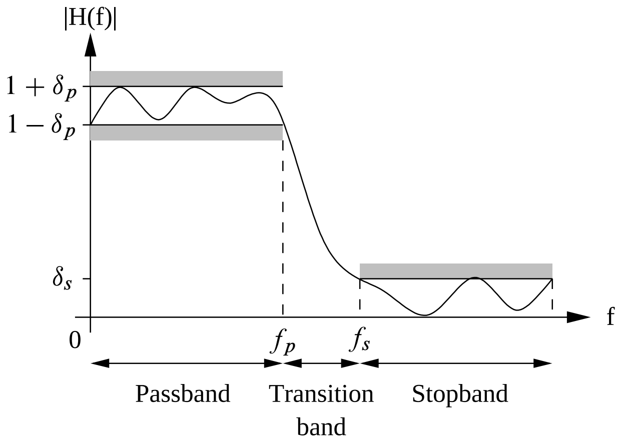
Filter design

1 Design considerations: a framework



The design of a digital filter involves five steps:

- **Specification:** The characteristics of the filter often have to be specified in the frequency domain. For example, for frequency selective filters (lowpass, highpass, bandpass, etc.) the specification usually involves tolerance limits as shown above.
- **Coefficient calculation:** Approximation methods have to be used to calculate the values $h[k]$ for a FIR implementation, or a_k, b_k for an IIR implementation. Equivalently, this involves finding a filter which has $H(z)$ satisfying the requirements.
- **Realisation:** This involves converting $H(z)$ into a suitable filter structure. Block or flow diagrams are often used to depict filter structures, and show the computational procedure for implementing the digital filter.

- **Analysis of finite wordlength effects:** In practice one should check that the quantisation used in the implementation does not degrade the performance of the filter to a point where it is unusable.
- **Implementation:** The filter is implemented in software or hardware. The criteria for selecting the implementation method involve issues such as real-time performance, complexity, processing requirements, and availability of equipment.

2 Finite impulse response (FIR) filter design

A FIR filter is characterised by the equations

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

The following are useful properties of FIR filters:

- They are always stable — the system function contains no poles. This is particularly useful for adaptive filters.
- They can have an exactly linear phase response. The result is no frequency dispersion, which is good for pulse and data transmission.
- Finite length register effects are simpler to analyse and of less consequence than for IIR filters.
- They are very simple to implement, and all DSP processors have architectures that are suited to FIR filtering.
- For large N (many filter taps), the FFT can be used to improve performance.

Of these, the linear phase property is probably the most important. A filter is said to have a generalised linear phase response if its frequency response can be expressed in the form

$$H(e^{j\omega}) = A(e^{j\omega})e^{-j\alpha\omega+j\beta}$$

where α and β are constants, and $A(e^{j\omega})$ is a real function of ω. If this is the case, then

- If A is positive, then the phase is

$$\sphericalangle H(e^{j\omega}) = \beta - \alpha\omega.$$

  If A is negative, then

$$\sphericalangle H(e^{j\omega}) = \pi + \beta - \alpha\omega.$$

In either case, the phase is a linear function of ω.

It is common to restrict the filter to having a real-valued impulse response $h[n]$, since this greatly simplifies the computational complexity in the implementation of the filter.

A FIR system has linear phase if the impulse response satisfies either the even symmetric condition

$$h[n] = h[N-1-n],$$

or the odd symmetric condition

$$h[n] = -h[N-1-n].$$

The system has different characteristics depending on whether N is even or odd. Furthermore, it can be shown that all linear phase filters must satisfy one of these conditions. Thus there are exactly four types of linear phase filters.

Consider for example the case of an odd number of samples in $h[n]$, and even symmetry. The frequency response for $N = 7$ is

$$\begin{aligned} H(e^{j\omega}) &= \sum_{n=0}^{6} h[n]e^{-j\omega n} \\ &= h[0] + h[1]e^{-j\omega} + h[2]e^{-j2\omega} + h[3]e^{-j3\omega} + h[4]e^{-j4\omega} \\ &\qquad + h[5]e^{-j5\omega} + h[6]e^{-j6\omega} \\ &= e^{-j3\omega}(h[0]e^{j3\omega} + h[1]e^{j2\omega} + h[2]e^{j\omega} + h[3] + h[4]e^{-j\omega} \\ &\qquad + h[5]e^{-j2\omega} + h[6]e^{-j3\omega}). \end{aligned}$$

