M.Tech. credit seminar report,Electronic Systems Group, EE Dept, IIT Bombay, submitted November2002

FIR Filter Design Techniques

Arojit Roychowdhury (Roll No: 02307424)
Supervisor: Prof P.C. Pandey

Abstract

This report deals with some of the techniques used to design FIR filters. In the beginning, the windowing method and the frequency sampling methods are discussed in detail with their merits and demerits. Different optimization techniques involved in FIR filter design are also covered, including Rabiner's method for FIR filter design. These optimization techniques reduce the error caused by frequency sampling technique at the non-sampled frequency points. A brief discussion of some techniques used by filter design packages like Matlab are also included.

Introduction

FIR filters are filters having a transfer function of a polynomial in z^{-} and is an all-zero filter in the sense that the zeroes in the *z*-plane determine the frequency response magnitude characteristic. The *z* transform of a *N*-point FIR filter is given by

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

FIR filters are particularly useful for applications where exact linear phase response is required. The FIR filter is generally implemented in a non-recursive way which guarantees a stable filter. FIR filter design essentially consists of two parts
(i) approximation problem
(ii) realization problem

The approximation stage takes the specification and gives a transfer function through four steps. They are as follows:
(i) A desired or ideal response is chosen, usually in the frequency domain.
(ii) An allowed class of filters is chosen (e.g. the length *N* for a FIR filters).
(iii) A measure of the quality of approximation is chosen.
(iv) A method or algorithm is selected to find the best filter transfer function.

The realization part deals with choosing the structure to implement the transfer function which may be in the form of circuit diagram or in the form of a program.

There are essentially three well-known methods for FIR filter design namely:

(1) The window method
(2) The frequency sampling technique
(3) Optimal filter design methods

The Window Method

In this method, [Park87], [Rab75], [Proakis00] from the desired frequency response specification $H_d(w)$, corresponding unit sample response $h_d(n)$ is determined using the following relation

$$h_d(n)=\frac{1}{2\pi}\int_{-\pi}^{\pi} H_d(w)\, e^{jwn} dw \qquad (2)$$

where $$H_d(w) = \sum_{n=-\infty}^{\infty} h_d(n)\, e^{-jwn} \qquad (3)$$

In general, unit sample response $h_d(n)$ obtained from the above relation is infinite in duration, so it must be truncated at some point say $n= M$-1 to yield an FIR filter of length M (i.e. 0 to M-1). This truncation of $h_d(n)$ to length M-1 is same as multiplying $h_d(n)$ by the rectangular window defined as

$$w(n) = \begin{cases} 1 & 0 \le n \le M\text{-}1 \\ 0 & \text{otherwise} \end{cases} \qquad (4)$$

Thus the unit sample response of the FIR filter becomes

$$\begin{aligned} h(n) &= h_d(n)\, w(n) \\ &= h_d(n) \qquad 0 \le n \le M\text{-}1 \\ &= 0 \qquad \text{otherwise} \end{aligned} \qquad (5)$$

Now, the multiplication of the window function $w(n)$ with $h_d(n)$ is equivalent to convolution of $H_d(w)$ with $W(w)$, where $W(w)$ is the frequency domain representation of the window function

$$W(w) = \sum_{n=0}^{M-1} w(n) e^{-jwn} \qquad (6)$$

Thus the convolution of $H_d(w)$ with $W(w)$ yields the frequency response of the truncated FIR filter

$$H(w) = \frac{1}{2\pi}\int_{-\pi}^{\pi} H_d(v) W(w-v) dw \qquad (7)$$

The frequency response can also be obtained using the following relation

$$H(w)= \sum_{n=0}^{M-1} h(n) e^{-jwn} \qquad (8)$$

But direct truncation of $h_d(n)$ to M terms to obtain $h(n)$ leads to the Gibbs phenomenon effect which manifests itself as a fixed percentage overshoot and ripple before and after an approximated discontinuity in the frequency response due to the non-uniform convergence of the fourier series at a discontinuity.Thus the frequency response obtained by using (8) contains ripples in the frequency domain. In order to reduce the ripples, instead of multiplying $h_d(n)$ with a rectangular window $w(n)$, $h_d(n)$ is multiplied with a window function that contains a taper and decays toward zero gradually, instead of abruptly as it occurs in a rectangular window. As multiplication of sequences $\mathrm{h_d(n)}$ and $w(n)$ in time domain is equivalent to convolution of $H_d(w)$ and $W(w)$ in the frequency domain, it has the effect of smoothing $H_d(w)$.

