CHAPTER 5

DIGITAL FILTERS I: FREQUENCY-SELECTIVE FILTERS

Engineering is the art of doing that thing well with one dollar, which any bungler can do with two after a fashion.

Arthur M. Wellington, *The Economic Theory of the Location of Railways* (1887).

5.1 Introduction: Digital Filters in General

Given a discrete-time sequence $\{x_n\}$ we often wish to perform some kind of operation on it. The DFT is one such operation; many others go by the name of filters. There are many different things that filters may be used to do; and for each of these actions there are many different ways to design filters that are "good". You should remember that filter design is engineering, not science; our aim is to get a good enough, and economical, answer, for the situation we face.

The list of tasks that filters can be used for includes:

A. Select out certain frequencies. This is the most common operation in the kind of analysis done in geophysics: especially removing high frequencies, to allow decimation of a sequence to make it smaller; or low frequencies, to reduce large low-frequency changes that hide smaller fluctuations.

B. Differentiate or integrate, at least approximately; since these are continuous-time operations, they only make sense for sequences properly sampled from a continuous-time function. The design of filters for these operations is quite similar to the design of frequency-selective ones.

C. Simulate what a continuous-time linear system would do to the series, again supposing the sequence to have been sampled from a continuous-time function. This goal leads to another way of designing filters, through simulating a linear system represented by an ordinary differential equation. In the signal processing literature, these methods are used to design frequency-selective filters by simulating filter designs developed for analog systems. For most geophysical data, such simulation filters are more important for reproducing a synthetic instrument through which one wants to pass synthetic data, for comparison with a real dataset recorded by a real instrument.

D. Predict future values, which is an major subfield all by itself: it can be economically very valuable, but it is also used a lot in such matters as autopilots for airplanes, where knowledge of where the plane is about to be is very important. Many methods exist, in part because prediction depends on what model we take to characterize the sequence. Quite a few methods model the sequence probabilistically: for example, the **Kalman filter** operates in the time domain to infer future values from past ones. A rapidly-growing part of this field connects to the burgeoning subject of nonlinear dynamics, since if a time series that appears to be random can actually be described as a deterministic but chaotic system, better predictions may be possible.

E. Detect a signal, defined as some kind of anomalous behavior. This goal is closely related to the previous one: if the series is not as predicted, then there might be an anomaly. Anomaly detection can be described using the methods of statistical hypothesis testing, and has developed a large literature of its own, again because of its technological importance in radar and sonar.

F. Assuming that the sequence has been produced as the result of a convolution of two other sequences w_n and y_n, determine something about one or both of these other two. This is known as **deconvolution**, and covers a wide variety of techniques, which depend on what we think we know about the convolving sequence w_n. One example would be the reverse of (C): we have the result of a sequence (or a function) being passed through a known linear system, and want to determine the original. Two geophysical examples of this are correcting data for instrument response, and downward continuation of magnetic or gravity data measured above its source (the height acts as a filter). Another type

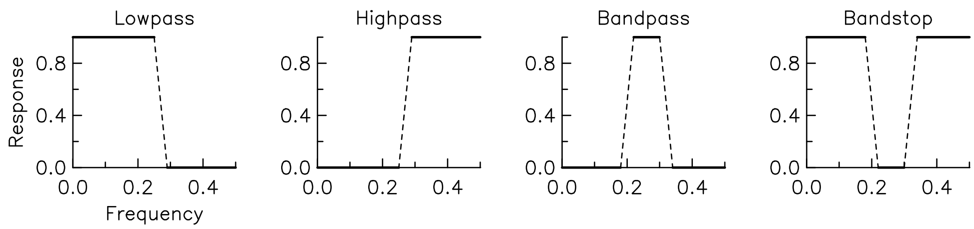

Figure 5.1: Ideal frequency-selective filters: the four commonest types. The parts of the frequency band for which the response is 1 are called the **passband**(s); where the response is 0 is the **stopband**(s). The dashed lines show the region in which (for most design methods) the response goes from 1 to 0 (not necessarily linearly as shown here), which is called the **transition band**. Another name for the bandstop filter is **notch filter**.

of deconvolution (used in exploration geophysics and image processing) occurs when we have only partial knowledge about the two sequences (e.g., some of their statistical properties).

