Techniques for Designing Finite-Duration Impulse-Response Digital Filters

LAWRENCE R. RABINER, MEMBER, IEEE

Abstract—Several new techniques for designing finite-duration impulse-response digital filters have become available in the past few years. The motivation behind three design techniques that have been proposed are reviewed here, and the resulting designs are compared with respect to filter characteristics, ease of design, and methods of realization. The design techniques to be discussed include window, frequency-sampling, and equiripple designs.

I. INTRODUCTION

INTEREST in design techniques for finite-duration impulse-response (FIR) digital filters has been renewed in the past few years because of the application of powerful optimization algorithms to the design problem. Although closed-form solutions to the approximation problem cannot, in general, be obtained explicitly, iterative techniques can be made to yield optimum solutions. Two optimization techniques have been proposed recently [1]-[4] which, along with the classical window design method [5], [6], provide the user with several possibilities for approximating filters with arbitrary frequency-response characteristics. In this paper we will discuss the general theory behind windowing and two optimization techniques—frequency-sampling and equiripple designs—and then compare these methods with respect to several practical and theoretical criteria.

II. TERMINOLOGY

Before discussing the design issues, it is important to distinguish the various types of digital filters that can be designed and to separate the way in which a filter is realized from the filter characteristics themselves. The following terms will be used throughout this paper.

1) Finite-duration impulse-response (FIR): This term means that the duration of the filter impulse response $h_n$ is finite; i.e.,

$$h_n = 0, \quad n > N_1 < \infty$$

$$h_n = 0, \quad n < N_2 > -\infty$$

and

$$N_1 > N_2.$$

2) Infinite-duration impulse-response (IIR): This term means that the duration of the filter impulse response $h_n$ is infinite; i.e., there exists no finite values of either $N_1$ or $N_2$ such that (1) is satisfied.

3) Recursive realization: This term describes the way a filter (either IIR or FIR) is realized. It means that the current filter output $y_n$ is obtained explicitly in terms of past filter outputs $y_{n-1}, \cdots$ as well as in terms of past and present filter inputs $x_{n-1}, \cdots$. Thus the output of a recursive realization can be written as

$$y_n = F(y_{n-1}, y_{n-2}, \cdots, x_n, x_{n-1}, \cdots).$$

4) Nonrecursive realization: This term means that the current filter output $y_n$ is obtained explicitly in terms of only past and present inputs; i.e., previous outputs are not used to generate the current output. The representation on a nonrecursive realization can be written as

$$y_n = F(x_n, x_{n-1}, \cdots).$$

The motivation behind this terminology is that it has been shown [7], [8] that FIR filters as well as IIR filters can be realized both nonrecursively and recursively. (It should be noted that, in general, recursive realizations of IIR filters and nonrecursive realizations of FIR filters are most efficient and are usually used.) Thus a term describing filter characteristics should be distinct from a term describing how the filter is realized. This is not how the terminology has traditionally been used.

III. SOME ADVANTAGES OF FIR FILTERS

Although our aim is to describe and compare design techniques for FIR filters, it is of interest to first discuss some reasons why FIR filters are of importance. These include the following.

1) FIR filters can easily be designed to approximate a prescribed magnitude frequency response to arbitrary accuracy with an exactly linear phase characteristic. In addition, FIR filters can approximate arbitrary frequency characteristics (both magnitude and phase), but IIR filters can also do this.

2) FIR filters can be realized efficiently both nonrecursively (using direct convolution, or high-speed convolution by using the fast Fourier transform (FFT) [9], [10]) and recursively (using a comb filter and a bank of resonators [11]).

3) An FIR filter realized nonrecursively is always stable. FIR filters realized nonrecursively contain only zeros in the finite z plane, and hence are always stable.

4) Quantization and roundoff problems inherent in recursive realizations of IIR filters are generally negligible in nonrecursive realizations of FIR filters.
5) The coefficient accuracy problems inherent in sharp cutoff IIR filters can often be made less severe for realizations of equally sharp FIR filters.