The several effects of windowing the Fourier coefficients of the filter on the result of the frequency response of the filter are as follows:

(i) A major effect is that discontinuities in $H(w)$ become transition bands between values on either side of the discontinuity.

(ii) The width of the transition bands depends on the width of the main lobe of the frequency response of the window function, $w(n)$ i.e. $W(w)$.

(iii) Since the filter frequency response is obtained via a convolution relation , it is clear that the resulting filters are never optimal in any sense.

(iv) As M (the length of the window function) increases, the mainlobe width of $W(w)$ is reduced which reduces the width of the transition band, but this also introduces more ripple in the frequency response.

(v) The window function eliminates the ringing effects at the bandedge and does result in lower sidelobes at the expense of an increase in the width of the transition band of the filter.

Some of the windows [Park87] commonly used are as follows:

1. Bartlett triangular window:

$$W(n) = \frac{2(n+1)}{N+1} \qquad n = 0,1,2,\ldots\ldots.,(N\text{-}1)/2 \tag{9}$$

$$= 2 - \frac{2(n+1)}{N+1} \qquad n = (N\text{-}1)/2,\ldots\ldots,N\text{-}1$$

$$= 0 , \qquad \text{otherwise}$$

-

2-5. Generalized cosine windows
(Rectangular, Hanning, Hamming and Blackman)

$$W(n) = \mathrm{a} - \mathrm{b}\cos(2\mathrm{p}(n+1)/(N+1)) + \mathrm{c}\cos(4\mathrm{p}(n+1)/(N+1)) \qquad n = 0,1\ldots.N\text{-}1 \tag{10}$$

$$= 0 \qquad \text{otherwise}$$

6. Kaiser window with parameter ß :

$$W(n) = \frac{I_o(\beta\sqrt{1-(2(n+1)/(N+1))^2}\,)}{I_o(\beta)} \qquad n = 0,1,\ldots.,N\text{-}1 \tag{11}$$

$$= 0 \qquad \text{otherwise}$$

The general cosine window has four special forms that are commonly used. These are determined by the parameters *a,b,c*

TABLE I
Value of coefficients for a,b and c from [Park87]

Window	a	b	c
Rectangular	1	0	0
Hanning	0.5	0.5	0

Hamming	0.54	0.46	0
Blackman	0.42	0.5	0.08

The Bartlett window reduces the overshoot in the designed filter but spreads the transition region considerably. The Hanning, Hamming and Blackman windows use progressively more complicated cosine functions to provide a smooth truncation of the ideal impulse response and a frequency response that looks better. The best window results probably come from using the Kaiser window, which has a parameter ß that allows adjustment of the compromise between the overshoot reduction and transition region width spreading.

The major advantages of using window method is their relative simplicity as compared to other methods and ease of use. The fact that well defined equations are often available for calculating the window coefficients has made this method successful.

There are following problems in filter design using window method:

(i) This method is applicable only if $H_d(w)$ is absolutely integrable i.e only if (2) can be evaluated. When $H_d(w)$ is complicated or cannot easily be put into a closed form mathematical expression, evaluation of $h_d(n)$ becomes difficult.

(ii) The use of windows offers very little design flexibility e.g. in low pass filter design, the passband edge frequency generally cannot be specified exactly since the window smears the discontinuity in frequency. Thus the ideal LPF with cut-off frequency fc, is smeared by the window to give a frequency response with passband response with passband cutoff frequency f_1 and stopband cut-off frequency f_2.

(iii) Window method is basically useful for design of prototype filters like lowpass,highpass,bandpass etc. This makes its use in speech and image processing applications very limited.

The Frequency Sampling Technique

In this method, [Park87], [Rab75], [Proakis00] the desired frequency response is provided as in the previous method. Now the given frequency response is sampled at a set of equally spaced frequencies to obtain N samples. Thus , sampling the continuous frequency response $H_d(w)$ at N points essentially gives us the N-point DFT of $H_d(2pnk/N)$. Thus by using the IDFT formula, the filter co-efficients can be calculated using the following formula

$$h(n)=\frac{1}{N}\sum_{n=0}^{N-1} H(k)e^{j(2\pi n/N)k} \qquad (12)$$

Now using the above N-point filter response, the continuous frequency response is calculated as an interpolation of the sampled frequency response. The approximation error would then be exactly zero at the sampling frequencies and would be finite in frequencies between them. The smoother the frequency response being approximated, the smaller will be the error of interpolation between the sample points.