In this chapter we discuss filters for purpose (A): certainly the commonest application in geophysics. In the next chapter we discuss (C), partly so we can introduce some new ways of looking at digital filters. Purpose (B), along with some other filtering topics, will be placed in a third chapter. Purposes (D), (E), and (F), all of which combine filtering and statistics, will not be covered in this course.

5.2 FIR Filters for Frequency Selection

Even just for purpose (A), we have many choices in how to design filters. We start by introducing the ideal filter response $W_d(f)$; for frequency-selective filters W_d is either 1 or 0 over particular frequency bands. Figure 5.1 shows some typical designs, and their names.

In this chapter we restrict ourselves to filters that consist of a finite sequence $\{w_n\}$, usually called the **weights**. These filters are used by convolving the weight sequence $\{w_n\}$ with the data sequence $\{x_n\}$ to produce a filtered sequence $\{y_n\} = \{w_n\} * \{x_n\}$. In the signal-processing literature this type of filter is called a **finite impulse response (FIR)** filter; in the statistics literature the usual name is **moving average**, usually abbreviated as MA. Another term is **nonrecursive filter**.

The frequency response of a FIR filter is just the Fourier transform of the weight sequence:

$$W(f) = \sum_{n=0}^{N-1} w_n e^{-2\pi i f n} = \sum_{n=-\infty}^{\infty} w_n e^{-2\pi i f n} \qquad \text{for } 0 \le f \le 1 \quad (\text{or } -½ \le f \le ½) \tag{5.1}$$

where N is the number of weights. It is useful to consider the weights to be an infinite sequence, though one that happens to be zero outside the finite range from 0 to $N-1$.

We next impose three more restrictions:

- The weights are real.
- The number of weights, N, is odd.
- The weights are "symmetric": that is, for $N = 2M+1$,

$$w_{M-k} = w_{M+k} \quad k = 0, \dots M$$

The consequence of all this is that $w_n = w_{2M-n}$, making the frequency response

$$\begin{aligned} W(f) &= \sum_{n=0}^{2M} w_n e^{-2\pi i f n} = \sum_{n=0}^{M-1} w_n (e^{-2\pi i f n} + e^{-2\pi i f (2M-n)}) + w_M e^{-2\pi i f M} \\ &= e^{-2\pi i f M} \left[w_M + \sum_{n=0}^{M-1} w_n (e^{-2\pi i f (n-M)} + e^{-2\pi i f (M-n)}) \right] \\ &= e^{-2\pi i f M} \left[w_M + 2 \sum_{n=0}^{M-1} w_n \cos 2\pi (M-n) f \right] \end{aligned} \tag{5.2}$$

The part in square brackets is purely real; like any real quantity viewed using the amplitude and phase form for complex numbers, its phase is zero. The shift theorem shows that the initial exponential is equivalent to an M point shift, which is to say a time delay. Ignoring this, a symmetric FIR filter thus does not shift the phases of different frequencies: this minimizes distortion of the input, other than by removing energy at certain frequencies. If the part in square brackets were one, we would have just an M-term delay, which obviously creates no distortion at all. Because the phase shift is linear in frequency, symmetric FIR filters are usually termed **linear phase**.

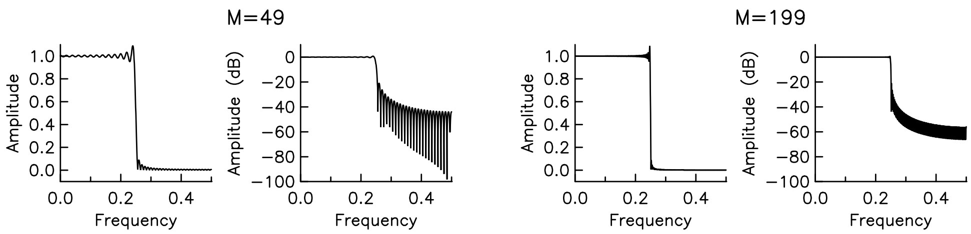

Figure 5.2: Frequency response (shown in both linear and log amplitude) for lowpass filters designed by simple truncation of the sequence of ideal weights. The ideal response $W_d(f)$ has a passband ($W_d = 1$) from 0 to 0.25, and a stopband ($W_d = 0$) from 0.25 to 0.5. The left-hand pair of plots show the response for $M = 49$ (99 weights total), and the right-hand pair for $M = 99$ (199 weights total).