IV. THEORY OF DESIGN

Since one of the most important reasons for designing FIR filters is that they can be designed with an exactly linear phase, we will restrict our discussion to this type. The general characteristics of the frequency response of a digital filter whose impulse-response coefficients are real are

\[ H[\exp(j\omega T)] = H[\exp(j(\omega + n \frac{2\pi}{T}) T)], \quad n = 0, \pm 1, \pm 2, \ldots \]  

(4)

\[ |H[\exp(j\omega T)]| = |H[\exp(-j\omega T)]|, \quad 0 \leq \omega \leq \pi \]  

(5)

\[ \theta(\omega) = -\theta(-\omega), \quad 0 \leq \omega \leq \pi \]  

(6)

where

\[ H[\exp(j\omega T)] = |H[\exp(j\omega T)]| \exp[j\theta(\omega)] \]  

(7)

where \( H[\exp(j\omega T)] \) is the frequency response of the filter and \( T \) is the sampling period. Equation (4) shows the periodicity of sampled-data systems in frequency. Equations (5) and (6) show the symmetry of the magnitude function and asymmetry of the phase function for filters with real impulse responses. In the remainder of this section we will discuss three techniques for approximating the desired magnitude response characteristics, assuming linear phase, with an FIR filter.

A. Windowing

Since the frequency response of a digital system is periodic, it can be expanded in a Fourier series. The coefficients of this Fourier series are the filter impulse-response coefficients. Generally, there are an infinite number of nonzero Fourier-series coefficients. To obtain an FIR filter which approximates the original frequency response, the Fourier series must be truncated. Direct truncation of the series, however, leads to the well-known Gibbs phenomenon, i.e., a fixed percentage overshoot and ripple before and after an approximated discontinuity, making this technique unsatisfactory for approximating many standard types of filters. In order to control the convergence of the Fourier series a weighting function is used to modify the Fourier coefficients. This time-limited weighting function is called a window. Since the multiplication of Fourier coefficients by a window corresponds to convolving the original frequency response with the Fourier transform of the window, a design criterion for windows is to find a finite window whose Fourier transform has relatively small sidelobes. Fig. 1 shows two choices of windows and their respective frequency responses. The rectangular window at the top left corresponds to direct truncation of the Fourier series. The Fourier transform of the rectangular window, shown at the lower left, has a narrow center lobe, but has sidelobes which contain a large part of the total energy and which decay quite slowly. Another window, the triangular window, is shown in the upper right of Fig. 1. Its Fourier transform, shown at the lower right, has a main lobe twice the width of the rectangular window main lobe, but has much less energy in the sidelobes.

The search for windows to meet the criterion previously mentioned, i.e., a finite window with most of its energy in the main lobe of its Fourier transform, has led to several useful and in some sense optimum designs. The Hamming window, which is of the form:

\[ w_n = 0.54 + 0.46 \cos \left( \frac{2\pi n}{N} \right), \quad \frac{-N}{2} \leq n \leq \frac{N}{2} \]  

(8)

has 99.96 percent of its energy in its main lobe, with the peak amplitude of the sidelobes down more than 40 dB from the peak. The width of the main lobe is twice the width of the rectangular window's main lobe. The Blackman window, which is

\[ w_n = 0.42 + 0.5 \cos \left( \frac{2\pi n}{N} \right) + 0.08 \cos \left( \frac{4\pi n}{N} \right), \quad \frac{-N}{2} \leq n \leq \frac{N}{2} \]  

(9)
further reduces peak sidelobe ripple to less than 0.0001 of the main-lobe peak at the expense of a main lobe whose width is triple the width of the rectangular window main lobe. Optimum window designs have been proposed by Kaiser [5] and Helms [6]. The Kaiser window is an approximation to the prolate spheroidal wave functions whose band-limiting properties are well known [12]. By adjusting a parameter of the window, a tradeoff can be obtained between peak sidelobe ripple and the width of the main lobe. The main disadvantage of these windows is that one must compute Bessel functions to get the window coefficients. Helms [6] has proposed the Dolph-Chebyshev window, which is optimum in the sense that the main-lobe width is as small as possible for a given peak ripple. The main disadvantage of this window is that inverse hyperbolic cosines must be evaluated to determine the window coefficients.

