencies possibly without introducing any distortion, such a system is called *digital filter*. After filtering, the same input sequence can be recovered using the *inverse* discrete-time Fourier transform in the form of Fourier integral given by

$$x[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{\jmath\omega}) e^{\jmath\omega} d\omega$$

In order to further investigate the concept of filtering, a real coefficient LTI discrete-time system can be characterized by a magnitude function

$$|H(e^{\jmath\omega})| \simeq \begin{cases} 1, & |\omega| \leq \omega_c, \\ 0, & \omega_c < \omega \leq \pi \end{cases} \tag{A.9}$$

Where the ω_c defines the cut off frequency of filter which also known as pass band frequency of the filter, the frequencies between ω_c and π are in the stop band region.

If an input with two sinusoidal tones applied to such a LTI discrete-time filter such as this input $x[n] = \mathrm{A}\ cos\omega_1 n + \mathrm{B}\ cos\omega_2 n$ where $0 < \omega_1 < \omega_c < \omega_2 < \pi$, because of linearity property of the LTI system, it follows that the output of the this filter is given by

$$y[n] \simeq A\,|H(e^{\jmath\omega_1})| \cos(\omega_1 n + \theta(\omega_1)),$$

showing that the LTI discrete-time system performs as a low pass filter, the output single tone with frequency of ω_1 has magnitude dictated by $|H(e^{\jmath\omega_1})|$ and its phase determined by the argument of $H(e^{\jmath\omega_1})$.

A.3 FIR Structure

The LTI discrete-time system can be classified based on its impulse response length or the method by which the input-output relation is determined. If impulse response has a finite length, i.e, $h[n] = 0$ for $n < N_1$ and $n > N_2$ if $N_1 < N_2$ this is known as *finite impulse resonse* (FIR) discrete-time system, and the convolution sum reduces to the following form

$$y[n] = \sum_{k=N_1}^{k=N_2} h[k]x[n-k]. \tag{A.10}$$

Looking at A.10 which is a finite convolution sum, this equation can be used to calculate the $y[n]$ using simple addition and multiplication. The main advantage of the FIR digital filter is its linear phase property. In an special case, with non-causal and symmetrical impulse response, this filter yields zero phase filter, that is the transfer function is completely real with no phase shift. However, due to the non-causal property of zero phase filter, it is non realizable. To make it realizable, one can shift the impulse response to the right (delay) by certain number of samples to make it causal. The symmetry property of FIR filter implies that $h[n] = h[n-N]$ for $0 \leq n \leq N$ considering filter order N even, it follows [47]

$$H(e^{\jmath\omega}) = e^{-\jmath N\omega/2} \left\{ \sum_{n=0}^{N/2} a[n] \cos(\omega n) \right\}$$

Where $a[0] = h[\frac{n}{2}]$, $a[n] = 2h[\frac{N}{2} - n]$, $1 \leq n \leq \frac{N}{2}$ for N even.

Similarly for antisymmetry impulse response with N even, $h[n] = -h[N-n]$ for $0 \leq n \leq N$, implying a constant group delay of $\frac{N}{2}$ samples. The expression for frequency response is

$$H(e^{\jmath\omega}) = e^{-\jmath N\omega/2} e^{\jmath\pi/2} \left\{ \sum_{n=1}^{N/2} c[n] \sin(\omega n) \right\}$$

Where $c[n] = 2h[\frac{N}{2} - n]$, $1 \leq n \leq \frac{N}{2}$ for N even.

The very important symmetry property of the impulse response of the linear phase FIR filter, give rise to another important property of this type of filter which makes the location of its zeros reciprocal, because the transfer function $H(z)$ can be written as

$$H(z) = \sum_{n=0}^{N} h[n]z^{-n} = \sum_{n=0}^{N} h[N-n]z^{-n}, \tag{A.11}$$

Using the symmetry condition $h[n] = h[N-n]$ and changing the variable $m = N - n$, one can write the rightmost expression in equationA.11 as

$$H(z) = \sum_{n=0}^{N} h[m]z^{-N+m} = z^{-N} \sum_{n=0}^{N} h[m]z^{m} = z^{-N}H(z^{-1}), \tag{A.12}$$

From equation A.12 it is inferred that if $z = \xi_0$ is a zero of $H(z)$, so is $z = \frac{1}{\xi_0}$. Further more for an FIR filter with real impulse response, the zeros occur in complex conjugate pairs. This implies that a zero at $z = \xi_0$ is associated with a zero at $z = \xi_0^*$. Thus, a complex zero that is not located on the unit circle, is associated with a set of four zeros determined by $z = re^{\pm \jmath\phi}, z = \frac{1}{r}e^{\pm \jmath\phi}$.

However this is not true for the case of linear phase FIR digital filter with non real coefficients. In such a case, each zero is paired only with its reciprocal in magnitude, and there is no conjugate pair, which yields complex coefficients linear phase FIR filter. The complex coefficient FIR filter is the one that will be useful in the application in hand, in which a non symmetrical impulse response, and frequency response, is desirable in order to compensate the non symmetrical distortions caused by RF filter. The zero locations in z-plane for the real coefficient FIR filter (conjugate pairs) is shown in the Figure A.1.

Further more it is understood that for FIR filter, from equation A.8, the denominator polynomial is one, and FIR filter is defined by only the numerator polynomial

coefficients, implying that all poles located at origin, hence no stability concerns.

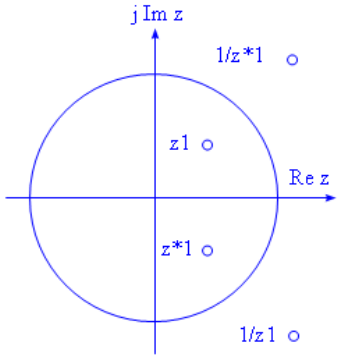

Figure A.1 Location of conjugate zeros and their reciprocals for real coefficient linear phase FIR

A.4 IIR Structure

In the infinite impulse response (IIR) filter, the impulse response length is not finite, this implies that there is feed back from output to input of this structure, as well as the feed forward path from the input to the output as it was for the FIR case. Referring to the equation A.8, the transfer function of the IIR filter can be shown as [41]

$$H(z) = \frac{Y(z)}{X(z)} = \frac{p_0 + p_1 z^{-1} + p_2 z^{-2} + ... + p_M z^{-M}}{d_0 + d_1 z^{-1} + d_2 z^{-2} + ... + d_N z^{-N}} \tag{A.13}$$

This is a rational function in z^{-1} and is the ratio of two polynomials in z^{-1}, multiplying the numerator and denominator by z^M and z^N, respectively, the transfer function in equation A.13 can be defined as a rational function in z

$$H(z) = \frac{Y(z)}{X(z)} = z^{N-M} \frac{p_0 z^M + p_1 z^{M-1} + p_2 z^{M-2} + ... + p_M}{d_0 z^N + d_1 z^{N-1} + d_2 z^{N-2} + ... + d_N} \tag{A.14}$$

Another way to represent the z-transfer function of the LTI discrete-time system, is to factor out the denominator and numerator of equations A.13, it follows

$$H(z) = \frac{p0}{d0} \frac{\prod_{k=1}^{M}(1 - \xi_k z^{-1})}{\prod_{k=1}^{N}(1 - \lambda_k z^{-1})} \tag{A.15}$$

Or for the transfer function in equation A.14, yields

$$H(z) = \frac{p0}{d0} z^{N-M} \frac{\prod_{k=1}^{M}(z - \xi_k)}{\prod_{k=1}^{N}(z - \lambda_k)} \tag{A.16}$$

Where in equations A.15, and A.16, the ξ_1, ξ_2, ..., ξ_M are the finite zeros, and λ_1, λ_2, ..., λ_N are the finite poles of the transfer unction. If $N > M$, (N-M) additional zeros located at origin $z = 0$, and if $N < M$, $(M - N)$ additional poles located at origin $z = 0$.