If we do not mind having the filter be acausal (the output for a step input precedes the step), we can make it symmetrical around $n = 0$, defining $w_n = w_{-n}$ for $n = -M, \ldots M$. The frequency response is then

$$W(f) = w_0 + 2\sum_{n=1}^{M} w_n \cos 2\pi n f \tag{5.3}$$

which has zero phase shift; in this form the FIR filter is often called a **zero-phase filter**. Because such a filter is acausal it can be implemented only after the data have been collected, but in digital systems this can always be the case – if only because we are free to relabel the absolute time of the terms in the series as we see fit. In general a lack of causality is not a problem, though seismic data can be an exception to this: if our interest is in the exact time of arrival of bursts of energy, we do not want the filtering to put any energy before it actually arrives.[1] In Chapter 7 we will see how to make FIR filters that are better suited to this application.

Given all the restrictions so far applied, the filter design problem becomes how to choose the $M+1$ weights $w_0, w_1, \ldots w_{M+1}$, to best approximate a given frequency response $W_d(f)$. As we will see, the meaning of "best" is not unique. Indeed it is characteristic of filter design that the best answer is not always the same: what we want from a filter will depend on what we are trying to do.

A naive approach would be to just take the inverse Fourier transform of the ideal filter response: since $W_d(f)$ is real and symmetric, it might appear

[1] A good example of exactly this problem confounding the interpretation of the initial rupture in earthquakes is given by Scherbaum and Bouin (1997).

that we can use the inverse transform for sequences to get

$$w_n = \int_{-1/2}^{1/2} W_d(f) e^{2\pi i f n}\, df = 2\int_0^{1/2} W_d(f) \cos 2\pi f n \, df$$

However, this will not in general give a sequence of finite length – something we certainly need for a practical filter. We need to be more sophisticated.

5.2.1 Designs Using Tapers

One method of filter design is to construct, from the ideal response, the (infinite) sequence of ideal weights

$$w_n^d = 2\int_0^{1/2} W_d(f) \cos 2\pi f n \, df$$

and then create a finite sequence from this by multiplying by a **taper** (or **window**) sequence $\{a_n\}$ to create the final weights

$$w_n = a_n w_n^d$$

The taper sequence must have two properties. The first is that $a_n = 0$ for $n > M$, so that all but a finite section of $\{w_n\}$ is zero, to give a finite-length filter. The second is that the taper must also be symmetric about $n = 0$, to keep the final filter weights symmetric.

We might expect that this time-domain multiplication would be equivalent to a convolution in the frequency domain; and so it is:

$$\begin{aligned} W(f) &= \sum_{n=-\infty}^{\infty} a_n w_n^d e^{-2\pi i f n} = \sum_{n=-\infty}^{\infty} a_n e^{-2\pi i f n} \int_{-1/2}^{1/2} W_d(u) e^{2\pi i f n u}\, du \\ &= \int_{-1/2}^{1/2} W_d(u) \sum_{n=-\infty}^{\infty} a_n e^{-2\pi i n (f-u)}\, du = \int_{-1/2}^{1/2} W_d(u) A(f-u)\, du \end{aligned}$$

where in the final convolution integral both functions are assumed to be periodic outside the range $[-1/2, 1/2]$.

This means that the closer $A(f)$, the transform of the taper sequence, is to a delta-function, the closer the filter response will be to the ideal $W_d(f)$ Of course, what we mean by "close to" will, again, depend on what we think is important: in thinking about closeness to a delta-function, is it more