One disadvantage of the window design technique is that one must be able to compute Fourier-series coefficients for the periodic frequency response being approximated. Generally it is not trivial to determine closed-form expressions for these coefficients. The solution to this problem is found by approximately obtaining the Fourier-series coefficients as the discrete Fourier transform (DFT) of a sampled version of the continuous frequency response. By sampling the frequency response at a number of frequencies $M$ much larger than the number of Fourier-series coefficients under the window $N$, one can obtain fairly good approximations to the first $N$ Fourier coefficients. Fig. 2 presents a summary of the window design procedure. Fig. 2(a) shows the desired frequency response. Fig. 2(b) shows the same frequency response sampled at $M$ equispaced frequencies, as well as a continuous interpolation in frequency. Through the use of the DFT formula [Fig. 2(c)] the $M$-point impulse response $h_m$ is obtained and is shown below the formula. An $N$-point window $w_n$ [Fig. 2(d)] with the Fourier transform $W(e^{j\omega T})$ [Fig. 2(e)] weights the impulse response to yield the $N$-point sequence $a_n$ [Fig. 2(f)]. The sequence $a_n$ is the filter impulse response and its Fourier transform [Fig. 2(g)] shows the final approximation to the desired response.

**Frequency Sampling**

A second technique for approximating a filter with given frequency-response specifications is to sample the desired frequency response at $N$ equispaced frequencies, where $N$ is the number of samples in the filter impulse response. By setting these frequency samples to be the DFT coefficients of the filter impulse response, one can derive an approximation to any desired continuous frequency response. For many types of filters, such as low-pass, bandpass, and high-pass filters, one can optimize the values of the frequency samples in transition bands to

---

1 Kaiser has a simple 12-line Fortran 4 program which computes a power-series expansion (up to 25 terms) of the Bessel function $I_n(x)$. 

---

**Fig. 2.** Step by step realization of windowing. (a) Desired frequency response. (b) Sampled frequency response, $N \ll M$. (c) Impulse response obtained using DFT. (d) Typical window function. (e) Window frequency response. (f) Windowed impulse response. (g) Frequency response corresponding to windowed impulse response.
optimize the filter design. The resulting filter designs will be shown to be quite efficient.

In Fig. 3 frequency-response specifications for a desired filter are shown (the solid curves) along with the frequency samples of this filter (the heavy points). The frequency samples are defined by the relation

\[ \sin \left( \frac{\omega T}{2} \right) \frac{\sin \left( \frac{\omega N T}{2} / \sin \left( \frac{\omega'}{2} \right) \right) \right] \]

which, in the design of standard filters such as a low-pass filter, provide good ripple cancellation. Thus, by allowing variable frequency samples in a transition band between the in- and out-of-band regions, one can choose the frequency samples in this band to provide optimum ripple cancellation for either the out-of-band region, inband region, or a combination of the two. As the number of samples in the transition band increases, ever finer ripple cancellation is possible. Fig. 4 shows an example of a low-pass filter with an impulse-response duration of 256 samples, and 3 variable frequency samples in the transition band. Thirty-two inband frequency samples were set to 1.0, and the out-of-band samples set to 0.0. An optimization program determined the transition samples to minimize peak out-of-band ripple. As seen in Fig. 4(a), a peak out-of-band ripple of \(-88\, \text{dB}\) was obtained. An expanded view of the frequency response of the filter is shown in Fig. 4(b).

Fig. 5 shows a frequency-sampling design for an ideal full-band differentiator. In this figure the filter impulse response (of 256-sample duration), the magnitude response, and the magnitude error are plotted. The peak error is less than 0.1 percent (the average error is considerably less than this) and there is no phase error at all.

It should be noted that the mathematical solution of the optimization problem is straightforward because of the linearity of the frequency response with respect to the unconstrained variables.