The specified symmetry property means that $h[0] = h[6]$, $h[1] = h[5]$, and $h[2] = h[4]$, so

$$\begin{aligned} H(e^{j\omega}) &= e^{-j3\omega}(h[0](e^{j3\omega} + e^{-j3\omega}) + h[1](e^{j2\omega} + e^{-j2\omega}) \\ &\qquad + h[2](e^{j\omega} + e^{-j\omega}) + h[3]) \\ &= e^{-j3\omega}(2h[0]\cos(3\omega) + 2h[1]\cos(2\omega) + 2h[2]\cos(\omega)) \\ &= e^{-j3\omega}\sum_{n=0}^{3} a[n]\cos(\omega n), \end{aligned}$$

where $a[0] = h[3]$, and $a[n] = 2h[3-n]$ for $n = 1, 2, 3$. The resulting filter clearly has a linear phase response for real $h[n]$. It is quite simple to show that in general for odd values of N the frequency response is

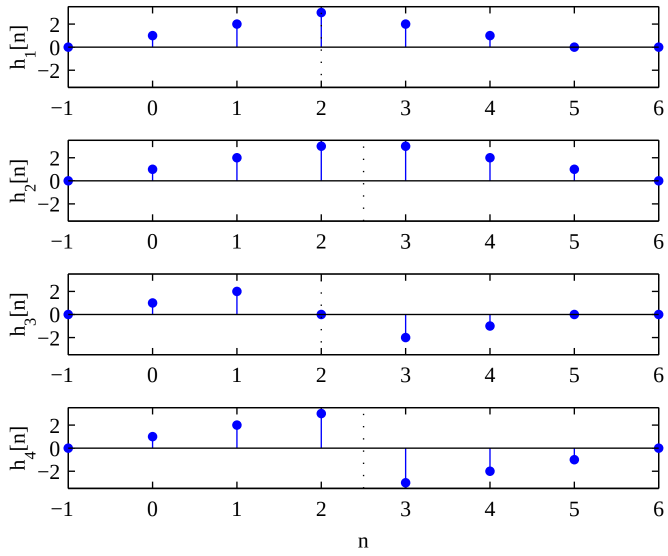
$$H(e^{j\omega}) = e^{-j\omega(N-1)/2}\sum_{n=0}^{(N-1)/2} a[n]\cos(\omega n),$$

for a set of real-valued coefficients $a[0], \ldots, a[(N-1)/2]$. As different values for $a[n]$ are selected, different linear-phase filters are obtained.

The cases of N odd and $h[n]$ antisymmetric are similar to that presented, and the frequency responses are summarised in the following table:

Symmetry	**N**	$H(e^{j\omega})$	**Type**
Even	Odd	$e^{-j\omega(N-1)/2} \sum_{n=0}^{(N-1)/2} a[n]\cos(\omega n)$	1
Even	Even	$e^{-j\omega(N-1)/2} \sum_{n=1}^{N/2} b[n]\cos(\omega(n-1/2))$	2
Odd	Odd	$e^{-j[\omega(N-1)/2-\pi/2]} \sum_{n=0}^{(N-1)/2} a[n]\sin(\omega n)$	3
Odd	Even	$e^{-j[\omega(N-1)/2-\pi/2]} \sum_{n=1}^{N/2} b[n]\sin(\omega(n-1/2))$	4

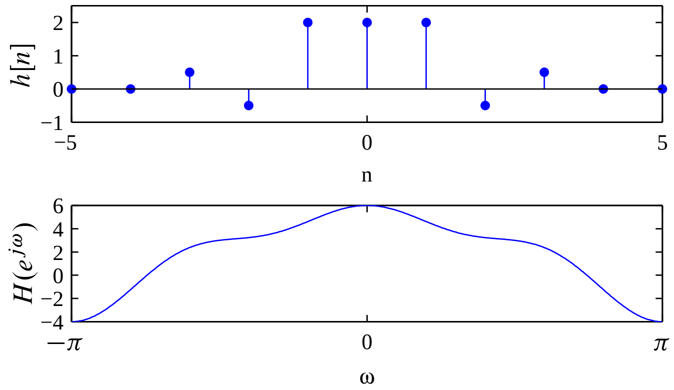
Recall that even symmetry implies $h[n] = h[N-1-n]$ and odd symmetry $h[n] = -h[N-1-n]$. Examples of filters satisfying each of these symmetry conditions are:



The center of symmetry is indicated by the dotted line.