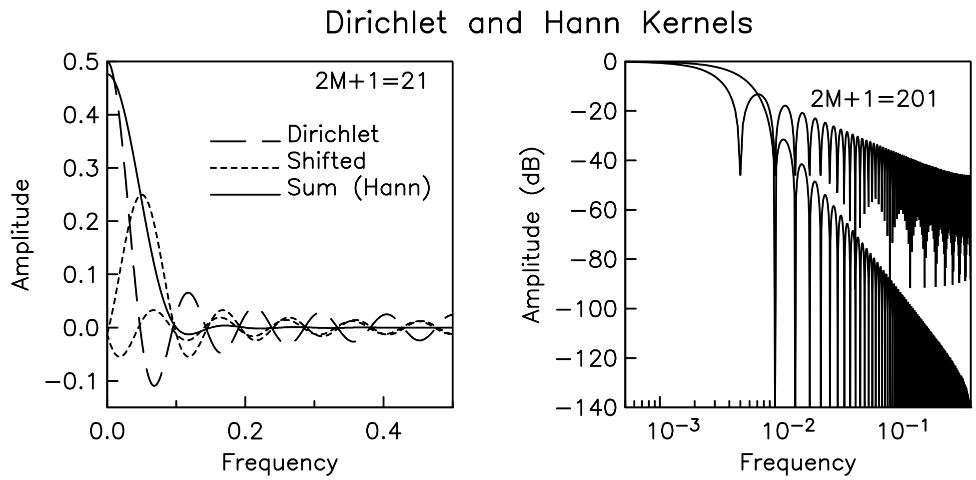

Figure 5.3: The three Dirichlet kernels used in forming the response of the von Hann taper (left, dashed), and that response itself (solid). The right plot shows $A(f)$ for the Dirichlet and von Hann kernels on a log-log scale, for a larger value of M.

important that the peak of $A(f)$ be narrow or the values away from the peak be small? Our answer to this will depend on what we want the filter to do, which is why there are no universal rules for filter designs.

The simplest taper is the rectangular taper, with $a_n = 1$ for $|n| \le M$; this corresponds to simply truncating the sequence of ideal weights. We can see that this is not a particularly good solution if we look at the corresponding $A(f)$:

$$A(f) = \sum_{n=-M}^{M} e^{-2\pi i f n} = (2M+1)D_{2M+1}(f)$$

where $D_N(f)$ is the Dirichlet kernel discussed in Chapter 3. As described there, successive maxima of $|D_N(f)|$ do decrease, but only as f^{-1}: if what we want is small values away from the peak, this $A(f)$ is not a good approximation to a delta-function. Convolving this $A(f)$ with the ideal response causes much of the passband response to “leak” into the stopband, and vice-versa. Figure 5.2 shows this effect clearly, and also shows that simply increasing the number of weights M does not diminish its importance.

If you are familiar with Fourier series theory, another way to view this result is to realize that in taking the rectangular taper we are simply finding partial sums of the Fourier series expression for $W_d(f)$; such partial sums always suffer from Gibbs’ phenomenon, with poor convergence near

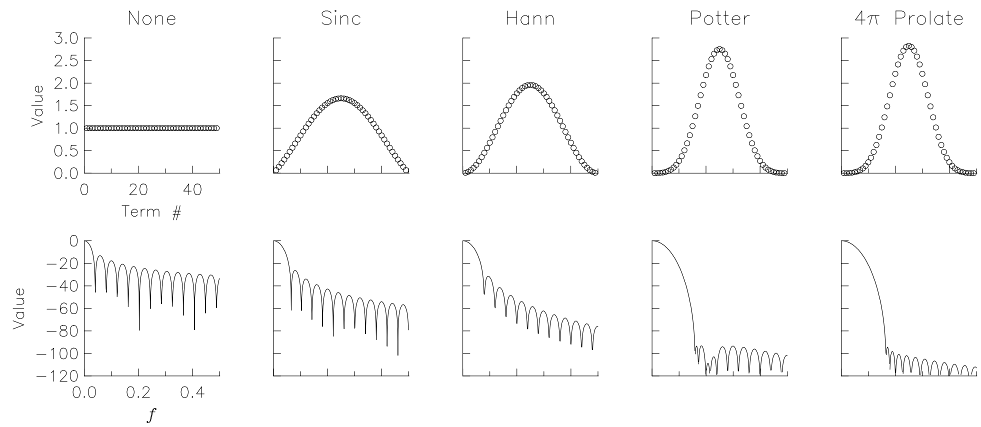

Figure 5.4: A collection of data tapers (top plots) and their Fourier transforms $A(f)$, for $2M+1 = 52$ (the transforms have been interpolated for plotting, and normalized to 1 at zero frequency). The sinc taper is given by $(\pi n/M)^{-1}\sin(\pi n/M)$, and the Potter taper by $\sum_{k=0}^{3} a_k \cos(\pi kn/(2M+1))$ with $a_0 = 0.3557702$, $a_1 = 0.4873966$, $a_2 = 0.1442299$, and $a_3 = 0.0126033$.