**C. Equiripple Designs**

A third technique for designing FIR filters solves a system of nonlinear equations to generate a filter with an equiripple approximation error [3], [4], [14]. In this method the unknown quantities are both the \((N + 1)/2\) coefficients in the impulse response (assuming \(N\) odd, and a symmetrical impulse response) and a set of \((N - 3)/2\) frequencies at which extrema of the approximation error occur. By writing constraint equations on the extrema and on the derivatives, a system of \((N - 1)\) nonlinear equations in \((N - 1)\) unknowns can be obtained. Standard nonlinear optimization techniques are used to solve these equations.

For simplicity, we assume the filter impulse response \(h_n\) is symmetric and exists from \(n = -(N - 1)/2\) to \(n = (N - 1)/2\), where \(N\) is odd. Thus the filter frequency response can be written as

\[ H[\exp(j\omega T)] = \sum_{n=-(N-1)/2}^{(N-1)/2} h_n \exp(-j\omega n T) \]

which can easily be evaluated using the FFT algorithm.

The basic ideas behind frequency sampling can be seen in (13). The filter frequency response is seen to be linearly related to the frequency samples, and thus linear optimization techniques can be used to optimally select values of several, or all, the frequency samples to give the best approximation to the desired filter. Furthermore,
Fig. 4. (a) Low-pass filter with three transition samples and impulse-response duration of 256 samples, designed by frequency-sampling techniques. (b) Expanded view of filter frequency response.

Fig. 5. 256-point impulse response full-band differentiator designed by frequency-sampling technique.

Fig. 6. (a) 41-point equiripple low-pass filter. (b) Plot of z-plane positions of zeros of filter.
The derivative of the frequency response can be evaluated as

\[ H'[\exp(j\omega T)] = -\sum_{n=1}^{(N-1)/2} 2h_n \sin(\omega n T). \] (16)

At each extremum of the approximation error, an equation can be written relating the frequency response to the desired value and the allowed error, i.e.,

\[ H[\exp(j\omega T)] = H[\exp(j\omega T)] \pm \delta_i \] (17)

where \( H[\exp(j\omega T)] \) is the desired value at the frequency \( \omega_i \), \( \delta_i \) is the allowable approximation error, and \( \omega_i \) is an unknown frequency. A second equation can be written because the value of the frequency response at frequency \( \omega_i \) is an extremum of the error, i.e.,

\[ H'[\exp(j\omega T)] = 0. \] (18)

A series of equations of the form of (17) and (18) can be solved using nonlinear optimization techniques to give the filter coefficients and the frequencies of the extrema. Fig. 6 shows typical results for the design of a low-pass filter with equiripple error both inband and out-of-band. The impulse-response duration is 41 samples long. The inband ripple in this case is 0.01, and the out-of-band ripple is \( 10^{-4} \). The lowest set of curves show the results for equiripple designs for various values of \( \delta_2 \). Since equiripple designs are optimum, i.e., they have the smallest width of transition band for fixed \( \delta_1, \delta_2 \), the data for this case falls below the data for the other two cases. The percentage difference in normalized bandwidth between equiripple designs and frequency-sampling designs, for fixed \( \delta_1, \delta_2 \), is only about 30 percent. Since the suboptimal frequency-sampling designs are not much less efficient than the optimal equiripple designs, there may be circumstances in which it is preferable to use the suboptimal design. We shall discuss such cases in the next section.

\[ D(\delta_1,\delta_2) = NT[F_H(\delta_1,\delta_2) - F_L(\delta_1,\delta_2)] \] (19)

where

- \( F_H(\delta_1,\delta_2) \) upper cutoff frequency in Hz
- \( F_L(\delta_1,\delta_2) \) lower cutoff frequency in Hz
- \( \delta_1 \) inband peak ripple
- \( \delta_2 \) out-of-band peak ripple.