In an IIR structure the denominator coefficient is not equal to one, hence the filter contains poles and zeros. If the poles and zeros appear in conjugate pair, the

polynomial coefficients will be in the form of real numbers, hence real IIR filter. Otherwise, if either of the pole and zeros appear in non conjugate pairs, the filter coefficients will be in complex form. For the stability concerns, the magnitude of the poles has to be less than one; in other words, the poles have to be located inside the unit circle. This requirement can be discussed in different form, an LTI digital filter is *Bounded* In, *Bounded*Out (BIBO) stable if its impulse response sequence $h[n]$ is absolutely summable, i.e.,

$$S = \sum_{n=-\infty}^{\infty} |h[n]| < \infty \tag{A.17}$$

A.5 Stability of Complex IIR Filter

Since a complex coefficients IIR filter will be designed in this thesis, it is beneficial to extend the stability criterion of the real coefficient IIR to the complex IIR form, also using the proof for the real impulse response, one can establish the proof for the BIBO condition for complex impulse response as follows:

$$|y[n]| = \left| \sum_{k=-\infty}^{\infty} h[k]x[n-k] \right| < \sum_{k=-\infty}^{\infty} |h[k]||x[n-k]|, \tag{A.18}$$

Since the input is bounded, hence $0 \leq |x[n]| \leq B_x$, using equation A.17, it follows $|y[n]| \leq B_x S$, thus $y[n]$ is also bounded.

To prove for the complex coefficient filter, one can use the reverse order from output to input and show that if a bounded output is produced by bounded input, the complex impulse response is bounded too. Consider the following bounded input defined by

$$x[n] = \frac{h^*[-n]}{|h[-n]|}$$

It follows:

$$y[0] = \sum_{k=-\infty}^{\infty} \frac{h^*[k]h[k]}{|h[k]|} = \sum_{k=-\infty}^{\infty} |h[k]| = S \tag{A.19}$$

The equation A.19 shows that for the bounded input to the LTI system with complex coefficient, the output is also bounded, so the system is BIBO if and only if

$$\sum_{k=-\infty}^{\infty} |h[k]| = S \leq \infty$$

Complex Filter implementation

In order for the compensation filter to be realizable, the system function of this filter has to be converted to the real form with real coefficients. To derive a relationship to satisfy this purpose, a single pole transfer function will be analyzed and the results can be extended to any transfer function with any number of poles.

The impulse response of a single pole can be represented by $h[n] = p^n u[n]$, also the transfer function in z-domain can be written as

$$H(z) = \sum_{n=0}^{\infty} p^n z^{-n} = \sum_{n=0}^{\infty} (pz^{-1})^n = \frac{1}{1 - pz^{-1}} \tag{A.20}$$

If nominator and denominator of the transfer function $H(z)$ multiplied by $(1-p^*z^{-1})$, this will yield

$$H(z) = \frac{1}{1 - pz^{-1}} = \frac{1 - p^*z^{-1}}{(1 - pz^{-1})(1 - p^*z^{-1})} = \frac{1 - \Re\{p\} + \jmath\,\Im\{p\}\, z^{-1}}{1 - 2\Re\{p\}\, z^{-1} + |p|^2 z^{-2}} \tag{A.21}$$

Where, p^* denotes the conjugate of the complex pole, $\Re\{p\}$ represents the *real* part of the pole, and $\Im\{p\}$ represents *imaginary* part of the pole, and both are real quantities. Therefore, the system function of a single complex pole, can be decomposed into two real functions that are completely *uncoupled*. If the two uncoupled portions of a single complex pole represented by $H_{re}(z)$ and $H_{im}(z)$, respectively, the following expressions hold

$$H_{re}(z) = \frac{1 - \Re\{p\}\, z^{-1}}{1 - 2\Re\{p\}\, z^{-1} + |p|^2 z^{-2}} \tag{A.22}$$

$$H_{im}(z) = \frac{\Im\{p\}\, z^{-1}}{1 - 2\Re\{p\}\, z^{-1} + |p|^2 z^{-2}} \tag{A.23}$$

One can extend the single complex pole relationship in equations A.22 and A.23 to the case of two complex poles as

$$H(z) = [H_{re1}(z) + \jmath H_{im1}(z)]\,[H_{re2}(z) + \jmath H_{im2}(z)]$$

where

$$H_{re1}(z) = \frac{1 - \Re\{p_1\} z^{-1}}{1 - 2\Re\{p_1\} z^{-1} + |p_1|^2}$$

$$H_{im1}(z) = \frac{\Im\{p_1\} z^{-1}}{1 - 2\Re\{p_1\} z^{-1} + |p_1|^2}$$

$$H_{re2}(z) = \frac{1 - \Re\{p_2\} z^{-1}}{1 - 2\Re\{p_2\} z^{-1} + |p_2|^2}$$

$$H_{im2}(z) = \frac{\Im\{p_2\} z^{-1}}{1 - 2\Re\{p_2\} z^{-1} + |p_2|^2}$$

After some manipulation, it follows

$$H(z) = [H_{re1}(z)H_{re2}(z) - H_{im1}(z)H_{im2}(z)] + \imath[H_{re1}(z)H_{im2}(z) - H_{im1}(z)H_{re2}(z)] \tag{A.24}$$

The expression in Equation (A.24) shows a relationship by which the transfer function for two complex poles can be defined in terms of the uncoupled real filters with symmetrical coefficients, the block diagram of this representation is shown in Figure A.2. The expression in Equation (A.24) can be extended to a general form of complex filter with multiple poles to derive the transfer function expression for any number of poles as a composition of two real functions for the real and imaginary parts respectively. This means that the complex filter comprising of complex poles is precisely realizable utilizing individual real filters with real coefficients.

Similar reasoning can be used for the complex FIR case for a single complex zero which is trivial, then extended to any number of zeros. Thus, any complex recursive and nonrecursive filter can be realized using two separate real filters with real coefficients. Therefore, in general, the following expression for the coefficients of the complex filter holds

$$C_{complex} = C_{real} + \jmath\, C_{imaginary}$$

Knowing that the complex coefficient FIR filter is comprised from two individual *real* coefficient FIR filters that are completely independent, and referred to as *uncoupled*

FIR filters, it is conceivable that each of these real FIR filters can be implemented individually, using the techniques mentioned previously. 5

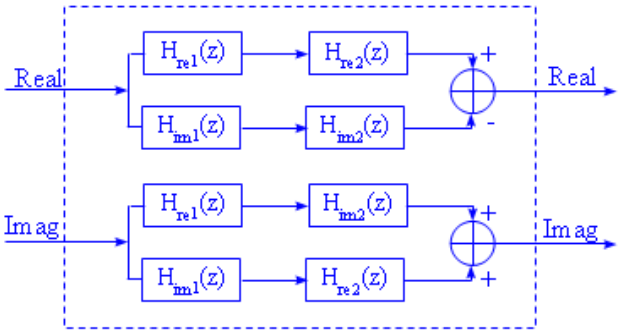

Figure A.2 Overall transfer function of Two complex poles represented by individual real and imaginary parts of each complex pole

B. Optimization Algorithms

It was stated before that the optimization method is a viable option in order to design the compensation filters. In this section a detailed description for the design of non-recursive and recursive filters will be presented. The pivoting topic in this section is the idea of *optimization* algorithm. Both the non-recursive and recursive filters to be design in this section, has two common attributes, both may have a complex coefficient, and they both use the notion of *reciprocal* zeros and poles.

In section 2.4 it was explained that the band pass distortions created by RF filter, once translated into the base band, will manifest themselves as a base band complex distortion effect, on the QAM signal. This complex nature of the distortions in base band gives rise to the fact that the compensation filters are likely low pass filter having complex response, and complex coefficients, hence complex zeros and poles.

For this reason each pole and zero in the positive and negative portion of the z-plane has to be dealt with individually. For instance, during an optimization algorithm, in the case of non-recursive FIR filter, the location of zeros in the positive and negative portion of z-plane will be adjusted individually in an iterative manner.

It is obvious that in the case of complex filter, both the positive and negative sides of the frequency response of the low pass complex filter are important, and has to be taken into account during the optimization process. It is conceivable that the main task of an optimization algorithm in this application is to adjust the position of zeros and poles until the desirable response obtained, and the difference between the frequency response of the filter under optimization become close to that of the specified response.

