2003

A class of linear phase IIR digital filters design.

Tao. Dai
University of Windsor

A Class of Linear Phase IIR Digital Filters Design

By

Tao Dai

A Thesis

Submitted to the Faculty of Graduate Studies and Research

through the Department of Electrical and Computer Engineering

in partial Fulfillment of the Requirements for

the Degree of Master of Applied Science at the

University of Windsor

Windsor, Ontario, Canada

2003

Your file Votre référence
ISBN: 0-612-82862-X
Our file Notre référence
ISBN: 0-612-82862-X

Abstract

Digital IIR filters find wide applications in fields such as speech processing, image processing and noise/echo cancellation. In recent years, the design of linear phase IIR filter becomes a very hot issue of research interest.

As we know, FIR filters designed to approximate a magnitude response that has a narrow transition band between the passband and the stopband usually require a large number of multipliers and they have a large delay also. Hence we prefer to IIR digital filters that have less number of multipliers than their corresponding FIR counterparts but one of disadvantages is that an IIR filter cannot reach an exactly linear phase. The importance of the phase response linearity of a digital filter was recognized in early years' research and Finite Impulse Response (FIR) filters provided a perfect solution to this requirement. However, due to fundamental incompetence of FIR filters stated before, approximately linear phase IIR filters became the focus of research as a compromise between the implementation cost and the linearity of the phase response.

About twenty years ago, most of researches on digital filters and signal processing, when and if they discussed the design of digital filter, they treated mainly the approximation of the magnitude response, or else, dealt with magnitude and phase responses separately. For example, as we know, the one of the most well known methods for solving the problem of IIR filter

phase linearity was based on the application of cascading the prototype IIR filter with an allpass phase equalizer.

In recent years, new methods have been proposed for the simultaneous approximation of both the magnitude and group delay of IIR digital filter. They have less number of multipliers than their corresponding FIR counterparts and yet provide the required phase response. This topic is covered all through this thesis.

Publications

*H.K.Kwan, and T.Dai, "Equiripple IIR Digital Filter Design With Constant Group Delay", IEEE ELECTRO/INFORMATION TECHNOLOGY (EIT) CONFERENCE, INDIANAPOLIS, June, 2003.

**Dedicated with love to my wife Yi Lin, my father GuoLu Dai,
my mother XingLan Wang and all my family members
for their constant love and support**

Acknowledgements

I would like to express my sincere gratitude to my thesis advisor Dr. H. K. Kwan for the financial support as his research assistant, for his invaluable guidance, suggestions, comments, constant encouragement, and generous help throughout the study on the work as reported in this thesis.

I would like to acknowledge Dr. Kwan for introducing me the field of digital filter design and suggesting linear phase IIR digital filter design as the theme of my thesis. In the thesis, Dr. Kwan has suggested and introduced me to the study of the following design strategies: (a) the combination of the Thiran's all-pole filter for approximating constant group delay, and the mirror image numerator transfer function designed using the Remez exchange algorithm for approximating magnitude specifications. In particular, Dr. Kwan has suggested me in this study the simultaneous use of equiripple passband and equiripple stopband as a design strategy for reducing filter-order and transition bandwidth. (b) Linear phase IIR digital filter design using parallel allpass filter sections. (c) IIR digital filter design meeting simultaneous magnitude and group delay responses specifications using non-linear optimization techniques.

I would also like to give my sincere appreciations to my internal reader, Dr. C. Chen and my external reader, Dr. R. Caron for their comments and suggestions in enhancing this thesis.

Finally, many thanks to all my classmates in the ISPLab who have given me help and pleasure during my study and research.

Table of Contents:

==========

List of Abbreviations:

DSP	Digital Signal Processing
FIR	Finite Impulse Response
IIR	Infinite Impulse Response
PCLS	Peak Constrained Least Squares
RF	Radio Frequency
MFM	Maximally Flat Magnitude
MFD	Maximally Flat Delay
SOS	Second Order Sections
GME	Generalized Multiple Exchange
RGME	Recursive Generalized Multiple Exchange
LTI	Linear Time-Invariant
SSE	Sum of Squared Error
GP	General Problem
LP	Linear Programming
QP	Quadratic Programming
NP	Nonlinear Programming
KT	Kuhn-Tucker (equation)
SQP	Sequential Quadratic Programming
BFGS	Broyden, Fletcher, Goldfarb, and Shanno
WLS	Weighted Least Squares

List of Figures:

List of Tables:

Chapter 1

Introduction

1.1 Background

The world of science and engineering is filled with *signals*: images from remote space probes, voltages generated by the heart and brain, radar and sonar echoes, seismic vibrations, and countless other applications. Digital Signal Processing is the science of using computers to understand these types of data. This includes a wide variety of goals: filtering, speech recognition, image enhancement, data compression, neural networks, and much more.

In signal processing, one of main functions of a filter is to remove unwanted parts of the signal or to extract useful parts of the signal. An analog filter uses analog electronic circuits made up from components such as resistors, capacitors and op amps to produce the required filtering effect for continuous-time signal processing such as voltage and current. A digital filter is a digital system that can be used to filter discrete-time signals. It is only a formula for going from one digital signal to another. It may exist as an equation on paper, as a small loop in a computer subroutine, or as a

handful of integrated circuit chips properly interconnected. Generally a digital filter uses a digital processor to perform numerical calculations on sampled values of the signal. The processor may be a general-purpose computer such as a PC, or a specialized DSP chip.

Compared with analog filters, digital filters have many advantages:

- Easily designed, tested, implemented and programmable;

  Its operation is determined by a program stored in the processor's memory. This means the digital filter can easily be changed without affecting the circuitry (hardware). An analog filter can only be changed by redesigning the filter circuit.

- Stable with respect to time and temperature;

  The characteristics of analog filter circuits (particularly those containing active components) are subject to drift and are dependent on temperature. Digital filters do not suffer from these problems.

- Suitable for low frequency signals accurately;

  Unlike their analog counterparts, digital filters can handle low frequency signals accurately.

- Flexible and versatile;

  Digital filters are very much more versatile in their ability to process signals in a variety of ways; this includes the ability of some types of digital filter to adapt to changes in the characteristics of the signal.

However, digital filters also have shortcomings compared with analog filters such as relatively narrower applicable frequency band. But as the speed of DSP technology continues to increase, digital filters are being applied to high frequency signals in the RF (radio frequency) domain, which in the past was the exclusive preserve of analog technology.

Digital filter design is the process of deriving the filter's function that satisfies filter's prescriptions such as magnitude response, phase response, stability and so on. Normally digital filters can be classified to two types, FIR (Finite Impulse Response) and IIR (Infinite Impulse Response). In filter design literatures, the problem of designing linear phase FIR filters with desired magnitude characteristics has been well studied. The design of IIR filters with linear phase in the passband has been considered by many researchers in the last two decades.

1.2 Some Methods of IIR Digital Filter Design

The conventional technique is to first design an IIR filter that meets the magnitude specification and then to employ allpass equalizers to linearize the phase response. At present, the problem most attractive is the simultaneous approximation of both magnitude and phase characteristics. Benefited from researchers' dedications, tremendous numbers of creative approaches come out and some of them are popularly applied in digital filter design. For the problem of linear phase IIR filter design, there are several

typical and effective ones adopted in this thesis such as: Thiran's all pole filter, Remez algorithm, Nonlinear optimization and allpass filters' applications.

1.2.1 About Thiran's All Pole Filter

At 1971, Thiran [1] developed an analytical method for deriving the all-pole transfer function of the digital filter that approximates a constant group delay in the maximally flat sense. Thiran's expression for the digital transfer function that approximates a maximally flat group delay is given by:

$$H(z^{-1}) = \frac{\left\{ \frac{2n!}{n!} \frac{1}{\prod_{i=n+1}^{2n} (2\tau + i)} \right\}}{\sum_{k=0}^{n} \left[(-1)^k \binom{n}{k} \prod_{i=0}^{n} \frac{2\tau + i}{2\tau + k + i} \right] z^{-k}} \qquad (1.1)$$

Thiran also proved that for all finite positive values of τ, the resulting filter would always be stable.

Thiran [2] had investigated the conditions under which the all pole transfer functions approximated a constant group delay in the equiripple sense, but the conditions gave rise to nonlinear equations that could not be solved in general. Based on the achievement of Thiran, Deczky [3] gave a way of approximating a constant group delay in the equiripple sense by using Remez Exchange Algorithm.

The authors of [4] started with Thiran's all pole transfer function that has a constant group delay response in the maximally flat sense. They proposed new methods for choosing the zeros of the numerator polynomial such that the augmented transfer function exhibits a maximally flat magnitude response in the passband and an equiripple response in the stopband.

Recent applications such as, Vesa Valimaki [5] proposed a novel method for designing fractional delay allpass filters in his paper that implements Thiran's filters by truncating the coefficient vectors of them. The new design formula for a fractional delay allpass filter is slightly modified for the Thiran's filter:

$$a_k = (-1)^k \binom{M}{k} \prod_{n=0}^{M} \frac{\tau + n}{\tau + k + n} \qquad \text{k=1,2...,N} \qquad \text{M>N(filter order)} \qquad (1.2)$$

The main advantages of the new method are thus the ease of the design using closed-form formulas and the possibility to design high-order filters. R. A. Gopinath [6] proposed an approach based on Thiran filters for group delay flatness and Herrmann filters for magnitude flatness. *D. Economou, C. Mavroidis* and I. *Antoniadis* [7] designed a preconditioning approach, based on the proper design of conventional low-pass IIR digital filters using the method proposed by Thiran. At 2001, in one paper of *Makundi, M.; Valimaki, V.; Laakso, T.I* [8], polynomial coefficients were obtained in closed form using the Thiran allpass filter design method with modifications which completely eliminated the division operations.

1.2.2 About Remez Algorithm

The Remez exchange algorithm (Remez 1957) was first studied by Parks and McClellan (1972). Also denoted as Parks-McClellan method (PM), it is not only the most widely used FIR filter design method, but also a popular IIR filter design approach. It is an iteration algorithm that accepts filter specifications in terms of passband and stopband frequencies, passband ripple, and stopband attenuation. The fact that designers can directly specify all the important filter parameters and experience has shown that the algorithm converges very fast, compared with other algorithms such as linear programming, make this method very popular.

The algorithm can mainly be described as two steps.

1. The determination of candidate filter coefficients from candidate "alternation frequencies", which involves solving a set of linear equations.
2. The determination of candidate alternation frequencies from the candidate filter coefficients.

A description emphasizing the mathematical foundations rather than digital signal processing applications is given by Cheney (1999).

Selesnick, I.W.; Burrus, C.S [9] gave a complement for the Remez algorithm for linear phase FIR filter design. It describes an exchange algorithm for the frequency domain design of equiripple linear phase FIR filter where the Chebyshev error in each band is specified. This algorithm combines several algorithms including Remez algorithm. One contribution is that it modifies

the usual Remez algorithm so that it achieves a specified Chebyshev error in one band and minimizes it in the other band. This is done by imposing an affined relationship between δ_p and δ_s and introducing two new parameters η_p, η_s, obtaining a filter satisfying the following affine relationship between δ_p and δ_s:

$$\delta_p = K_p\delta + \eta_p$$
$$\delta_s = K_s\delta + \eta_s \qquad (1.3)$$

In their paper, they also presented a flowchart and an illustration of iteration procedures of the modified Remez algorithm.

Hegde, R and Shenoi, B.A [10], present two solutions to the problem of designing linear phase FIR filters with a flat passband and specified bandwidth. First, by deriving conditions to attain desired degrees of flatness at $\omega=0$ and/or $\omega=\pi$, they obtain an analytical solution. They then present an iterative design procedure to obtain simultaneously a magnitude response that is flat in the passband and equiripple in the stopband. An IIR filter design is provided in another paper of theirs [11]. They also provided an MFM-MFD filter [12] with equiripple in stopband in order to decrease the transition band and obtain an equiripple magnitude response in the stopband of the filter, while the flat magnitude and group delay response in the passband are maintained. This is done by increasing the order of the mirror image polynomial by purposely adding some zeros in the stopband region on the unit circle.

Similar with the upper method, there are many methods that have something in common but differ in the optimization criteria. All of them decouple the design problem of simultaneously approximating the magnitude and group delay by first generating an all pole transfer function that approximates the group delay, in the maximally flat or equiripple sense or least pth sense, then adding a numerator that is chosen to approximate the magnitude either in the same sense of a different sense. For example, Hinamoto and Maekawa [13] first optimize the coefficients of the all pole Transfer function

$$H(z^{-1}) = \frac{1}{D(z^{-1})}$$

such that it approximates the constant group delay in the least pth sense. Then they augment it by adding a mirror image polynomial as the numerator and optimize its coefficients such that

$$H(z^{-1}) = \frac{N(z^{-1})}{D(z^{-1})}$$

approximates the prescribed magnitude also in the least pth sense.

1.2.3 About Nonlinear Optimization

A.G.Deczky [14] considered a general transfer function and decomposed it into a cascade of second-order sections (SOS structure). He developed an algorithm for minimizing an error function that contains a weighted sum of the error in the magnitude as well as in the group delay. Using the minimum

p-error criterion, this error function is successfully solved using the Fletcher-Powell algorithm. Also in this paper, an important theorem guaranteeing the existence of a stable optimum for a large class of synthesis problems is stated and necessary modifications to the Fletcher-Powell algorithm to assure stability are considered.

J.L.Sullivan and J.W.Adams [15] adopted new algorithm with simultaneous optimization of the frequency response magnitude and the group delay and obtained a dramatic improvement in the solution of this classic IIR digital filter design problem. In this paper, the nonlinear optimization problem is solved using the GME (generalized multiple exchange) algorithm combined with the recursive quadratic programming concept. This new combination is called the recursive generalized multiple exchange (RGME) algorithm. With the same filter structure and the same specifications, they lowered the group delay ripple significantly.

1.2.4 About Allpass Filters' Applications

As stated before, at the beginning, allpass filters are used as 'phase equalizer' by cascading with a filter that meets magnitude requirements. In recent years, a number of digital filter structures composed of allpass subfilters have been developed for various applications. Figure 1.1 shows one of those applications composed of two allpass subfilters in parallel.

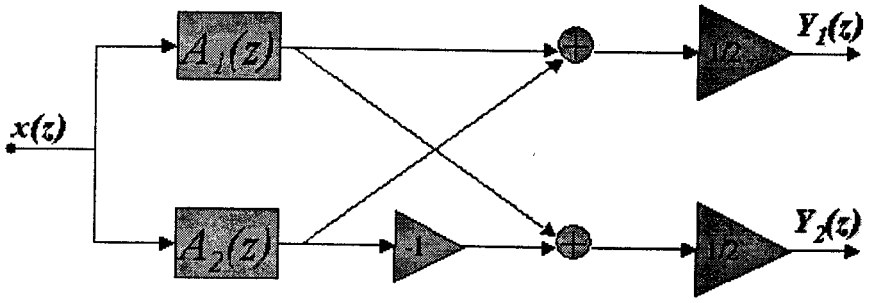


Figure 1.1 Sum and difference of two allpass filters

The transfer function $A_1(z)$ and $A_2(z)$ are allpass functions and from this structure, two transfer functions can be obtained as:

$$H_1(z) = \frac{Y_1(z)}{X(z)} = \frac{1}{2}[A_1(z) + A_2(z)]$$
$$H_2(z) = \frac{Y_2(z)}{X(z)} = \frac{1}{2}[A_1(z) - A_2(z)] \qquad (1.4)$$

Two main advantages of using a parallel connection of allpass filters are:

- Low sensitivity of filter characteristics to some parameters varying;
- The complementary filter can be obtained from the original one with ease.

M.Renfors and T. Saramaki [16] gave an example of a lowpass IIR filter composed of two allpass filters and proposed that one of the allpass network be chosen as a pure delay network. The pure delay term ensures a good phase performance for the overall filter in the passband. This is due to "the

property that with small passband variation, the phase of the allpass section is forced to closely follow the linear phase of the delay branch".

Stancic, G and Djuric, B [17] proposed a method for synthesis of digital IIR filters given by parallel connection of two allpass networks. This filter shows low sensitivity in passband. The magnitude sensitivity in stopband is considerably bigger than it in passband. However, it can be decreased by duplicating a pole of allpass network transfer function whose phase angle is in transition region between passband and stopband. Finally the problem is to minimize an error function where the weight function is constant for all frequencies.

$$E = \sum_{j=1}^{M} W(\omega_j)[\sum_{i=1}^{n/2} \tau_{2i}(\omega_j) - n + 1]^2 \qquad (1.5)$$

Artur Krukowski and Izzet Kale [18] presented in their paper a weighted least square algorithm in which a flexible weighting function is adopted. Choosing constant weights for the passband and stopband led to stopband ripples decreasing monotonically with frequency while passband ripples were monotonically increasing. Therefore an iterative method was applied which was changing the weighting function according to the shape of the envelop of the passband/stopband group delay ripples at every iteration. At the beginning of this algorithm, the weight vector is initiated to be unity for all frequencies and iteratively update it by an equation:

$$W(f) = [1 + W(f)] \cdot [1 + \tau^*(f)] - 1 \qquad (1.6)$$

This paper also remarked that it's important to monitor the group delay ripples both in the filter passband and its stopband. This ensured that both small passband ripples and high stopband attenuation are achieved. In this paper, authors also showed us effect caused from different coefficients word lengths.

1.3 Motivations

Digital Signal Processing is one of the most powerful technologies that will shape science and engineering in the twenty-first century. Revolutionary changes have already been made in a broad range of fields: communications, medical imaging, radar & sonar, high fidelity music reproduction, and oil prospecting, to name just a few.

Digital filters are a very important part of DSP. In fact, their extraordinary performance is one of the key reasons that DSP has become so popular.

IIR (Infinite Impulse Response) digital filters have significant advantages over FIR (Finite Impulse Response) digital filters on resource economization and computational speed. However, IIR filters are not used as widely as FIR filters at present because of some problems such as phase linearity and stability. Therefore, designing approximately linear phase IIR filters has already become one of the hottest topics due to their valuable merits.

1.4 Objectives of the Thesis

This thesis looks at the IIR digital filter design problems with both magnitude response and phase response at the same time. Appropriate algorithms and approaches (Remez algorithm, Thiran's method, Nonlinear optimization methods, PCLS, Adaptive weighting function and so on) are applied to tackle different problems successfully. The work presented is trying to creatively research and implement IIR digital filters in different ways and figure out their various characteristics and appropriate usages.

1.5 Organization of the Thesis

In the following chapters, we illustrate three typical and efficient approaches for linear phase IIR filter design. In chapter I, we introduce an analytical design approach for magnitude response that is in maximally flat sense for group delay response and equiripple sense for magnitude response both in passband and stopband. The filter's phase linearity and stability are guaranteed because of adopting Thiran's filter. In chapter II, one robust structure IIR filter that is composed of allpass subfilters in parallel gives another efficient approach for IIR filter design and shows many appealing characteristics such as low sensitivity to coefficient quantization. In chapter III, a general form nonlinear optimization algorithm is presented that gives

designers more flexibility for designing filter magnitude and phase response simultaneously.

Chapter 2

Maximally Flat Magnitude & Maximally Flat Delay (MFM-MFD) Equiripple IIR filter

2.1 Introduction

This design is an analytical method for directly designing IIR filters. In this chapter, we consider the approximation of a constant group delay first and then magnitude response of a one-dimensional IIR digital filter. We will describe a method of designing an all-pole transfer function that approximates the prescribed group delay in the maximally flat sense. Then by augmenting it with a numerator polynomial such that it changes the group delay of the all-pole transfer function only by a constant but changes the magnitude such that the overall transfer function approximates the prescribed magnitude.

2.2 Design Theory

2.2.1 Maximally Flat Group Delay Filters

Thiran [1] developed an analytical method for deriving the all-pole transfer function of the digital filter that approximates a constant group delay in the maximally flat sense.

Let τT be the prescribed group delay, and let the all-pole transfer function be chosen as

$$H(z^{-1}) = \frac{\sum_{i=0}^{n} a_i}{\sum_{i=0}^{n} a_i z^{-i}} \qquad (2.1)$$

Then the error in the phase is

$$\delta(\omega t) = -\omega\tau - \tan^{-1} \frac{\sum_{i=0}^{n} a_i \sin(i\omega t)}{\sum_{i=0}^{n} a_i \cos(i\omega t)} \qquad (2.2)$$

Another form of the error derived from (2.2) is given as

$$\varepsilon(\omega t) = \tan(\delta(\omega t)) = -\tan(\omega\tau) - \frac{\sum_{i=0}^{n} a_i \sin(i\omega t)}{\sum_{i=0}^{n} a_i \cos(i\omega t)} \qquad (2.3)$$

which can be rewritten as

$$\varepsilon(\omega t) = \frac{-\sin(\omega\tau)\sum_{i=0}^{n} a_i \cos(i\omega t) - \cos(\omega\tau)\sum_{i=0}^{n} a_i \sin(i\omega t)}{\cos(\omega\tau)\sum_{i=0}^{n} a_i \cos(i\omega t)} \quad (2.4)$$

Assuming that in the sequel the sampling period T is normalized to one second so that $\omega_s=2\pi$ and τ denotes the delay which is the number of sampling periods. Hence ω will be the normalized frequency. Then (2.4) reduces to

$$\varepsilon(\omega) = -\frac{\sum_{i=0}^{n} a_i \sin(i+\tau)\omega}{\cos(\omega\tau)\sum_{i=0}^{n} a_i \cos(i\omega)} \quad (2.5)$$

The numerator is an odd function and the denominator is an even function and therefore their expansion in power series gives

$$\varepsilon(\omega) = \frac{\sum_{k=0}^{\infty} p_k \omega^{2k+1}}{\sum_{k=0}^{\infty} q_k \omega^{2k}} \quad (2.6)$$

Since $\varepsilon(\omega)$ is an odd function, its Taylor series contains only odd powers of ω and hence is in the form,

$$\varepsilon(\omega) = \sum_{k=0}^{\infty} c_k \omega^{2k+1} \quad (2.7)$$

where the coefficient c_k is the k[th] derivative of $\varepsilon(\omega)$ evaluated at ω=0. The coefficients can also be generated from the recursive relation

$$c_k = \frac{1}{q_0}\left[p_k - \sum_{j=1}^{k} c_{k-j} q_j \right] \quad (2.8)$$

For getting a maximally flat approximation of a constant group delay τ, we need to make the first n derivatives of $\varepsilon(\omega)$ at ω=0 to be zero i.e. c_k=0 for k=0,1,…,(n-1). From (2.6), (2.7), (2.8) we see that the equivalent condition to be satisfied is p_k=0 for 0≤ k ≤n-1. Using the Taylor series expansion

$$\sin x = \sum_{k=0}^{\infty} (-1)^k \frac{x^{2k+1}}{(2k+1)!} \quad (2.9)$$

on (2.5) we get

$$\varepsilon(\omega) = \sum_{k=0}^{\infty} (-1)^k \left[\frac{\sum_{i=0}^{n} a_i (i+\tau)^{2k+1}}{(2k+1)!} \right] \omega^{2k+1} \quad (2.10)$$

Hence the coefficient

$$c_k = (-1)^k \left[\frac{\sum_{i=0}^{n} a_i (i+\tau)^{2k+1}}{(2k+1)!} \right] \quad (2.11)$$

is zero when the numerator is zero. From the condition that the coefficients c_k in (2.8) are zero for k=0,1,...,(n-1) (with a_0=1), the condition for maximally flat approximation of a constant group delay becomes

$$\tau^{2k+1} + \sum_{i=1}^{n} a_i (i+\tau)^{2k+1} = 0$$

$$\Downarrow \qquad (2.12)$$

$$\sum_{i=1}^{n} a_i (i+\tau)^{2k+1} = -\tau^{2k+1} \qquad for\ k = 0,1,\ldots,n-1$$

Solving these linear equations for the coefficients a_i in terms of two Vandermonde determinants, Thiran shows that the coefficients are given by

$$a_k = (-1)^k \binom{n}{k} \prod_{i=0}^{n} \frac{2\tau + i}{2\tau + k + i} \qquad (2.13)$$

Using the Gamma functions, he also shows that the polynomial $P_n(z^{-1},\tau) = \sum_{k=0}^{n} a_k z^{-k}$ in the denominator of the maximally flat delay filter function (2.1) can be expressed as

$$P_n(z^{-1},\tau) = \sum_{k=0}^{n} \frac{\Gamma(-n+k)}{\Gamma(-n)} \frac{\Gamma(2\tau+n+1)}{\Gamma(2\tau)} \frac{\Gamma(2\tau+k)}{\Gamma(2\tau+k+n+1)} \frac{z^{-k}}{k!} \qquad (2.14)$$

The author derives the numerator of (2.1) from its denominator evaluated on |z|=1 and shows that

$$P_n(1,\tau)=\sum_{i=0}^{n}a_i=\frac{2n!}{n!}\frac{1}{\prod_{i=n+1}^{2n}(2\tau+i)} \qquad (2.15)$$

Finally, *Thiran*'s expression for the digital transfer function that approximates a maximally flat group delay is give by:

$$H(z^{-1})=\frac{\left\{\frac{2n!}{n!}\frac{1}{\prod_{i=n+1}^{2n}(2\tau+i)}\right\}}{\sum_{k=0}^{n}\left[(-1)^k\binom{n}{k}\prod_{i=0}^{n}\frac{2\tau+i}{2\tau+k+i}\right]z^{-k}} \qquad (2.16)$$

$n \Rightarrow$ *the order of the filter*

$\tau \Rightarrow$ *the desired group delay*

Thiran has also shown that the above transfer function is stable for all finite positive values of τ. Using the above formula, the coefficients of the denominator polynomials of $\boldsymbol{H(z^{-1})}$ can be obtained for the purpose of the next step's filter design.