One way to reduce the error is to increase the number of frequency samples [Rab75]. The other way to improve the quality of approximation is to make a number of frequency samples specified as unconstrained variables. The values of these unconstrained variables are generally optimized by computer to minimize some simple function of the approximation error e.g. one might choose as unconstrained variables the frequency samples that lie in a transition band between two frequency bands in which the frequency response is specified e.g. in the band between the passband and the stopband of a low pass filter.

There are two different set of frequencies that can be used for taking the samples. One set of frequency samples are at $f_k = k/N$ where $k = 0,1,\ldots.N-1$. The other set of uniformly spaced frequency samples can be taken at $f_k = (k + ½)/N$ for $k = 0,1,\ldots.N-1$.

The second set gives us the additional flexibility to specify the desired frequency response at a second possible set of frequencies. Thus a given band edge frequency may be closer to type-II frequency sampling point that to type-I in which case a type-II design would be used in optimization procedure.

In a paper by Rabiner and Gold [Rabi70], Rabiner has mentioned a technique based on the idea of frequency sampling to design FIR filters. The steps involved in this method suggested by Rabiner are as follows:

(i) The desired magnitude response is provided along with the number of samples,N . Given N, the designer determines how fine an interpolation will be used.

(ii) It was found by Rabiner that for designs they investigated, where N varied from 15 to 256, 16N samples of $H(w)$ lead to reliable computations, so 16 to 1 interpolation was used.

(iii) Given N values of H_k , the unit sample response of filter to be designed, $h(n)$ is calculated using the inverse FFT algorithm.

(iv) In order to obtain values of the interpolated frequency response two procedures were suggested by Rabiner. They are

(a) $h(n)$ is rotated by $N/2$ samples(N even) or $(N-1)/2$ samples for N odd to remove the sharp edges of impulse response, and then 15N zero-valued samples are symmetrically placed around the impulse response.

(b) $h(n)$ is split around the $N/2^{nd}$ sample, and 15N zero-valued samples are placed between the two pieces of the impulse response.

(v) The zero augmented sequences are transformed using the FFT algorithm to give the interpolated frequency responses.

Merits of frequency sampling technique

(i) Unlike the window method, this technique can be used for any given magnitude response.
(ii) This method is useful for the design of non-prototype filters where the desired magnitude response can take any irregular shape.

There are some disadvantages with this method i.e the frequency response obtained by interpolation is equal to the desired frequency response only at the sampled points. At the other points, there will be a finite error present.

Optimal Filter Design Methods

Many methods are present under this category. The basic idea in each method is to design the filter coefficients again and again until a particular error is minimized. The various methods are as follows:

(i) Least squared error frequency domain design
(ii) Weighted Chebyshev approximation
(iii) Nonlinear equation solution for maximal ripple FIR filters
(iv) Polynomial interpolation solution for maximal ripple FIR filters

Least squared error frequency domain design

As seen in the previous method of frequency sampling technique there is no constraint on the response between the sample points, and poor results may be obtained.

The frequency sampling technique is more of an interpolation method rather than an approximation method. This method [Rab75], [Parks87] controls the response between the sample points by considering a number of sample points larger than the order of the filter.The purpose of most filters is to separate desired signals from undesired signals or noise. As the energy of the signal is related to the square of the signal, a squared error approximation criterion is appropriate to optimize the design of the FIR filters.

The frequency response of the FIR filter is given by (8) for a N-point FIR filter. An error function is defined as follows

$$\text{Error}(E) = \Sigma\ |(H(w_k)\text{-}H_d(w_k)|^2 \tag{13}$$

where w_k= $(2\pi k)/L$ and $H_d(w_k)$ are L samples of the desired response, which is the error measure as a sum of the squared differences between the actual and desired frequency response over a set of L frequency samples. The method consists of the following steps:

(i) First 'L' samples from the continuous frequency response are taken, where L>N(length of the impulse response of filter to be designed)

(ii) Then using the following formula

$$h(n)= \frac{1}{N}\sum_{n=0}^{L-1} H(k)e^{j(2\pi n/N)k} \tag{14}$$

the L-point filter impulse response is calculated.