In this figure the inband ripple \( \delta_1 \) is shown as an additional parameter of the curves. The upper dashed curve shows the results for the Kaiser window, which, because of its time-limiting and band-limiting properties, tends to have the smallest normalized bandwidth. For an out-of-band ripple of \( -80 \) dB (\( \delta_2 = 10^{-4} \)) the normalized bandwidth is 5.01, whereas for a \( -40 \) dB ripple (\( \delta_2 = 10^{-5} \)) the normalized bandwidth is about 2.22. For these data points the inband peak ripple \( \delta_1 \) is equal to \( \delta_2 \).\(^2\)

The middle dashed curve shows the results for frequency-sampling designs. For an out-of-band ripple of \( -86 \) dB, the normalized width of the transition band is 4, or about \( \frac{1}{2} \) that of the Kaiser window. For values of \( \delta_2 \) of \( -66 \) and \( -46 \) dB, the transition width decreases to 3 and 2 for these designs. The design tradeoff between the Kaiser window and equiripple designs is the larger values of \( \delta_1 \) for the latter case versus the larger value of transition width for the former case.

The lowest set of curves show the results for equiripple designs for various values of \( \delta_2 \). Since equiripple designs are optimum, i.e., they have the smallest width of transition band for fixed \( \delta_1, \delta_2 \), the data for this case falls below the data for the other two cases. The percentage difference in normalized bandwidth between equiripple designs and frequency-sampling designs, for fixed \( \delta_1, \delta_2 \), is only about 30 percent. Since the suboptimal frequency-sampling designs are not much less efficient than the optimal equiripple designs, there may be circumstances in which it is preferable to use the suboptimal design. We shall discuss such cases in the next section.

\[ D(\delta_1,\delta_2) = NT[F_H(\delta_1,\delta_2) - F_L(\delta_1,\delta_2)] \] (19)

where

- \( F_H(\delta_1,\delta_2) \) upper cutoff frequency in Hz
- \( F_L(\delta_1,\delta_2) \) lower cutoff frequency in Hz
- \( \delta_1 \) inband peak ripple
- \( \delta_2 \) out-of-band peak ripple.

A. Transition Bandwidth for Low-Pass Filters

One of the most basic ways of comparing low-pass filters is to compare the width of the transition band for different values of peak out-of-band and inband ripple. Fig. 7 shows a comparison of this type for the Kaiser window, frequency-sampling, and equiripple designs. In this figure, the normalized width of the transition band as a function of the out-of-band ripple is plotted. The normalized width of the transition band is defined as follows:\(^3\)

\[ D(\delta_1,\delta_2) = NT[F_H(\delta_1,\delta_2) - F_L(\delta_1,\delta_2)] \] (19)

where

- \( F_H(\delta_1,\delta_2) \) upper cutoff frequency in Hz
- \( F_L(\delta_1,\delta_2) \) lower cutoff frequency in Hz
- \( \delta_1 \) inband peak ripple
- \( \delta_2 \) out-of-band peak ripple.

\[^2\] The multiplication by \( NT \) in (19) makes \( D(\delta_1,\delta_2) \) a dimensionless quantity, independent of \( NT \).

\[^3\] Low-pass filter designs, using the Kaiser window, have transition bandwidths that are within about 10 percent of the transition bandwidth of optimum equiripple filters.

B. Ease of Design

An important issue to any filter designer or user is how easy it is to design a new filter to meet some particular
specifications. For example, a user may desire an approximation to a frequency response which is not one for which the coefficients have already been precataloged. The following issues then arise.

1) What types of filters have been cataloged and can be referenced readily?

2) How easily can the design technique be applied to arbitrary magnitude and phase specifications?

3) Can filters with long impulse responses, i.e., large values of $N$, be designed to obtain sufficiently sharp transition ratios to meet the most stringent of design specifications?

In this section we shall try to answer these questions with respect to each of the three design techniques.

Since the window design technique is so straightforward to apply there is no need to catalog a large set of reference designs, and hence this has not been done. The window technique is readily amenable to approximating any set of filter specifications with no limitations on length of impulse response. The problems inherent in the window technique are the necessity to have Fourier-series coefficients of the desired periodic frequency response and the computation required for using optimum windows such as the Kaiser window or the Dolph–Chebychev window. The first problem is solved by the approximate procedure outlined in Fig. 2. Programs for computing Bessel functions for the Kaiser window, or inverse hyperbolic cosines for the Dolph–Chebychev window, are necessary. These are generally available as library subroutines on most computers. Thus the window technique tends to be relatively easy to use in the general design case.