Furthermore, the minimum number of the required zeros or poles has to be determined adaptively. For the case of FIR filter, the important question is what would be the initial layout for the zeros in the z-plane, to achieve the desired response as quickly as possible. Since the amplitude response of the RF filter, after being translated into the base band, are more or less close to a symmetrical response with respect to the center of the filter pass band, a zero configuration that yields an ideal symmetrical amplitude response would be a good starting point. Then adjusting the magnitude/phase of zeros to achieve the desired response. The similar technique can be used for the IIR filter.

Adjusting the location of the poles and zeros in an adaptive manner, can be done using the optimization method. Two optimization methods *Grid search* and *Gradient* algorithms, will be presented and their performance will be compared. In this context, some fundamental concepts and parameters needs to be defined.

B.1 Background

This section is presented to give an overall background about optimization technique, One can refer to the text books for optimization [48] [36] [49] to acquire more insight into this topic. Most of the topics presented here are inspired from these references.

B.2 Objective function

The optimization algorithm seeks to minimize or maximize an *objective* or *cost* function. This function is usually in the form of an error function as a difference of a specified response and the problem in hand.

B.3 Convex set and function

Convex sets are used in the formulation of optimization problems [48]. A convex set defines a vector space such that all points between the two end points located on

a line are also included in this vector space. Its definition is:

$$x = \lambda a + (1 - \lambda)b, \quad 0 \leq \lambda \leq 1$$

This expression defines a convex set S that any point between a and b, is also a member of the set S including points a and b.

The convex function is important in that, usually the objective or cost function(or part of it) is in the form of *Convex* or *Concave* function, and the optimization algorithm is trying to find the minimum or maximum point of this function.

B.4 Mean Value Theorem

If f is a differentiable function in the closed interval $[a, b]$, there exist a point (c) in $[a, b]$ such that

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

The mean value theorem implies that the slope of $f(x)$ at point c has the same slope as the line connecting between points a and b. This theorem also implies that if point c is found, then no other derivative higher than $f'(x)$ would be required to evaluate $f(b)$, this concept can be extended to the higher order of derivatives assuming the $f(x)$ itself is the n^{th} derivative of another function. Then c can be defined as the convex combination of a and b as follows:

$$c = \lambda a + (1 - \lambda)b, \quad 0 \leq \lambda \leq 1$$

B.5 Minimum of Convex Function

The necessary condition for the minimum of a convex function f at point x_0 is $f'(x_0) = 0$. Moreover, the sufficient condition for this point, follows from approximating the function using the truncated *Taylor* series

$$f(x_0+h)-f(x_0)=\frac{h^2}{2}f''(\lambda x_0+(1-\lambda)(x_0+h))>0 \quad 0\leq\lambda\leq 1$$

Since h^2 is always positive, also from the continuity of $f''(x_0)$, if $f''(\lambda x_0+(1-\lambda)(x_0+h))>0$ it follows that $f''(x_0)>0$ meaning the point x_0 is a minimum point. This conclusion was made assuming the $f(x)$ and its first n derivatives are continuous, then $f(x)$ has a relative maximum or minimum if and only if n is even, where n is the order of first non-vanishing derivatives at x_0.

A generalization of the above formulation can be considered in an n-dimensional case, where $\bar{x}=(x_1,x_2,...,x_n)$ represents a point in an *Euclidian* space R^n, in this context $f(\bar{x})$ will represent $f(x_1,x_2,...,x_n)$ and $f(\bar{x})$ is convex over convex set $\mathbf{X}$ in $\mathbf{R}^n$ if for any two points $\mathbf{x}_1$, $\mathbf{x}_2$ in $\mathbf{X}$ and for all λ, $0\leq\lambda\leq 1$, there exists

$$f[\lambda\bar{x}_1+(1-\lambda)\bar{x}_2]\leq\lambda f(\bar{x}_1)+(1-\lambda)f(\bar{x}_1)$$

Note that the sum of convex functions is also a convex function. Using the truncated *Taylor* series and extend it over n-dimensional case, if $f(\bar{x})$ is continuous and has continuous first and second order partial derivatives over an open convex set $\mathbf{X}$ in $\mathbf{R}^n$, then for any two points $\bar{x}_1$, and $\bar{x}_2=\bar{x}_1+h$ in $\mathbf{X}$, there exist a λ, $0\leq\lambda\leq 1$, such that for a quadratic function yields

$$f(\bar{x}_2)\simeq f(\bar{x}_1)+\bar{\nabla}^T f(\bar{x}_1)h+\frac{1}{2}h^T H[\lambda(\bar{x}_1)+(1-\lambda)(\bar{x}_2)]h \tag{B.1}$$

In equation B.1, $\bar{\nabla}f=(\frac{\partial f}{\partial x_1},\frac{\partial f}{\partial x_2},...,\frac{\partial f}{\partial x_n})$ and H is *Hessian* matrix of $f(\bar{x})$, that is, the square matrix of the second partial derivatives of $f(\bar{x})$ evaluated at $\bar{x}_1$.

$$H=\begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1\partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1\partial x_n} \\ \frac{\partial^2 f}{\partial x_2\partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 f}{\partial x_n\partial x_1} & \frac{\partial^2 f}{\partial x_n\partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \tag{B.2}$$

In the case of multi variable function, for minimum at $\bar{x} = \bar{x}_0$, there is $\frac{\partial f(\bar{x}_0)}{\partial x_i} = 0$ where $i = 1, 2, ..., n$

Assuming the existence and continuity of the second partial derivatives, $\frac{\partial^2 f}{\partial x_i \partial x_j}$ will have the same sign of $\frac{\partial^2 f}{\partial x_i \partial x_j}[\lambda(\bar{x}_0) + (1-\lambda)(\bar{x}_0 + h)]$, therefore, if $h^T H[\bar{x}_0]h$ is negative, $f(\bar{x}_0 + h) - f(\bar{x}_0)$ will also be negative, which means the point $\bar{x}_0$ is minimum in this multi variable convex function.

A quadratic form $h^T H[\bar{x}_0]h$ is negative if and only if the Hessian is a negative definite matrix. This criterion will be further used during the steepest descent optimization algorithm for FIR filter design in section 3.3. For example, one method for the negative definite test of the Hessian matrix, entails the calculation of the *eigenvalues* of the Hessian matrix [50] Once they are all positive, Hessian is said to be positive definite, hence pointing to the minimum point in the convex function.

B.6 Multivariate Grid search

In multivariate grid search optimization method, there is a multi dimensional convex objective function that could have several minimum or maximum points, among them, one point is a *global* maximum or minimum, and the rest are *local* maximum or minimums.

The goal is to minimize the convex objective function such that the difference between the current value of the problem in hand and the specified value become minimum, hence seeking possibly the global minimum point of the objective function.

In this method, the region of the variation of variables will be divided into a grid structure or mesh. The objective function is to be evaluated at each node of this grid. By moving from one node to the next point, the value of function can be increased or decreased. Although this may not be an efficient way of optimization, however it is simple and straight forward. The inefficiency grows specially when there are several independent variables in the function. The algorithm can be defined as:

1. Divide the whole region into a grid with grid size of $\triangle x_i$, $i = 1, 2, ..., n$ in each variable x_i in the n-dimensional region. And for $a_i \leq x_i \leq b_i$ over which we optimize the function $f(\bar{x})$

2. Choose a starting point on the grid

3. Evaluate function $f(\bar{x})$ at $3^n - 1$ neighboring points.

4. Select the point for which the function $f(\bar{x})$ has minimum value(for maximization is reverse)

5. Repeat steps three and four until the central point become the minimum of $f(\bar{x})$

6. For better accuracy after some number of iterations, reduce the grid size, i.e., by half until the error becomes less than the pre-defined tolerance.