(iii) Then the obtained filter impulse response is symmetrically truncated to desired length N.

(iv) Then the frequency response is calculated using the following relation

$$H(w)= \sum_{n=0}^{N-1} h(n)e^{-jwn} \tag{15}$$

(v) The magnitude of the frequency response at these frequency points for w_k = $(2\pi k)/L$ will not be equal to the desired ones , but the overall least square error will be reduced effectively this will reduce the ripple in the filter response.

To further reduce the ripple and overshoot near the band edges, a transition region will be defined with a linear transfer function. Then the L frequency samples are taken at w_k = $(2\pi k)/L$ using which the first N samples of the filter are calculated using the above method.Using this method, reduces the ripple in the interpolated frequency response.

Weighted Chebyshev Approximation

In this method, [Rab75] following terms are defined

$H_d(w)$ = the desired (real) frequency response of the filter
$H(w)$= the frequency response of the designed filter
$W(w)$= the frequency response of the weighting function

The weighting function enables the designer to choose the relative size of the error in different frequency bands. The frequency response of linear phase filters for four different types can be written as follows

$$H(\mathrm{w}) = \mathrm{e}^{-jw(N\text{-}1)/2}\, e^{j(\mathrm{p}/2)L}\, H^*(w) \tag{16}$$

TABLE II
Different expressions for $H^*(e^{jw})$ for different types of filter from [Rab75]

	L	$H^*(e^{jw})$
Case 1- *N* odd Symmetrical impulse Response	0	$\sum_{n=0}^{(N-1)/2} a(n)\cos(wn)$
Case 2- N even Symmetrical impulse Response	0	$\sum_{n=1}^{N/2} b(n)\cos(w(n-1/2))$
Case 3- *N* odd Anti-symmetrical impulse Response	1	$\sum_{n=1}^{(N-1)/2} c(n)\sin(wn)$
Case 4- *N* even Anti-symmetrical impulse Response	1	$\sum_{n=1}^{N/2} d(n)\sin(w(n-1/2))$

Now each of the expressions for $H^*(w)$ can be written as a product of a fixed function of w,$Q(w)$ and a term that is a sum of cosines, $P(w)$. The expressions for $P(w)$ and $Q(w)$ are as follows:

TABLE III
Expressions for $P(w)$ and $Q(w)$ for different types of filter from [Rab75]

	$Q(w)$	$P(w)$
Case 1	1	$\sum_{n=0}^{(N-1)/2} \tilde{a}(n)\cos(wn)$
Case 2	Cos(w/2)	$\sum_{n=0}^{(N/2)-1} \tilde{b}(n)\cos(wn)$
Case 3	sin(w)	$\sum_{n=0}^{(N-3)/2} \tilde{c}(n)\cos(wn)$
Case 4	sin(w/2)	$\sum_{n=0}^{(N/2)-1} \tilde{d}(n)\cos(wn)$

The weighted error of approximation E(w) is by definition

$$E(w)= W(w)[\, H_d\,(w) - H^*(w)] \tag{17}$$

$$E(w)= W(w)[\, H_d\,(w) - P(w)\, Q(w)] \tag{18}$$

As $Q(w)$ is a fixed function of frequency, we can factor out $Q(w)$

$$E(w) = W(w)Q(w)[\, H_d\,(w)/Q(w)\ -\ P(w)] \tag{19}$$

Then two more terms are defined

$$\hat{W}(w) = W(w)Q(w) \tag{20}$$

$$H_d{}^*(w) = H_d(w)/Q(w) \tag{21}$$

The error function can now be written as

$$E(w) = \hat{W}(w)\,[\, H_d{}^*(w) - P(w)\, Q(w)] \tag{22}$$

Thus the Chebyshev approximation consists of finding the set of coefficients $\tilde{a}(n)$ to $\tilde{d}(n)$ so as to minimize the maximum absolute value of $E(w)$ over the frequency bands in which the approximation is being performed. The Chebyshev approximation problem may be stated mathematically as

$$|E(w)| = \min\,[\max|E(w)|] \tag{23}$$

The solution to this problem is given by Parks and McClellan [Proakis00] who applied a theorem in theory of Chebyshev approximation called the alternation theorem.