The process of linear-phase filter design involves choosing the $a[n]$ values to obtain a filter with a desired frequency response. This is not always possible, however — the frequency response for a type II filter, for example, has the property that it is *always* zero for $\omega = \pi$, and is therefore not appropriate for a highpass filter. Similarly, filters of type 3 and 4 introduce a 90° phase shift, and have a frequency response that is always zero at $\omega = 0$ which makes them unsuitable for as lowpass filters. Additionally, the type 3 response is always zero at $\omega = \pi$, making it unsuitable as a highpass filter. The type I filter is the most versatile of the four.

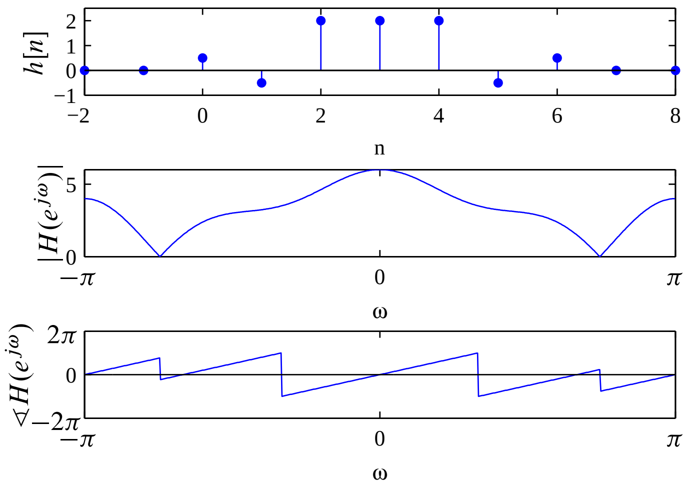
Linear phase filters can be thought of in a different way. Recall that a linear phase characteristic simply corresponds to a time shift or delay. Consider now a real FIR filter with an impulse response that satisfies the even symmetry condition $h[n] = h[-n]$:



Recall from the properties of the Fourier transform this filter has a real-valued frequency response $A(e^{j\omega})$. Delaying this impulse response by $(N-1)/2$ results in a causal filter with frequency response

$$H(e^{j\omega}) = A(e^{j\omega})e^{-j\omega(N-1)/2}.$$

This filter therefore has linear phase.



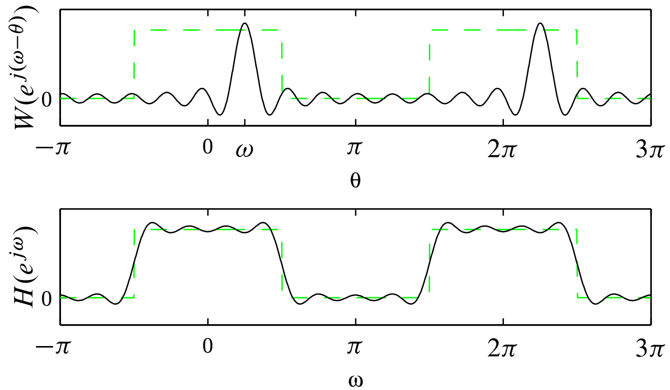
2.1 Window method for FIR filter design

Assume that the desired filter response $H_d(e^{j\omega})$ is known. Using the inverse Fourier transform we can determine $h_d[n]$, the desired unit sample response. In the window method, a FIR filter is obtained by multiplying a window $w[n]$ with $h_d[n]$ to obtain a finite duration $h[n]$ of length N. This is required since $h_d[n]$ will in general be an infinite duration sequence, and the corresponding filter will therefore not be realisable. If $h_d[n]$ is even or odd symmetric and $w[n]$ is even symmetric, then $h_d[n]w[n]$ is a linear phase filter.

Two important design criteria are the *length* and *shape* of the window $w[n]$. To see how these factors influence the design, consider the multiplication operation in the frequency domain: since $h[n] = h_d[n]w[n]$,

$$H(e^{j\omega}) = H_d(e^{j\omega}) * W(e^{j\omega}).$$

The following plot demonstrates the convolution operation. In each case the dotted line indicates the desired response $H_d(e^{j\omega})$.