2.2.2 Use of Mirror Image Polynomial

Mirror image polynomial *($\boldsymbol{N(z)=z^{-p}N_a(z)}$, $N_a(z)$ is defined in (2.18))* is adopted as the numerator polynomial in this design. The mirror image polynomial of even order, used as the numerator, is a real function of ω. Such a polynomial has zeros inside the unit circle |z|=1 and also the reciprocal of these zeros which are outside the unit circle. It may also have zeros on the unit circle. So the overall transfer function is not a minimum

phase transfer function (An LTI filter H(z)=B(z)/A(z) is said to be minimum phase if all its poles and zeros are inside the unit circle |Z|=1).

Consequently, there are fewer restrictions between the magnitude and the phase response. In other words, there is more flexibility for simultaneously shaping the magnitude and phase responses. Therefore there exist many different choices that have been proposed for solving this general problem.

2.3 Design Procedure

2.3.1 Maximally Flat Magnitude and Maximally Flat Delay (MFM-MFD) Filter with Equiripple in Stopband [12]

2.3.1.1 Design Theory

The most general form, the transfer function of a one-dimensional IIR filter is of the form:

$$H(z)=\frac{\sum_{i=0}^{M} b_i z^{-i}}{\sum_{i=0}^{N} a_i z^{-i}}=\frac{N(z)}{D(z)} \xleftrightarrow{z=e^{jw}} H(e^{jw})=\frac{\sum_{i=0}^{M} b_i e^{-ijw}}{\sum_{i=0}^{N} a_i e^{-ijw}} \quad (2.17)$$

In this equation, in order to approximate a constant group delay in the maximally flat sense, we choose the "*Thiran's polynomial*" as ***D(z).*** The coefficient vector "***den***" has been obtained from (2.16), then set

$N(z)=z^{-p}N_a(z)$, where $N_a(z)$ is a mirror image polynomial given in the form:

$$N_a(z) = b_0 + b_1\left(\frac{z+z^{-1}}{2}\right) + b_2\left(\frac{z^2+z^{-2}}{2}\right) + \ldots + b_p\left(\frac{z^p+z^{-p}}{2}\right) \quad (2.18)$$

substituting $z=e^{j\omega}$ in the above (2.18), $N(e^{j\omega})= e^{-j\omega p} N_a(e^{j\omega})$ where

$$N_a(e^{j\omega}) = b_0 + b_1\cos\omega + b_2\cos(2\omega) + \ldots + b_p\cos(p\omega) \quad (2.19)$$

The numerator $N(e^{j\omega})$ adds a pure delay of p samples to that of the *Thiran's* all-pole filter because $N_a(e^{j\omega})$ is a real valued function. The coefficients b_i have to be found such that $H(e^{j\omega})$ has a magnitude response with the desired design requirements, besides having maximally flat group delay characteristics.

Theorem 1:

$$If \quad \left|H(e^{j\omega})\right|_{\omega=0} = \left|\frac{N_a(e^{j\omega})}{D(e^{j\omega})}\right|_{\omega=0} = 1 \quad (2.20)$$

$$and \quad \left.\frac{d^k\left|D(e^{j\omega})\right|}{d\omega^k}\right|_{\omega=0} = \left.\frac{d^k N_a(e^{j\omega})}{d\omega^k}\right|_{\omega=0} \quad (2.21)$$

$$Then \quad \left.\frac{d^k\left|H(e^{j\omega})\right|}{d\omega^k}\right|_{\omega=0} = 0 \quad (2.22)$$

Theorem 2:

$$If \quad \left|H(e^{j\omega})\right|_{\omega=\pi} = \left|\frac{N_a(e^{j\omega})}{D(e^{j\omega})}\right|_{\omega=\pi} = 0 \qquad (2.23)$$

$$and \quad \left.\frac{d^k N_a(e^{j\omega})}{d\omega^k}\right|_{\omega=\pi} = 0 \qquad (2.24)$$

$$Then \quad \left.\frac{d^k \left|H(e^{j\omega})\right|}{d\omega^k}\right|_{\omega=\pi} = 0 \qquad (2.25)$$

The equation *2.19* becomes *2.26* as below:

$$\begin{aligned}\left|D(e^{j\omega})\right|_{\omega=0} &= N_a(e^{j\omega})\Big|_{\omega=0} \\ &= b_0 + b_1 + b_2 + \cdots + b_p \end{aligned} \qquad (2.26)$$

and the equation *2.21* becomes 2.*27*:

$$\begin{aligned}\left.\frac{d^k \left|D(e^{j\omega})\right|}{dw^k}\right|_{w=0} &= \left.\frac{d^k N_a(e^{j\omega})}{d\omega^k}\right|_{\omega=0} \\ &= \begin{cases} (-1)^{k/2}\left[b_1 + b_2(2)^k + \cdots + b_p(p)^k\right] & k \quad even \\ 0 & k \quad odd \end{cases}\end{aligned} \qquad (2.27)$$

We require that these coefficients satisfy another specification: the magnitude of $\boldsymbol{H(e^{j\omega})}$ at a specified bandwidth ω_b is 3 dB below the 0 dB magnitude at ω=0. So

$$\because \quad \left|H(e^{j\omega})\right\|_{\omega=\omega_b} = \left|\frac{N_a(e^{j\omega})}{D(e^{j\omega})}\right\|_{\omega=\omega_b} = 0.7079$$

$$\therefore \quad b_0 + b_1\cos\omega_b + b_2\cos 2\omega_b + \cdots + b_p\cos p\omega_b = 0.7079\left|D(e^{j\omega_b})\right| \qquad (2.28)$$

In order to decrease the transition band and obtain an equiripple magnitude response in the stopband of the filter, while the flat magnitude and group delay response in the passband are maintained, we increase the order of the mirror image polynomial by purposely adding some zeros in the stopband region on the unit circle.

Let the stopband region of the modified filter be denoted by ωs<ω<π, where ωs >ωb. The desired magnitude response to be approximated by the modified filter, over this stopband region, is given by

$$\left|H_d(e^{j\omega})\right| = 0 \qquad \omega_s < \omega < \pi \qquad (2.29)$$

In order to obtain an equiripple magnitude response over this stopband region, the magnitude response ought to satisfy the following set of equations:

$$\left[\left|H_d(e^{j\omega_i})\right| - \left|H(e^{j\omega_i})\right|\right] = (-1)^{i+1}\delta \qquad i = 1,2,\ldots,m$$

$$where \qquad \omega_s \le \omega_1 < \omega_2 < \cdots < \omega_m \le \pi \qquad (2.30)$$

m is the amount of desired extrema in the stopband region. From $H(z) = \frac{z^{-p}N_a(z)}{D(z)}$, and the equations *(2.29)* and *(2.30)*, we get

$$\left|N_a(e^{j\omega_i})\right| - (-1)^i \delta \left|D(e^{j\omega_i})\right| = 0 \qquad i = 1,2,...,m \qquad (2.31)$$

According to *(2.19)* and *(2.31)*, we get

$$b_0 + b_1 \cos\omega_i + b_2 \cos(2\omega_i) + ... + b_p \cos(p\omega_i) + (-1)^{i+1} \delta \left|D(e^{j\omega_i})\right| = 0$$
$$i = 1,2,...,m \qquad (2.32)$$

Hence the equation *(2.28)*, *(2.26)*, *(2.27)* and *(2.32)* can be expressed in a matrix form *(2.33)*.

$$Ab = d \qquad (2.33)$$

matrix A and vectors b and d are as shown below:

$$A=\begin{bmatrix} 1 & \cos\omega_b & \cos 2\omega_b & \cdots & \cos p\omega_b & 0 \\ 1 & 1 & 1 & \cdots & 1 & 0 \\ 0 & -(1^2) & -(2^2) & \cdots & -(p^2) & 0 \\ 0 & (1^4) & (2^4) & \cdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & (-1)^{M/2}(1)^M & (-1)^{M/2}(2)^M & \cdots & (-1)^{M/2}(p)^M & 0 \\ 1 & \cos\omega_1 & \cos 2\omega_1 & \cdots & \cos p\omega_1 & D_1 \\ 1 & \cos\omega_2 & \cos 2\omega_2 & \cdots & \cos p\omega_2 & D_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \cos\omega_m & \cos 2\omega_m & \cdots & \cos p\omega_m & D_m \end{bmatrix} \quad (2.34)$$

$$b=\begin{bmatrix} b_0 & b_1 & b_2 & \cdots & b_p & \delta \end{bmatrix}^T \quad (2.35)$$

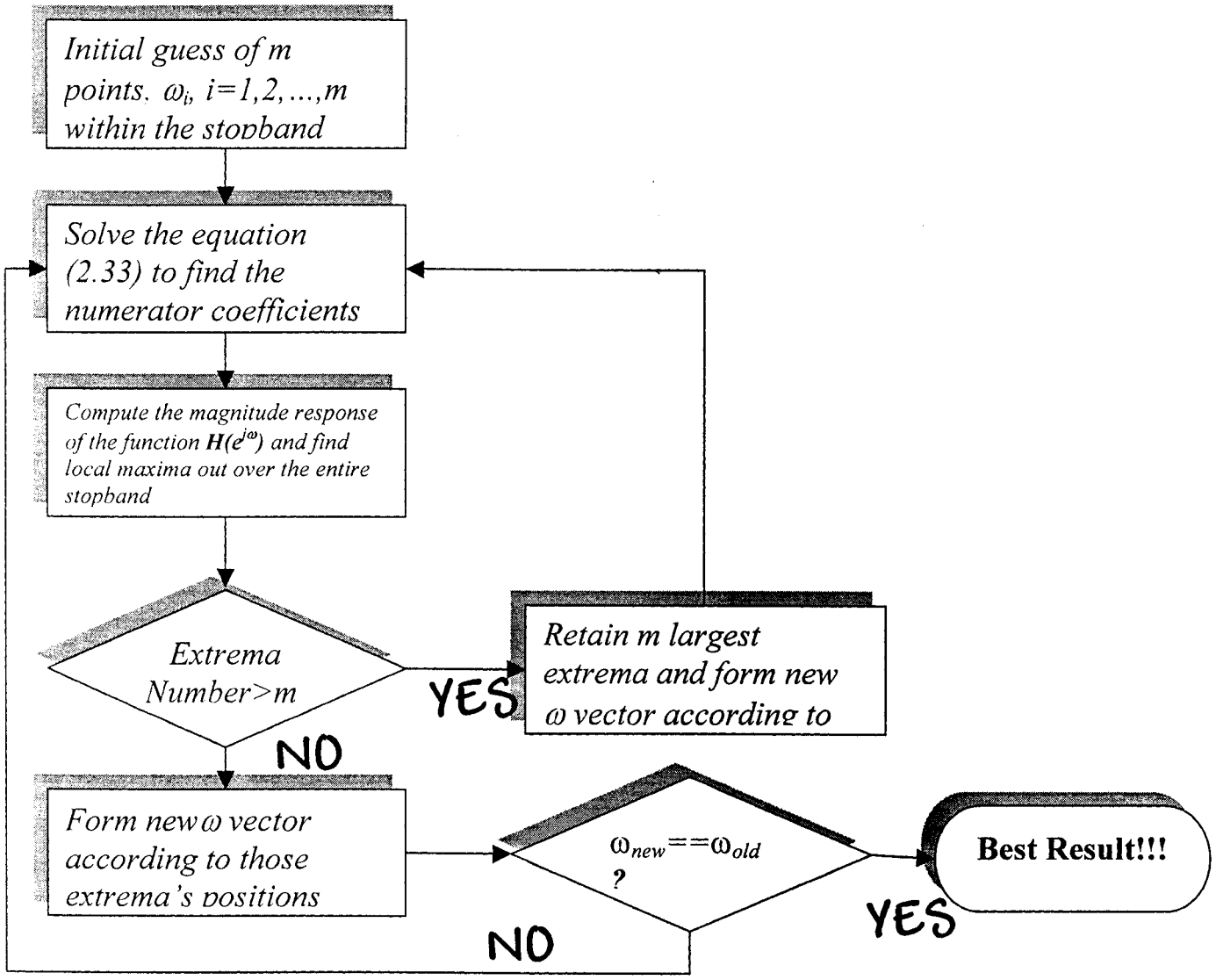
$$d=\begin{bmatrix} D_{\omega_b} & D^{(0)} & D^{(2)} & D^{(4)} & \cdots & D^{(M)} & 0 & 0 & \cdots & 0 \end{bmatrix}^T \quad (2.36)$$

where $D_i=(-1)^{i+1}\left|D(e^{jw_i})\right| \qquad D^{(0)}=\left|D(e^{j0})\right|$

$$D_{w_b}=0.7079\left|D(e^{jw_b})\right| \qquad D^{(k)}=\left.\frac{d^k\left|D(e^{jw})\right|}{dw^k}\right|_{w=0}$$

We see that there are *((L+2m+3)/2)* equations in *(2.33)* which are linear equations in the *(p+2)* variables ***b 0, b 1,...,b$_p$,δ***. By choosing *p =((L+2m-1)/2)*, we can obtain a unique solution to *(2.33)*. So we get values of the unknown variables ***b$_0$,b$_1$,...,b$_p$***. According to *N(z)=z^{-p}N$_a$(z)* and *(2.18)*, there comes out the numerator coefficients. But, notice that the values of ω are still initialized ones, not exactly the final optimal results at this time. The locations of maximum deviations or extrema occur are not known in advance. Hence the set of equations *(2.33)* have to be solved recursively, using the ***Remez Exchange Algorithm*** (which will be stated in detail later) by starting with an initial guess for extrema points: ω_1, ω_2... ω_m.

2.3.1.2 Remez Exchange Algorithm



2.3.1.3 Example:

A lowpass filter with passband flatness L=11 at ω=0, and a bandwidth ω_b=0.24π, stopband cutoff frequency ω_s=0.32π, and an equiripple response with a minimum attenuation of 30 dB in the stopband. The denominator is chosen to be a 7th order polynomial which provides a group delay τ=5. There are 10 ripples within the stopband (m=10).

By using the *Remez Exchange Algorithm*:

The ω vector becoming convergent means that the ***SSE*** (sum of squared error) between itself and its previous version ω' *(SSE=sum((ω'-ω).^2))* is less than a specified value, for instance, 10^-6, or else, equals to zero while they are exactly identical.

Loop	Extrema Number	SSE(Sum of Square Error)
1	10	6.41619549323e-002
2	10	2.91181054022e-002
3	10	6.00922161085e-003
4	10	9.19178538126e-005
5	10	1.47068566100e-007
6	10	0
7	10	0
...	...	...

Table 2.1: The ω vector's convergence

How the **SSE** of the ω vector approaches its best estimation is illustrated by Table 1. After the 6^{th} loop, the ω vector becomes a constant vector and positions of extrema have been fixed. At this time, the magnitude response is illustrated at figure 2.1. Figure 2.2 is the enlarged version of its stopband. Attenuation of stopband is about 31dB. Now the stopband is exactly the "equiripple". Figure 2.3 and 2.4 are Zero-Pole plot and group delay response of this filter, respectively.

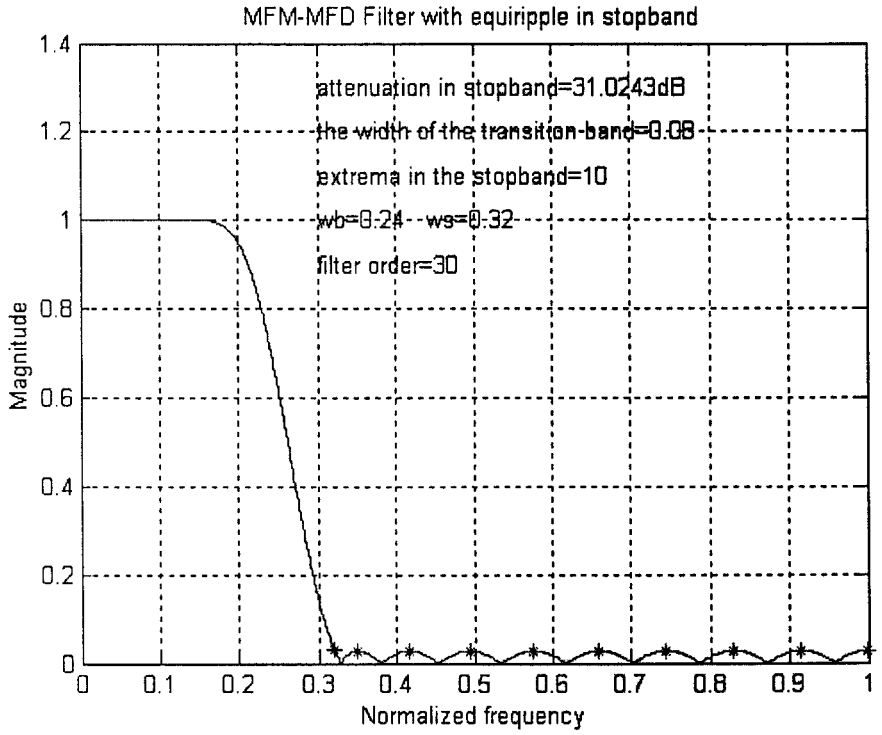


Figure 2.1 Magnitude response of MFM-MFD Filter with equiripple in stopband

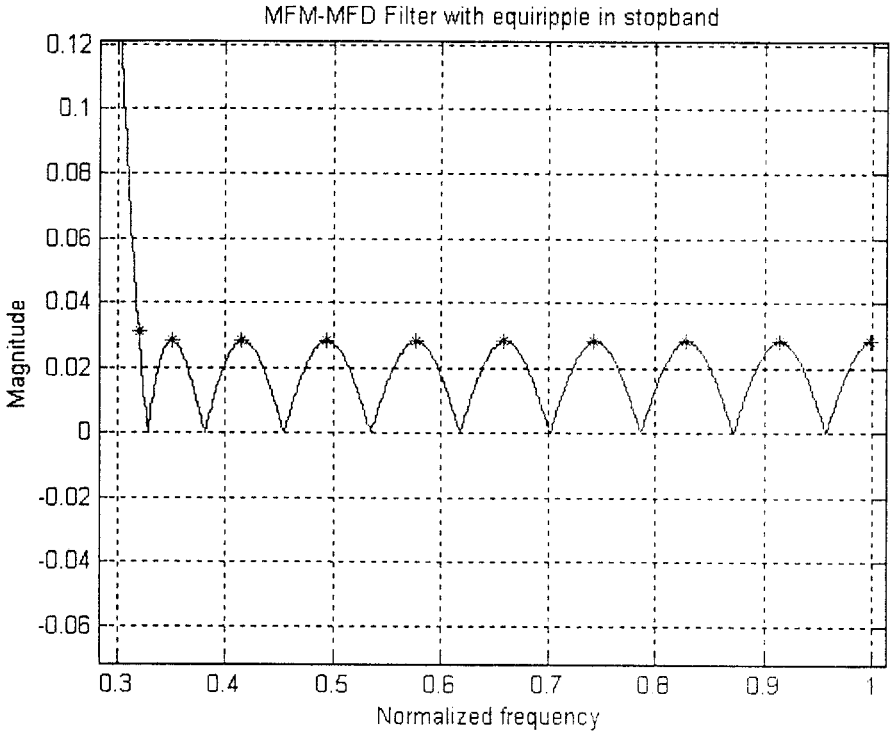


Figure 2.2 Magnitude response (stopband)

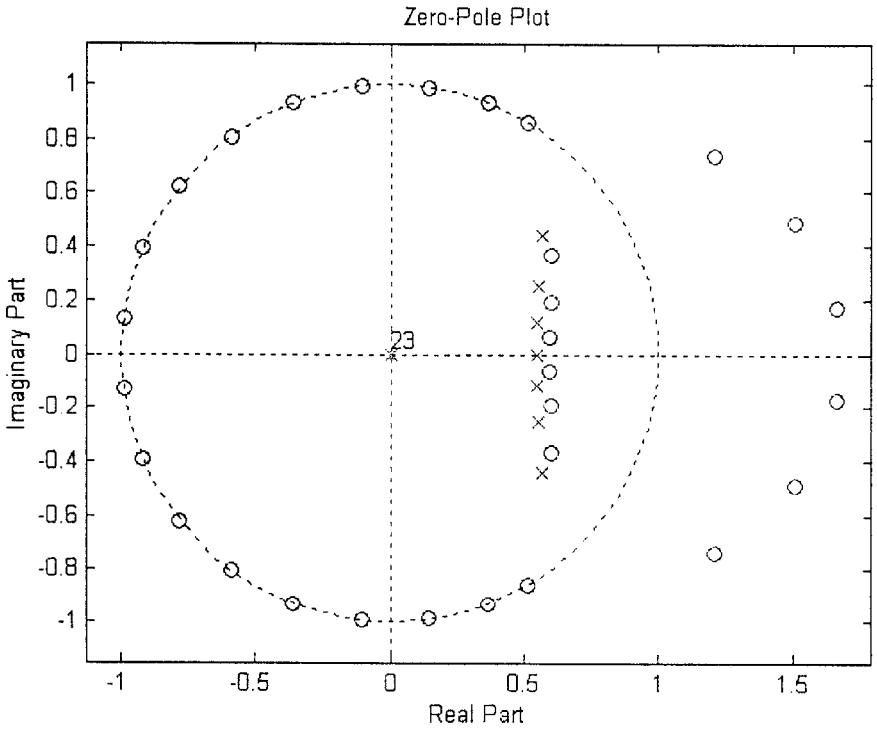


Figure 2.3 Zero-Pole of MFM-MFD IIR Equiripple filter

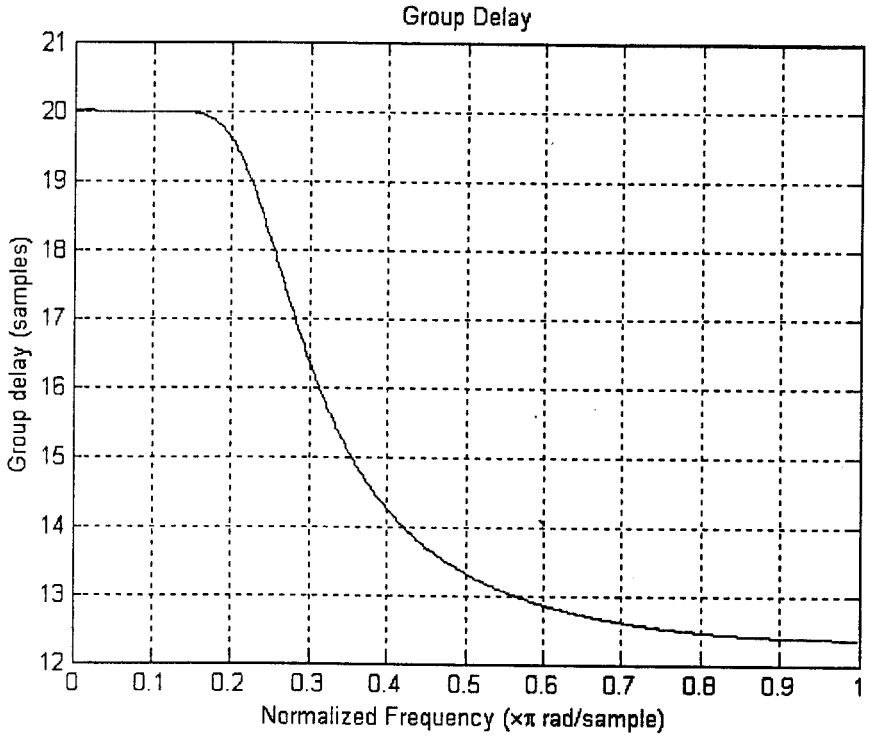


Figure 2.4 Group Delay of MFM-MFD IIR Equiripple filter

2.3.2 Maximally Flat Delay Filter (MFD) with Equiripple both in Passband and Stopband

2.3.2.1 Design Theory

In the above design theory we specify only equiripple in stopband. But actually with the same target of inserting equiripple in stopband, we can choose inserting ripples into both passband and stopband while the maximally flat delay response is maintained.

By purposely adding some ripples in the passband region, we can narrow the transition band, or else, with the same prescriptions, tremendously decrease the order of the digital filter.

Let the passband region of the modified filter be denoted by $0<\omega<\omega_b$. The desired magnitude response to be approximated by the modified filter is given by

$$\left|H_d(e^{j\omega})\right| = 1 \qquad 0 < \omega < \omega_b \qquad (2.37)$$

In order to obtain an equiripple magnitude response over this passband region, we require that magnitude response in passband satisfies following conditions:

$$\left|\left|H_d(e^{j\omega_i})\right| - \left|H(e^{j\omega_i})\right|\right| = (-1)^{i+1}\delta \qquad i = 1,2,\ldots,n$$
$$\text{where} \qquad 0 \le \omega_1 < \omega_2 < \cdots < \omega_n \le \omega_b \qquad (2.38)$$

n is the number of desired extrema in the passband region. In (2.38), n is even. The right part of this equation (2.38) becomes $(-1)^i\delta$ provide that n is odd. Here assuming the first situation is default.