A slightly different variation of multivariate search is univariate search which reduces the computation burden, in that only one variable at a time will be changed and the other variables remain fixed. Starting at some arbitrary point, one variable will be changed until a maximum of $f(\bar{x})$ in that direction is reached, then switch to different variable to find the maximum of $f(\bar{x})$ in the other direction, and repeat this process on each of the n coordinates of the function. The algorithm is like this:

1. start at some point $\bar{x}_0$ within a reasonable interval

2. Find the maximum of $f(\bar{x})$ in one direction using a line search, until reach next point $\bar{x}_1$, i.e., $\bar{x}_1 = \bar{x}_0 + \lambda_1 \bar{a}_1$ Where $\bar{a}_1 = [1, 0, ..., 0]^T$ and λ_1 is an scalar such that $f(\bar{x}_0 + \lambda_1 \bar{a}_1)$ is minimized.

3. The step corresponding to the k^{th} variable is to find the next point $\bar{x}_k$ performing the maximization with respect to the first variable, i.e., $\bar{x}_k = \bar{x}_{k-1} + \lambda_k \bar{a}_k$ such that $f(\bar{x}_{k-1} + \lambda_k \bar{a}_k)$ is maximized.

4. Find the n^{th} point by maximizing the function with respect to the n^{th} variable.

5. repeat steps 2,3 and 4 until the $|\lambda_k|$ is less than tolerance.

B.7 Gradient Steepest Descent

The previous method grid/univariate grid search may not be as efficient as required, so an algorithm with much faster convergence rate is desirable. One good candidate that widely used in the literature such as *Himmelblau* [36], would be the gradient method using *directional derivatives* and calculating the steepest descent. This unconstrained optimization method, can be traced back to the popular mathematician *Cauchy*. When dealing with multidimensional functions it is important to know the direction of the maximum rate of change of the function. This can be obtained using the concept of directional derivatives. The directional derivative of a function $f(\bar{x})$ at $\bar{x}_0$ in the direction $\bar{u}$ is defined by

$$D_{\bar{u}} f(\bar{x}_0) = \lim_{\lambda \to 0} \frac{f(\bar{x}_0 + \lambda \bar{u}) - f(\bar{x}_0)}{\lambda}$$

The derivative of $f(\bar{x})$ with respect to the direction $\bar{u}$ can be written in terms of partial derivatives as

$$D_{\bar{u}} f(\bar{x}_0) = \bar{\nabla} f(\bar{x}_0) \bar{u}, \quad |\bar{u}| = 1$$

$$D_{\bar{u}} f(\bar{x}_0) = \sum_{j=1}^{n} \frac{\partial f(\bar{x}_0)}{\partial x_j} \bar{u}_j, \quad |\bar{u}| = 1$$

To find the direction $\bar{u}$ by which the rate of change of $f(\bar{x})$ at point $\bar{x}_0$) ia maximum, one has to maximize

$$\sum_{j=1}^{n} \frac{\partial f(\bar{x}_0)}{\partial x_j} \bar{u}_j$$

This is a *constrained* maximization subject to $g(\bar{u}) = \sum_{j=1}^{n} u_j^2 = 1$, stating the

Lagrangian function for this constraint maximization as

$$F(\bar{u},\lambda) = \sum_{j=1}^{n} \frac{\partial f(\bar{x}_0)}{\partial x_j}\bar{u}_j + \lambda \quad 1 - \sum_{j=1}^{n} u_j^2 \Bigg) \tag{B.3}$$

To maximize the equation B.3, one can take the derivative of this function with respect to u_j and λ and equate it to zero, after some manipulations it follows

$$\bar{u} = \frac{\pm\bar{\nabla} f(\bar{x}_0)}{|\bar{\nabla} f(\bar{x}_0)|}$$

which gives the direction of maximum increase/decrease of the function. Plus sign gives the direction of the maximum increase and minus sign gives maximum decrease. Now that the direction of the maximum rate of changes is known, in other word, the gradient vector in the direction of greatest local increase/decrease is determined. One can proceed in the direction of steepest descent at point $\bar{x}^k$ where the direction of decrease is:

$$\hat{\bar{s}}^k = \frac{\bar{\nabla} f(\bar{x}^k)}{||\bar{\nabla} f(\bar{x}^k)||}$$

Where

1. $\bar{s}^k$ is a vector in the direction of steepest descent.
2. $\hat{\bar{s}}^k$ is a unit vector in the direction of steepest descent.
3. $\bar{\nabla} f(\bar{x}^k)$ is the gradient vector of $f(\bar{x})$ at $\bar{x}^k$

Therefore, in the steepest descent algorithm the transition from one point to the next point, i.e., from $\bar{x}^k$ to $\bar{x}^{k+1}$ is defined by

$$\bar{x}^{k+1} = \bar{x}^k + \lambda^k \frac{\bar{\nabla} f(\bar{x}^k)}{||\bar{\nabla} f(\bar{x}^k)||} \tag{B.4}$$

In the equation B.4 it is beneficial to know what would be the best value for the optimum λ^k in order to get to the minimum point with the least number of iterations.

To answer to this question, one can take the derivative of $f(\bar{x}^k)$ with respect to λ and from the solution of

$$\frac{df(\bar{x}^k + \lambda\hat{\bar{s}}^k)}{d\lambda} = 0$$

To be more specific, suppose that $f(\bar{x})$ is a quadratic function, substituting $(\hat{\bar{s}}^k)$ in equation B.1 for the h, then λ can be expressed as:

$$\lambda^k = -\frac{[\bar{\nabla}^T f(\bar{x}^k)](\hat{\bar{s}}^k)}{(\hat{\bar{s}}^k)^T H(\hat{\bar{s}}^k)}$$

Now that the step size λ is known, the equation B.4 represents the solution to the optimization problem using the gradient steepest descent method. This method with some variation, will be used during the design of non-recursive, and recursive compensation digital filters in sections 3.4.3 and 3.5.

C. Modulation Error Ratio

The modulation error rate (MER) is an alternative definition for the signal to noise ratio(SNR) in the digital complex base band sense and can be used interchangeably. The MER is a direct measure of the quality of the modulated signal and it defines how well each constellation point of the modulated signal presents an ideal constellation point.

For example, consider a 16 QAM modulated signal in Figure C.1, the constellation points indicated by a cross and vector in solid line, are the ideal constellation points, and those indicated in dotted line are related to the modulated constellation points. The difference between the modulated vectorsV_m and ideal vectorsV_i represent the MER error for each point n as follow

$$ModulationError(n) = V_m(n) - V_i(n) \tag{C.1}$$

If all these landing points for each symbol are plotted over time, there will be a cloud of symbols around each ideal constellation point as indicated in Figure C.1, then the modulation error ratio is the ratio of average symbol power to average error power

$$MER(dB) = 10log_{10}(\frac{average\ symbol\ power}{average\ error\ power}) \tag{C.2}$$

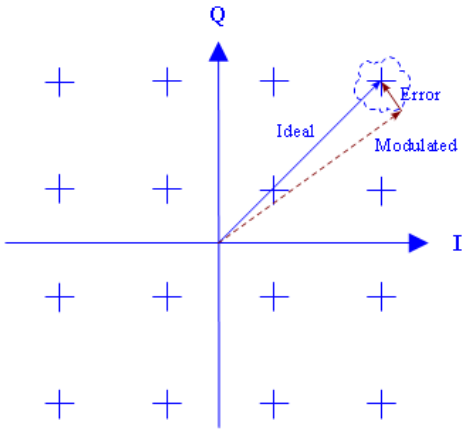

Figure C.1 The ideal and modulated constellation points, along with the modulation error vector

A more precise mathematical form of MER can be shown as

$$MER(dB) = 10log_{10}\left|\frac{\sum_{j=1}^{N}(I_j^2 + Q_j^2)}{\sum_{j=1}^{N}(\delta I_j^2 + \delta Q_j^2)}\right| \tag{C.3}$$

where I and Q represent the ideal constellation points, and δI and δQ represent in-phase and quadrature parts of the modulation error vector. In this equation it is assumed that a large number of symbols with equal probability of occurrence is used.