From this, note that

- The *mainlobe* width of $W(e^{j\omega})$ affects the *transition* width of $H(e^{j\omega})$. Increasing the length N of $h[n]$ reduces the mainlobe width and hence the transition width of the overall response.

- The *sidelobes* of $W(e^{j\omega})$ affect the passband and stopband tolerance of $H(e^{j\omega})$. This can be controlled by changing the shape of the window. Changing N does not affect the sidelobe behaviour.

Some commonly used windows for filter design are

- **Rectangular:**

$$w[n] = \begin{cases} 1 & 0 \le n \le N \\ 0 & \text{otherwise} \end{cases}$$

- **Bartlett (triangular):**

$$w[n] = \begin{cases} 2n/N & 0 \le n \le N/2 \\ 2 - 2n/N & N/2 < n \le N \\ 0 & \text{otherwise} \end{cases}$$

- **Hanning:**

$$w[n] = \begin{cases} 0.5 - 0.5\cos(2\pi n/N) & 0 \le n \le N \\ 0 & \text{otherwise} \end{cases}$$

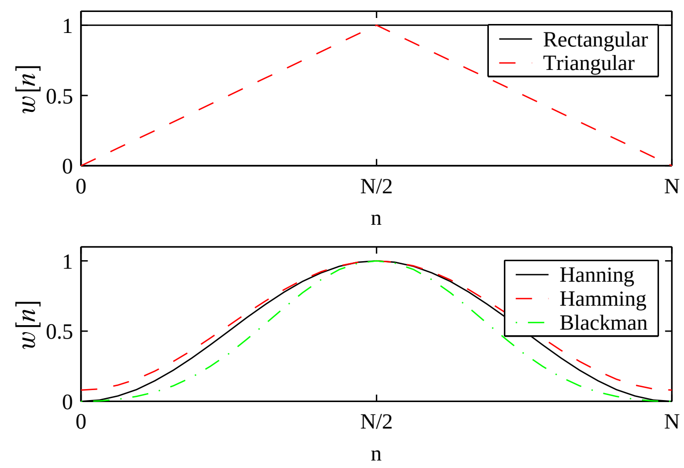
- **Hamming:**

$$w[n] = \begin{cases} 0.54 - 0.46\cos(2\pi n/N) & 0 \le n \le N \\ 0 & \text{otherwise} \end{cases}$$

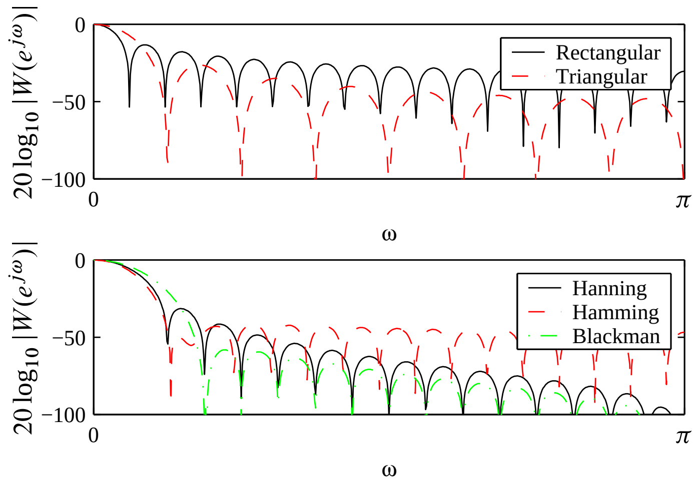
- **Kaiser:**

$$w[n] = \begin{cases} I_0[\beta(1 - [(n-\alpha)/\alpha]^2)^{1/2}] & 0 \le n \le N \\ 0 & \text{otherwise} \end{cases}$$

Examples of five of these windows are shown below:



All windows trade off a reduction in sidelobe level against an increase in mainlobe width. This is demonstrated below in a plot of the frequency response of each of the windows:



Some important window characteristics are compared in the following table:

Window	Peak sidelobe amplitude (dB)	Mainlobe transition width	Peak approximation error (dB)
Rectangular	-13	$4\pi/(N+1)$	-21
Bartlett	-25	$8\pi/N$	-25
Hanning	-31	$8\pi/N$	-44
Hamming	-41	$8\pi/N$	-53

The Kaiser window has a number of parameters that can be used to explicitly tune the characteristics.