From $H(z) = \dfrac{z^{-p}N_a(z)}{D(z)}$, and the equations *(2.37)* and *(2.38)*,

$$\left|\left|N_a(e^{j\omega_i})\right| + (-1)^{i+1}\delta\left|D(e^{j\omega_i})\right|\right| = \left|D(e^{j\omega_i})\right| \qquad i = 1,2,\ldots,n \qquad (2.39)$$

According to *(2.19)* and *(2.39)*, we get

$$b_0 + b_1 \cos\omega_i + b_2 \cos(2\omega_i) + \ldots + b_p \cos(p\omega_i) + (-1)^{i+1}\delta \left| D(e^{j\omega_i}) \right| = \left| D(e^{j\omega_i}) \right| \qquad i = 1,2,\ldots,n \qquad (2.40)$$

Hence the equation *(2.28)*, *(2.26)*, *(2.40)* and *(2.32)* can be expressed as:

$$Ab = d \qquad (2.41)$$

matrix A and vectors b, d are shown below:

$$A = \begin{bmatrix} 1 & \cos\omega_b & \cos 2\omega_b & \cdots & \cos p\omega_b & 0 & 0 \\ 1 & 1 & 1 & \cdots & 1 & 0 & 0 \\ 1 & \cos\omega_1' & \cos 2\omega_1' & \cdots & \cos p\omega_1' & 0 & D_1' \\ 1 & \cos\omega_2' & \cos 2\omega_2' & \cdots & \cos p\omega_2' & \vdots & D_2' \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & \cos\omega_n' & \cos 2\omega_n' & \cdots & \cos p\omega_n' & 0 & D_n' \\ 1 & \cos\omega_1 & \cos 2\omega_1 & \cdots & \cos p\omega_1 & D_1 & 0 \\ 1 & \cos\omega_2 & \cos 2\omega_2 & \cdots & \cos p\omega_2 & D_2 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & 0 \\ 1 & \cos\omega_m & \cos 2\omega_m & \cdots & \cos p\omega_m & D_m & 0 \end{bmatrix} \qquad (2.42)$$

$$b = \begin{bmatrix} b_0 & b_1 & b_2 & \cdots & b_p & \delta_s & \delta_p \end{bmatrix}^T \qquad (2.43)$$

$$d = \begin{bmatrix} D_{\omega_b} & D^{(0)} & D_1' & \cdots & D_n' & 0 & 0 & \cdots & 0 \end{bmatrix}^T \qquad (2.44)$$

where $D_i = (-1)^{i+1}\left|D(e^{j\omega_i})\right| \qquad D^{(0)} = \left|D(e^{j0})\right|$

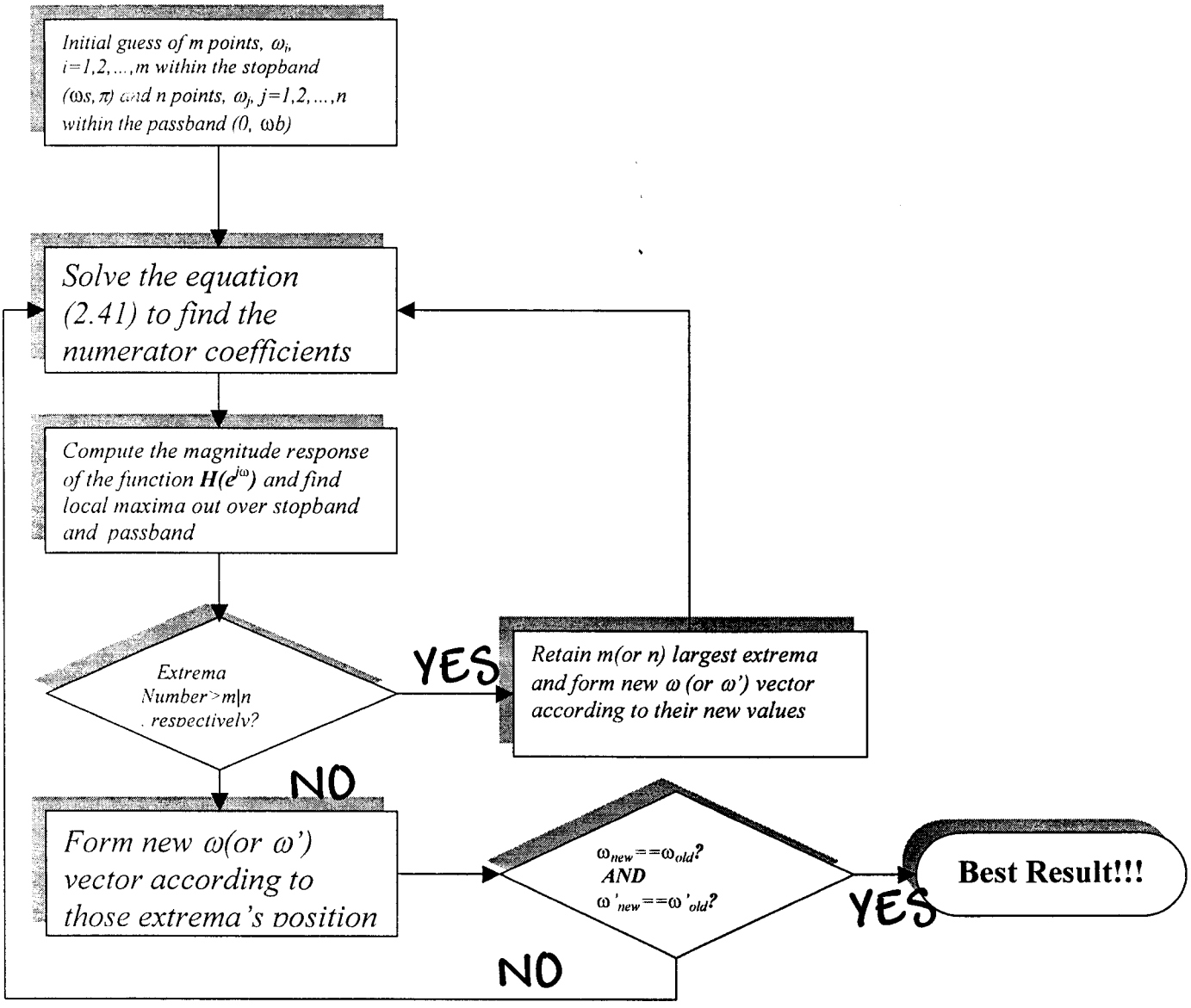
$D_{\omega_b} = 0.7079\left|D(e^{j\omega_b})\right| \qquad D_i' = (-1)^{i+1}\left|D(e^{j\omega'_i})\right|$

The symbol ω and ω' denotes stopband frequencies and passband frequencies, respectively.

We see that there are *(m+n+2)* equations in *(2.41)* which are linear equations in the *(p+3)* variables $\boldsymbol{b0, b1,\ldots,b_p,\delta_s,\ \delta_p}$. By choosing $p=(m+n-1)$, we can obtain a unique solution to *(2.41)*, unknown variables $\boldsymbol{b_0, b1,\ldots,b_p}$. According to $\boldsymbol{N(z)=z^{-p}N_a(z)}$ and *(2.18)*, there comes out the numerator coefficients. But, similar to the previous design, the values of ω at which the maximum deviations or extrema occur are not known in advance. Hence the set of equations *(2.41)* have to be solved recursively, using the ***Remez Exchange Algorithm*** by starting with an initial guess for the extrema points:$\omega_1,\ \omega_2\ldots\omega_n$ in the passband and: $\omega_1,\ \omega_2\ldots\omega_m$ in the stopband.

This method of obtaining equiripple both in the passband and stopband is similar with but more complex than previous design during optimization because both ripples in passband and stopband have to be processed simultaneously.

2.3.2.2 Remez Exchange Algorithm



2.3.2.3 Example:

A lowpass filter with 4 ripples (n=4) in the passband and other specifications are same as example 2.3.1.3. Bandwidth $\omega_b=0.24\pi$, stopband cutoff frequency $\omega_s=0.32\pi$, and an equiripple response with a minimum attenuation of 30 dB in the stopband. The denominator is chosen to be a 20^{th} order polynomial which provides a group delay $\tau=5$. There are 7 ripples within the stopband (m=7).

Loop	Extrema Number in Passband	Extrema Number in Stopband	SSE(Sum of Square Error)
1	4	7	0.25544448675435
2	4	7	0.21778383887520
3	4	7	0.09763470311403
4	4	7	0.02982785830217
5	4	7	0.00101653792888
6	4	7	1.882477646081789e-005
7	4	7	0
...	...	...	...

Table 2.2: ω and ω' vector's convergence

Letting SSE=sum((ω_{new}-ω_{old}).^2)+sum((ω'_{new}-ω'_{old}).^2), table 2.2 illustrates ω *and* ω' becoming convergent gradually. Figure 2.5, 2.6 and 2.7 are results for this example.

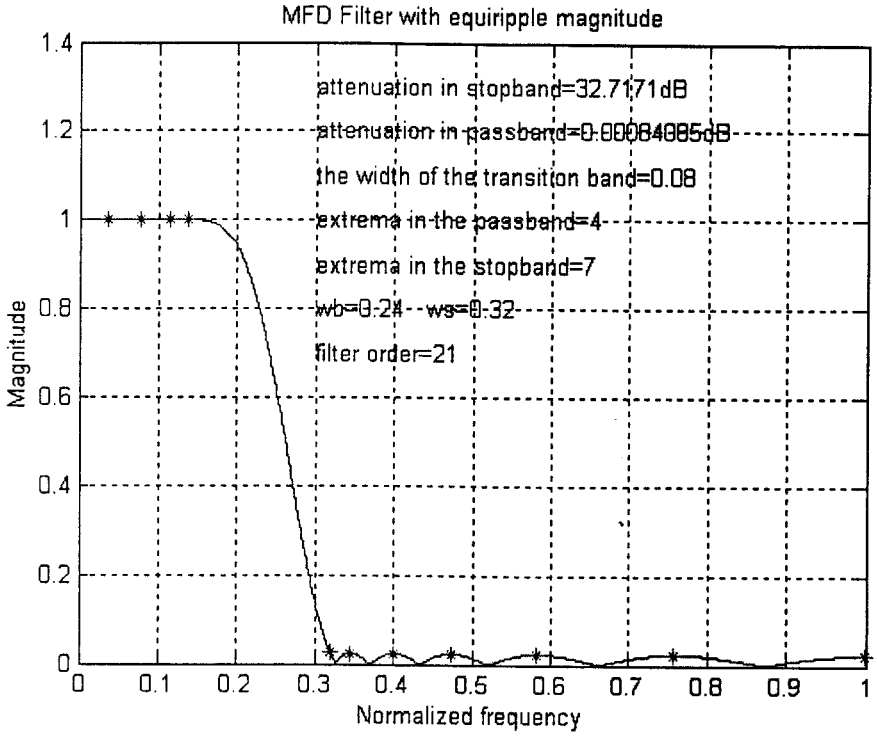


Figure 2.5 Magnitude Response of MFD filter with equiripple in pass and stop band

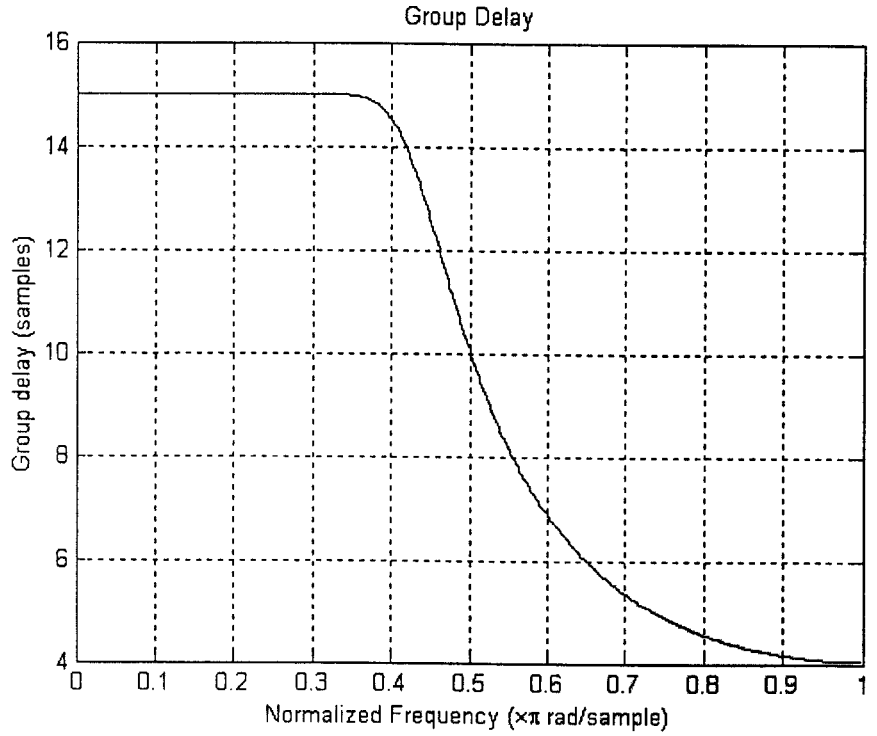


Figure 2.6 Group Delay of MFD filter with equiripple in pass and stop band

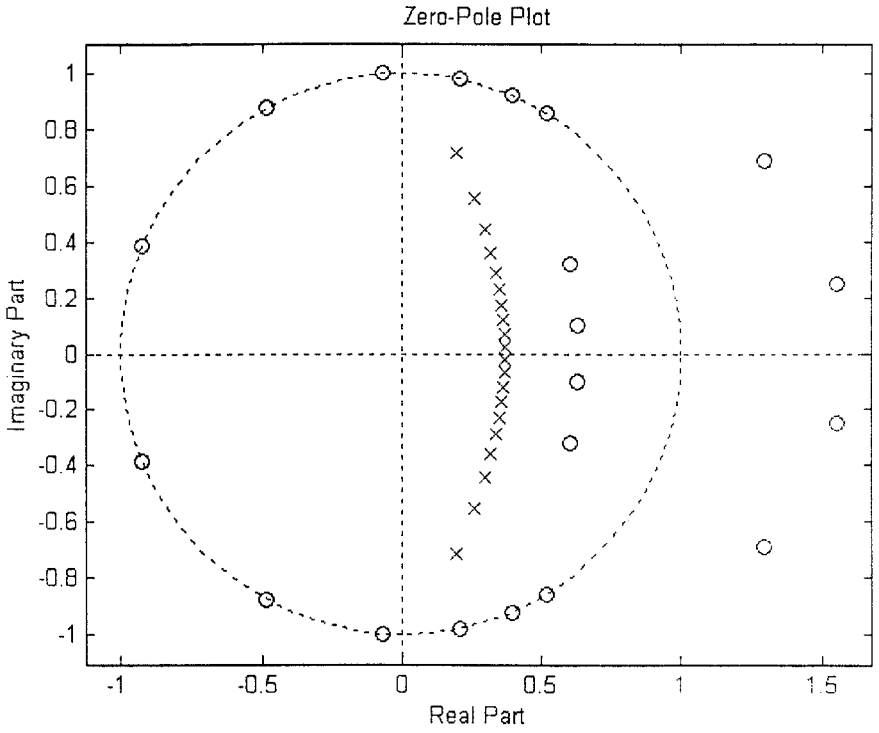


Figure 2.7 Zero-Pole Plot of MFD filter with Equiripple in pass and stop band

2.3.3 Comparisons between Two Kinds of Filters

Compared Figure 2.5 with Figure 2.1, by inserting several ripples into the passband, we obtained almost the same characteristics as only inserting ripples into stopband, for some aspects even better. For example, Figure 2.10 shows that the flat band (passband) in this example is wider than its corresponding one which has an unsatisfactory ripple at the passband edge, although at $\omega=0$, it is perfectly flat. In most applications, no matter for magnitude or group delay responses, the most attractive characteristic is the average flatness over passband/stopband since partial perfection is meaningless. In Figure 2.11, we find that the group delay, unlike that of

MFD and equiripple in stopband filter, is state-of-the-art flat all through the entire passband. At the same time, the delay introduced by the numerator polynomial is also decreased (from 15 to 10) because of the reduction of the filter order.

The most inspiring achievement is a tremendously decreased filter order (from 30 to 20). This advantage is most important and valuable in the modern industrial applications. As we stated many time in this thesis, the main purpose of trying to research and develop IIR filters is pursuing a much lower order than FIR filters.

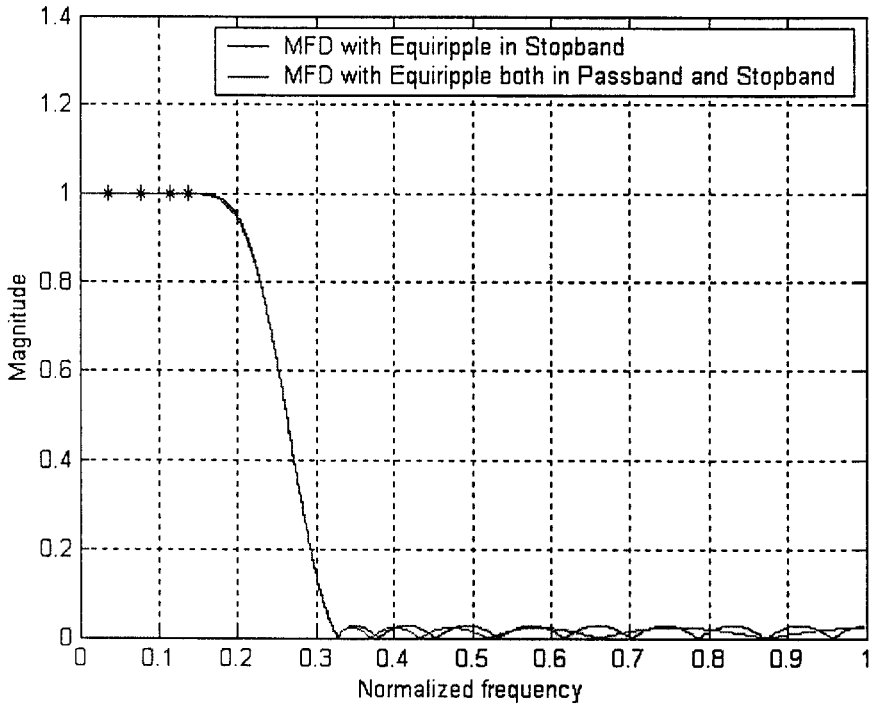


Figure 2.8 Magnitudes comparison

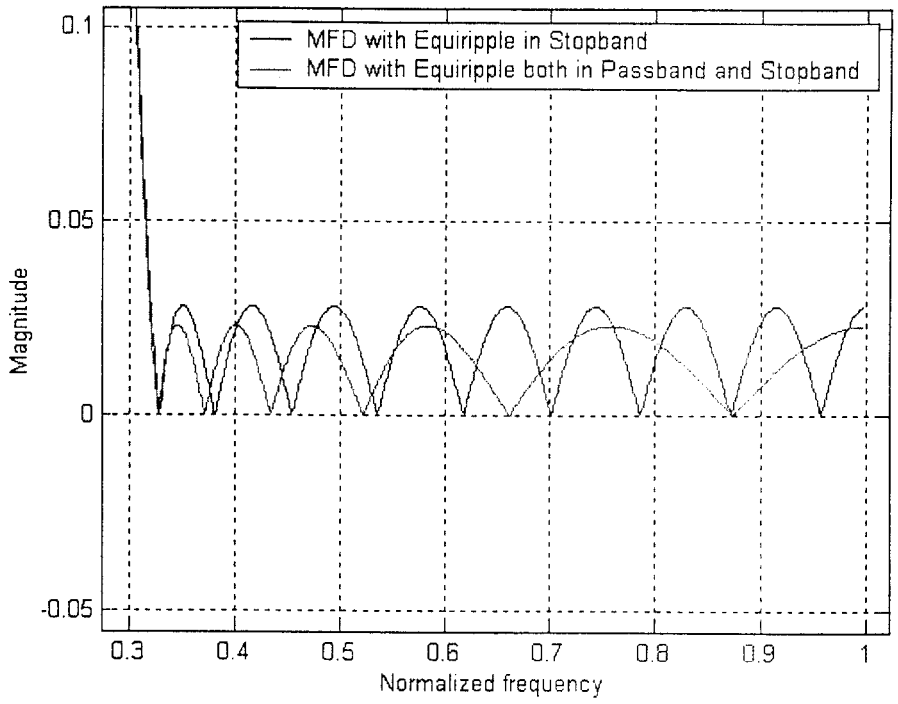


Figure 2.9 Stop band magnitudes comparison

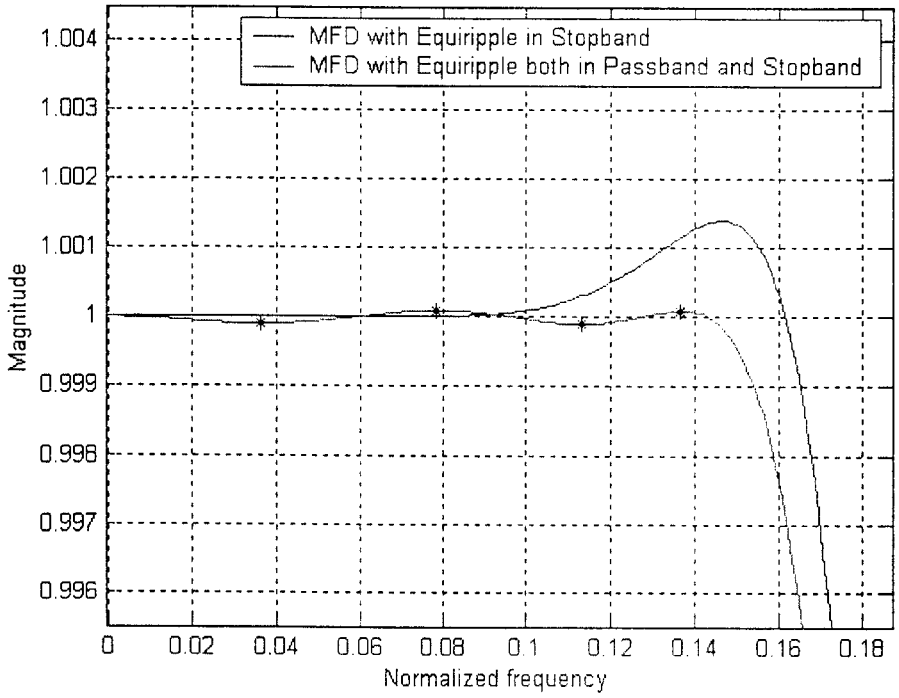


Figure 2.10 Passband magnitudes comparison

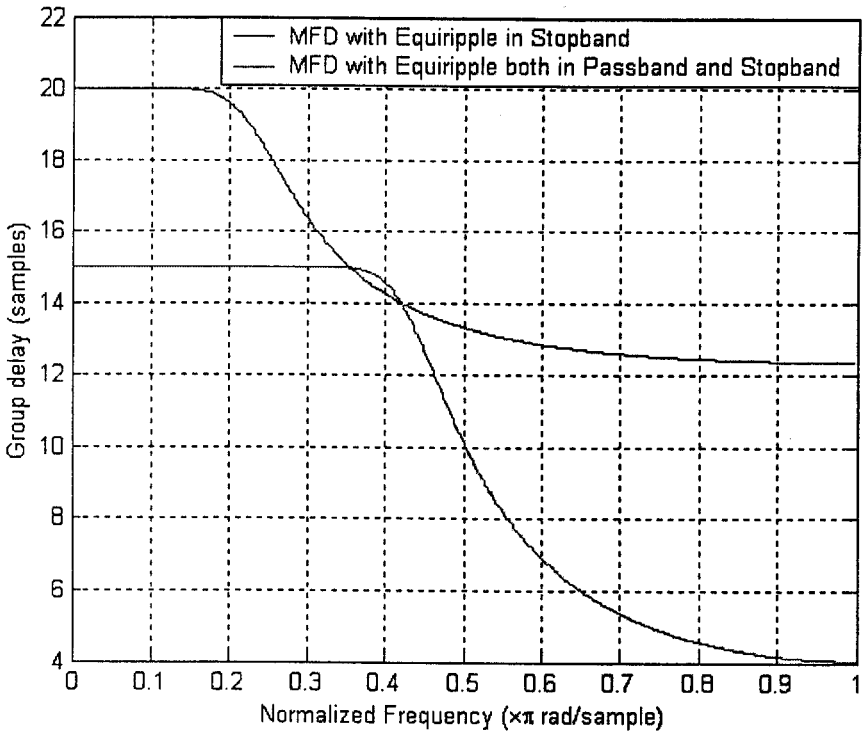


Figure 2.11 Group delay comparison

2.4 Conclusion

This project is to design a *Maximally Flat Magnitude-Maximally Flat Delay IIR filter with equiripple in stopband* and *a Maximally Flat Delay IIR filter with equiripple in stopband and passband.* Its procedure is mostly analytical and its implementation is explicit. By purposely inserting several zeros into the MFM-MFD filter's stopband and properly settling those zeros' position, we can compress the width of transition band efficiently. By trying different values for L and *ws, wb, the number of extrema*, we are able to get a lot of high quality IIR equiripple filters. When the variable z^{-1} is replaced by $-z^{-1}$,

the magnitude of the lowpass filter with a passband cutoff frequency *wb* changes to that of a highpass filter with a cutoff frequency $wb'=\pi-wb$. This method can tremendously enlarge the application range of this project.

The second filter is extension of the first one give in [12]. By replacing the flat passband by an equiripple one, the performance of the IIR filter is remarkably improved, for example, a narrower transition band and/or a much lower filter order. It is a wise choice to make numerator and denominator have identical length for the purpose of attaining better filter performance without an undesirable filter order increase.

There always exists tradeoff among all those requirements, such as transition bandwidth, filter order, group delay flatness, attenuation in the passband and/or stopband, and so on. For instance, in example given in 2.3.2.3, reducing transition band from 0.08 to 0.06 results in degradations both in passband and stopband (see Figure 2.12). Contrarily, better passband and stopband attenuations cost a broader transition band from 0.08 to 0.10 (Figure 2.13).