D. Effect of distortions on bit Error Rate

In order to characterize the performance degradation of a typical CATV digital communication system, it is necessary to specify a system performance measure. One of the most commonly used performance measures in digital communications system is bit error rate (BER). BER is a suitable parameter in order to characterize the degradations caused by RF filer distortions, in that, these distortions cause symbol dispersion and ISI error. Depending on the modulation scheme, this distortion can result in significant *Probability* of *error*, P_e, versus S/N performance degradation, hence BER performance degradation of the system versus amplitude and group delay distortion.

Here some background theory for the calculation of the probability of error is presented. In general, the modulated QAM signal can be presented by [9]:

$$s(t) = \Re\left[\sum (I_n + jQ_n)\, g(t - nT_s)\, e^{j2\pi f_o t}\right]$$

where Re [ ] denotes real part

1. f_o is the carrier frequency
2. $1/T_s$ is the symbol rate; for 256QAM, $T_s = 8T_b$
3. $1/T_b$ is the bit rate
4. $g(t)$ is a pulse defined by

And $g(x)$ is the pulse shaping function defined as:

$$g(x) = \begin{cases} 1, & \text{for } 0 \leq t \leq T_s \\ 0, & \text{for elsewhere} \end{cases} \tag{D.1}$$

I_n and $Q_n = \pm 1, \pm 3, ... \pm 15$ are the sampled values of in-phase and quadrature components. For a particular white Gaussian noise power at the detector input, the error probability of the I^{th} symbol for the in-phase channel is given by:

$$Ps_i^I = \begin{cases} \frac{1}{2} erfc(\frac{|\bar{S}_i - Th1_i|}{\sqrt{2}\sigma}), & \text{if } I_i = \pm 15 \\ \frac{1}{2} erfc(\frac{|\bar{S}_i - Th1_i|}{\sqrt{2}\sigma}) + \frac{1}{2} erfc(\frac{|Th2_i - \bar{S}_i|}{\sqrt{2}\sigma}), & \text{if } I_i = \pm 1, ..., \pm 13 \end{cases} \tag{D.2}$$

where I_i is the in-phase channel transmitted at I_{th} sample, $\bar{S}_i$ is the magnitude of I_th received in-phase channel sample,

$$Th1_i = |I_i| - 1$$

$$Th2_i = |I_i| + 1$$

and

$$erfc(x) \triangleq \frac{2}{\sqrt{\pi}} \int_x^{\infty} \quad e^{-t^2} dt$$

And obviously σ^2 is the received white Gaussian noise power at the input of the threshold detector. The similar equations can be used for the quadrature component Qs_i^I from equation (D.2). In each I and Q channel, the average symbol error rate P_s over a series of N symbols is as follows:

$$P_s = \frac{1}{N} \sum \frac{1}{2} (Ps_i^I + Ps_i^Q)$$

E. Minmax and Remez Algorithms

E.0.1 Minmax Algorithm Review

A widely used method for finding the coefficients of an FIR filter is known as the *weighted Chebyshev* method. In this algorithm, an error function is formulated for the desired filter as a linear combination of cosine functions, then the error function is minimized by using an efficient multivariable optimization algorithm called the *Remez exchange algorithm*. Once the convergence is achieved in the optimization process, the error function becomes equiripple as in other types of Chebyshev solutions. The amplitude of the error function in various frequency bands in the pass band is controlled by applying weighting to the error function.

The weighted-Chebyshev method is very flexible and can be used to design different type of filters, such as differentiators, band pass, low pass, and filters with arbitrary amplitude response. Furthermore, this method is considered computationally heavy, in that, it requires a large amount of calculations, however, as the cost of computation is becoming cheaper, this disadvantage is not serious.

The fundamental concept of the weighted Chebyshev method was published by Herrmann [12]. Improvements were made by Hofstetter, Oppenheim [13] and later on by Parks, McClellan, Rabiner, and others [14–19]. The latter work led to the popular computer program introduced by McClellan-Parks-Rabiner [20]. Finally the work was further promoted by Remez, known as Remez Exchange Algorithm [21]. Here a brief review of basic concept of this algorithm is presented.

A nonrecursive filter can be characterized by its system function

$$H(z) = \sum_{n=0}^{N-1} h[n] z^{-n}$$

where h(t) is impulse response sampled at nT, T is the time between samples. If N is odd, the impulse response is symmetrical, and $\omega_s = 2\pi$. Because $T = \frac{2\pi}{\omega_s} = 1 second$, The frequency response can be stated as

$$H(e^{\jmath\omega}) = e^{-\jmath\omega c} P_c(\omega)$$

where

$$P_c(\omega) = \sum_{k=0}^{c} a_k \cos k\omega \tag{E.1}$$

and $a_0 = h(c)$, $c = \frac{N-1}{2}$, and $a_k = 2h[c-k]$ for $k = 0, 1, 2, ..., c$. The important assumption in this formulation is that the zeros of the filter is presumed to be in conjugate pairs which yields the real polynomial, hence real filter coefficients.

If a desired frequency response is denoted by $e^{-\jmath c\omega} D(\omega)$ and a weighting function denoted by $W(\omega)$, then an error function $E(\omega)$ can be defined as

$$E(\omega) = W(\omega)[D(\omega) - P_c(\omega)] \tag{E.2}$$

It is conceivable that the error function $E(\omega)$ could be minimized such that $|E(\omega)| \leq \delta_p$, with respect to some compact sub bands in the frequency interval $[0, \pi]$, so called Ω, a filter is obtainable in which

$$|E(\omega)| = [D(\omega) - P_c(\omega)] \leq \frac{\delta_p}{|W(\omega)|} \quad for\ \omega \in \Omega \tag{E.3}$$

In the case of the low pass filter with its amplitude response shown in Figure E.1, with the pass band and stop band ripples δ_p, δ_s and pass band and stop band frequency

edges ω_p, ω_s, therefore, the requirement is

$$D(\omega) = \begin{cases} 1, & 0 \leq \omega \leq \omega_p, \\ 0, & \omega_s < \omega \leq \pi \end{cases} \tag{E.4}$$

with

$$E_0(\omega) = \begin{cases} \delta_p, & 0 \leq \omega \leq \omega_p, \\ \delta_s, & \omega_s < \omega \leq \pi \end{cases} \tag{E.5}$$

Thus, from equation E.3 and E.5, the weighting function will be obtained as

$$W(\omega) = \begin{cases} 1, & 0 \leq \omega \leq \omega_p, \\ \delta_p/\delta_s, & \omega_s < \omega \leq \pi \end{cases} \tag{E.6}$$

The design of such a filter using the optimization technique, entails the solution of minmax problem

$$\min_{\mathbf{x}} \{ \max_{\omega} |E(\omega)| \} \tag{E.7}$$

Where $\mathbf{x} = [a_0 \; a_1 \; ... \; a_c]^T$. The solution of minmax problem stated in equation E.7 can be achieved by utilizing the *alternation* theorem [51]. This theory states that if $P_c(\omega)$ is a linear combination of $r = c + 1$ cosine functions of the form

$$P_c(\omega) = \sum_{k=0}^{c} a_k \cos(k\omega)$$

Then a necessary and sufficient condition that $P_c(\omega)$ be unique, also best weighted Chebyshev approximation to the continuous function $D(\omega)$ on Ω, where Ω is a compact subset of the frequency interval $[0, \pi]$, is that the weighted error function $E(\omega)$ exhibits at least $r+1$ extremal frequencies in Ω, that is, there must exists at least $r+1$ points $\hat{\omega}_i$ in Ω such that $\hat{\omega}_0 < \hat{\omega}_1 < ... < \hat{\omega}_r$ and $E(\hat{\omega}_i) = -E(\hat{\omega}_{i+1})$ for $i = 0, 1, ..., r-1$ and

$$|E(\hat{\omega}_i)| = \max_{\omega \in \Omega} |E(\omega)|$$

referring to the alternation theorem and equation E.2, one can conclude the expression

$$E(\hat{\omega}_i) = W(\hat{\omega}_i)[D(\hat{\omega}_i) - P_c(\hat{\omega}_i)] = (-1)^i\delta \tag{E.8}$$

Where, $i = 0, 1, ..., r$ and δ is constant. This system of equations can be presented in matrix form

$$\begin{pmatrix} 1 & \cos(\hat{\omega}_o) & \cos(2\hat{\omega}_o) & \dots & \cos(c\hat{\omega}_o) & \frac{1}{W(\hat{\omega}_o)} \\ 1 & \cos(\hat{\omega}_1) & \cos(2\hat{\omega}_1) & \dots & \cos(c\hat{\omega}_1) & \frac{-1}{W(\hat{\omega}_1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \cos(\hat{\omega}_r) & \cos(2\hat{\omega}_r) & \dots & \cos(c\hat{\omega}_r) & \frac{-1^r}{W(\hat{\omega}_r)} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_c \\ \delta \end{pmatrix} = \begin{pmatrix} D(\hat{\omega}_0) \\ D(\hat{\omega}_1) \\ \vdots \\ D(\hat{\omega}_r) \end{pmatrix} \tag{E.9}$$

If the extremal frequencies are known, as well as the coefficients a_k, hence the frequency response of the filter using equation E.1 can be calculated. The $(r+1)(r+1)$ matrix is nonsingular [51], therefore, this system of equation always has a solution.