In practice, the window shape is chosen first based on passband and stopband tolerance requirements. The window size is then determined based on transition width requirements. To determine $h_d[n]$ from $H_d(e^{j\omega})$ one can sample $H_d(e^{j\omega})$ closely and use a large inverse DFT.

2.2 Frequency sampling method for FIR filter design

In this design method, the desired frequency response $H_d(e^{j\omega})$ is sampled at equally-spaced points, and the result is inverse discrete Fourier transformed.

Specifically, letting

$$H[k] = H_d(e^{j\omega})\Big|_{\omega=\frac{2\pi k}{N}}, \qquad k = 0, \ldots, N-1,$$

the unit sample response of the filter is $h[n] = \text{IDFT}(H[k])$, so

$$h[n] = \frac{1}{N}\sum_{k=0}^{N-1} H[k]e^{j2\pi nk/N}.$$

The resulting filter will have a frequency response that is exactly the same as the original response at the sampling instants. Note that it is also necessary to specify the *phase* of the desired response $H_d(e^{j\omega})$, and it is usually chosen to be a linear function of frequency to ensure a linear phase filter. Additionally, if

a filter with real-valued coefficients is required, then additional constraints have to be enforced.

The *actual* frequency response $H(e^{j\omega})$ of the filter $h[n]$ still has to be determined. The z-transform of the impulse response is

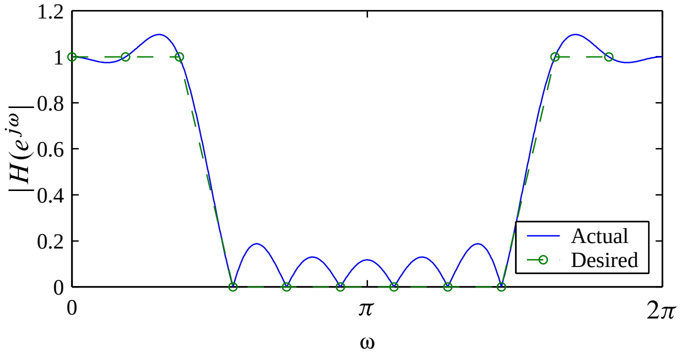
$$\begin{aligned} H(z) &= \sum_{n=0}^{N-1} h[n]z^{-n} = \sum_{n=0}^{N-1}\left[\frac{1}{N}\sum_{k=0}^{N-1} H[k]e^{j2\pi nk/N}\right]z^{-n} \\ &= \frac{1}{N}\sum_{k=0}^{N-1} H[k]\sum_{n=0}^{N-1} e^{j2\pi nk/N}z^{-n} \\ &= \frac{1}{N}\sum_{k=0}^{N-1} H[k]\left[\frac{1-z^{-N}}{1-e^{j2\pi k/N}z^{-1}}\right]. \end{aligned}$$

Evaluating on the unit circle $z = e^{j\omega}$ gives the frequency response

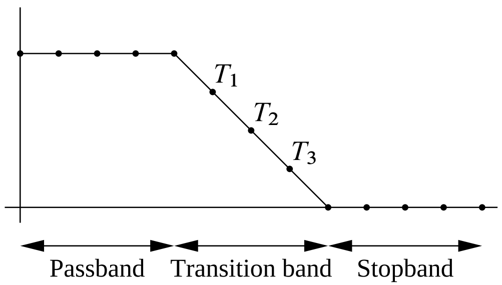
$$H(e^{j\omega}) = \frac{1-e^{-j\omega N}}{N}\sum_{k=0}^{N-1}\frac{H[k]}{1-e^{j2\pi k/N}e^{-j\omega}}.$$

This expression can be used to find the actual frequency response of the filter obtained, which can be compared with the desired response.