By carefully adjusting those parameters of this analytical algorithm, we can always obtain a top-ranking filter that meets reasonable requirements.

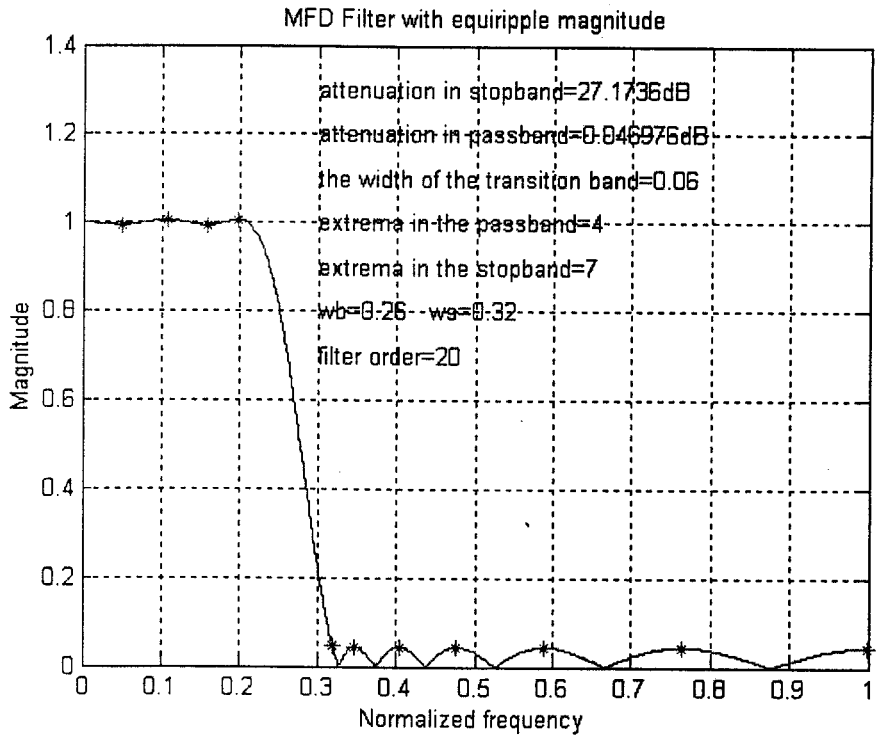


Figure 2.12 Magnitude degradations in passband and stopband caused by a narrower transition band

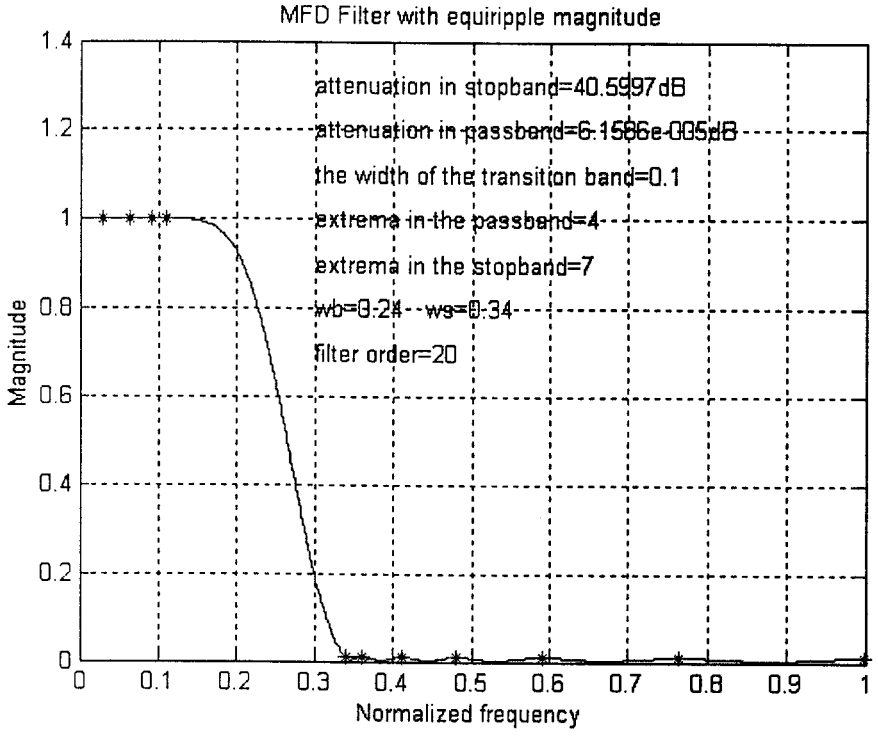


Figure 2.13 Magnitude improvements in passband and stopband caused by a broader transition band

By using Symbolic Math Toolbox of Matlab, the quality and accuracy of this program have been promoted evidently.

Chapter 3

Design of Approximately Linear Phase IIR Digital Filter Using Allpass Sections in Parallel

3.1 Introduction

In recent years, a number of digital filters composed of allpass subfilters have been developed for various applications. An important application of allpass networks is in the realization of a class of approximately linear phase IIR filters proposed by Renfors and Saramaki [16]. They showed how a simple parallel connection of an allpass digital filter with a chain of delay elements of certain length can be used for designing filters that possess, inherently, an approximately linear phase in frequency-selective designs. Compared to linear phase FIR digital filters the filter can generate only an approximately linear phase but with a shorter signal delay. This network, called *approximately linear phase filter composed of allpass filters in parallel* in this chapter, belongs to the major family of allpass-based IIR digital filters consisting of two allpass networks.

Allpass-based digital filters are characterized by numerous attractive features that may even seem to be of contradictory nature in view of other available structures. Listed below are the most important ones.

1. Parallelism of the structure can be utilized to speed up the processing rate.

2. In allpass-based realization of an IIR filter, if possible, the poles are distributed between two allpass subfilters. Assume that the two subfilters are real and of orders M and N. Since the numerator of the original system function should be symmetric, we need approximately 1.5(M+N) multipliers for the direct-form realization. On the other hand, a real allpass filter, can be canonically realized with the same number of multipliers as its order. Therefore, the parallel allpass realization requires only M+N multipliers

3. Low coefficients sensitivity. Quantizing filter coefficients can have serious effects on the performance of digital filters. As a result of coefficient quantization, the frequency response of the filter with quantized coefficients can be significantly different from the desired filter without quantized coefficients. In some cases, the performance of the quantized filter can make it unsuitable. Figure 3.1, 3.2 and 3.3 give an illustration for this problem. Figure 1 is an elliptic IIR lowpass filter implemented in direct form II structure. After being quantized with 16-bit word length, its performance has

been too bad to applicable (Figure 2). Moreover, zero-pole shifting might give risks of instability (Figure 3.3).

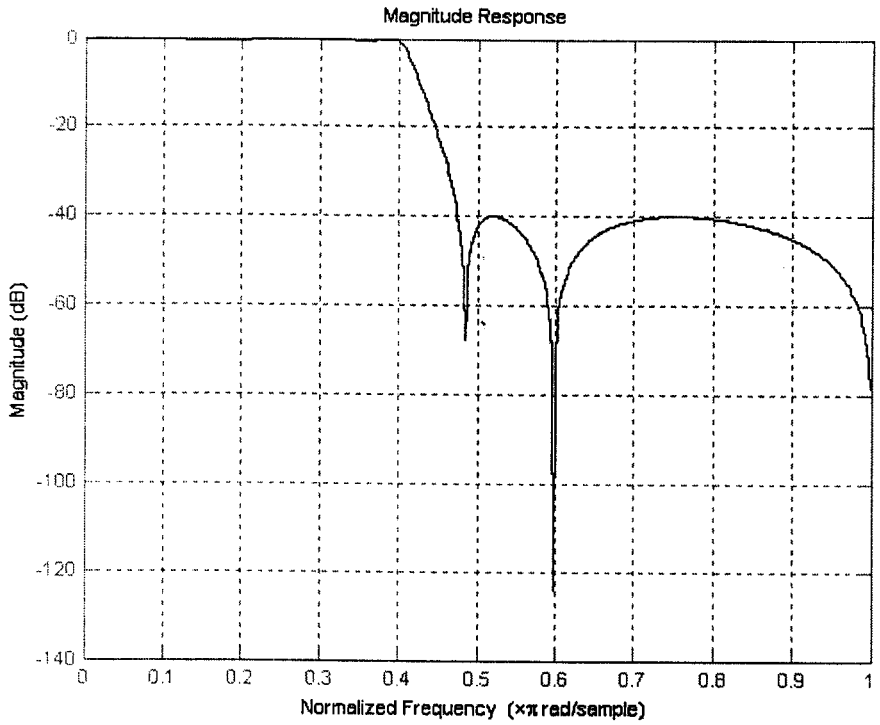


Figure 3.1 Infinite coefficients accuracy filter

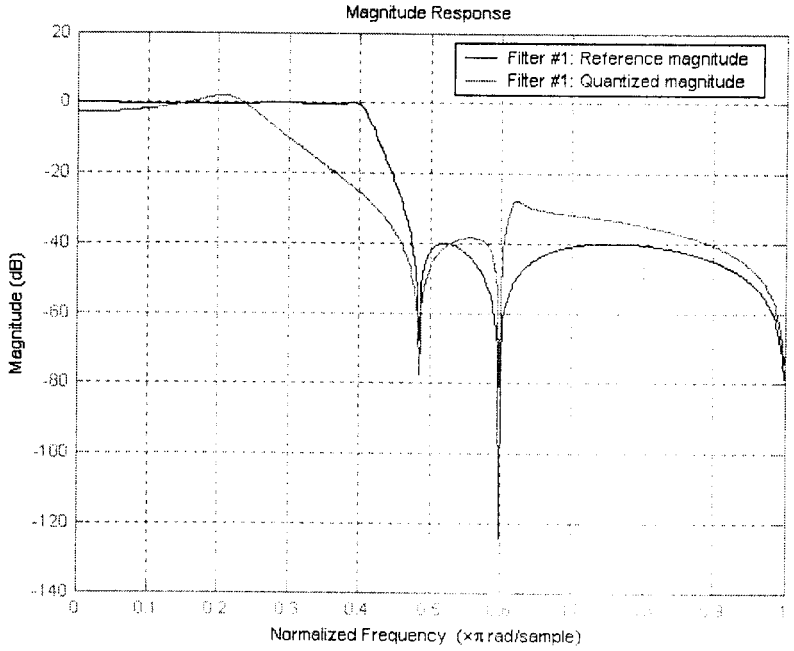


Figure 3.2 Magnitude comparison between infinite accuracy filter and quantized filter (Direct Form)

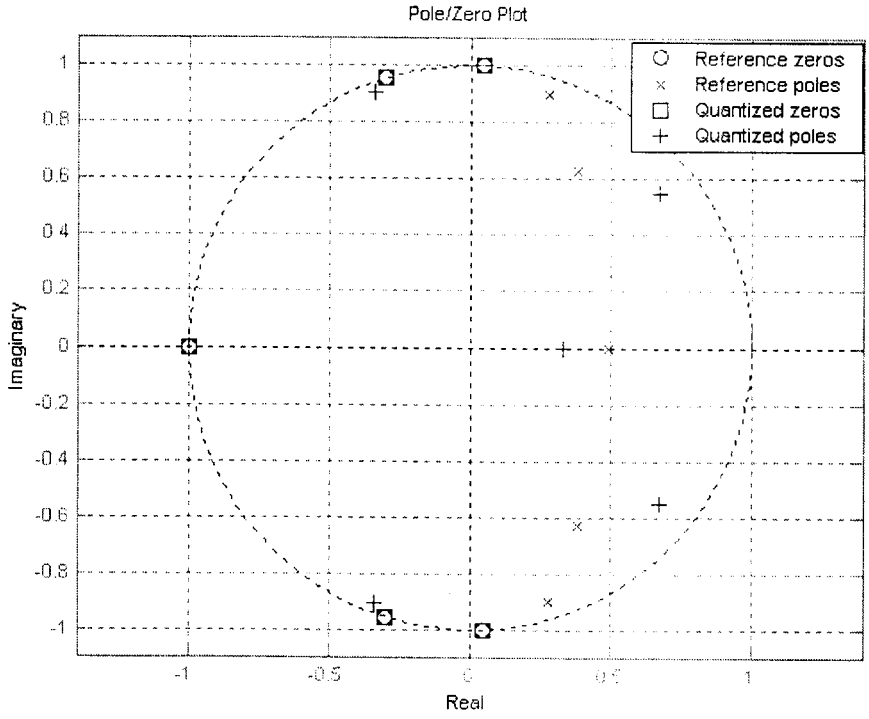


Figure 3.3 Zero-Pole comparison between infinite accuracy filter and quantized filter

Low sensitivity filter architectures, or robust architectures as they are sometimes called, are interesting because they can reduce the effects of coefficient quantization. By being inherently less sensitive to coefficient quantization, these filter architectures withstand the quantization process and result in filters that retain the performance of the original filter. This is illustrated in Figure 3.4. Converting structure from direct form to lattice coupled-allpass form, the word length effect has been weakened to a neglectable level.

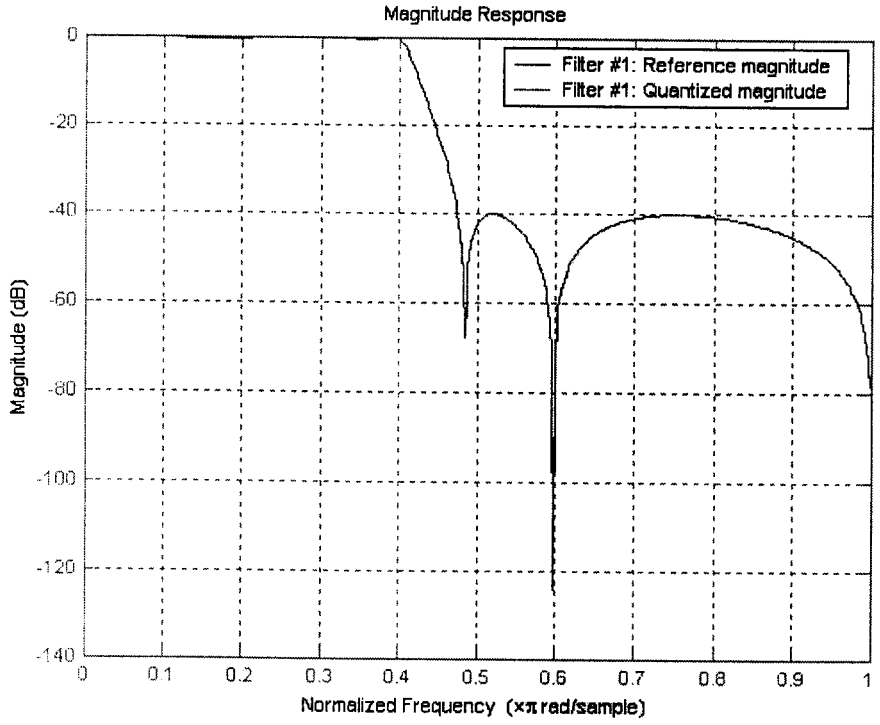


Figure 3.4 Magnitude comparison between infinite accuracy filter and quantized filter (Lattice Coupled-Allpass Form)

Generally, filter architecture sensitivity ranges from high to low as below:

- Direct forms
- Lattice forms
- Allpass forms

This is because allpass subfilters can be realized structurally lossless. Thus, they will remain allpass in spite of multiplier quantization and their sum or difference is a bounded real filter. Consequently, the overall structure becomes of very low sensitivity in the passband and reasonably low sensitivity in the stopband. This means that we need fewer bits per multiplier coefficient and the economic advantage is twofold.

4. The complementary filter is obtained from the original one by simply changing the sign of one of the allpass sections. Therefore, a complementary filter pairs, for example, lowpass/highpass, bandstop/bandpass, can be implemented from the unique structure.

In this project, a digital filter that consists of two allpass filters in parallel is considered. One of them is a pure delay that ensures a good phase performance for the overall filter in the passband. Remez algorithm is used on different objective functions (error functions) for minimizing them in equiripple sense. The first design minimizes an error function of the direct phase response of the objective allpass filter. The second error function is an indirect phase response of that allpass filter. The general idea of those

two procedures is identical but from results, we still can tell some differences.

Because either magnitude or phase response can be regarded as minimizing objective in this application, using different combinational analyses of them on different frequency bands is reasonable for meeting various requirements.

3.2 Design Theory

Let's consider a digital filter that consists of two allpass filters in parallel shown in figure 3.5.

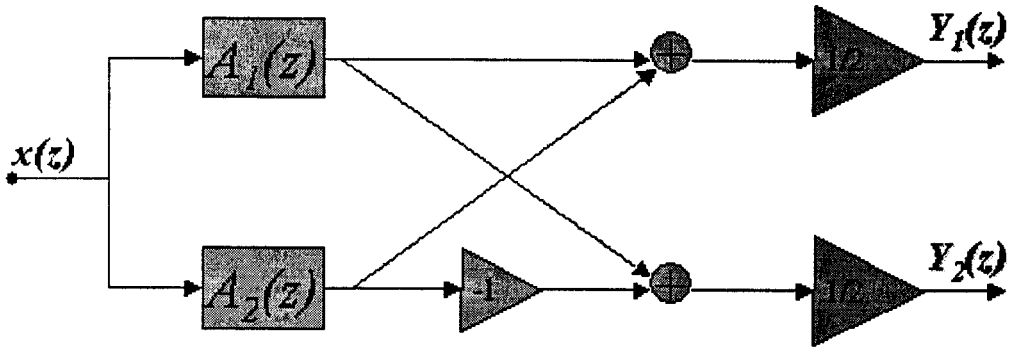


Figure 3.5 Two Allpass Filters in Parallel

The transfer function $A_1(z)$ and $A_2(z)$ are allpass functions and from them, two transfer functions can be obtained as:

$$G(z) = \frac{Y_1(z)}{X(z)} = \frac{1}{2}\left[A_1(z) + A_2(z)\right] \qquad (3.1)$$

$$H(z) = \frac{Y_2(z)}{X(z)} = \frac{1}{2}\left[A_1(z) - A_2(z)\right] \qquad (3.2)$$

Since $A_1(z)$ and $A_2(z)$ are assumed to be allpass functions with real coefficients, they can be written as

$$A_1(z) = z^{-(N-r)}\frac{D_1(z^{-1})}{D_1(z)} \qquad (3.3)$$

and

$$A_2(z) = z^{-r}\frac{D_2(z^{-1})}{D_2(z)} \qquad (3.4)$$

Therefore

$$G(z) = \frac{1}{2}\left[\frac{z^{-(N-r)}D_1(z^{-1})D_2(z) + z^{-r}D_1(z)D_2(z^{-1})}{D_1(z)D_2(z)}\right] \qquad (3.5)$$

$$H(z) = \frac{1}{2}\left[\frac{z^{-(N-r)}D_1(z^{-1})D_2(z) - z^{-r}D_1(z)D_2(z^{-1})}{D_1(z)D_2(z)}\right] \qquad (3.6)$$

If we represent them as

$$G(z) = \frac{P(z)}{D(z)} = \frac{\sum_{n=0}^{N} p_n z^{-n}}{D(z)} \qquad (3.7)$$

$$H(z) = \frac{Q(z)}{D(z)} = \frac{\sum_{n=0}^{N} q_n z^{-n}}{D(z)} \qquad (3.8)$$

Then we can show that the following conditions are satisfied.

Property (1) $P(z^{-1})=z^{N}P(z)$. Hence $p_n=p_{N-n}$. The coefficients of P(z) are symmetric.

Property (2) $Q(z^{-1})=-z^{N}Q(z)$. Hence $q_n=-q_{N-n}$. The coefficients of Q(z) are anti-symmetric.

Property (3) $P(z)P(z^{-1})+Q(z)Q(z^{-1})=D(z)D(z^{-1})$. Hence $|G(e^{j\omega})|^2+|H(e^{j\omega})|^2=1$. G(z) and H(z) are said to form a power complementary pair.

Property (4) $\left|G(e^{j\omega})\right| = \frac{1}{2}\left|e^{j\theta_1(\omega)} + e^{j\theta_2(\omega)}\right| = \frac{1}{2}\left|1 + e^{j(\theta_1(\omega)-\theta_2(\omega))}\right| \le 1$

Assuming that the above four conditions are satisfied and the derivation of (3.1) and (3.2) can be obtained below.

Consider $P(z)P(z^{-1})+Q(z)Q(z^{-1})=D(z)D(z^{-1})$. Using Properties (1) and (2) we get

$$P(z)z^{N}P(z) - z^{N}Q(z)Q(z) = D(z)D(z^{-1}) \qquad (3.9)$$

$$P^2(z) - Q^2(z) = D(z)z^{-N}D(z^{-1}) \qquad (3.10)$$

$$\left[P(z)+Q(z)\right]\left[P(z)-Q(z)\right] = z^{-N}D(z)D(z^{-1}) \qquad (3.11)$$

Since $\left[P(z^{-1})+Q(z^{-1})\right]=z^{N}\left[P(z)-Q(z)\right]$, we get

$$\left[P(z)+Q(z)\right]z^{-N}\left[P(z^{-1})+Q(z^{-1})\right]=z^{-N}D(z)D(z^{-1}) \qquad (3.12)$$

and also the result that the zeros of [P(z)-Q(z)] are reciprocals.

We shall assume that G(z) is asymptotically stable. Hence from Property (4) we infer that G(z) has no poles on the unit circle. In other words, the zeros of D(z) are within the unit circle and the zeros of $D(z^{-1})$ are outside the unit circle. So also the zeros of [P(z)+Q(z)] are not on the unit circle. Let us assume that D(z) has r zeros z_k (k=1,2,...,r) that are outside the unit circle.

Thus we will assume the polynomial D(z) in the form

$$D(z)=\prod_{k=1}^{r}\left(1-z^{-1}z_k\right)\prod_{k=r+1}^{N}\left(1-z^{-1}z_k^{-1}\right) \qquad (3.13)$$

then we can also derive

$$\left[P(z)+Q(z)\right]\left[P(z)-Q(z)\right]=z^{-N}D(z)D(z^{-1})=$$
$$\prod_{k=1}^{r}\left(1-z^{-1}z_k\right)\prod_{k=r+1}^{N}\left(1-z^{-1}z_k^{-1}\right)\prod_{k=1}^{r}\left(z^{-1}-z_k\right)\prod_{k=r+1}^{N}\left(z^{-1}-z_k^{-1}\right) \qquad (3.14)$$

Thus we identify

$$[P(z)+Q(z)]=\alpha\prod_{k=1}^{r}\left(1-z^{-1}z_k\right)\prod_{k=r+1}^{N}\left(1-z^{-1}z_k^{-1}\right) \qquad (3.15)$$

$$[P(z)-Q(z)]=\frac{1}{\alpha}\prod_{k=1}^{r}\left(1-z^{-1}z_k\right)\prod_{k=r+1}^{N}\left(z^{-1}-z_k^{-1}\right) \qquad (3.16)$$

then

$$G(z)+H(z)=\frac{P(z)+Q(z)}{D(z)}=\alpha\prod_{k=r+1}^{N}\frac{z^{-1}-z_k^{-1}}{\left(1-z^{-1}z_k^{-1}\right)}=\alpha A_1(z) \qquad (3.17)$$

$$G(z)-H(z)=\frac{P(z)-Q(z)}{D(z)}=\frac{1}{\alpha}\prod_{k=r+1}^{N}\frac{z^{-1}-z_k}{\left(1-z^{-1}z_k\right)}=\frac{1}{\alpha}A_2(z) \qquad (3.18)$$

From the power complementary property, we must have $\alpha^2=1$. Therefore, $\alpha=1$ so that

$$G(z)=\frac{1}{2}[A_1(z)+A_2(z)] \qquad (3.19)$$

$$H(z)=\frac{1}{2}[A_1(z)-A_2(z)] \qquad (3.20)$$

So here proved that when the four Properties listed above are satisfied, we can synthesize G(z) as the sum of two allpass functions. Indeed, it has been shown that the four Properties are both necessary and sufficient conditions [19].

3.3 Design Procedure

3.3.1 Allpass Sections in Parallel

When $A_2(z)$ is chosen as a pure delay [16] network function, we obtain

$$H_1(z) = \frac{1}{2}\left[A_1(z) + z^{-M}\right] \qquad (3.21)$$

$$H_2(z) = \frac{1}{2}\left[A_1(z) - z^{-M}\right] \qquad (3.22)$$

Where $A_1(z)$ is an allpass transfer function

$$A_1(z) = z^{-N}\frac{\sum_{n=0}^{N} a_n z^n}{\sum_{n=0}^{N} a_n z^{-n}} \qquad (3.23)$$

In this project, we consider a lowpass filter (3.21) in which we select $N=M+1$.

$$H(z) = \frac{1}{2}\left[A_1(z) + z^{-M}\right] \qquad (3.24)$$

The complementary highpass filter is obtained just simply changing the sign of the pure delay filter in (3.24).