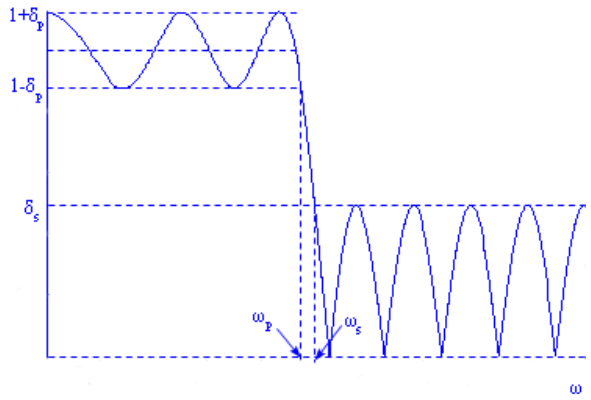

Figure E.1 Amplitude response of an equiripple lowpass filter

E.0.2 Remez Exchange Algorithm Review

Starting with the minmax problem formulation, the *Remez* exchange algorithm is an optimization multivariate algorithm which is suited for the solution of the minmax problem stated in equation E.7. This algorithm is based on the second optimization algorithm proposed by *Remez* [21], the comprehensive and detailed description of the algorithm is not declared here, rather an overall review will be given which includes the following steps:

1. Initialize extremals $\hat{\omega}_0, \hat{\omega}_1, ..., \hat{\omega}_r$ ensuring an extremal i assigned at each band edge

2. Locate the frequencies $\widehat{\omega}_0, \widehat{\omega}_1, ..., \widehat{\omega}_r$ at which $|E(\omega)|$ is maximum and $|E(\widehat{\omega})| \geq \delta$. These frequencies are potential extremals for the next iteration.

3. Compute the convergence factor

$$Q = \frac{\max|E(\widehat{\omega}_i)| - \min|E(\widehat{\omega}_i)|}{\max|E(\widehat{\omega}_i)|}$$

   where, $i = 0, 1, ..., \rho$

4. Reject $\rho - r$ superfluous extremals $\widehat{\omega}_i$ after finding an appropriate rejection criterion and renumber the remaining $\widehat{\omega}_i$ sequentially, then $\hat{\omega}_i = \widehat{\omega}_i$ for $i = 0, 1, ..., r$.

5. If $Q > \epsilon$, where ϵ is the convergence tolerance repeat from step 2, otherwise continue to step 6.

6. Calculate $P_c(\omega)$ using the last set of extremals, then deduce $h[n]$, which is the impulse response of the required filter.

F. Equalization Theory

F.1 Introduction

The aim of this section is to give some background on equalization theory. To this end, the topic relevant to this research are covered. More sophisticated treatment can be found in literature Haykin [30], Farhang [43], Sayed [31], Proakis [44].

F.2 Optimum Filtering

A filter is linear if the output sequence of this filter can be expressed as the linear combination of the input sequence. For the solution of linear filtering problem, a statistical approach is used with the aid of a statistical metrics such as *mean* or *correlation* of the noisy input sequence to minimize the effect of ISI/noise at the output of the filter [30].

One approach to this optimization process for the design of linear optimum filtering can be accomplished by minimizing *mean square error* signal defined by the difference between some *desired* response and the actual filter output. For a more straightforward solution, some assumptions can be made, i.e., if the input sequence is presumed to be statistically *stationary*[1] [45], then the solution is known as *optimum* in *mean square error sense*, the resulting filter designed by this approach is referred to as Winner filter. The error signal is a multi-dimensional function of the filter coefficients, if plotted against the filter coefficients, yields an error *performance* surface

[1] An stochastic process is said to be stationary if its statistics do not depend on the time origin. In addition, if its autocorrelation function only depends on the time difference (τ), it is Wide Sense Stationary(WSS) as well.

in the form of a *convex* function which its minimum represents the Winner solution.

In this section a brief explanation on the design of optimum filter will be given, this description involves some intuitive observations from the vector space point of view, followed by some more rigorous treatment of the problem. Refereing to the Figure F.1, the input sequence $x(n)$ is applied to the filter input with the tap coefficients $\omega(0), \omega(1), ..., \omega(N)$, the resulting output $y(n)$ is subtracted from the desired response $d(n)$ to generate the error signal $e(n)$.

The goal for the optimum filter design is to obtain the filter weights $\omega(0), \omega(1), ..., \omega(N)$ based on the statistical properties embedded in the input sequence and the desired response, such that the error signal $e(n)$ be minimized in the mean square sense. The input sequence $x(n)$, the desired signal $d(n)$, and the output sequence $y(n)$ are stationary with zero *mean*. This implies that the error signal is also a sequence of zeros mean random variables. This allows the mean power of the error signal to be used as the performance measure. That is to say the filter coefficients that minimize the power in the error yields an output $y(n)$ that approaches $d(n)$.

The reason for this approach is that in real applications, the transmitted signal is corrupted by noise or the channel distortions, causing ISI error on the signal. The linear optimum filter is in fact a feed forward structure[2]that can be used to replicate the inverse of the channel response (in absence of noise), compensating the destructive effect of the channel on the signal. Using the original signal as desired signal during the optimum filter design, the output of the filter will approach to the desired signal minimizing ISI and noise.

Another way to look at the optimum linear filter design process, is from a geometrical point of view and vector space analysis. From this stand point, during the optimum filter design, the desired sequence vector $d(n)$ will be expressed based

[2]The transmission channel can be thought of as an FIR structure, and the equalizer is to generate the inverse of channel response($1/H(z)$) using a Feed Forward structure, this entails using an FIR with infinite length. This, however is not realizable, hence a certain tap number will be used to yield acceptable accuracy.

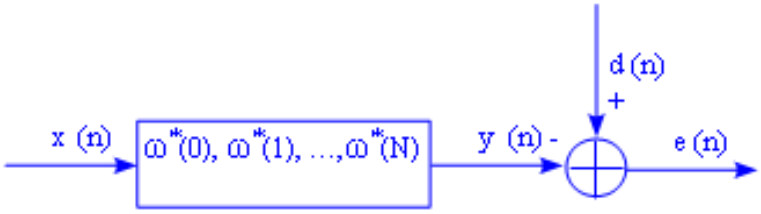

Figure F.1 The block diagram for the optimum linear filter design

on the input random variable vector $x(n)$ as a linear combination of its components $x(1), x(1), ..., x(N)$. In other word, using the random variables defined by vector $x(n)$, another random variable $\hat{d}(n)$ will be estimated, such that the difference between the desired vector $d(n)$ and estimated vector $\hat{d}(n)$ has a minimum norm square. Geometrically this means that the vector $d(n)$ will be projected on the multidimensional coordinates defined by $x(1), x(1), ..., x(N)$, the minimum error will be obtained, provided, the projection of vector $d(n)$ on these coordinates is *orthogonal*. In other word, if this vector projects orthogonally on the vector space $x(n)$ the resulting error will have minimum norm square. Furthermore, the error itself also is orthogonal to the vector space defined by $x(n)$, meaning that the correlation between $e(n)$ and $x(n)$ is zero. This concept will be elaborated using mathematical expressions which follows.