the discontinuities in the frequency response.[2]

There is a taper that is almost as simple, but that has much smaller values away from the central peak: this is the von Hann taper, also called the hanning, $\cos^2$, or $1+\cos$ taper:

$$a_n = \tfrac{1}{2}[1+\cos(\pi n/M)] = \cos^2(\pi n/2M) \qquad \text{for } n = -M, \dots M$$

which has the transform:

$$\begin{aligned} A(f) &= \sum_{n=-M}^{M} a_n e^{-2\pi i f n} = \tfrac{1}{2}\sum_{n=-M}^{M} e^{-2\pi i f n} + \tfrac{1}{4}\sum_{n=-M}^{M}[e^{\pi i n/M} + e^{-\pi i n/M}]e^{-2\pi i f n} \\ &= \tfrac{1}{2}\sum_{n=-M}^{M} e^{-2\pi i f n} + \tfrac{1}{4}\sum_{n=-M}^{M} e^{-2\pi i(f-(M/2))n} + \tfrac{1}{4}\sum_{n=-M}^{M} e^{-2\pi i(f+(M/2))n} \\ &= (2M+1)[\tfrac{1}{2}D_{2M+1}(f) + \tfrac{1}{4}D_{2M+1}(f-(M/2)) + \tfrac{1}{4}D_{2M+1}(f+(M/2))] \end{aligned}$$

The three Dirichlet kernels and their sum are shown in Figure 5.3: it is clear that, as f increases, the combination approaches zero much more rapidly than does the original Dirichlet kernel, a point made even more dramatically by a log-log plot.

[2] For a good introduction, both historical and mathematical, to Gibbs' phenomenon, see Hewitt and Hewitt (1979).

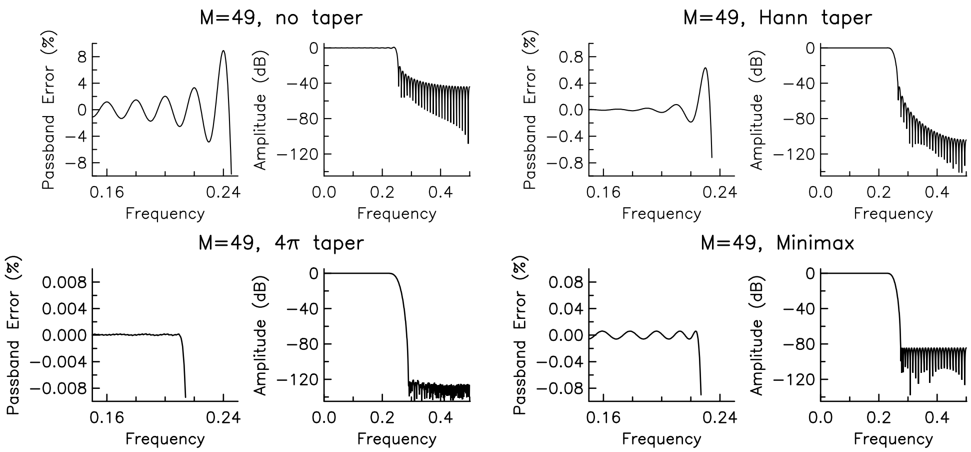

Figure 5.5: Frequency responses of lowpass filters for three data tapers, and also for the minimax design procedure described in Section 5.2.2 For each, the error (departure from unity) is shown over part of the passband (note the different scales), and the complete response is shown in log form.

There are many taper sequences available;[3] the frequency responses of these can vary considerably, but inevitably have a tradeoff between width of the central peak and smallness of the fluctuations (these are called the **sidelobes**) away from it. Figure 5.4 gives a small selection of tapers, including two with especially small sidelobes: the Potter taper (an empirically-designed one) and the 4π prolate spheroidal taper. The last, as we will see in Chapter **??**, is the function for which $A(f)$ is most concentrated within a frequency band around zero, and is one of several that are important when estimating a power spectrum.