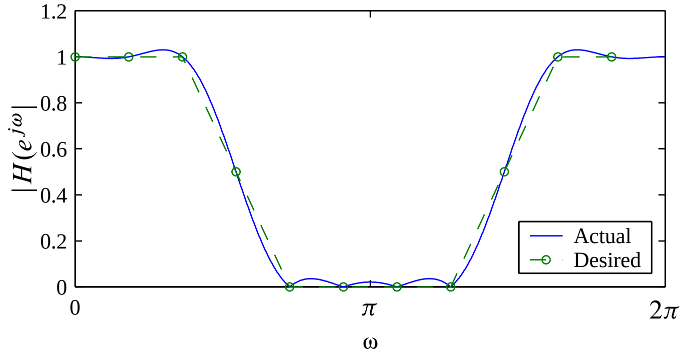
The method described only guarantees correct frequency response values at the points that were sampled. This sometimes leads to excessive ripple at intermediate points:



One way of addressing this problem is to allow **transition samples** in the region where discontinuities in $H_d(e^{j\omega})$ occur:



This effectively increases the transition width and can decrease the ripple, as observed below:



By leaving the value of the transition sample unconstrained, one can to some extent optimise the filter to minimise the ripple. Empirically, with three transition samples a stopband attenuation of 100dB is achievable. Recall however that for $h[n]$ real we require even or odd symmetry in the impulse response, so the values are not entirely unconstrained.

2.3 Optimum approximations of FIR filters

This method of filter design attempts to find the filter of length N that optimises a given design objective. In this case the objective is chosen to be the minimisation of

$$\max_{0 \leq \omega \leq 2\pi} |E(e^{j\omega})|$$

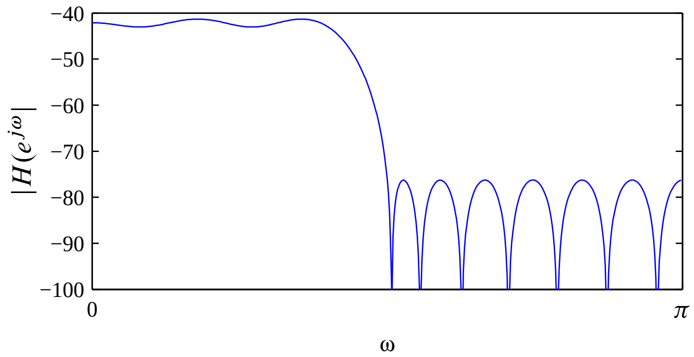
where $E(e^{j\omega})$ is a weighted error function

$$E(e^{j\omega}) = W(e^{j\omega})[H_d(e^{j\omega}) - H(e^{j\omega})].$$

The minimisation is performed over the filter coefficients $h[n]$.

In practice, the design problem can be specified as follows: given δ_p, δ_s, f_p, and f_s, determine $h[n]$ such that the design specification is satisfied with the smallest possible N. The optimal (or minimax) design method therefore yields the shortest filter that meets a required frequency response over the entire frequency range. It is widely used in practice.

Solutions to this optimisation problem have been explored in the literature, and many implementations of the method are available. It turns out that when $\max|E(e^{j\omega})|$ is minimised, the resulting filter response will have equiripple passband and stopband, with the ripple alternating in sign between two equal amplitude levels:



The maxima and minima are known as extrema. For linear phase lowpass filters, for example, there are either $r+1$ or $r+2$ extrema, where $r=(N+1)/2$ (for type 1 filters) or $r=N/2$ (for type 2 filters).

For a given set of filter specifications, the locations of the extremal frequencies, apart from those at band edges, are not known a priori. Thus the main problem in the optimal method is to find the locations of the extremal frequencies. Numerous algorithms exist to do this. Once the locations of the extremal frequencies are known, it is simple to specify the actual frequency response, and hence find the impulse response for the filter.

3 Infinite impulse response (IIR) filter design

An IIR filter has nonzero values of the impulse response for all values of n, even as $n \to \infty$. To implement such a filter using a FIR structure therefore requires an infinite number of calculations.

However, in many cases IIR filters can be realised using LCCDEs and computed recursively.

Example:
A filter with the infinite impulse response $h[n]=(1/2)^n u[n]$ has z-transform

$$H(z)=\frac{1}{1-1/2z^{-1}}=\frac{Y(z)}{X(z)}.$$

Therefore, $y[n]=1/2y[n-1]+x[n]$, and $y[n]$ is easy to calculate.

IIR filter structures can therefore be far more computationally efficient than FIR filters, particularly for long impulse responses.

FIR filters are stable for $h[n]$ bounded, and can be made to have a linear phase response. IIR filters, on the other hand, are stable if the poles are inside the unit circle, and have a phase response that is difficult to specify. The general approach taken is to specify the magnitude response, and regard the phase as acceptable. This is a disadvantage of IIR filters.

IIR filter design is discussed in most DSP texts.