3.3.2 Use of Direct Phase Error Function

3.3.2.1 Design Description

Assuming amplitude tolerances in passband and stopband are δ_p and δ_s, respectively. Illustrated as Figure 3.6, this lowpass filter's magnitude specifications can be expressed as:

$$1-\delta_p \le \left|H(e^{j\omega})\right| \le 1 \qquad \text{for } \omega \in [0,\omega_p] \qquad (3.25)$$

$$\left|H(e^{j\omega})\right| \le \delta_s \qquad \text{for } \omega \in [\omega_s,\pi] \qquad (3.26)$$

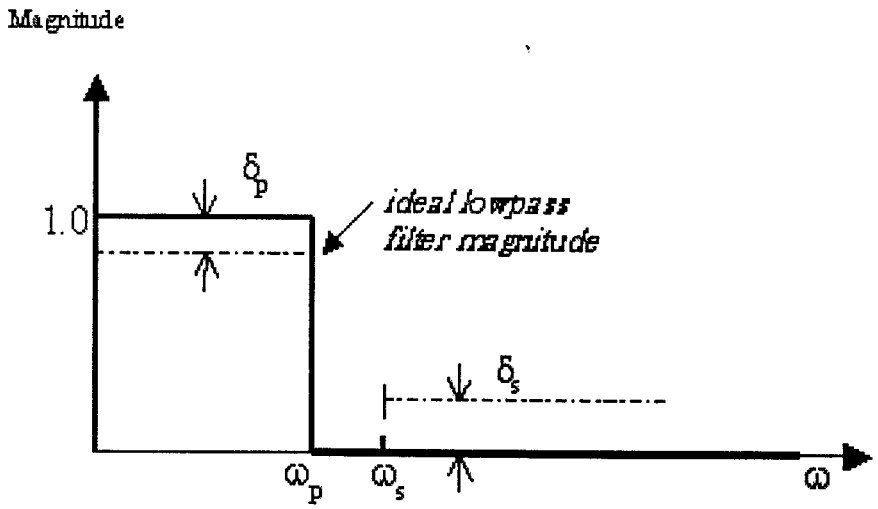


Figure 3.6 Magnitude Response Tolerances

Let $A_1(z)$ be denoted as

$$A_1(z) = z^{\varphi(\omega)} \qquad \varphi(\omega) \text{ is the phase response of } A_1 \qquad (3.27)$$

From (3.22), the filter's magnitude is derived as steps below:

$$H(z) = \frac{1}{2}\left[A_1(z) + z^{-M}\right]$$
$$= \frac{1}{2}\left[z^{\varphi(\omega)} + z^{-M}\right]$$
$$= \frac{1}{2}\left[\cos[\varphi(\omega)] + j * \sin[\varphi(\omega)] + \cos(M\omega) - j * \sin(M\omega)\right]$$

$$|H(z)| = \frac{1}{2}\left|\cos[\varphi(\omega)] + \cos(M\omega) + j * \{\sin[\varphi(\omega)] - \sin(M\omega)\}\right|$$
$$= \frac{1}{2}\sqrt{2 + 2 * (\cos[\varphi(\omega)]\cos(M\omega) - \sin[\varphi(\omega)]\sin(M\omega))}$$
$$= \left|\cos\frac{1}{2}[\varphi(\omega) + M\omega]\right| \qquad (3.28)$$

According to figure 3.6 and (3.28), we notice that, approximately, $\varphi(\omega) + Mw$ equals to 0 in passband and -π in stopband, respectively. Also from **property (4),**

$$|H(e^{j\omega})| = \frac{1}{2}\left|e^{j\theta_1(\omega)} + e^{j\theta_2(\omega)}\right| = \frac{1}{2}\left|1 + e^{j(\theta_1(\omega) - \theta_2(\omega))}\right|$$
$$= \begin{cases} 1 & \theta_1(\omega) \approx \theta_2(\omega) \qquad \text{in passband} \\ 0 & \theta_1(\omega) - \theta_2(\omega) = (2k+1)\pi \quad k = 0, \pm 1, \pm 2, \ldots \qquad \text{in stopband} \end{cases} \qquad (3.29)$$

For an allpass filter described by (3.23), the phase can be derived as:

$$\varphi(\omega) = -N\omega + 2\tan^{-1}\frac{\sum_{i=0}^{N} a_i \sin(i\omega)}{\sum_{i=0}^{N} a_i \cos(i\omega)} \qquad (3.30)$$

Construct an error function to be minimised:

$$E(\omega) = \varphi(\omega) - D(\omega) \qquad (3.31)$$

the desired phase response is

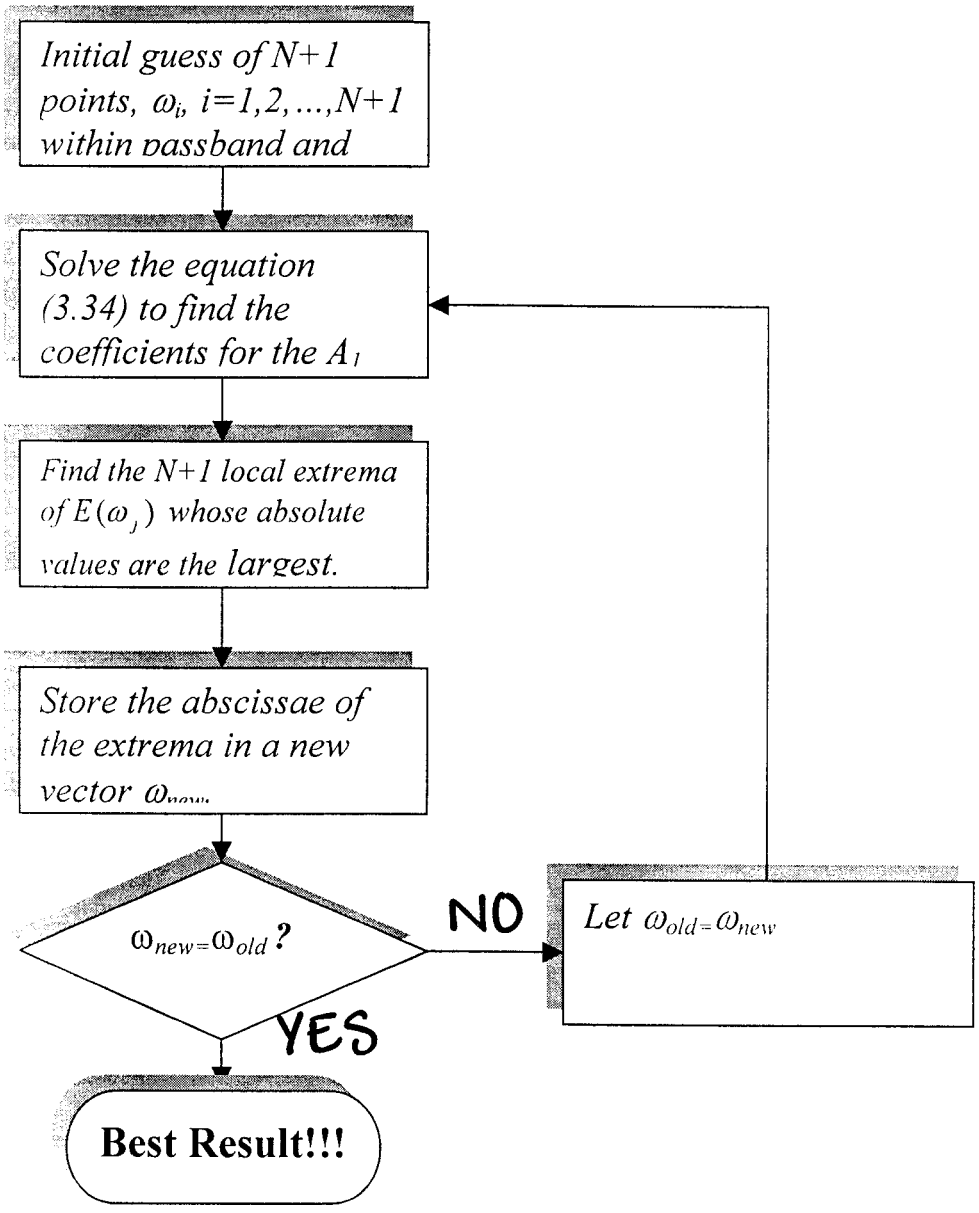
$$D(\omega) = \begin{cases} -M_1\omega & \text{for } \omega \in [0, \omega_p] \quad (3.32) \\ -M_1\omega - \pi & \text{for } \omega \in [\omega_s, \pi] \quad (3.33) \end{cases}$$

Then the approximation target is to minimize the maximum of absolute values of the error function E(ω). It is able to optimize both the magnitude and phase response simultaneously. This target can be fulfilled using ***Remez Algorithm***.

$$E(\omega_j) = \varphi(\omega_j) - D(\omega_j) = (-1)^j \delta \qquad (3.34)$$

δ is ripple height:δp in passband and δs in stopband

3.3.2.2 Design Procedure



3.3.2.3 Example 1:

A lowpass filter with passband bandwidth $\omega_p=0.6\pi$, stopband cutoff frequency $\omega_s=0.75\pi$, and an equiripple response with a minimum attenuation of 40 dB in the stopband. Allpass filter A_l's order is N=11 and choose M=N-1.

By using the *Remez Exchange Algorithm*:
Defining ω vector gets convergence when the maximum difference between the new extrema frequencies vector and the old one is less than ε (for instance, 0.001. We can also implement either a strict or loose convergent condition by using a smaller or larger ε. The result would be different according to different convergent requirements)

Magnitude response and group delay response are plotted as figure 3.7 and 3.8, respectively.

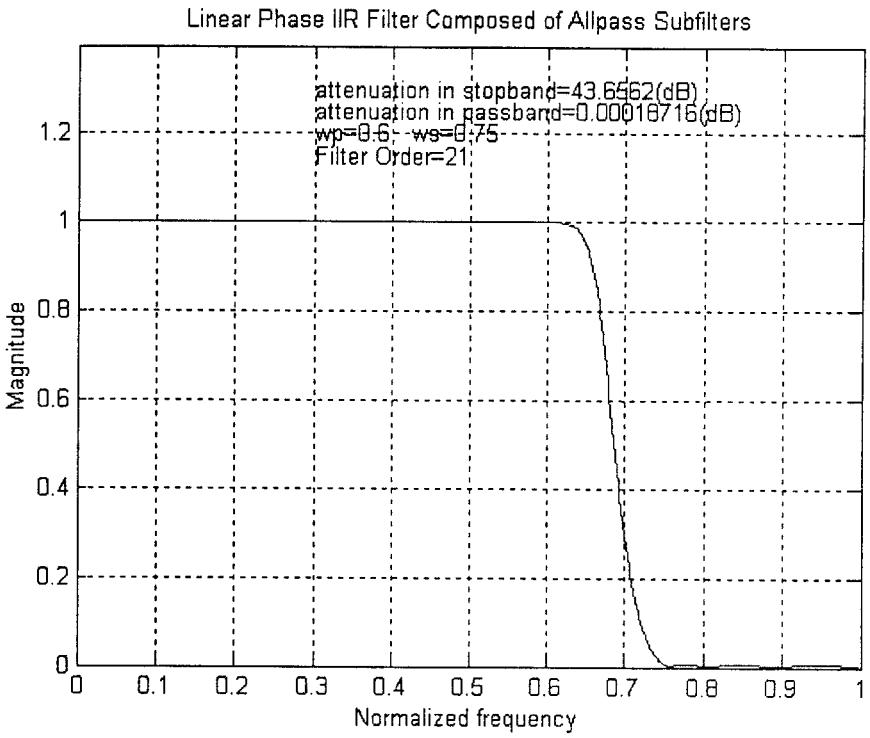


Figure 3.7 Magnitude Response of Example 1

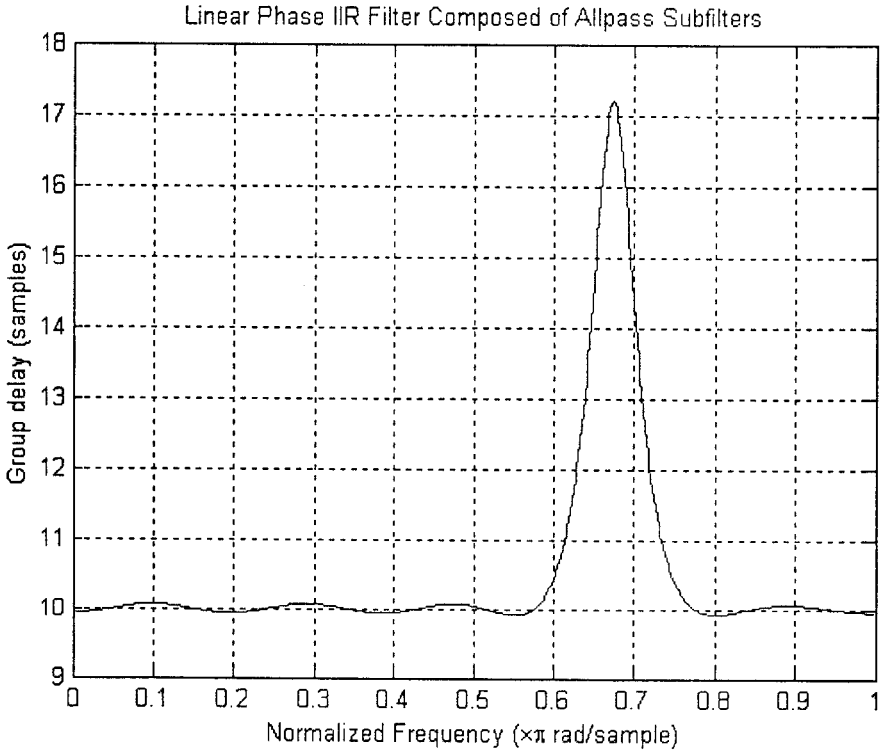


Figure 3.8 Group Delay Response of Example 1

3.3.3 Use of Indirect Phase Error Function

3.3.3.1 Design Description

Instead of approximating the error function in (3.34), we can formulate an equivalent problem as described below. Consider

$$F(z) = \frac{A_1(z)}{z^{-M}} = z^{M-N} \frac{\sum_{n=0}^{N} a_n z^n}{\sum_{n=0}^{N} a_n z^{-n}} \qquad (3.35)$$

$$F(e^{j\omega}) = e^{j(M-N)\omega} \frac{\sum_{n=0}^{N} a_n e^{j\omega n}}{\sum_{n=0}^{N} a_n e^{-j\omega n}} = e^{j2\varphi} \qquad (3.36)$$

where

$$\varphi = \tan^{-1} \frac{\sum_{n=0}^{N} a_n \sin(n + \frac{M-N}{2})\omega}{\sum_{n=0}^{N} a_n \cos(n + \frac{M-N}{2})\omega} = \tan^{-1}\Phi(\omega) \qquad (3.37)$$

We note that φ is one half of the phase difference between $A_1(z)$ and pure delay z^{-M}. As discussed in **3.3.2,** approximately, this difference equals to 0 in passband and $-\pi$ in stopband, respectively. Therefore the desired value of

φ is 0 in the passband and $-\pi/2$ in the stopband. Hence the design problem is to find the coefficients a_n such that the following conditions are satisfied.

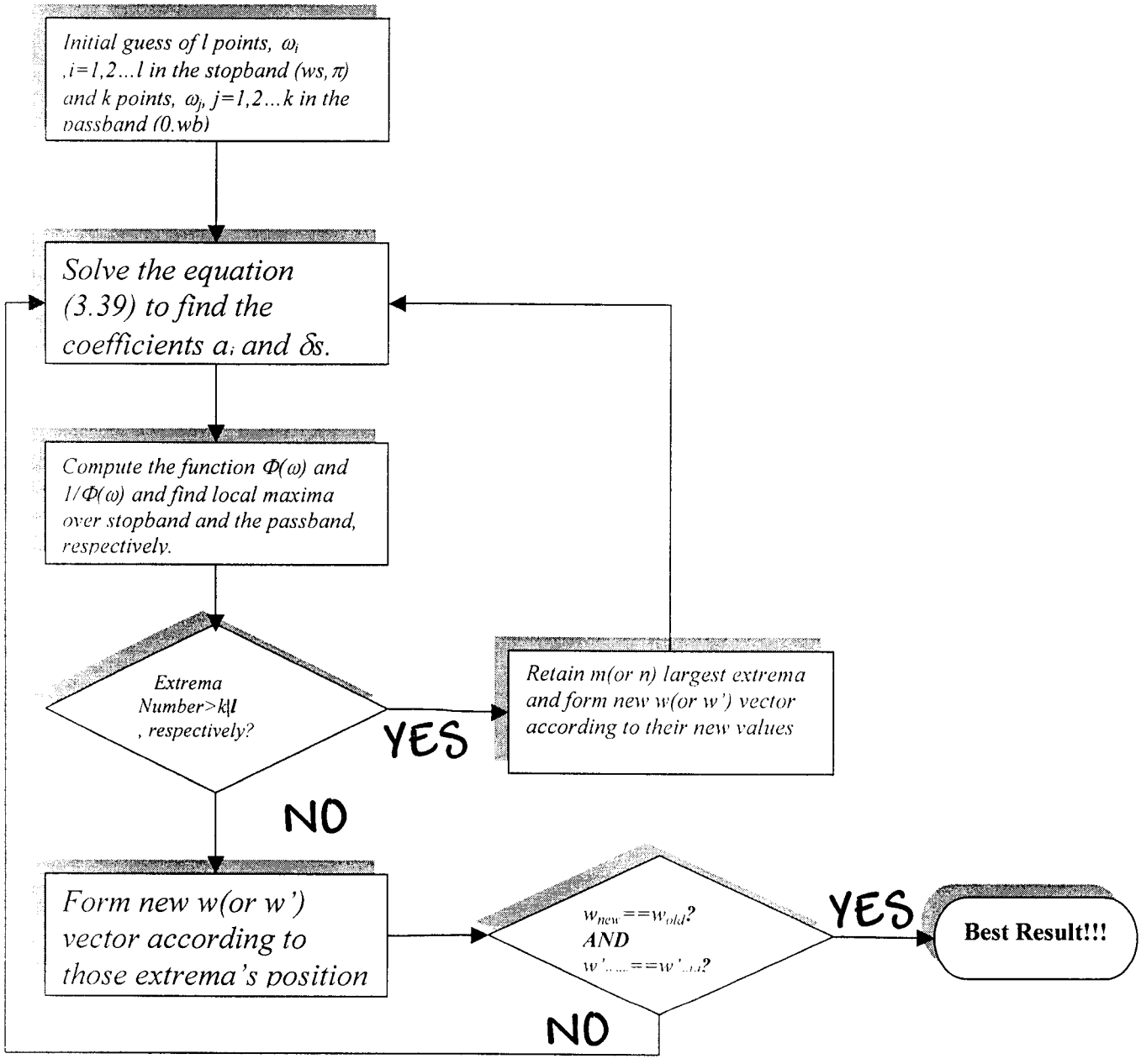
$$\begin{cases} \Phi(\omega) = 0 & \text{in passband} \\ \dfrac{1}{\Phi(\omega)} = 0 & \text{in stopband} \end{cases} \tag{3.38}$$

Then using Remez algorithm to realize the minimization of (3.38)

$$\begin{cases} \Phi(\omega_i) = \dfrac{\sum_{n=0}^{N} a_n \sin(n-\frac{1}{2})\omega_i}{\sum_{n=0}^{N} a_n \cos(n-\frac{1}{2})\omega_i} = (-1)^i \delta_p & \text{in passband} \\ \dfrac{1}{\Phi(\omega_j)} = \dfrac{\sum_{n=0}^{N} a_n \cos(n-\frac{1}{2})\omega_j}{\sum_{n=0}^{N} a_n \sin(n-\frac{1}{2})\omega_j} = (-1)^j \delta_s & \text{in stopband} \end{cases} \tag{3.39}$$

δ_p and δ_s are ripple heights in the passband and stopband respectively.

3.3.3.2 Design Procedure



3.3.3.3 Example 2

Use same specifications as the Example 1. The magnitude and group delay response are shown as below.

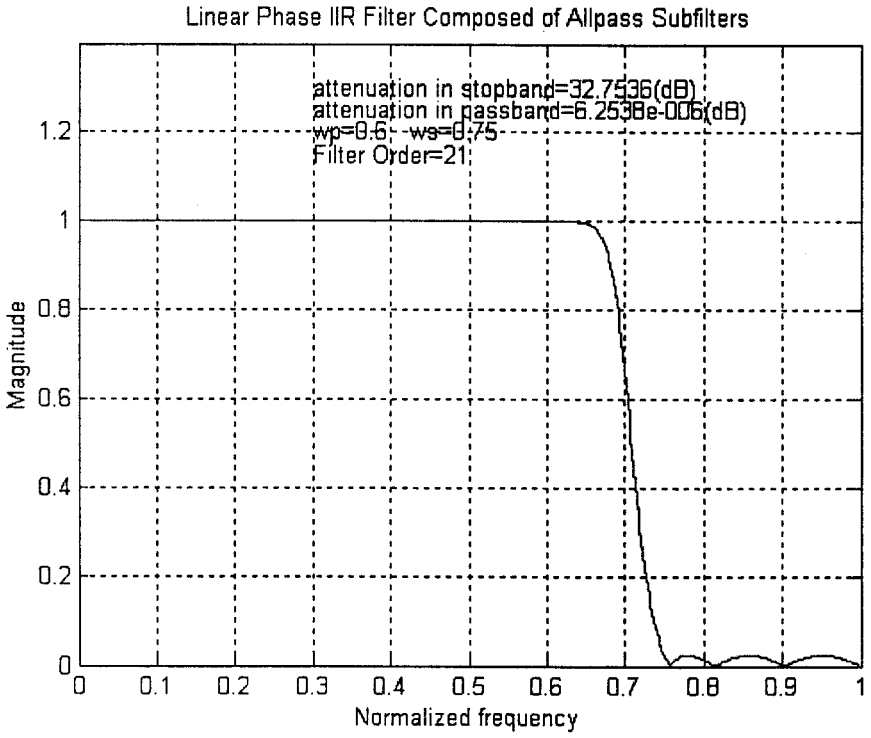


Figure 3.9 Magnitude Response of Example 2

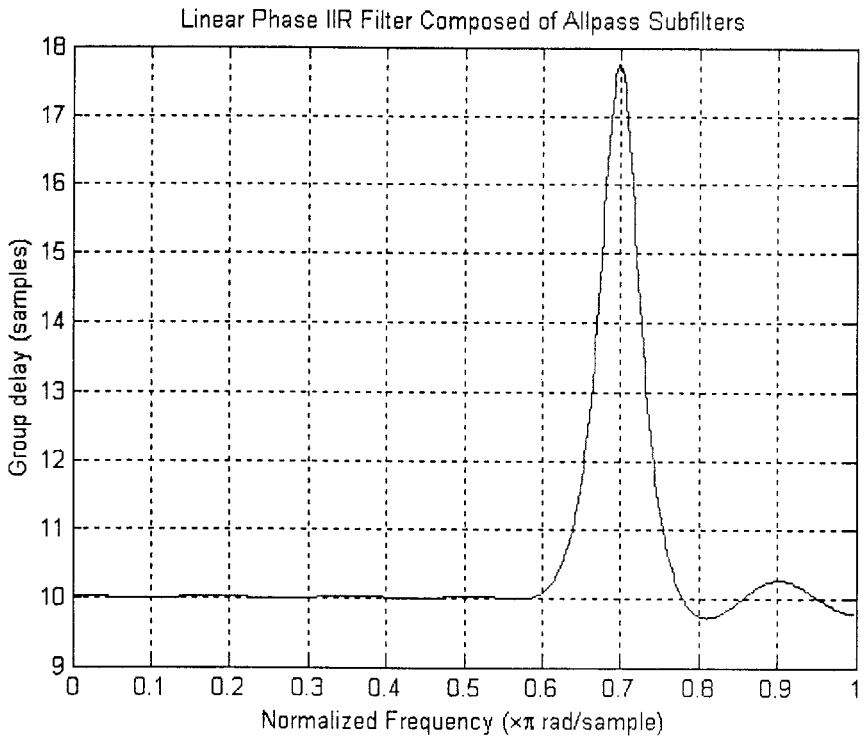


Figure 3.10 Group Delay response of Example 2

3.3.4 Use of Magnitude Response Approximation

3.3.4.1 Design Description

In this structure, we have known that, by approximating phase response of subfilter A1 in order to satisfy

$$\varphi(\omega) = -M\omega \quad \text{in passband}$$
$$\varphi(\omega) = -M\omega - \pi \quad \text{in stopband} \qquad (3.40)$$

we are able to obtain not only approximate phase linearity (both in passband and stopband, although stopband phase linearity is unnecessary), but also approximately satisfactory magnitude response described below:

$$
\begin{aligned}
&|H(\omega)| = 1 \qquad \text{in passband} \\
&|H(\omega)| = 0 \qquad \text{in stopband} \qquad\qquad (3.41)
\end{aligned}
$$

This procedure is reversible according to (3.28). In another word, if only the subfilter A1 is properly designed such that the resultant filter magnitude response fits magnitude response condition (3.41), the filter's phase response satisfies linearity condition (3.40).