The effect of using different tapers for filters is shown in Figure 5.5. As expected, the use of a Hann taper gives much lower sidelobes than does a rectangular taper; the 4π prolate spheroidal taper gives even lower ones, with sidelobe amplitudes due in part to roundoff error in computing the DFT. But filter design using tapers is somewhat inflexible: given a particular taper the tradeoff between sidelobe level and the width of the transition between passband and stopband is fixed. A somewhat more flexible ap-

[3] Harris (1976) is a kind of bestiary of every taper then thought of, with plots of the time and frequency responses. Most of these tapers deserve to be forgotten, since much better ones, notably the prolate spheroidal tapers, have been developed since. It is worth noting that the 'Kaiser-Bessel' tapers described in this paper closely approximate the prolates.

proach has been developed by Kaiser and Reed (1977) and Kaiser and Reed (1978), using an adjustable taper and some empirical rules for setting it (and the filter length) to meet a given sidelobe level and transition width; this method offers a convenient design that is adequate for many applications. However, even more flexibility is possible using different techniques, one of which we now describe.

5.2.2 Design by Optimal Fitting of the Frequency Response

A more powerful technique in designing filters is to look at the difference between the ideal response $W_d(f)$ and the actual response $W(f)$, and choose the filter weights to minimize this difference in some way. What seems like the most obvious approach does not however work very well, this being to minimize the mean-square misfit. Minimizing this is just like the least-squares criterion, though in this case expressed as an integral over frequency: the quantity to minimize is

$$\epsilon^2 = \int_{-½}^{½} |W_d(f) - W(f)|^2 df$$

The difference between the ideal and actual responses, in terms of the weights, is

$$W_d(f) - W(f) = \sum_{n=-\infty}^{\infty} (w_n^d - w_n) e^{-2\pi i f n}$$

where the actual weights w_n are extended to infinity by adding zeroes at both ends. But then we may apply Parseval's theorem for infinite sequences to get:

$$\epsilon^2 = \sum_{n=-\infty}^{\infty} |w_n^d - w_n|^2 = \sum_{n=-\infty}^{-M-1} |w_n^d|^2 + \sum_{n=M+1}^{\infty} |w_n^d|^2 + \sum_{n=-M}^{M} |w_n^d - w_n|^2$$

and since only the last term is adjustable, we minimize ϵ^2 by setting $w_n = w_n^d$. But this is no different from using a rectangular taper on the ideal sequence – and we have already seen that this produces filters with large sidelobes. Least squares, for all its merits, is not a panacea.

What is more useful is to minimize the maximum value that the misfit $|W_d(f) - W(f)|$ attains over a range of frequencies. This is computationally more complicated than anything we have discussed so far, and we will not describe how it is done; suffice it to say that it is possible to find weights that will do this. Specifically, given K non-overlapping frequency bands $f_{L_k} < f \le f_{H_k}$, for $k = 1, 2 \ldots K$, we can find the M weights which minimize

$$b_k \max_{[f_L, f_H]} |W_d(f) - W(f)|$$

over all the bands; b_k is a weight applied to each band, which allows the fit to the ideal response to be tighter or looser. The actual amount of misfit is largely a function of the filter length M; what this misfit actually is, can only be determined after the optimal weights are found. Usually some trial and error is needed to decide on the appropriate tradeoff between length and misfit. This approach applies the L_∞ norm, whereas the mean-square-error criterion applies the L_2 norm; and in general gives very good designs, with more flexibility than the tapering method. The resulting filters are termed **equiripple** filters, because the frequency response shows a uniform level of departure from the ideal.

A filter designed using this method is shown in the fourth example in Figure 5.5: by sacrificing low sidelobes at high frequency, the equiripple design can have a narrower transition band than any of the tapered designs. The procedure for designing equiripple filters is called the Parks-McClellan algorithm; it is included in and in MATLAB as routine `firpm`. The underlying algorithm is called the Remez exchange method.

There are many other variants in this method of filter design; we may, for example, require, as well as a good fit to W_d, that the derivative dW/df be of one sign in the passband: this will completely eliminates ripple there.[4] In general, most such refinements, and how to design filters where the word length is short and roundoff a problem, are rarely important in geophysical data analysis.

5.2.3 A Filtering Example

We close with an example that illustrates how frequency-selective filters can be used, and also illustrates why there are no fixed rules for designing them – what filter you choose depends on the problem you want to solve. The data to be filtered are measurements of strain at Piñon Flat Observatory, made at one-second intervals with a long-base laser strainmeter. For

[4] Steiglitz *et al.* (1992) present a fairly general program for including different kinds of constraints using linear programming: relatively slow, but very flexible. See `http://www.cs.princeton.edu/~ken/meteor.html` for source code and examples.