Nonlinear Equation solution for maximal ripple FIR filters

The real part of the frequency response of the designed FIR filter can be written as $\Sigma a(n)\cos(wn)$ [Rab75] where limits of summation and $a(n)$ vary according to the type of the filter. The number of frequencies at which $H(w)$ could attain an extremum is strictly a function of the type of the linear phase filter i.e. whether length N of filter is odd or even or filter is symmetric or anti-symmetric. At each extremum, the value of $H(w)$ is predetermined by a combination of the weighting function $W(w)$, the desired frequency response, and a quantity δ that represents the peak error of approximation distributing the frequencies at which $H(w)$ attains an extremal value among the different frequency bands over which a desired response was being approximated. Since these filters have the maximum number of ripples, they are called maximal ripple filters.

This method is as follows:

1.At each of the N_e unknown external frequencies, $E(w)$ attains the maximum value of either $\pm\,\delta$ and $E(w)$ or equivalently $H(w)$ has zero derivative. Thus two N_e equations of the form

$$H(\mathrm{w_i}) = \pm\delta/W(w)\quad +\ D(w_i) \tag{24}$$

$$\mathrm{d/dw}\ \{H(w_{i)}\}\ \text{at}\ w = w_i\ = 0 \tag{25}$$

are obtained.

These equations represent a set of $2N_e$ nonlinear equations in two N_e unknowns, N_e impulse response coefficients and N_e frequencies at which $H(w)$ obtains the extremal value. The set of two N_e equations may be solved iteratively using nonlinear optimiation procedure.

An important thing to note is that here the peak error (δ) is a fixed quantity and is not minimized by the optimization scheme. Thus the shape of $H(w)$ is postulated apriori and only the frequencies at which $H(w)$ attains the extremal values are unknown.

The disadvantage of this method is that the design procedure has no way of specifying band edges for the different frequency bands of the filter. Thus the optimization algorithm is free to select exactly where the bands will lie.

Polynomial Interpolation Solution for Maximal Ripple FIR filters

This algorithm [Rab75] is basically an iterative technique for producing a polynomial $H(w)$ that has extrema of desired values. The algorithm begins by making an initial estimate of the frequencies at which the extrema in $H(w)$ will occur and then uses the well-known Lagrange interpolation formula to obtain a polynomial that alternatively goes through the maximum allowable ripple values at these frequencies. It has been experimentally found that the initial guess of extremal frequencies does not affect the ultimate convergence of the algorithm but instead affects the number of iterations required to achieve the desired result.

Let us consider the case of design of a low pass filter using the above method.

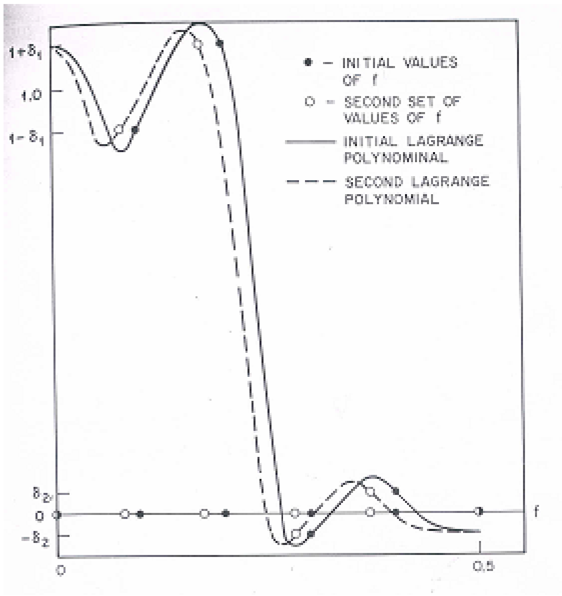

Fig. 1. Iterative solution for a maximum ripple lowpass filter from [Rab75]

The Fig. 1 shows the response of a lowpass filter with N = 11.The number of extremal frequencies i.e. the frequencies where ripples occur are 6 in this case. They are divided into 3 passband extrema and 3 stopband extrema. The filled dots indicate the initial guess as to the extremal frequencies of $H(w)$. The solid line is the initial Lagrange polynomial obtained by choosing polynomial coefficients so that the values of the polynomial at the guessed set of frequencies are identical to the assigned extreme values. But this polynomial has extrema that exceeds the specified maxima values. The next stage of the algorithm is to locate the frequencies at which the extrema of the first Lagrange interpolation occur. These frequencies are now used as the new frequencies for which the extrema of the filter response occur. This second set of frequencies are indicated by open dots in Fig. 1. Now similarly the new set of frequencies are taken as those frequencies where the maximum exceeds the specified maxima. Thus the method is completely iterative in nature.