According to (3.24),

$$
\begin{aligned}
H(z) &= \frac{1}{2}\left[A_1(z) + z^{-M}\right] \\
&= \frac{1}{2}\left[z^{-N}\frac{\sum_{n=0}^{N} a_n z^{n}}{\sum_{n=0}^{N} a_n z^{-n}} + z^{-M}\frac{\sum_{n=0}^{N} a_n z^{-n}}{\sum_{n=0}^{N} a_n z^{-n}} \right] = \frac{1}{2}\left[\frac{\sum_{n=0}^{N} a_n z^{n-N} + \sum_{n=0}^{N} a_n z^{-n-M}}{\sum_{n=0}^{N} a_n z^{-n}} \right] \\
&= \frac{1}{2}\left[\frac{\sum_{n=0}^{N} a_n \cos[(n-N)\omega] + j\sum_{n=0}^{N} a_n \sin[(n-N)\omega] + \sum_{n=0}^{N} a_n \cos[(n+M)\omega] - j\sum_{n=0}^{N} a_n \sin[N+M)\omega]}{\sum_{n=0}^{N} a_n \cos(n\omega) + j\sum_{n=0}^{N} a_n \sin(n\omega)} \right] \qquad (3.42)
\end{aligned}
$$

$$|H(z)| = \frac{1}{2}\left(\frac{u^2+v^2}{x^2+y^2}\right)^{1/2} \qquad (3.43)$$

where

$$u = \sum_{n=0}^{N} a_n \{\cos[(n-N)\omega] + \cos[(n+M)\omega]\}$$

$$v = \sum_{n=0}^{N} a_n \{\sin[(n-N)\omega] - \sin[(n+M)\omega]\}$$

$$x = \sum_{n=0}^{N} a_n \cos(n\omega)$$

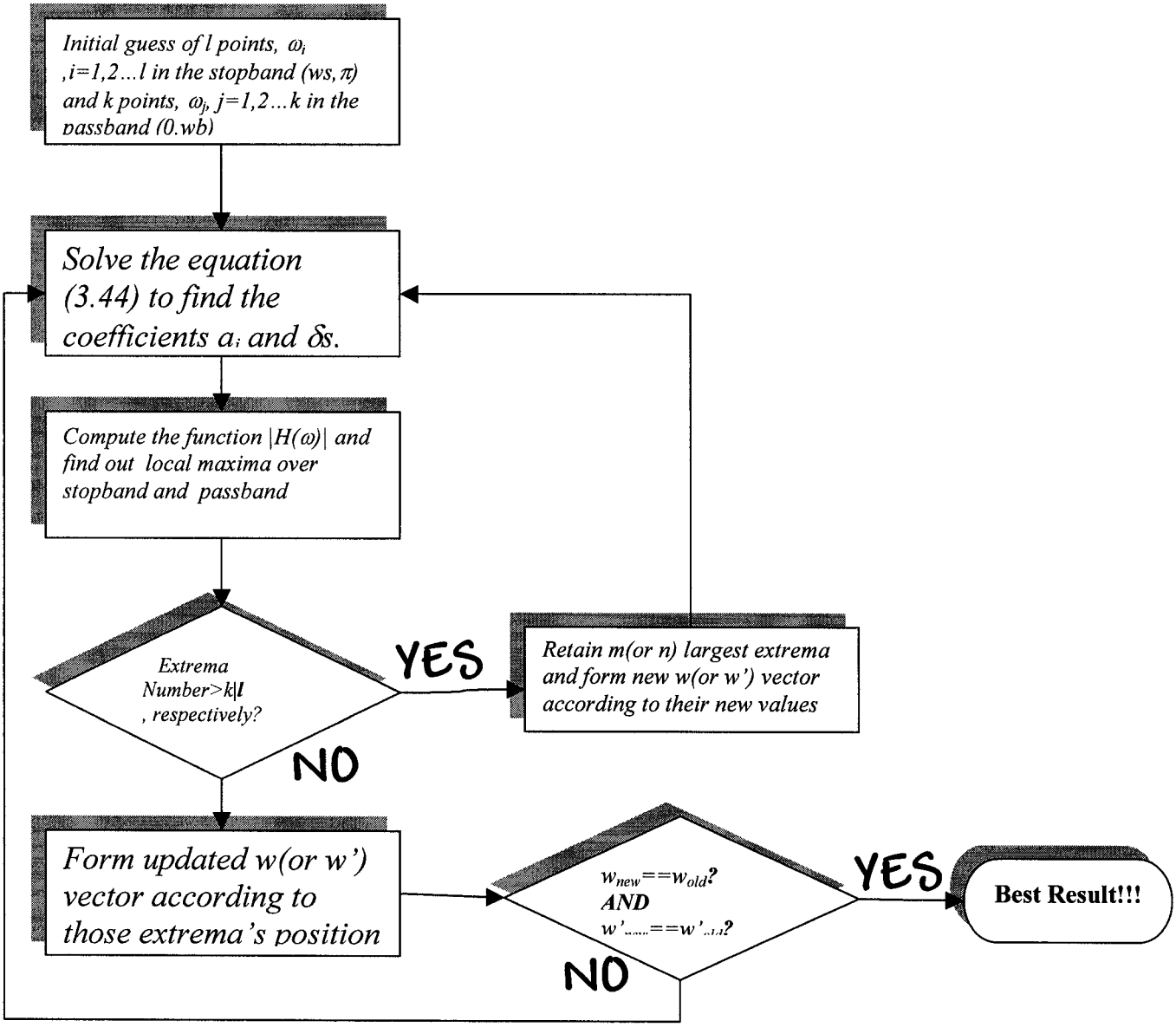
$$y = \sum_{n=0}^{N} a_n \sin(n\omega)$$

The design problem becomes finding the coefficients a_n such that (3.41) is satisfied.

This problem can be solved also using the Remez Exchange Algorithm. We evaluate |H(ω)| and use Remez Exchange Algorithm on (3.44).

$$\begin{cases} 1-|H(\omega_i)| = (-1)^i \delta_p & \text{in the passband} \\ |H(\omega_i)| = (-1)^i \delta_s & \text{in the stopband} \end{cases} \qquad (3.44)$$

3.3.4.2 Design Procedure



3.3.4.3 Example 3

Use same specifications as the Example 1. The magnitude and group delay response are shown as below.

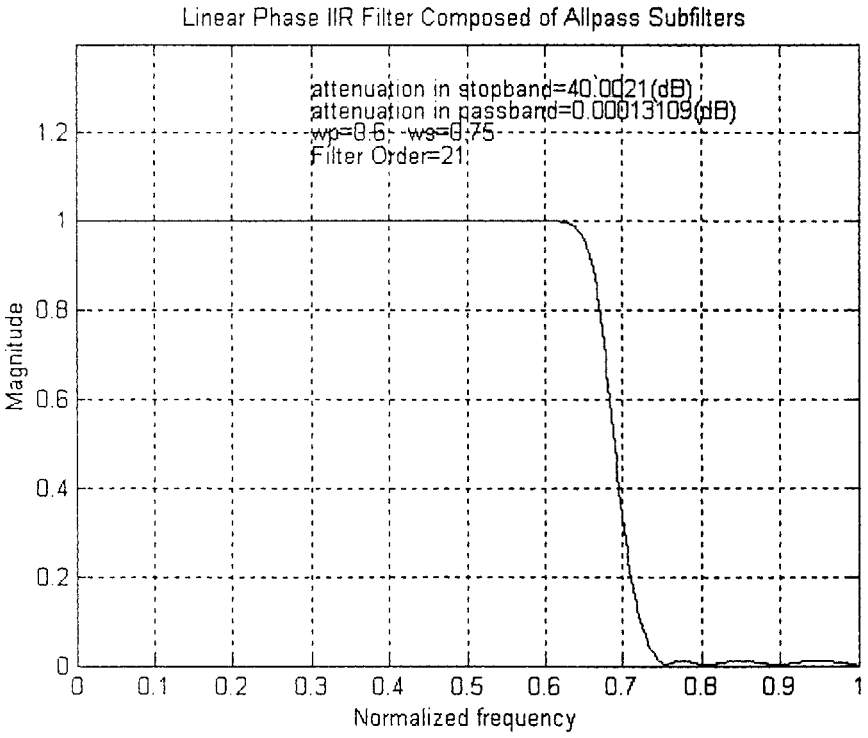


Figure 3.11 Magnitude response of example 3

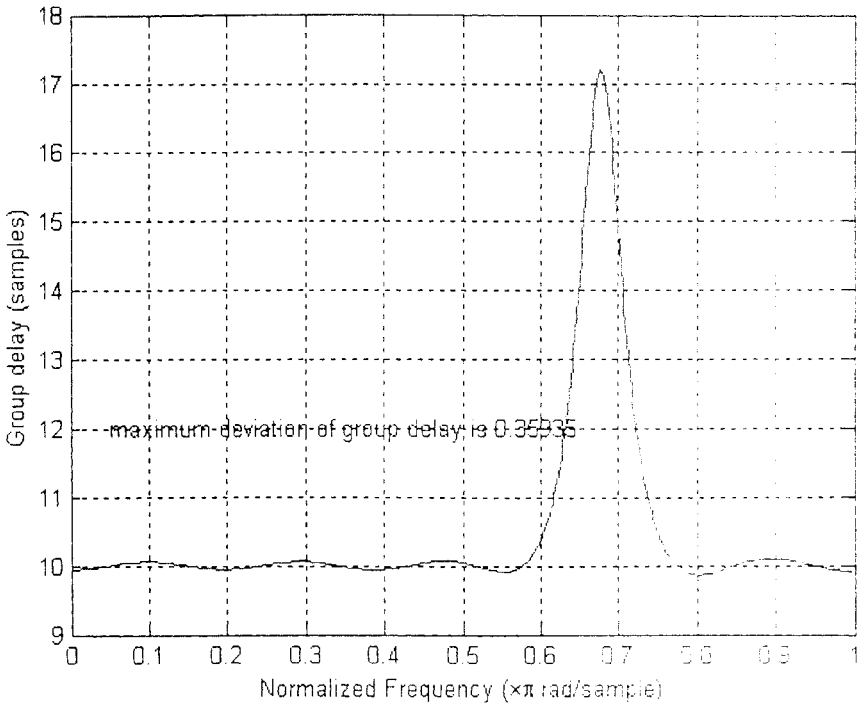


Figure 3.12 Group delay response of example 3

3.3.5 Examples for Combinational Magnitude Response and Phase Response Approximations

It is reasonable to split filter's responses as two individual parts, passband and stopband, in spite of magnitude response or phase response, and using either magnitude approximation described in 3.3.3 or phase approximation described in 3.3.2 to carry out filter optimizations. There are four combinational possibilities:

	Approximation used in passband	Approximation used in stopband	Example No in this chapter
1	Phase	Phase	2
2	Magnitude	Magnitude	3
3	Phase	Magnitude	4
4	Magnitude	Phase	5

Table 3.1 Possible approximation combinations

The first and the second combinations have been illustrated in example 2 and example 3, respectively. For the third method, combining passband part of (3.39) and stopband part of (3.44) together creates (3.45) and similarly, for the forth method, (3.46) comes from combination of stopband part of (3.39) and passband part of (3.44).

$$\begin{cases} \Phi(\omega_i) = \dfrac{\sum_{n=0}^{N} a_n \sin(n-\frac{1}{2})\omega_i}{\sum_{n=0}^{N} a_n \cos(n-\frac{1}{2})\omega_i} = (-1)^i \delta_p & \text{in the passband} \\ |H(\omega_i)| = (-1)^i \delta_s & \text{in the stopband} \end{cases} \tag{3.45}$$

$$\begin{cases} 1-|H(\omega_i)| = (-1)^i \delta_p & \text{in the passband} \\ \dfrac{1}{\Phi(\omega_j)} = \dfrac{\sum_{n=0}^{N} a_n \cos(n-\frac{1}{2})\omega_j}{\sum_{n=0}^{N} a_n \sin(n-\frac{1}{2})\omega_j} = (-1)^j \delta & \text{in the stopband} \end{cases} \tag{3.46}$$

Followed same flow chart and specifications of example 2 and 3, by replacing objective equations to be solved with the corresponding one, we

are able to come up with other filters illustrated as figure 13.14 (example 4), and figure 15, 16 (example 5).

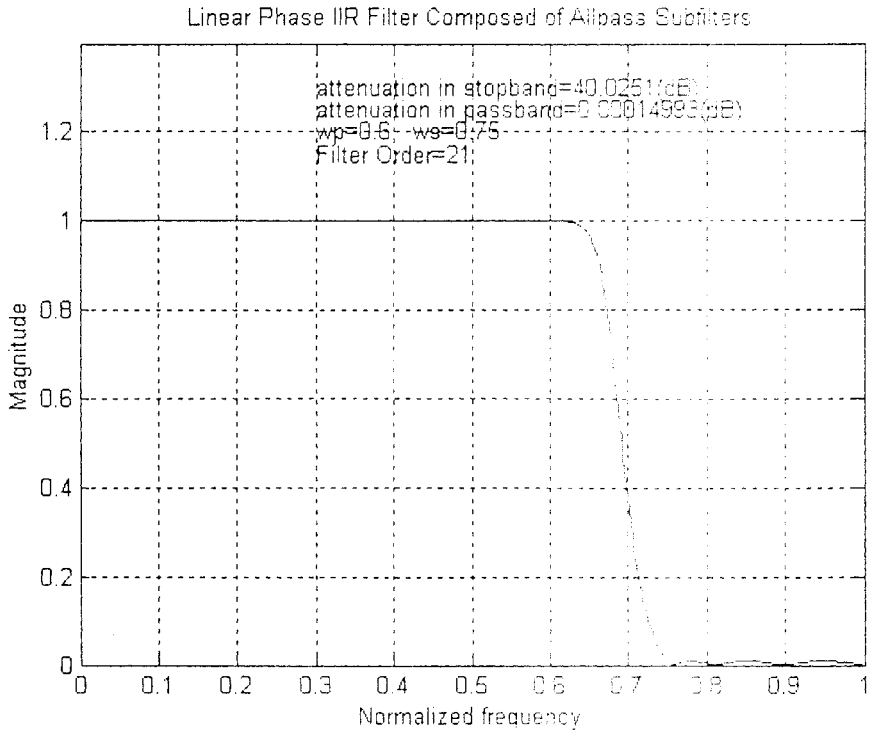


Figure 3.13 Magnitude response of example 4

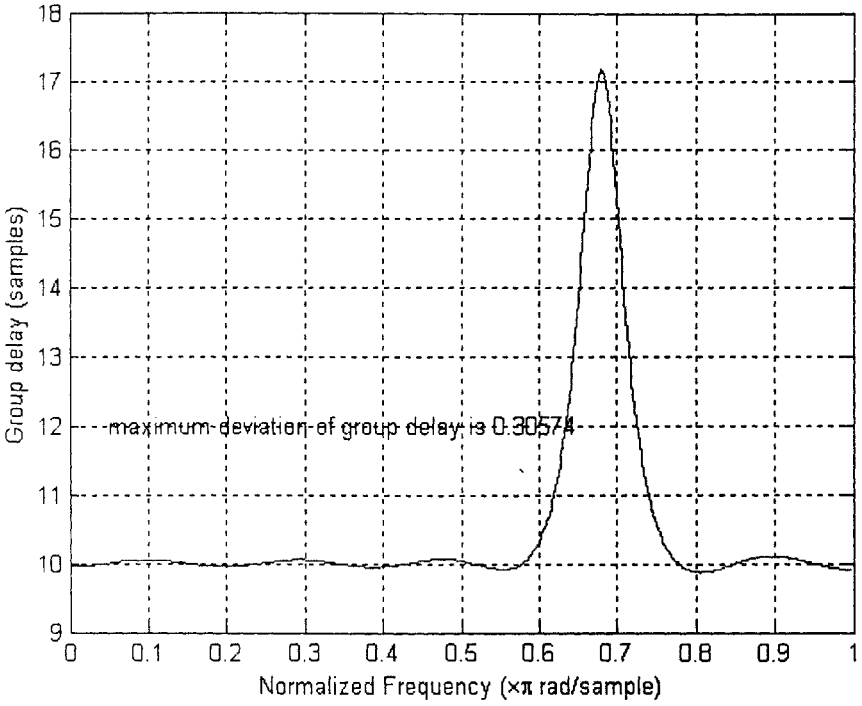


Figure 3.14 Group delay response of example 4

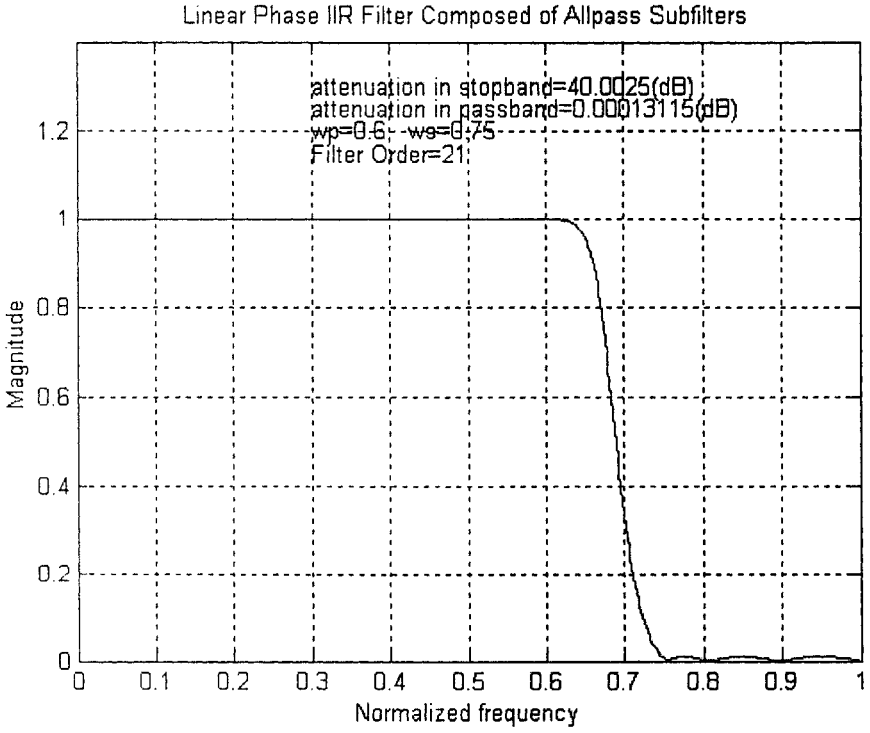


Figure 3.15 Magnitude response of example 5

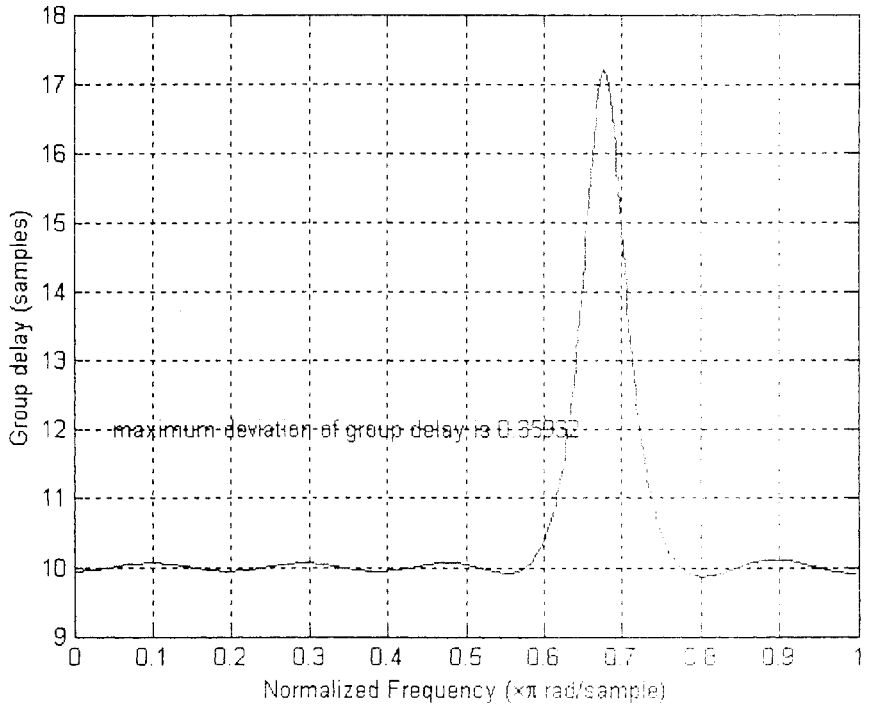


Figure 3.16 Group delay response of example 5

There are two interrelated parameters that need to be optimized in above designs, δp and δs. Normally, for a fixed specification-set, the optimized δp-δs pair is unique. For the purpose of increasing design flexibility, it is recommended to specify one of them and optimize the other one together with filter coefficients. The resultant filter could satisfy different Magnitude-Groupdelay requirement combinations.

Taking example 4, phase-magnitude approximation for passband and stopband respectively, as an illustration:

Let δs ranges from 0.01 to 0.018 with step 0.001. Trend curves of attenuation in passband (magnitude response) and group delay are illustrated as figure 3.17.

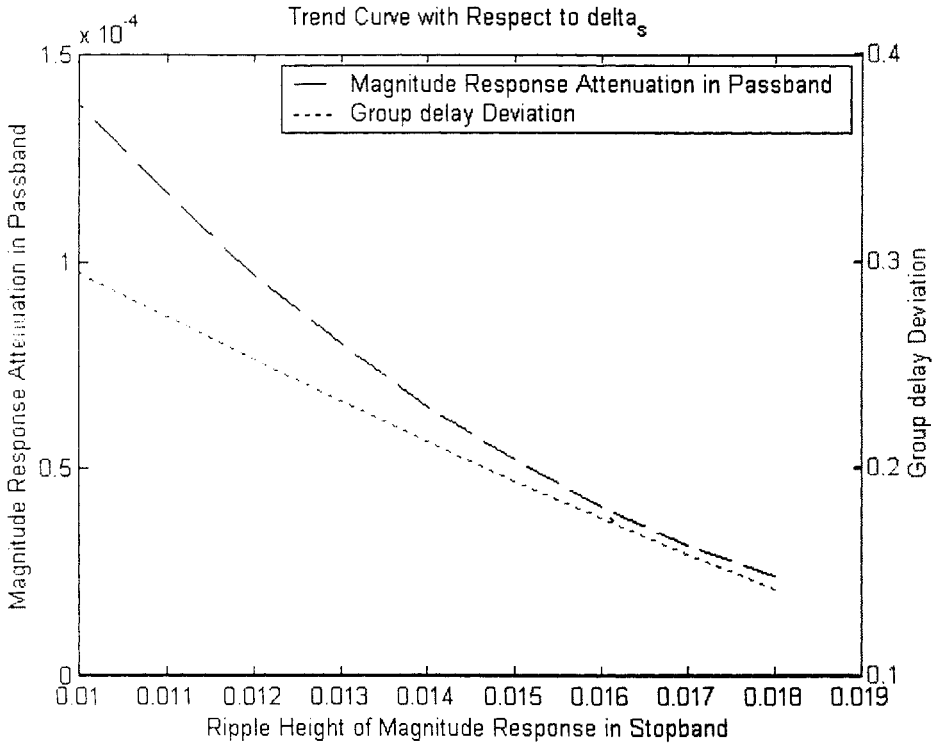


Figure 3.17 Trend curves for passband attenuation and group delay with respect to delta_s

By specifying attenuation in stopband is 40dB, for example, let $\delta s=0.01$ in example 3 and 4, and comparing attenuations in passband and group delay responses for example 2,3,4,5 in figure 3.18, it is evident that those using phase approximation are able to get better phase response but worse magnitude response; in the same way, using magnitude approximation can obtain better magnitude response but worse phase response.

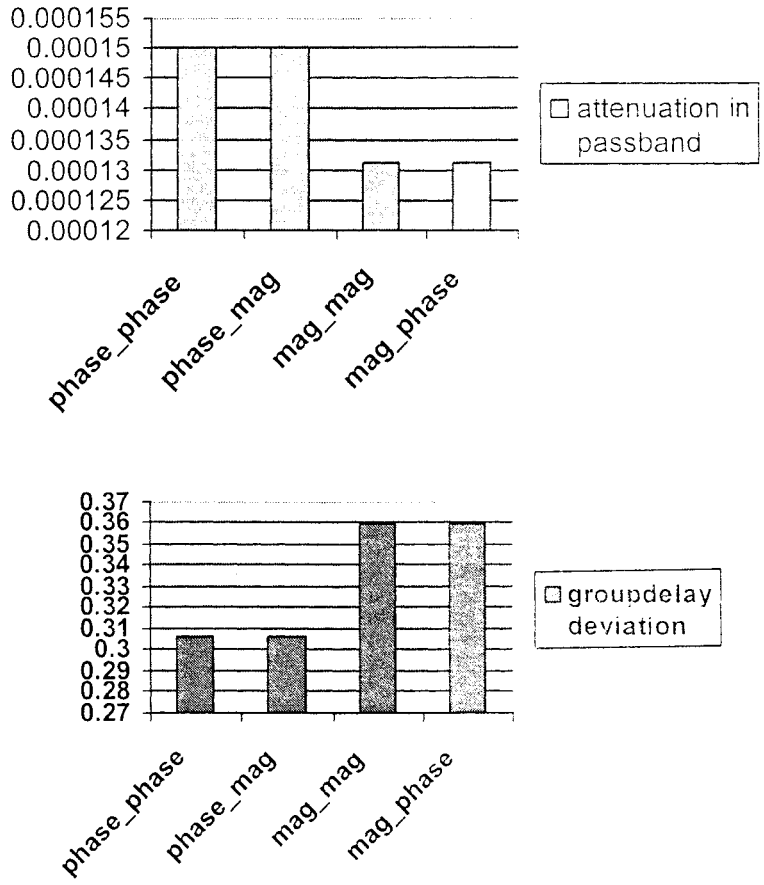


Figure 3.18 Comparisons for example 2.3.4.5 (attenuation in stopband =40dB)

3.4 Conclusion

In this project, several methods for synthesis of linear phase digital IIR filters, which are composed of allpass subfilters, are described and illustrated. By modifying the sign of one of allpass subfilters, we can get the complement highpass/lowpass filter pair. This kind allpass filter structure is convenient for research and development because within certain frequency section, designers only need to pay attention to either phase or magnitude optimization because of their correlativity. As long as one of them reaches its optimization objective, the other one is simultaneously optimized. Moreover, because both magnitude approximation and phase approximation (direct or indirect) can be used in this design, designers are able to flexibly adopt different combinations on different frequency sections in order to meet various design requirements.

Chapter 4

Non-linear Optimization with Simultaneous Magnitude and Group Delay Response Specifications in IIR Filter Design

4.1 Introduction

The classical IIR digital filter design method is to find a transfer function that approximates the specified magnitude response and then find an allpass transfer function which when cascaded with that transfer function gives rise to an overall group delay characteristic that approximates the specified constant group delay.