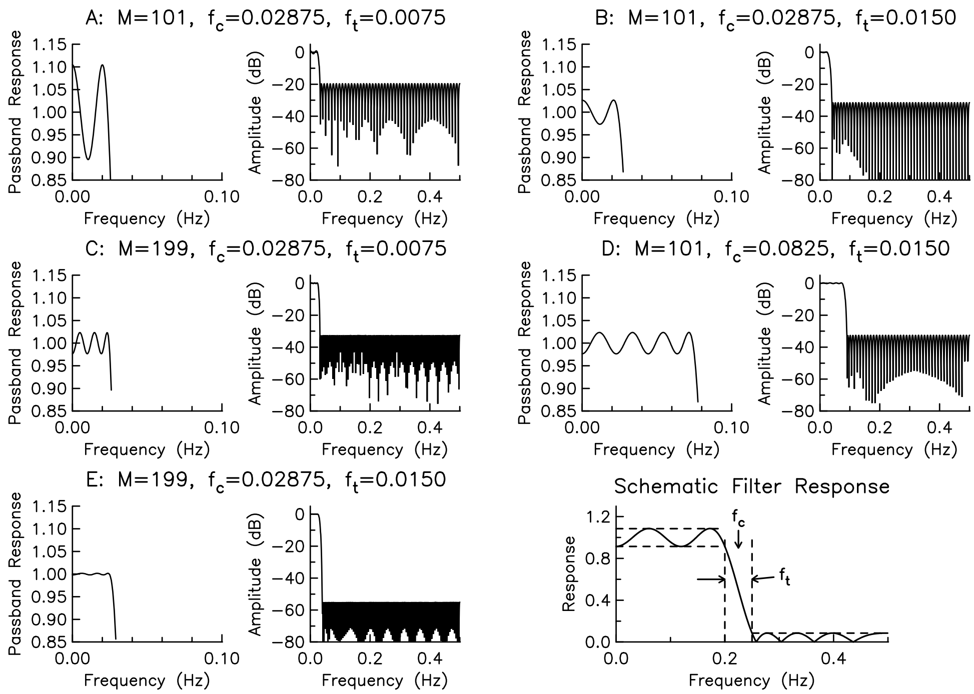

Figure 5.6: Equiripple filter designs. The lower right shows a schematic response, with the frequency parameters indicated. The different designs (A through E), have different frequency settings or lengths M. For each, the passband response is shown on a linear scale, the overall response in decibels.

the time period shown, these strain data show the signal from the 1994 Northridge earthquake, 203 km away. This signal, though dominated by the seismic waves from the earthquake, also contains the static change in strain caused by the elastic rebound of the rocks, which is what drives the faulting. The aim of the filtering will be to remove the "high-frequency" seismic signal so as to show the static change more clearly.

For maximum flexibility we use the equiripple design; Figure 5.6 shows some possible filter responses. The lower right corner of this figure shows parameters which we can choose: the transition frequency f_c and the width of the transition band f_t. Once we choose these parameters, we next pick a length for the filter; the program will then find the best design possible, with the smallest amount of ripple in both the passband and stopband. In this case we choose these amounts to be the same; while they can be in any proportion to each other, making one smaller makes the other larger.