In order to use the frequency-sampling technique, the user must program the linear optimization procedure, or use some available optimization program and adapt it to his specific needs. Once this has been accomplished, there is no problem designing filters with arbitrary magnitude and phase characteristics, or with any length impulse response. The linear nature of the problem guarantees convergence of the mathematical algorithms. For several standard filters there exists an extensive catalog of frequency-sampling designs [2], [13]. The types of filters included are low-pass and bandpass filters and wide-band (up to full-band) differentiators. From this catalog one could apply simple frequency transformations to obtain band-stop or high-pass filters [15].

In order to use the equiripple design techniques, the user must program the nonlinear optimization procedure, or use some available routine and again adapt it to his needs. The procedure can only design equiripple approximations to the magnitude response (assuming linear phase). It may be possible to modify it for arbitrary phase but this has not been done. The optimization technique used by Herrmann and Schuessler [3], [4] was capable of obtaining solutions only for small values of $N$ (on the order of 40). Recent work by Herrmann [16] and Hofstetter et al. [14] have solved the mathematical problems and filters can be designed with large values of $N$. (A small catalog of equiripple designs for low-pass filters is available from Herrmann.)

C. Realization of FIR Filters

Any FIR filter can be realized nonrecursively by direct convolution or fast convolution, or recursively using a cascade of a comb filter and a bank of parallel resonators. To use a nonrecursive realization requires the coefficients of the filter impulse response $h_n$. For direct convolution the output $y_n$ is determined explicitly from the inputs $x_n$ as

$$y_n = \sum_{m=0}^{N-1} h_m x_{n-m}.$$  (20)

For fast convolution a block of output samples is obtained from a block of input samples by Fourier transforming the input, multiplying the transform by the Fourier transform of the impulse response, and inverse Fourier transforming the product. Details of this technique are explained in [9], [10].

The way in which an FIR filter may be realized recursively is seen from (11). Instead of the impulse-response coefficients, the DFT of the impulse response, or, as we have called them, the frequency samples, can be used to realize the filter as a cascade of a comb filter $(1 - z^{-N})$ and a parallel bank of complex resonators. The significance of this realization is that for frequency-sampling designs, in many cases several, if not the majority, of the frequency samples would be 0.0. Hence this realization can be much more efficient than nonrecursive realizations. For filters designed by the other techniques, all the DFT coefficients will be nonzero, in general, and this realization will be much less efficient. Furthermore, for frequency-sampling designs where the majority of the frequency samples are 0.0 or 1.0, the effects of quantization are much less severe in the recursive realization than in the nonrecursive realization.

VI. Conclusion

As in the design of IIR digital filters, there are now several techniques available for designing FIR filters. The choice of technique depends heavily on the decision whether to compromise accuracy of approximation, ease of design, or a method of realization with a fixed quantization accuracy.

Acknowledgment

The author is grateful for the insights he has gained into filter-design problems from discussions of this material with Dr. R. W. Schafer, Dr. J. F. Kaiser, Dr. B. Gold, Dr. H. Helms, Prof. A. V. Oppenheim, Prof. O. Herrmann, and Prof. H. W. Schuessler.

O. Herrmann, Institut für Nachrichtentechnik, University of Erlangen–Nuremberg, 8520 Erlangen, Germany.
REFERENCES


[16] O. Herrmann, personal communication.

Lawrence R. Rabiner (S'62-M'67) was born in Brooklyn, N. Y., on September 28, 1943. He received the S.B. and S.M. degrees simultaneously in 1964, and the Ph.D. degree in electrical engineering in 1967, all from the Massachusetts Institute of Technology, Cambridge.

From 1962 through 1964, he participated in the cooperative plan in electrical engineering at Bell Telephone Laboratories, Inc., Whippany and Murray Hill, N. J., where he worked on digital circuitry, military communications problems, and problems in binaural hearing. He is presently engaged in research on speech communications and digital signal processing techniques at Bell Telephone Laboratories, Murray Hill.

Dr. Rabiner is a member of Eta Kappa Nu, Sigma Xi, Tau Beta Pi, and the Acoustical Society of America.