The most popular optimality criterion for digital filters is minimax (Chebyshev) in each band. In particular, the Remez (Parks-McClellan) algorithm is very popular for both FIR and IIR digital filter design. Minimax and least squares optimization problems are two special cases of a more general class of optimization problems. We refer to problems in this general class as Peak-Constrained Least Squares (PCLS) optimization problems [20]. In PCLS optimization we minimize the total squared error subject to constraints on the peak error.

In this article, we design an IIR digital filter using PCLS idea that has an "equalized" group delay without the use of allpass equalizer sections, and it can simultaneously optimize the magnitude response and group delay. This algorithm uses all of the filter coefficients available to optimize the filter.

4.2 Design Theory

4.2.1 Optimization Overview

Optimization techniques are used to find a set of design parameters, $x=\{x_1,x_2,...,x_n\}$, that can in some way be defined as optimal. In a simple case this may be the minimization or maximization of some system characteristic that is dependent on x. In a more advanced formulation the objective function, f(x), to be minimized or maximized, may be subject to constraints in the form of equality constraints, or inequality constraints:

$$G_i(x) = 0 \quad (i = 1,...,m)$$
$$G_i(x) \le 0 \quad (i = m_e + 1,...,m) \qquad (4.1)$$

and/or parameter bounds, *xl, xu*.

A General Problem (GP) description is stated as

$$\underset{x \in \Re^n}{\text{minimize}} \; f(x) \qquad (4.2)$$

subject to

$$G_i(x) = 0 \quad (i = 1,...,m)$$
$$G_i(x) \leq 0 \quad (i = m_e + 1,...,m)$$
$$x_l \leq x \leq x_u$$

where x is the vector of design parameters, $f(x)$ is the objective function that returns a scalar value, and the vector function $G(x)$ returns the values of the equality and inequality constraints evaluated at x.

An efficient and accurate solution to this problem is not only dependent on the size of the problem in terms of the number of constraints and design variables but also on characteristics of the objective function and constraints. When both the objective function and the constraints are linear functions of the design variable, the problem is known as a Linear Programming (LP) problem. Quadratic Programming (QP) concerns the minimization or maximization of a quadratic objective function that is linearly constrained. For both the LP and QP problems, reliable solution procedures are readily available. More difficult to solve is the Nonlinear Programming (NP) problem in which the objective function and constraints may be nonlinear functions of the design variables. A solution of the NP problem generally requires an iterative procedure to establish a direction of search at each major iteration. This is usually achieved by the solution of an LP, a QP, or an unconstrained subproblem.

4.2.2 Constrained Optimization

Unlike FIR filter design, in IIR digital filter design, filter stability has to be thought much of. This design requires that all poles must locate in the unity circle so that all poles' radii must less than 1. Constrained optimization is adopted in this design for satisfying this requirement.

In constrained optimization, the general aim is to transform the problem into an easier subproblem that can then be solved and used as the basis of an iterative process. A characteristic of a large class of early methods is the translation of the constrained problem to a basic unconstrained problem by using a penalty function for constraints, which are near or beyond the constraint boundary. In this way the constrained problem is solved using a sequence of parameterized unconstrained optimizations, which in the limit (of the sequence) converge to the constrained problem. These methods are now considered relatively inefficient and have been replaced by methods that have focused on the solution of the Kuhn-Tucker (KT) equations. The KT equations are necessary conditions for optimality for a constrained optimization problem. If the problem is a so-called convex programming problem, that is, *f(x)* and *Gi(x)*, i=1,…,m, are convex functions, then the KT equations are both necessary and sufficient for a global solution point.

Referring to GP equation, the Kuhn-Tucker equations can be stated as

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda^*_i \cdot \nabla G_i(x^*) = 0$$
$$\lambda^*_i \cdot G_i(x^*) = 0 \qquad i = 1,...,m$$
$$\lambda^*_i \geq 0 \qquad i = m_e + 1,...,m \qquad (4.3)$$

The first equation describes a canceling of the gradients between the objective function and the active constraints at the solution point. For the gradients to be canceled, Lagrange Multipliers (λ_i, i=1,...,m) are necessary to balance the deviations in magnitude of the objective function and constraint gradients. Since only active constraints are included in this canceling operation, constraints that are not active must not be included in this operation and so are given Lagrange multipliers equal to zero. This is stated implicitly in the last two equations.

The solution of the KT equations forms the basis to many nonlinear programming algorithms. These algorithms attempt to compute directly the Lagrange multipliers. Constrained quasi-Newton methods guarantee superlinear convergence by accumulating second order information regarding the KT equations using a quasi-Newton updating procedure. These methods are commonly referred to as Sequential Quadratic Programming (SQP) methods since a QP subproblem is solved at each major iteration (also known as Iterative Quadratic Programming, Recursive Quadratic Programming, and Constrained Variable Metric methods).

4.2.3 Sequential Quadratic Programming (SQP)

SQP methods represent state-of-the-art in nonlinear programming methods. Schittowski [21], for example, has implemented and tested a version that outperforms every other tested method in terms of efficiency, accuracy, and percentage of successful solutions, over a large number of test problems.

Given the problem description in GP (4.2), the principal idea is the formulation of a QP subproblem based on a quadratic approximation of the Lagrangian function.

$$L(x,\lambda) = f(x) + \sum_{i=1}^{m} \lambda_i \cdot g_i(x) \qquad (4.4)$$

Here Eq 4.2 is simplified by assuming that bound constraints have been expressed as inequality constraints. The QP subproblem is obtained by linearizing the nonlinear constraints.

$$\begin{aligned} &\underset{d \in \Re^n}{\text{minimize}} \ \frac{1}{2} d^T H_k d + \nabla f(x_k)^T d \\ &\nabla g_i(x_k)^T d + g_i(x_k) = 0 \qquad i = 1,\ldots,m_e \\ &\nabla g_i(x_k)^T d + g_i(x_k) \le 0 \qquad i = m_e+1,\ldots,m \end{aligned} \qquad (4.5)$$

This subproblem can be solved using any QP algorithm (for instance, Quadratic Programming Solution). The solution is used to form a new iterate

$$x_{k+1} = x_k + \alpha_k d_k \qquad (4.6)$$

The step length parameter α_k is determined by an appropriate line search procedure so that a sufficient decrease in a merit function is obtained (see Updating the Hessian Matrix). The matrix H_k is a positive definite approximation of the Hessian matrix of the Lagrangian function (Eq. 4.4).

H_k can be updated by any of the quasi-Newton methods, although the BFGS method (see Updating the Hessian Matrix) appears to be the most popular.

A nonlinearly constrained problem can often be solved in fewer iterations than an unconstrained problem using SQP. One of the reasons for this is that, because of limits on the feasible area, the optimizer can make well-informed decisions regarding directions of search and step length.

4.2.4 SQP Implementation

The typical SQP implementation consists of three main stages, which are discussed briefly in the following subsections:

- Updating of the Hessian matrix of the Lagrangian function
- Quadratic programming problem solution
- Line search and merit function calculation

4.2.4.1 Updating of the Hessian Matrix

At each major iteration a positive definite quasi-Newton approximation of the Hessian of the Lagrangian function, H, is calculated using the BFGS method where λi (i=1,...,m) is an estimate of the Lagrange multipliers.

$$H_{k+1} = H_k + \frac{q_k q_k^T}{q_k^T s_k} - \frac{H_k^T H_k}{s_k^T H_k s_k} \qquad (4.7)$$

$$s_k = x_{k+1} - x_k$$

$$q_k = \nabla f(x_{k+1}) + \sum_{i=1}^{n} \lambda_i \cdot \nabla g_i(x_{k+1}) - \left(\nabla f(x_k) + \sum_{i=1}^{n} \lambda_i \cdot \nabla g_i(x_k) \right)$$

Powell [22] recommends keeping the Hessian positive definite even though it may be positive indefinite at the solution point. A positive definite Hessian is maintained providing $q_k^T s_k$ is positive at each update and that H is initialized with a positive definite matrix. When $q_k^T s_k$ is not positive, q_k is modified on an element by element basis so that $q_k^T s_k > 0$. The general aim of this modification is to distort the elements of q_k, which contribute to a positive definite update, as little as possible. Therefore, in the initial phase of the modification, the most negative element of $q_k^T s_k$ is repeatedly halved. This procedure is continued until $q_k^T s_k$ is greater than or equal to 1e-5. If after this procedure, $q_k^T s_k$ is still not positive, q_k is modified by adding a vector v multiplied by a constant scalar w, that is,

$$q_k = q_k + wv \qquad (4.8)$$

where

$$vi = \nabla g_i(x_{k+1}) \cdot g_i(x_{k+1}) - \nabla g_i(x_k) \cdot g_i(x_k)$$

$$if\ (q_k)_i \cdot w < 0\ and$$

$$(q_k)_i \cdot (s_k)_i < 0 (i = 1,...m)$$

$$v_i = 0\ otherwise$$

and w is systematically increased until $q_k^T s_k$ becomes positive.

4.2.4.2 Quadratic Programming Solution

At each major iteration of the SQP method a QP problem is solved of the form where A_i refers to the ith row of the m-by-n matrix A.

$$\begin{aligned} &\underset{d \in \Re^n}{\text{minimize}} \; q(d) = \frac{1}{2} d^T H d + c^T d \\ &A_i d = b_i \qquad i = 1,\ldots,m_e \\ &A_i d \le b_i \qquad i = m_e + 1,\ldots,m_e \end{aligned} \tag{4.9}$$

The solution procedure involves two phases: the first phase involves the calculation of a feasible point (if one exists); the second phase involves the generation of an iterative sequence of feasible points that converge to the solution. In this method an active set is maintained, $\overline{A_k}$, which is an estimate of the active constraints (i.e., which are on the constraint boundaries) at the solution point. Virtually all QP algorithms are active set methods. This point is emphasized because there exists many different methods that are very similar in structure but that are described in widely different terms.

$\overline{A_k}$ is updated at each iteration, k, and this is used to form a basis for a search direction $\hat{d}_k$. Equality constraints always remain in the active set, $\overline{A_k}$. The notation for the variable, $\hat{d}_k$, is used here to distinguish it from $\hat{d}_k$ in the major iterations of the SQP method. The search direction, $\hat{d}_k$, is calculated and minimizes the objective function while remaining on any active

constraint boundaries. The feasible subspace for $\hat{d}_k$ is formed from a basis, Z_k whose columns are orthogonal to the estimate of the active set $\overline{A_k}$ (i.e., $\overline{A_k}\ Z_k=0$). Thus a search direction, which is formed from a linear summation of any combination of the columns of Z_k, is guaranteed to remain on the boundaries of the active constraints.

The matrix Z_k is formed from the last m-l columns of the QR decomposition of the matrix $\overline{A_k}^{\mathrm{T}}$, where l is the number of active constraints and $l < m$. That is, Z_k is given by

$$Z_k = Q[:, l+1:m] \qquad (4.10)$$

where $Q_T \overline{A_k}^T = \begin{bmatrix} R \\ 0 \end{bmatrix}$

Having found Z_k, a new search direction $\hat{d}_k$ is sought that minimizes *q(d)* where $\hat{d}_k$ is in the null space of the active constraints, that is, $\hat{d}_k$ is a linear combination of the columns of Z_k: $\hat{d}_k = Z_k p$ for some vector p.

Then if we view our quadratic as a function of p, by substituting for $\hat{d}_k$, we have

$$q(p) = \frac{1}{2} p^T Z_k{}^T H Z_k p + c^T Z_k p \qquad (4.11)$$

Differentiating this with respect to p yields

$$\nabla q(p) = Z_k{}^T H Z_k p + Z_k{}^T c \qquad (4.12)$$

$\nabla q(p)$ is referred to as the projected gradient of the quadratic function because it is the gradient projected in the subspace defined by Z_k. The term $Z_k{}^T H Z_k$ is called the projected Hessian. Assuming the Hessian matrix H is positive definite (which is the case in this implementation of SQP), then the minimum of the function q*(p)* in the subspace defined by Z_k occurs when $\nabla q(p) = 0$, which is the solution of the system of linear equations

$$Z_k{}^T H Z_k p = -Z_k{}^T c \qquad (4.13$$

A step is then taken of the form

$$x_{k+1} = x_k + \alpha \hat{d}_k \qquad where \quad \hat{d}_k = Z_k{}^T p \qquad (4.14)$$

At each iteration, because of the quadratic nature of the objective function, there are only two choices of step length α. A step of unity along $\hat{d}_k$ is the exact step to the minimum of the function restricted to the null space of $\overline{A_k}$. If such a step can be taken, without violation of the constraints, then this is the solution to QP (Eq. 2.10). Otherwise, the step along $\hat{d}_k$ to the nearest constraint is less than unity and a new constraint is included in the active set at the next iterate. The distance to the constraint boundaries in any direction $\hat{d}_k$ is given by

$$\alpha = \min_i \left\{ \frac{-(A_i x_k - b_i)}{A_i \hat{d}_k} \right\} \qquad (i = 1,\ldots,m) \qquad (4.15)$$

which is defined for constraints not in the active set, and where the direction $\hat{d}_k$ is towards the constraint boundary, i.e., $A_i \hat{d}_k > 0, \quad i = 1,\ldots,m$

When n independent constraints are included in the active set, without location of the minimum, Lagrange multipliers, λ_k are calculated that satisfy the nonsingular set of linear equations

$$\bar{A}_k^T \lambda_k = c \qquad (4.16)$$

If all elements of λ_k are positive, x_k is the optimal solution of QP (Eq. 2.10). However, if any component of λ_k is negative, and it does not correspond to an equality constraint, then the corresponding element is deleted from the active set and a new iterate is sought.

The algorithm requires a feasible point to start. If the current point from the SQP method is not feasible, then a point can be found by solving the linear programming problem

$$\begin{aligned} &\underset{\gamma \in \Re, x \in \Re^n}{\text{minimize}} \quad \gamma \\ &A_i x = b_i \qquad i = 1,\ldots,m_e \\ &A_i x - \gamma \le b_i \qquad i = m_e + 1,\ldots,m \end{aligned} \qquad (4.17)$$

The notation A_i indicates the *ith* row of the matrix A. A feasible point (if one exists) to Eq. 2.17 can be found by setting x to a value that satisfies the equality constraints. This can be achieved by solving an under- or over-determined set of linear equations formed from the set of equality constraints. If there is a solution to this problem, then the slack variable γ is set to the maximum inequality constraint at this point.

The above QP algorithm is modified for LP problems by setting the search direction to the steepest descent direction at each iteration where g_k is the gradient of the objective function (equal to the coefficients of the linear objective function).

$$\hat{d}_k = -Z_k Z_k^T g_k \qquad (4.18)$$

If a feasible point is found using the above LP method, the main QP phase is entered. The search direction $\hat{d}_k$ is initialized with a search direction $\hat{d}_1$ found from solving the set of linear equations

$$H\hat{d}_1 = -g_k \qquad (4.19)$$

where g_k is the gradient of the objective function at the current iterate x_k (i.e., $Hx_k + c$).

If a feasible solution is not found for the QP problem, the direction of search for the main SQP routine $\hat{d}_k$ is taken as one that minimizes γ.

4.2.4.3 Line Search and Merit Function

The solution to the QP subproblem produces a vector d_k, which is used to form a new iteration

$$x_{k+1} = x_k + \alpha d_k \qquad (4.20)$$

The step length parameter α_k is determined in order to produce a sufficient decrease in a merit function. The merit function used by Han [23] and Powell [24] of the form below has been used in this implementation.

$$\Psi(x) = f(x) + \sum_{i=1}^{m_e} r_i \cdot g_i(x) + \sum_{i=m_e+1}^{m} r_i \cdot \max\{0, g_i(x)\} \qquad (4.21)$$

Powell recommends setting the penalty parameter

$$r = (r_{k+1})_i = \max_i \left\{ \lambda_i, \frac{1}{2}((r_k)_i + \lambda_i) \right\}, \qquad i = 1,\ldots,m \qquad (4.22)$$

This allows positive contribution form constraints that are inactive in the QP solution but were recently active. In this implementation, initially the penalty parameter r_i is set to

$$r_i = \frac{\|\nabla f(x)\|}{\|\nabla g_i(x)\|} \qquad (4.23)$$

where $\|\ \|$ represents the Euclidean norm.

This ensures larger contributions to the penalty parameter from constraints with smaller gradients, which would be the case for active constraints at the solution point.

4.2.5 Deczky's and Lawon's methods

Deczky [25] considered a general transfer function in the cascaded SOS (second-order section) form:

$$H(z) = k_0 \prod_{i=1}^{N} \frac{1 + a_{i1} z^{-1} + a_{i2} z^{-2}}{1 + b_{i1} z^{-1} + b_{i2} z^{-2}} \qquad (4.24)$$

and he also developed an algorithm for minimizing an error function that both contains a weighted sum of the error in the magnitude and in the group delay. This error function will be discussed below.

Lawon [26] showed that minimax (Chebyshev) optimization problems could be reformulated in terms of equivalent weighted least-squares (WLS) optimization problems. As an example we consider the following minimax approximation problem where *D(x)* is the desired function and *F(x)* is the approximating function.

$$Minimize : Maximum\{|F(x) - D(x)|; \quad a \le x \le b\} \qquad (4.25)$$

The problem in (4.25) could be reformulated as the following weighted least squares minimization problem.

$$Minimize: E = \int_0^{\phi} W(x) \mid F(x) - D(x) \mid^2 dx \qquad (4.26)$$

The trick here is to find the appropriate $W(x)$ that makes (4.25) equivalent to (4.26). The "Lawon algorithm" actually is an iterative procedure for determining $W(x)$.

In this chapter, we use Deczky's IIR filter design algorithm as an engine to design a constrained least-square IIR filter and determine the weighting functions needed for PCLS filters by using idea of Lawon's algorithm.

4.3 Design Procedure

4.3.1 Formulation of the PCLS Optimization Problem for IIR Filters

The Deczky's general IIR filters form (4.24) can also be expressed in zero-pole form:

$$H(z) = k_0 \prod_{i=1}^{N} \frac{z^2 - 2r_{oi}\cos(\phi_{oi})z + r_{oi}^2}{z^2 - 2r_{pi}\cos(\phi_{pi})z + r_{pi}^2} \qquad (4.27)$$

where

k_0 gain constant;

r_{oi} ith zero radius;

ϕ_{oi} ith zero angle;

r_{pi} ith pole radius;

ϕ_{pi} ith pole angle;

N number of quadratic sections.

The magnitude response is shown below:

$$|H(A,\phi)| = k_0 \sum_{i=1}^{N} \frac{\left\{1 - 2r_{oi}\cos(\phi - \phi_{oi}) + r_{oi}^2\right\}^{1/2}\left\{1 - 2r_{oi}\cos(\phi + \phi_{oi}) + r_{oi}^2\right\}^{1/2}}{\left\{1 - 2r_{pi}\cos(\phi - \phi_{pi}) + r_{pi}^2\right\}^{1/2}\left\{1 - 2r_{pi}\cos(\phi + \phi_{pi}) + r_{pi}^2\right\}^{1/2}} \qquad (4.28)$$

where A is the coefficients vector defined as

$$A = [r_{o1}, \phi_{o1}, r_{p1}, \phi_{p1}, \cdots, r_{oi}, \phi_{oi}, r_{pi}, \phi_{pi}, \cdots, k_0] \qquad (4.29)$$

The equation for the group delay is

$$\tau(A,\phi) = \sum_{i=1}^{N}\left[\frac{1 - r_{pi}\cos(\phi - \phi_{pi})}{1 - 2r_{pi}\cos(\phi - \phi_{pi}) + r_{pi}^2} + \frac{1 - r_{pi}\cos(\phi + \phi_{pi})}{1 - 2r_{pi}\cos(\phi + \phi_{pi}) + r_{pi}^2}\right.$$
$$\left. - \frac{1 - r_{oi}\cos(\phi - \phi_{oi})}{1 - 2r_{oi}\cos(\phi - \phi_{oi}) + r_{oi}^2} - \frac{1 - r_{oi}\cos(\phi + \phi_{oi})}{1 - 2r_{oi}\cos(\phi + \phi_{oi}) + r_{oi}^2}\right] \qquad (4.30)$$

Deczky developed an algorithm for minimizing an error function that contains a weighted sum of the error in the magnitude as well as in the group delay. The error function is given by

$$J(h) = (1-\lambda)\sum_{i=1}^{Q} W_i \left[\left| H(e^{j\omega_i}) \right| - \left| G(e^{j\omega_i}) \right| \right]^p$$
$$+ \lambda \sum_{i=1}^{Q} V_i \left[\tau_H(\omega_i) - \tau(\omega_0) - \tau_G(\omega_i) \right]^p \qquad (4.31)$$

Where ω_i represent a set of Q discrete frequencies, $\tau_{H(\omega i)}$ is the overall group delay of the designed filter, $\tau_{G(\omega i)}$ is the desired group delay and $\tau_{(\omega 0)}$ is called "the nominal delay" which is a constant value duly chose to reach a minimum value for the error function.

When p=2, this algorithm becomes the weighted least squares approximation and p=∞, minimax algorithm.

When λ=1 is chosen, the problem reduces to that of approximating the group delay only, without considering its magnitude distortion and vice versa.

This algorithm (4.31) can be used for PCLS optimization by finding appropriate weighting functions, as discussed in the next section.

4.3.2 Weighting Function and Tolerances Updates Strategy

At the beginning, we sample the objective filter with both magnitude and group delay response over several sampling frequency points. For each sampling frequency point, identical weight value is initialized. We consider the computation of the magnitude and group delay weighting functions,

$W(k+1,e^{j\omega})$ and $V(k+1,e^{j\omega})$, respectively, corresponding to iteration k+1, based on the results from iteration k. For the sake of concision, the update strategy for magnitude response weighting function is introduced here only. Similar approach can be derived for group delay.

At iteration k we apply the weighting function $W(k,e^{j\omega})$ to minimize the corresponding weighted squared error (4.31) and obtain the resulting $H(k,e^{j\omega})$. Assuming that the peak stopband error specification is δs and for passband, δp, and denoting the value of the i-th weighting function component in W at iteration k+1 as $W(k+1, i)$.

We update the weighting function according to the following general rules:

1) Make $W_s(k+1, i)$ larger than $W_s(k, i)$ if $H_s(k, i) > \delta s$;
2) Make $W_s(k+1,i)$ smaller than $W_s(k, i)$ if $H_s(k,i) < \delta s$, but the limitation is Cs.
3) No change for $W_s(k, i)$ if $H_s(k, i) = \delta s$.

The same rules applied on the passband:

4) Make $W_p(k+1, i)$ larger than $W_p(k, i)$ if $|H_p(k, I)-1| > \delta p$;
5) Make $W_p(k+1, i)$ smaller than $W_p(k, i)$ if $|H_p(k, i)-1| < \delta p$, but the limitation is Cp.
6) No change for $|H_p(k, i)-1| = \delta p$.

We then consider in detail about implementing the above rules. We first discuss the stopband updates of the weighting function.

$$W_s(k+1,i) = Bs\left[1+\left(\frac{H_s(k,i)-G-\delta_s}{\delta_s}\right)\right]\bullet W_s(k,i) \qquad (4.32)$$

The weight updating factor $\left[1+\left(\frac{H_s(k,i)-G-\delta_s}{\delta_s}\right)\right]$ is processed by a saturation function, minimum at "a", for the purpose of avoiding intensive alteration on weight value. Normally set a=0.5 from experiences.

Weighting function update approach for passband is similar:

$$W_p(k+1,i) = Bp\left[1+\left(\frac{H_p(k,i)-G-\delta_p}{\delta_p}\right)\right]\bullet W_p(k,i) \qquad (4.33)$$

For stopband G=0, passband G=1. This is the ideal magnitude response for a lowpass filter.

In order to constrain magnitude response within transition band, some points sampled in transition band are also taken into consideration. Therefore the overall designed magnitude response $H(\omega)$ in (4.31) includes not only passband and stopband sampling points, but also several points in transition band. The numbers of points are proportional to corresponding bandwidth. In another word, the wider this band, the more points are sampled in it.

For updating the group delay weighting function, only passband performance is taken into account. Assuming that the peak group delay error specification is δt and denoting the value of the i-th weighting value in V at iteration k+1 as $V(k+1, i)$. Its weighting function is updated according to:

$$V(k+1,i) = Bt\left[1+\left(\frac{\tau_{H}(k,i)-\tau_{o}-\delta_{t}}{\delta_{t}}\right)\right]\bullet V(k,i) \qquad (4.34)$$

Tolerance parameters δs, δp, and δt can be constants all through the whole optimization, or else they are subject to be updated loop by loop prior to updates of weighting functions. An efficient update strategy is critical for reducing algorithm convergence difficulties. Update strategies of this design can be stated as:

(1) $\delta = E/\alpha$

E : maximum deviation value

(2) $\delta = \text{avrg(D)}/\alpha$

D : deviation values set for sampling frequencies (4.35)

A carefully chosen α is crucial for optimization converges and filter performances balance.

Now basic steps of this algorithm can be summarized as:

1. Choose a set of Q discrete frequency points, ω_i, i=1,2...Q, over whole frequency axes (0=>1, normalized frequency);

2. Initialize weighting functions of magnitude and group delay such that all discrete weights have the same constant value.

3. Minimize the weighted square error defined in (4.31) and get results for magnitude response H and group delay response τ_H.

4. Compute tolerance parameters δp, δs, and δt by either method in (4.35), then update weighting functions according to the strategies described by (4.32), (4.33), and (4.34).

5. Compare the resultant weighting function with the previous one, stop iteration if their difference is small enough, otherwise go to step 3.