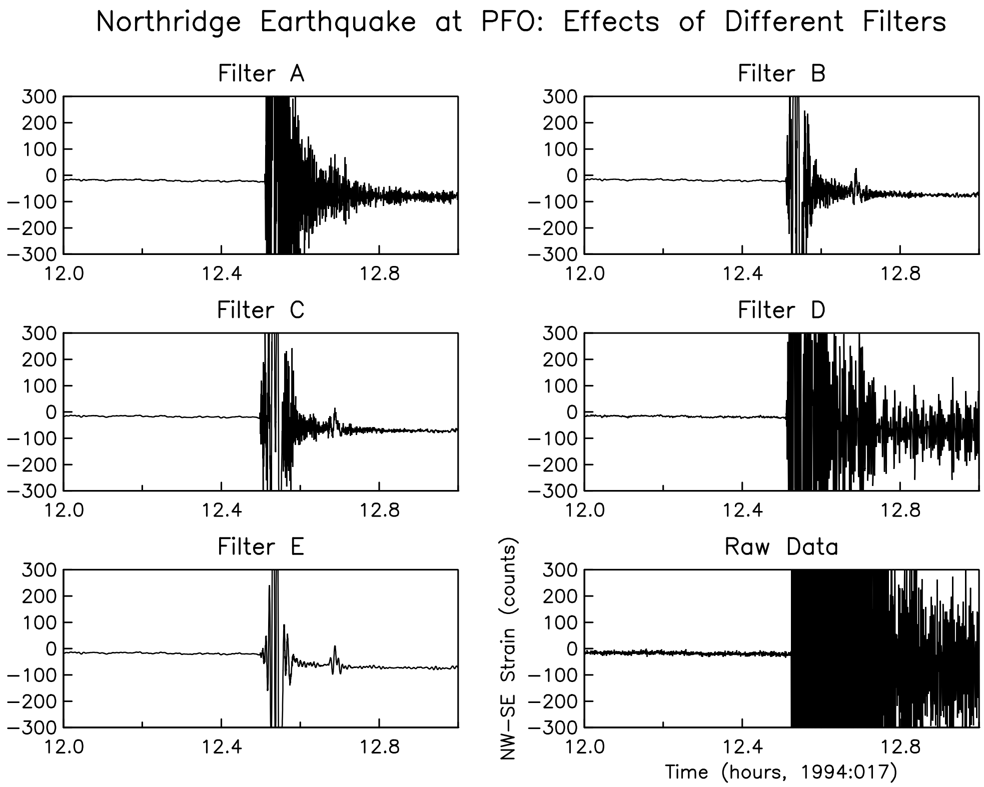

Figure 5.7: Strain variations from the Northridge earthquake recorded on the NW-SE laser strainmeter at Piñon Flat Observatory (PFO). The lower right plot is the original unfiltered data; this actually has a much larger range than is shown. Plots A through E show the results of lowpassing the data with the filters whose response is shown in Figure 5.6.

Our first design (A) has a narrow passband, and a narrow transition band – and the consequence is that the ripple is large, nearly 10%. This means that in the stopband any signal will be only 0.1 of what it was originally – we would say, 20 dB down, which is not a large reduction. To do better, we can try two approaches: (B) make the transition band wider, and (C) use more filter weights (199 instead of 101). Either one gives much less ripple; for these two examples we are trading computation time (filter length) against the width of the passband response. In (D) we try making the passband much larger by increasing f_c; in (E) we combine the larger f_t of (B) with the more numerous weights of (C) to get a response even closer to the ideal.

Figure 5.7 shows what these filters do to the data; without filtering the static offset is well hidden in the seismic coda. Filter A reduces this enough that the offset can be seen, but still leaves a lot of energy: looking at this, you would (or should) want to say that the "true" signal would be a smoothed version of this: a sure sign that the data need more filtering. Filters B and C provide more filtering: since their stopband levels are the same, the results look about the same as well; between the two, B is probably preferable because it is shorter. A much wider passband (D) turns out to be a poor idea, at least for showing the static offset. Filter E is perhaps "best" for this application.

CHAPTER 6

DIGITAL FILTERS II: RECURSIVE FILTERS AND SIMULATING LINEAR SYSTEMS

6.1 Introduction

We now turn to another kind of digital filter: one that will allow us to use a computer to imitate what some physical system does. We might need this when, for example, we want to model a seismogram. The first step would be to have a way of computing the the ground motion input to the seismometer. The next step would be to simulate what the output of the seismometer (a physical system) is for this ground motion. The FIR filters of the previous chapter are not well-suited for this, but other designs are, and it is these we will now describe.

But, to discuss these filters and how to design them we need to first spend some time introducing additional mathematics for the linear systems we discussed in Chapter 2. We will then return to the problem of designing digital filters, introducing what is known as a **recursive filter**; finally, we will show how to make such a digital filter accurately simulate an analog system.

6.2 Lumped-Parameter Systems

We saw in Chapter 2 that a linear time-invariant system could be characterized in three different ways:

- By its frequency response $\tilde{g}(f)$; this expresses the ratio (a complex number) between output and input when the input is a pure sinusoid, $e^{2\pi ift}$.