The input sequence $x(n)$ is a WSS random *process*, hence the output sequence $y(n)$ is also a WSS. The input sequence $x(n)$ and the desired response $d(n)$ are *jointly* stationary process in WSS sense, that is, the second moment of $x(n)$ and $d(n)$ with lag k , $E[x(n)d(n-k)]$ only depends on the lag k [45] and *independent* of sample (n). The autocorrelation function of input sequence is defined by Equation (F.1), and the *cross correlation* of desired response $d(n)$ and input $x(n)$ is defined by Equation (F.2).

$$R_{xx} = E[x(n)x^*(n)] \tag{F.1}$$

$$P = E[d(n)x(n-k)] \tag{F.2}$$

For the zero mean random variable $e(n)$, the power is given by the square of the random variable over observation of N samples.

$$\begin{aligned}\epsilon^2 = E[|e(n)|^2] = E[|(d(n) - y(n)|^2] \\ = E[\{d(n) - \omega^*(n)x(n)\}\{d(n) - \omega^*(n)x(n)\}^*]\end{aligned} \tag{F.3}$$

In Equation (F.3), using the definitions from Equations (F.1), and (F.2), and after some manipulation it follows

$$\epsilon^2 = E[|e(n)|^2] = \sigma_d^2 - \omega^H p - p^H \omega - \omega^H R\,\omega \tag{F.4}$$

In Equation (F.4) parameter σ_d^2 denotes the power of desired signal, ω denotes a row vector containing the filter tap weights, and the autocorrelation and cross correlation functions are denoted by R and P respectively. In minimizing the mean error power ϵ^2 some properties of the error function parameters become helpful. The matrix R is *Hermitian* and *positive definite* (full rank process) and P is a column vector, the product of $\omega^H p$ is first order term. The product of $\omega^H R\omega$ is a squared term in the form of *quadratic* function. This suggests that the error function is a multi dimensional function of the filter tap weights presented by vector ω, and if plotted against the tap weights, it yields an error performance function with a *convex* shape with its minimum representing the optimal tap weights ω_{opt}. Therefore, the minimum of this function can be found by calculating the derivative of this function with respect to the vector ω and equate it to zero. Since all the vectors in the error function are complex, a multidimensional derivative versus both real and imaginary parts is required

$$\nabla_\omega \epsilon^2 = -2P + 2R\,\omega \tag{F.5}$$

where $\nabla_\omega = [\frac{\partial}{\partial\omega_1}\ \frac{\partial}{\partial\omega_2}\ ...\ \frac{\partial}{\partial\omega_N}]^T$, by equating the derivative of error function to zero it follows

$$\omega_{opt} = R^{-1}P \tag{F.6}$$

The expression in Equation (F.6) represent the optimum filter coefficients based on

the autocorrelation and cross correlation matrices. If these two matrices are known the optimum Winner filter can be designed. However, often these two parameters are not known, the matrix R, the cross correlation of input sequence $x(n)$ is not known and depends on the channel impulse response which is also not known. The parameter P representing the cross correlation of the input $x(n)$ and desired sequence $d(n)$, in the case of pilot training $d(n)$ is known, but again $x(n)$ is not known due to the lack of information of the transmission channel. Thus, the filter coefficients calculation can be accomplished using different approach called *adaptive filtering*. Furthermore, for the optimum filter coefficient ω_{opt}, the error function $e(n)$ can be expressed as [43]

$$e(n) = d(n) - y(n) = d(n) - \omega^*(n)x(n) = d(n) - P^* R^{-1}\, x(n) \tag{F.7}$$

To prove the error vector $e(n)$ is orthogonal to vector space $x(n)$, the expectation of their product must be equal to zero

$$E[e(n)x^*(n)] = E[\left\{d(n) - P^* R^{-1} x(n)\right\} x^*(n)] \tag{F.8}$$

where $E[x(n)d(n)] = P$ and $R = E[x(n)x^*(n)]$ it follows

$$E[e(n)x^*(n)] = P^* - P^* R^{-1} R = 0 \tag{F.9}$$

Therefore, the vector $e(n)$ is orthogonal to $x(n)$.

F.3 Adaptive Filtering

The requirement for the Winner solution entails some prior statistical knowledge of the input sequence, such as the *autocorrelation* of the input signal and the *cross correlation* between the desired response and the input sequence. Employing such information, the optimum filter tap weights can be computed from the formulations for the Winner solution.

However, usually a priori knowledge is not provided, therefore the design of the optimum filter with optimum weights is not possible. In the case that the statistical information of the input sequence is not available, one solution would be the *adaptive* filter design. Adaptive filters use an iterative algorithm which starts from an initial point with the minimum available statistical information from the input signal. Provided the input signal is stationary, it is possible that the algorithm after some successive iterations converge toward the optimum Winner solution in some statistical sense, this method, however, may not yield the perfect Winner solution since the statistical data used during the optimization process were estimate of the actual data.

The consequence of the application of the recursive algorithm in which the parameters of the filter are updated from one iteration to the next iteration, is that the parameters will become *data dependent*. Some important attributes of the recursive adaptive algorithms which play a pivotal role on the selection of the appropriate algorithm for the design of adaptive filters are: Rate of convergence, misadjusting, tracking, robustness.

In order to choose an appropriate adaptive algorithm, these parameters become very helpful, depending the critical requirements of the application in hand. For example, for the design of the complex compensator for the this application, since there is no transmission channel involved, and the connection between the transmitter and receiver is made by a coaxial cable, the robustness and tracking is not a concern.

F.4 Linear LMS equalizer

F.4.1 Introduction

LMS equalizers use a recursive adaptive algorithm in which the necessary statistical information is obtained iteratively. In order to describe such an algorithm, a well know method for the minimization of the convex error function for finding the optimum filter tap weights $\omega(c)$ will be introduced [31] known as *steepest descent* method.

F.4.2 Steepest Descent Method

The error function $e(n)$ is a scalar quadratic convex function of complex vector $\omega(n)$, this implies that it has only one minimum which represents the optimum filter coefficients $\omega(n)$, the goal is to find this minimum point using the *gradient* of this function. The sign and slope of gradient of the error function can be used to find the direction of increase/decrease of the error function. For example, if the gradient is positive, in the next iteration, some value corresponding to the gradient will be subtracted from $\omega(n)$. On the other hand, if gradient is negative, in next iteration, some value corresponding to the gradient will be added to $\omega(n)$. Repeating this in an iterative manner, the minimum point will be found. The error function $e(n)$ is a function of complex vector $\omega(n)$, therefore the gradient must be calculated for both the real and imaginary parts of $\omega(n)$ as follows:

$$\begin{aligned} w_r(i+1) &= w_r(i) - \tfrac{\mu}{2}\nabla_{w_r}(\epsilon^2)|_{w=w(i)} \\ w_i(i+1) &= w_i(i) - \tfrac{\mu}{2}\nabla_{w_i}(\epsilon^2)|_{w=w(i)} \end{aligned} \tag{F.10}$$

where, μ represents the amount of increase/decrease proportional to the gradient value, referred to as step size. w_r and w_i represent the real and imaginary parts of the tap weight vector $w(n)$ respectively. The expressions in equation F.10 can be represented in the combined form as

$$w(i+1) = w(i) - \frac{\mu}{2}\nabla_w(\epsilon^2)|_{w=w(i)} \tag{F.11}$$

where ∇_w represents the multivariate gradient operator with respect to $w(n)$ defined as

$$\nabla_\omega = [\frac{\partial}{\partial\omega_1} \frac{\partial}{\partial\omega_2} ... \frac{\partial}{\partial\omega_N}]^T \tag{F.12}$$