4.3.3 Design Example

We now investigate the design of a 12^{th} order IIR filter to meet the following specifications: fp=0.5π, fs=0.7π, let λ=0.5, p=2 so that this is a least square error optimization problem,α in passband; stopband and transition band for magnitude response are 1.2, 1.0 and 1.0, respectively. α for group delay response is 1.8; The second method is used in (4.35).

At the beginning, we use the uniform weighting function for both the magnitude and group delay square error calculation. Figure 4.1 shows the resultant response and refreshed weighting function after the 1^{st} iteration.

Figure 4.2 and 4.3 are of the 3rd iteration and 6th iteration. According to these pictures, we demonstrated that with weight values of those peak error points increasing, both in magnitude and group delay responses, peak errors are constrained effectively. Weight peaks generally exist at band edges because, obviously, those maximal errors always present at band edges.

After several iterations, the objective filter could be hard to achieve further improvement without increasing the filter order. We have to increase order of this kind of filters in order to get a better performance.

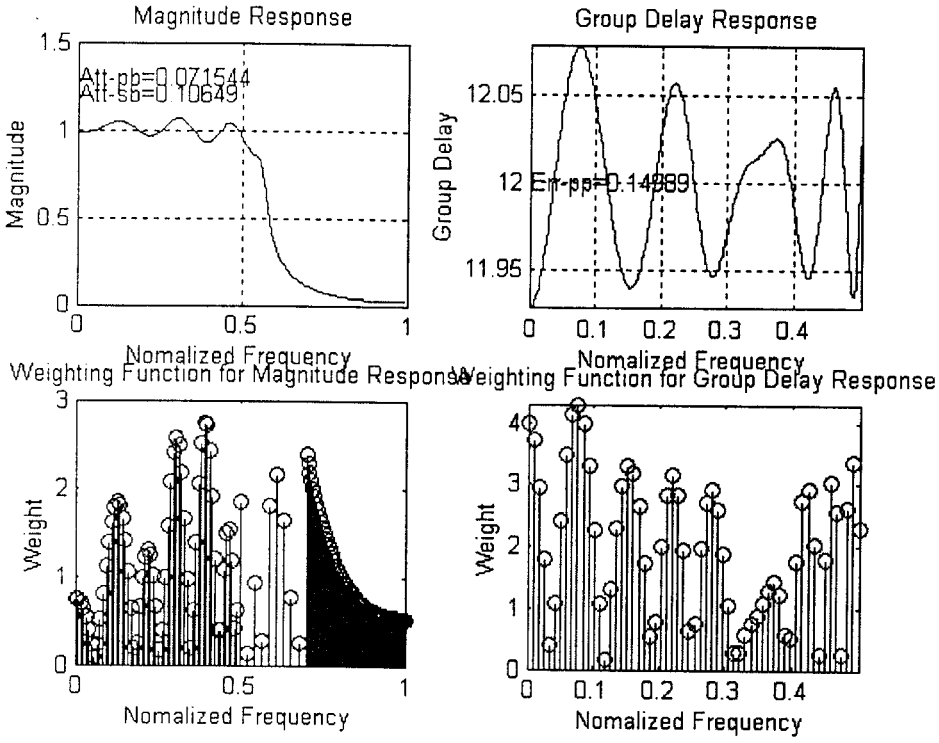


Figure 4.1 Optimization procedures (Iteration 1)

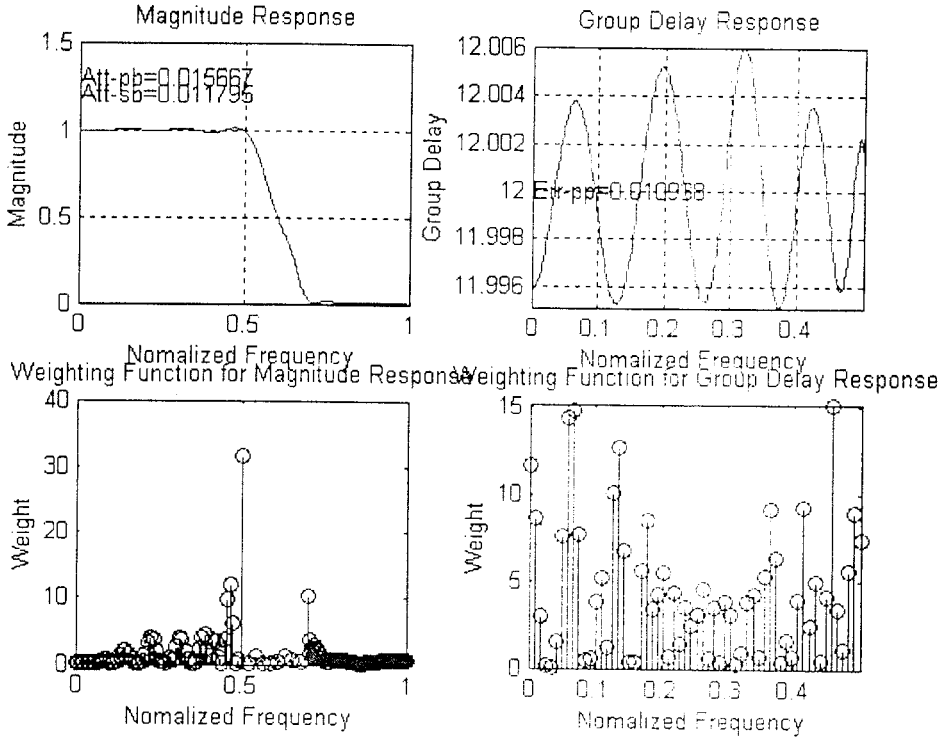


Figure 4.2 Optimization procedures (Iteration 3)

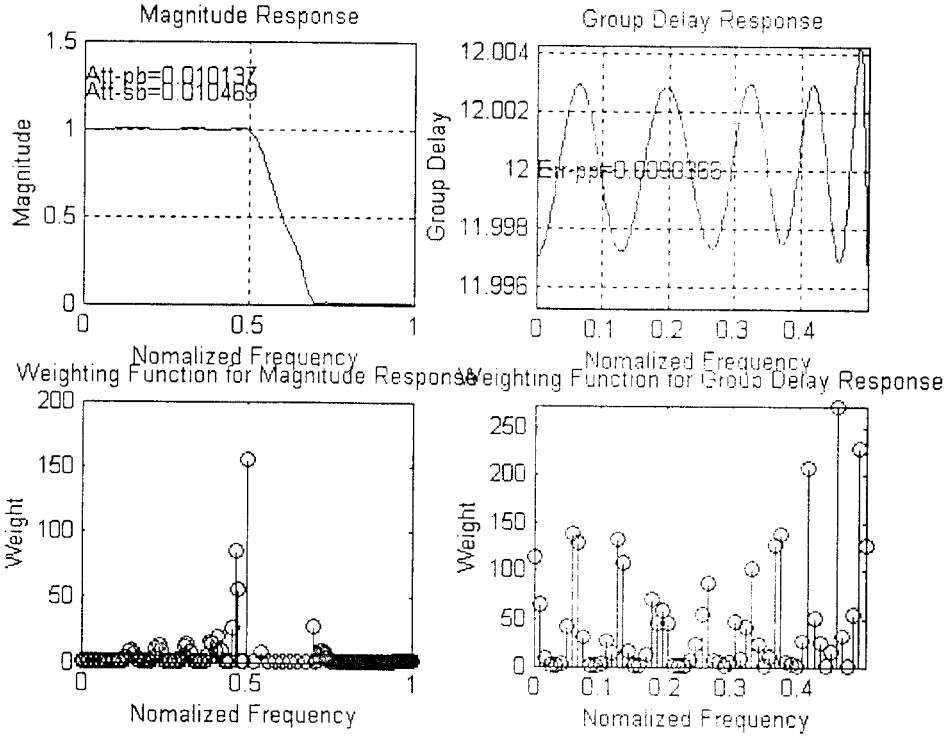


Figure 4.3 Optimization procedures (Iteration 6)

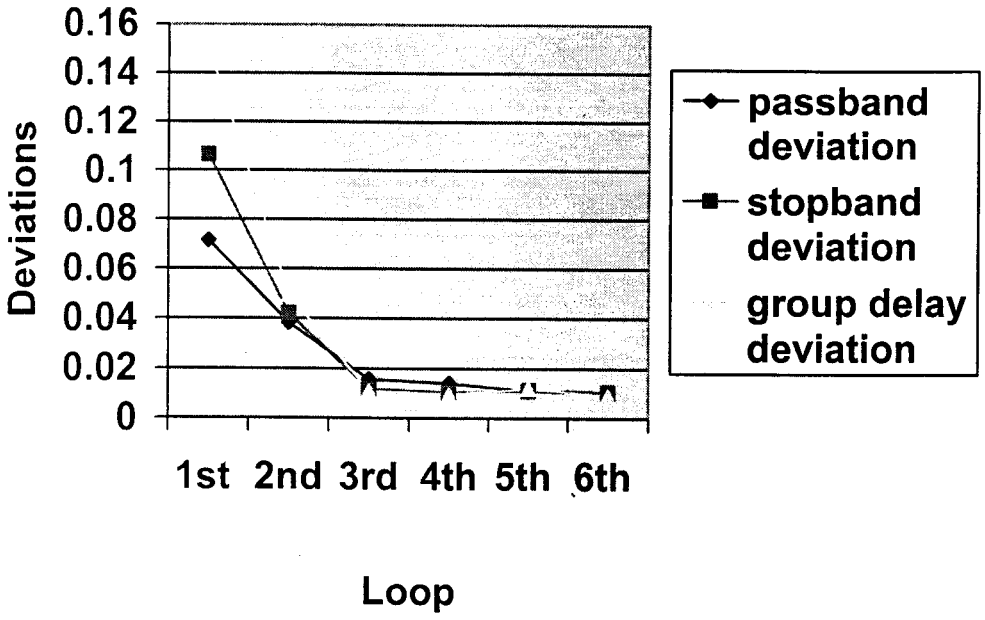


Figure 4.4 Optimization trend

As stated before, λ is the parameter for adjusting tradeoff between magnitude response and group delay response. How magnitude response and group delay response behave with respect to λ are illustrated by figure 4.4, 4.5 and 4.6.

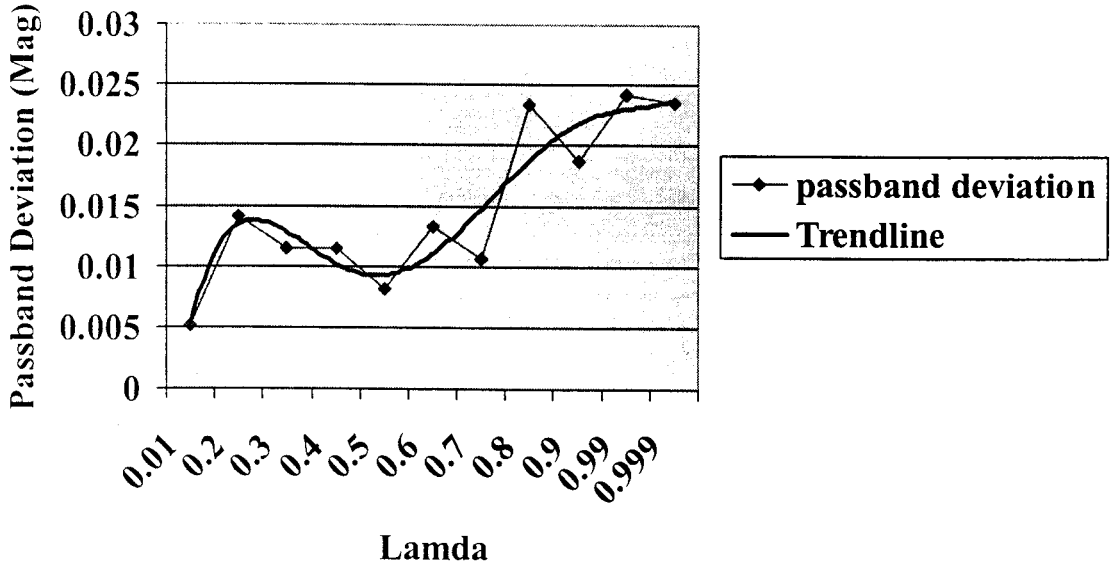


Figure 4.5 λ and magnitude response deviations in passband

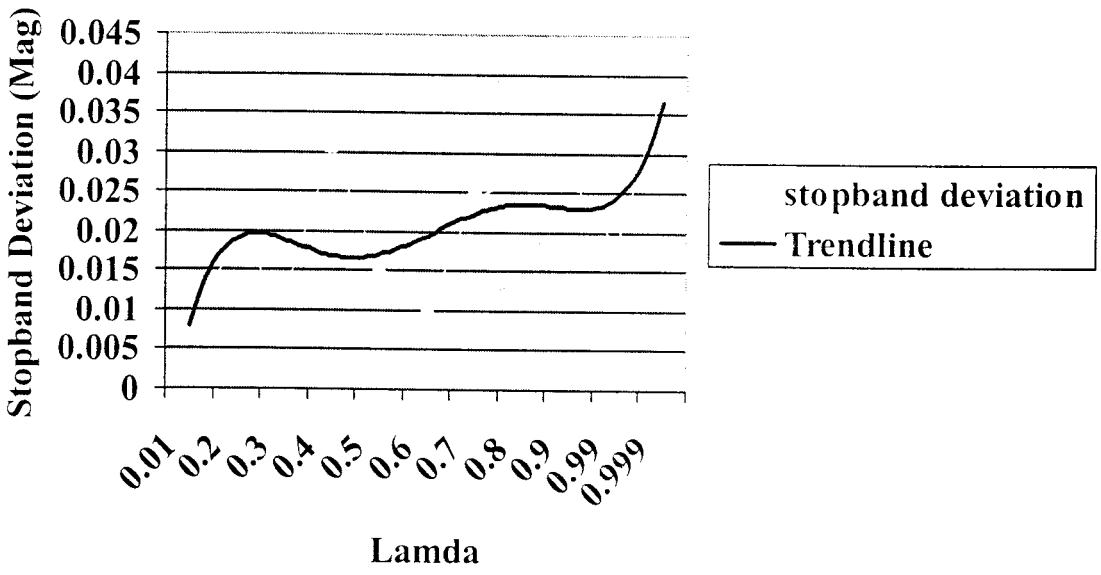


Figure 4.6 λ and magnitude response deviations in stopband

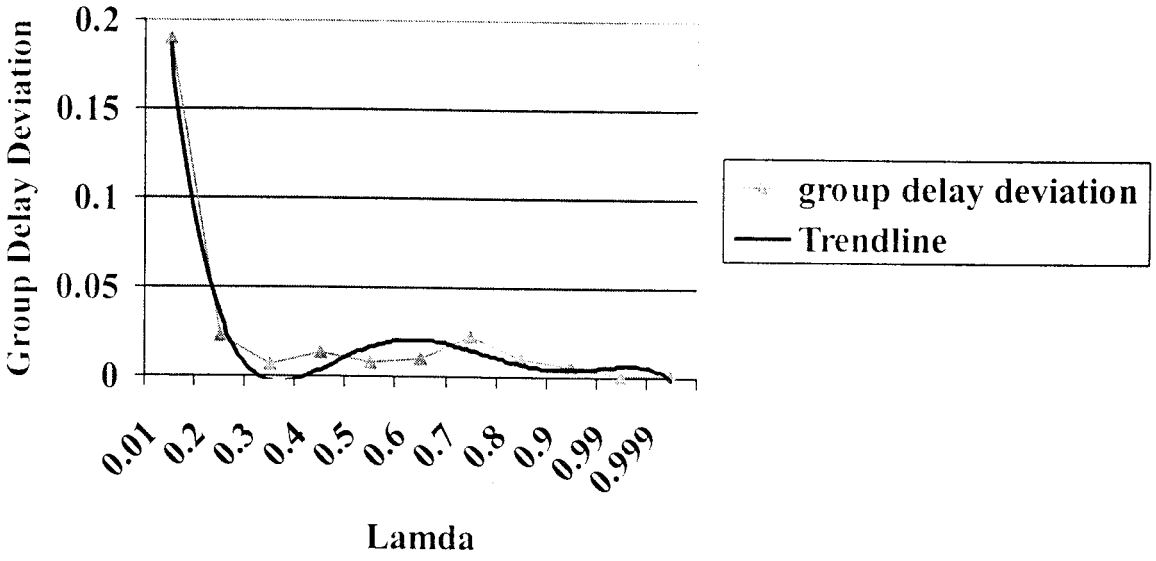


Figure 4.7 λ and group delay response deviations

4.4 Conclusion

In this chapter we showed how to design IIR digital filters according to the PCLS optimality criterion. This method is mainly based on nonlinear programming techniques, Deczky's method, to approximate given magnitude response and group delay simultaneously.

The method outlined in this chapter for the synthesis of IIR digital filters using a weighted minimum p-error criterion has been applied to a large class of problems. Thus filters having arbitrary magnitude, arbitrary group delay using allpass sections, and combined magnitude and group delay specifications, as well as allpass group delay equalizers, were synthesized using this method.

This method is highly flexible at balancing magnitude and group delay performances by modifying λ in (1.31) or relevant weighing function and tolerances updating strategy.

Chapter 5

Conclusions

This thesis explored and implemented a few methods for approximations of both linear phase and prescribed magnitude response of IIR filters. These methods based on nonlinear programming need more computation than those based on linear programming. For these research-purposed filters under a prerequisite of stability, their evaluations are mainly based on filter order, frequency response (magnitude and phase) and robustness.

By comparing two methods described in Chapter 2, we get the conclusion that introducing some ripples into MFD filters both in passband and stopband can remarkably increase filter's performance and efficiency. This is a kind of linear programming problem that can be solved without much computation time. From these points of view, it is very attractive and for this reason, they have been explored in greater details than others. It is a wise choice to make numerator and denominator have identical length because it boosts filter performance without an undesirable filter order increase.

Based on properties of paralleled allpass filters stated in Chapter 3, this design problem can be solved by different approaches mainly based on phase and magnitude response approximation and there are advantages and disadvantages in results with respect to different approximation

combinations. Among those examples, "phase optimization in passband, magnitude optimization in stopband (example 4)", and "magnitude optimization in passband, phase optimization in stopband (example 5)" are creatively introduced by this thesis. Besides their inherent low sensitivity, this and relative structures are widely applied in IIR filter design because of their high performance and efficiency.

The nonlinear optimization approach described in Chapter 4 is easily to get an almost perfect group delay response or linear phase with few filter order, but it is hard for attaining a satisfactory magnitude response compared with the former two methods even after weighting more on magnitude in error function. Therefore this kind of filters can be good candidate for some applications that are very strict with phase response but not magnitude response. In IIR digital filter design areas, this is a quite explicit approach that doesn't require much theoretic deduction and the future work should be focused on magnitude response improvements.

One of Remez Exchange Algorithm steps is the determination of candidate filter coefficients from candidate "alternation frequencies," which involves solving a set of linear equations (Chapter II) or nonlinear minimization problems (Chapter III and Chapter IV). Normally those alternation frequencies come from locations of extrema, under one constraint "whose magnitudes are larger than ripple height" (refer to [10], [11], [12]), over the frequency domain. However, this constraint has been removed for Remez Algorithm applications in this thesis because from lab experiences, I think it is not necessary, or even harmful for optimization procedures. The number

of interrupted cases caused by "ripples are less than specified number" is extremely reduced by eliminating this constraint.

References

1. *J.P.Thiran, "Recursive Digital Filters With Maximally Flat Group Delay," IEEE Trans. On Circuit Theory, CT-18, pp.659-664, Nov. 1971.*
2. *J.P.Thiran, "Equal Ripple Delay Recursive Digital Filters," IEEE Trans. On Circuit Theory, CT-18, pp.664-669, Nov. 1971.*
3. *A.G.Deczky, "Recursive Digital Filters Having Equiripple Group Delay," IEEE Trans. On Circuits and Systems, CAS-21, pp. 131-134, Jan 1974.*
4. *Hegde, R.; Shenoi, B.A. "Magnitude approximation of IIR digital filters with constant group delay response", Circuits and Systems, 1997*
5. *Makundi, M.; Valimaki, V.; Laakso, T.I. "Closed-form design of tunable fractional-delay allpass filter structures", 2001.*
6. *R. A. Gopinath, "Least Squared Error FIR Filters with Flat Amplitude and Group Delay Constraints",2002.*
7. *D. Economou, C. Mavroidis and I. Antoniadis, "robust residual vibration suppression using IIR digital filters", 2001*
8. *Makundi, M.; Valimaki, V.; Laakso, T.I. "Closed-form design of tunable fractional-delay allpass filter structures", 2001.*
9. *Selesnick, I.W.; Burrus, C.S. "Exchange algorithms that complement the Parks-McClellan algorithm for linear-phase FIR filter design" Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on , Volume: 44 Issue: 2 , Feb. 1997*

10. *Rajamohana Hegde, B,A,Shenoi, "Design of linear phase FIR filters with flat passband and flat or equiripple stopband magnitude response" IEEE, International Symposium on Circuits and Systems, June 9-12, 1997.*

11. *Rajamohana Hegde, Design of Digital Filters Using the Maximally Flat Criterion, M.S.Thesis, Wright State University, 1996.*

12. *Hegde, R.; Shenoi, B.A. "Magnitude approximation of IIR digital filters with constant group delay response" Proc. IEEE Int'l Sympo on Circuits and Systems, vol.IV,pp.2200-2203,1997.*

13. *T. Hinamoto and S. Maekawa, "Design of Two-dimensional recursive digital filters using mirror image polynomials", IEEE Trans. On Circuits and Systems, 1986.*

14. *A.G.Deczky, "Synthesis of recursive digital filters using the minimum p error criterion". IEEE Trans. On Audio and electroacoustics. AU-20, pp. 257-263, 1972.*

15. *James L. Sullivan and John W. Adams, "PCLS IIR Digital Filters with Simultaneous Frequency Response Magnitude and Group Delay Specifications", IEEE Transactios on signal processing, Vol. 46, No. 11, November 1998*

16. *M.Renfors and T. Saramaki, "A class of approximately linear phase filters composed of allpass subfilters", Proc. IEEE int'l Sympo. On Circuits and Systems, pp. 678-681, 1986.*

17. *Stancic, G.; Djuric, B. "Design of IIR digital filters using allpass networks", Telecommunications in Modern Satellite, Cable and Broadcasting Services, 4th International Conference on , Page(s): 145 -148 vol.1. Volume: 1 , 1999.*

18. *Artur Krukowski and Izzet Kale, "Almost linear phase polyphase IIR lowpass/hipass filter approach", 5th International Symposium on Signal Processing and its Applications (ISSPA99), Brisbane, Australia, August 22-25th, 1999.*

19. *T.Saramaki, "On the design of digital filters as a sum of two all-pass filters," IEEE "Trans. On Circuits and Systems, CAS-32, pp. 1191-1193, Nov 1985.*

20. *J.W.Adams, "FIR digital filters with Least Sqares Stopbands Subject to Peak-Gain Constraints," IEEE Transactions on circuits and systems, Vol, 39, No.4,April 1991.*

21. *Schittowski, K., "NLQPL: A FORTRAN-Subroutine Solving Constrained Nonlinear Programming Problems," Annals of Operations Research, Vol. 5, pp. 485-500, 1985.*

22. *Powell, M.J.D., "A Fast Algorithm for Nonlinearly Constrained Optimization Calculations," Numerical Analysis, G.A.Watson ed., Lecture Notes in Mathematics, Springer Verlag, Vol. 630, 1978.*

23. *Han, S.P., "A Globally Convergent Method for Nonlinear Programming," J. Optimization Theory and Applications, Vol. 22, p. 297, 1977.*

24. *Powell, M.J.D., "A Fast Algorithm for Nonlinearly Constrained Optimization Calculations," Numerical Analysis, G.A.Watson ed., Lecture Notes in Mathematics, Springer Verlag, Vol. 630, 1978.*

25. *A.G.Deczky, "Synthesis of recursive digital filters using the minimum p error criterion", IEEE Trans. On Audio and Electroacoustics. AU-20, pp. 257-263, 1972.*

26. *C.L.Lawson, "Contributions to the theory of linear least maximum approximation" Ph.D. dissertation, Dept. of Mathematics, University of California, Los Angeles, 1961.*

27. *John W. Adams, Qing Gao, and James L. Sullivan, "IIR Digital Filters with Peak-Constrained Least-Squared errors", IEEE 1058-6393/95.*

28. *"Digital Filters-Analysis, Design, and Applications", Second Edition, by Andreas Antoniou. University of Victoria. McGraw-Hill, Inc.*

29. *"Digital Signal Processing", Second Edition, by Sanjit K.Mitra. Department of Electrical and Computer Engineering University of California, Santa Barbara. McGraw-Hill Irwin.*

30. *"EE 4078 Digital Signal Processing", Prof. J. H. McClellan, Georgia Tech, 3-June-1998.*

31. *"Digital Filters-Analysis, Design, and Applications", Second Edition, by Andreas Antoniou. University of Victoria. McGraw-Hill, Inc.*

32. *"Magnitude and Delay Approximation of 1-D and 2-D Digital Filters", by B.A.Shenoi.Springer-Verlag Berlin Heidelberg 1999.*

VITA AUCTORIS

NAME: TAO DAI

PLACE OF BIRTH: SICHUAN, CHINA

DATE OF BIRTH: NOVEMBER, 1972

EDUCATION: B.A.Sc.
Department of Information & Electrical Engineering
University of ZheJiang
HangZhou, ZheJiang, China
1991-1995

M.A.Sc.
Department of Electrical & Computer Engineering
University of Windsor
Windsor, Ontario, Canada
2001-2003