Conclusions

The report has described the various techniques involved in the design of FIR filters. Every method has its own advantages and disadvantages and is selected depending on the type of filter to be designed.The window method is basically used for the design of prototype filters like the low-pass, high-pass, band-pass etc. They are not very suitable for designing of filters with any given frequency response. On the other hand, the frequency sampling technique is suitable for designing of filters with a given magnitude response. The ideal frequency response of the filter is approximated by placing appropriate frequency samples in the z^{-} plane and then calculating the filter co-efficients using the IFFT algorithm.

The disadvantage of the frequency sampling technique was that the frequency response gave errors at the points where it was not sampled. In order to reduce these erros the different optimization technique for FIR filter design were presented wherein the remaining frequency samples are chosen to satisfy an optimization criterion. An appendix consisting of the filter design methods used by the software package Matlab is also presented.

Appendix

Matlab [Mat02] is a software that is used in a number of applications like signal processing and control system. The Signal Processing Toolbox provides functions that support a range of filter design and implementation methodologies. Some of the techniques used by Matlab for FIR filter design are as follows:

1. Windowing
2. Multiband with Transition bands
3. Arbitrary response
4. Raised cosine

Windowing

Three different functions i.e. fir1,fir2 and kaiserord [Mat02] are used to design FIR filters. Fir1 function implements the classical method of windowed linear-phase FIR digital filter design. It is used for design of filters in standard lowpass, highpass, bandpass, and bandstop configurations.

Fir2 function is used for designing of frequency sampling-based digital FIR filters with arbitrarily shaped frequency response.

Kaiserord function returns a filter order n and beta parameter to specify a Kaiser window for use with the fir1 function. Given a set of specifications in the frequency domain, kaiserord estimates the minimum FIR filter order that will approximately meet the specifications. kaiserord converts the given filter specifications into passband and stopband ripples and converts cutoff frequencies into the form needed for windowed FIR filter design.

Multiband with Transition bands

Two different functions i.e. firls and remez [Mat02] are used to design FIR fitlers. The firls function is used to design a linear-phase FIR filter that minimizes the weighted, integrated squared error between an ideal piecewise linear function and the magnitude response of the filter over a set of desired frequency bands.

The remez function is used to design a linear-phase FIR filter using the Parks-McClellan algorithm. The Parks-McClellan algorithm uses the Remez exchange algorithm and Chebyshev approximation theory to design filters with an optimal fit between the desired and actual frequency responses.

Arbitrary response

This method uses the cremez [Mat02] function to design the filter. The cremez function allows arbitrary frequency-domain constraints to be specified for the design of a possibly complex FIR filter. The Chebyshev (or minimax) filter error is optimized, producing equiripple FIR filter designs.

Raised cosine

This method uses the firrcos [Mat02] function to design FIR filters. The function firrcos returns an order n lowpass linear-phase FIR filter with a raised cosine transition band.

References

[Park87] T.W. Parks and C.S. Burrus, *Digital Filter Design*. New York:Wiley,1987

[Rab75] L.R. Rabiner and B. Gold, *Theory and Applications of Digital Signal Processing.* New Jersey: Prentice-Hall, 1975

[Proakis00] J.G. Proakis and D.G. Manolakis, *Digital Signal Processing-Principles,Algorithms and Applications* New Delhi: Prentice-Hall, 2000

[Rabi70] L.R. Rabiner, B. Gold and C.A. McGonegal, “An approach to the Approximation Problem for Nonrecursive Digital Filters,” *IEEE Trans. Audio and Electroacoustics,* vol. AU-18, pp. 83-105,June 1970.

[Mat02] http://www.mathworks.com/access/helpdesk/help/toolbox/signal/signal.shtml Signal Processing Toolbox,The Mathworks,Inc., accessed October 25,2002