After substituting the parameters into the error function ϵ^2 and some manipulations, it follows

$$w(i+1) = w(i) + \mu[P - R\omega(i)] \tag{F.13}$$

In Equation (F.13) the parameters P and R are not known a *priori*, therefore, in each iteration start with the minimum amount of information available from the existing values of input sequence $x(n)$ and $d(n)$, by computing a poor estimate of the autocorrelation and cross correlation as

$$\begin{aligned} R &= x(n)x^H(n) \\ P &= x(n)d^H(n) \end{aligned} \tag{F.14}$$

Although these estimations are not an accurate representative of R and P, however if computed and averaged over a large number of iterations in real time, they approach asymptotically to the minimum of the convex function. Substituting the pertinent values of P and R into the Equation (F.14) and changing the variable i to n for the real time iteration, yields

$$w(n+1) = w(n) + \mu x(n)[d^H(n) - x^H(n)\omega(n)] \tag{F.15}$$

$$w(n+1) = w(n) + \mu x(n)e^*(n) \tag{F.16}$$

Equation (F.16) represents the iterative algorithm by which the minimum point of the convex error function $e(n)$ can be calculated, known as *LMS* algorithm. The detailed step by step procedure defined by LMS algorithm is

1. $\omega_0 = \omega_{init}$
2. For n = 1 to final

3. $y(n) = \omega^H(n)x(n)$

4. $e(n) = d(n) - y(n)$

5. $w(n+1) = w(n) + \mu x(n)e^*(n)$

6. end.

F.5 LMS equalizer Convergence

In LMS algorithm, the optimum Winner filter weights can not be achieved, since the starting point for the algorithm was an estimate of the parameters P and R not their exact values. Therefore, the descent direction in each iteration is not quite accurate, however, due to the stationarity property of the input sequence and the error function, despite the instantaneous fluctuations of the filter weights $\omega(n)$ in each iteration, the mean value of $\omega(n)$ will approach to the optimum filter weight ω_{opt} [31] [43] [30].

It is conceivable that the final filter weights achieved by LMS algorithm differ from the optimum filter weights. This differences denoted by $\Delta(n)$, can be represented by the following expression

$$\Delta(n) = \omega(n) - \omega_{opt} \tag{F.17}$$

Using Equations (F.16) and (F.17)

$$\Delta(n+1) = \Delta(n) + \mu x(n)e^*(n) \tag{F.18}$$

After substituting the parameters of the error function into equation F.18 it follows

$$\Delta(n+1) = \Delta(n) + \mu x(n)\left\{d^*(n) - x^H(n)[\omega_{opt} + \Delta(n)]\right\} \tag{F.19}$$

Taking the expectation of both sides, and using $E[\Delta(n)] = v(n)$ yields

$$v(n+1) = v(n) + \mu[P - R\omega_{opt}] - \mu E[x(n)x^H(n)\Delta(n)] \tag{F.20}$$

$$v(n+1) = v(n) - \mu E[x(n)x^H(n)\Delta(n)] \tag{F.21}$$

In order to simplify the right most side of the Equation (F.21), some assumptions have to be made. If the filter weights $\omega(n)$ in each iteration, is statistically independent[3]from the the input sequence $x(n)$ and desired sequence $d(n)$, the right most side of the equation F.21 can be expressed as

$$v(n+1) = v(n) - \mu E[x(n)x^H(n)]E[\Delta(n)] \tag{F.22}$$

with above assumption Equation (F.22) reduces to

$$v(n+1) = (I - \mu R)v(n) \tag{F.23}$$

The parameter R is the autocorrelation matrix of input signal, hence it is Hermitian and positive definite and can be expressed as $R = TDT^H$ with T being *unitary*. Substituting Equation (F.23) into the Equation (F.22), denoting $T^H v(n) = u(n)$ yields

$$u(n+1) = (I - \mu D)u(n) \tag{F.24}$$

The goal is

$$\lim_{n\to\infty} ||v(n)||^2 = 0 \tag{F.25}$$

which entails

$$0 < (1 - \mu\lambda_i)^2 < 1 \tag{F.26}$$

where, λ_i denotes the eigenvalues of R. Hence

$$0 < \mu < \frac{2}{\lambda_{max}} \tag{F.27}$$

[3]This assumption does not hold completely, since the filter weights $\omega(n)$ in each iteration is a function of $x(n)$, and most entries of vector $x(n)$ in each iteration is correlated to those of previous iteration, hence $\omega(n)$ is correlated to the input sequence, however, one can argue that by choosing the parameter μ small enough this correlation become very small and the assumption is loosely valid.

or in practice

$$0 < \mu < \frac{2}{tr(R)} \tag{F.28}$$

The parameter μ has a critical role on the convergence and the performance of the LMS algorithm. Had this parameter chosen without discretion, the algorithm may not converge, the expression in Equation (F.28) defines an upper bound on this parameter which guaranties the convergence of the algorithm.

F.6 LMS Mean Square Error

As it was stated before, in LMS algorithm the filter tap weights are fluctuating in each iteration, and its mean value will converge toward the optimum tap weights ω_{opt}. However, even after convergence, even though, after the mean value of the $\omega(n)$ is equal to the ω_{opt}, there is some variations in $\omega(n)$, which is referred to as the variance of $\omega(n)$, the mean square of this parameter needs to be minimizes for a better performance of adaptive filter. Therefore, a *Mean Square Error* analysis needs to be performed. The starting point for this analysis is the expression for the error function, which is

$$e(n) = d(n) - \omega^*(n)x(n) = d(n) - [\omega_{opt} + \Delta(n)]^* x(n) = e_o(n) - \Delta^*(n)x(n) \tag{F.29}$$

where $e_o(n)$ represents the optimum error for the optimum filter taps ω_{opt}. Taking the expectation of both sides, and more elaborations yields

$$E[e^2(n)] = E[e_o^2(n)] - 2E[\Delta^*(n)x(n)e_o(n)] + E[\Delta^*(n)x(n)^2] \tag{F.30}$$

In Equation (F.30), the term $E[e_o^2(n)]$ represents the power of optimum noise function and can be shown by ϵ_{min}^2, the second term on the rightmost side due to the statistical independence of the error function to the input sequence and its orthogonality with $x(n)$ is zero, then this equation reduces to

$$\epsilon^2 = E[e^2(n)] = \epsilon_{min}^2 - E[\{\Delta^*(n)x(n)\}\{x^*(n)\Delta(n)\}] \tag{F.31}$$

again using the statistical independence of $\Delta(n)$ with $x(n)$ after some manipulation[4] the expression in Equation (F.31) can be written as

$$\epsilon^2 = E[e^2(n)] = \epsilon^2_{min} - E[\Delta^*(n)E[x(n)x^*(n)]\Delta(n)] \tag{F.32}$$

where $E[x(n)x^*(n)]$ is the autocorrelation matrix R, more manipulation follows

$$\epsilon^2 = E[e^2(n)] = \epsilon^2_{min} - E[tr(\Delta(n)\Delta^*(n)R)] = \epsilon^2_{min} - tr(E[\Delta(n)\Delta^*(n)R]) \tag{F.33}$$

substituting $R = TDT^*$ and denoting $E[\Delta(n)\Delta^*(n)]$ with $k(n)$ and representing $T^*k(n)T$ with $k'(n)$

$$\epsilon^2 = E[e^2(n)] = \epsilon^2_{min} + tr(k'(n))D] \tag{F.34}$$

The second term in the rightmost side of the Equation (F.34) is the additional noise produced as the result of poor estimation of R and P in each iteration, and has to be bounded for the convergence of algorithm. This value is called *excess error* and is represented by ($\epsilon_{excess} = tr(k'(n))D$). By using some statistical analysis manipulations, assuming a very small convergence step size μ, and using the statistical approximations, this term can be further simplified to

$$\epsilon_{excess} \approx \mu \tag{F.35}$$

The precise value of excess error is already calculated in literature. A more sophisticated treatment of the MSE and other performance parameters of the adaptive filters can be found in the literature [31] [30].

[4]If two random variables x and y are independent i.e., $f_{xy}(x, y) = f(x)f(y)$ the following expression is true $E[xy] = E[x]E[y] = E[xE[y]]$ [45]