It is then seen that the signal estimate remains constant with random input frequency regardless of frequency redundancy whereas noise with frequency redundancy is decreased in accordance with a frequency sampling efficiency gain in Table I. This sampling efficiency is $(F_{(K)})/(F_{(\infty)}) = (F_{(K)})/(1.53)$. The effective number of integration samples with unity frequency redundancy is (1)/(1.53) providing 1.1 dB loss in post-detection integration gain as shown in Fig. 5 of [1]. Table I can also be converted to ripple loss versus frequency redundancy by entering the post detection integration gain curve of Fig. 3 with the normalized effective number of independent Fourier transforms of Table I. The results are shown in Fig. 6.

Conclusion

It has been shown that time redundancy can provide 0.9 dB gain in sensitivity when using the DFT. Further, frequency redundancy can provide 3.92 dB improvement over the maximum and 1.1 dB over the average ripple sensitivity loss experienced with the DFT.

References

[1] J. R. Williams and G. G. Ricker, "Signal detectability performance of optimum Fourier receivers," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 264-270, Oct. 1972.
[2] J. V. Difranco and W. L. Rubin, *Radar Detection*. Englewood Cliffs, N.J.: Prentice-Hall, 1965.
[3] C. N. Pryor, "Effect of finite sampling ratio on smoothing the output of a square law detector with narrowband input," Naval Ordnance Lab., White Oak, Md., NOLTR 71-29, Feb. 26, 1971.
[4] J. I. Marcum, "A statistical theory of target detection by pulsed radar," *IRE Trans. Inform. Theory*, vol. IT-6, pp. 59-144, Apr. 1960.
[5] J. R. Williams, "Fourier receiver sensitivity," in *Proc. 29th Navy Symp. Underwater Acoustics* (New London, Conn.), vol. 2, pp. 503-514, Oct. 31, 1972.
[6] G. D. Bergland, "A guided tour of the fast Fourier transform," *IEEE Spectrum*, vol. 6, p. 41, July 1969.
[7] B. Gold and C. Rader, *Digital Processing of Signals*. New York: McGraw-Hill, 1969.
[8] J. W. Cooley and J. W. Tukey, "An algorithm for the machine calculation of complex Fourier series," in *Math. Comput.*, vol. 19, 1965, pp. 297-301.
[9] L. R. Rabiner, R. W. Schafer, and C. M. Rader, "The chirp z-transform algorithm," *IEEE Trans. Audio Electroacoust.*, vol. AU-17, pp. 86-92, June 1969.
[10] P. D. Welch, "The use of the FFT for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. Audio Electroacoust.*, vol. AU-15, pp. 70-73, June 1967.
[11] A. H. Nuttall, "Spectral estimation by means of overlap FFT processing of window data," New London, Conn., NUSC Rep. 4169.
[12] R. B. Blackman and J. W. Tukey, *The Measurement of Power Spectra*. New York: Dover, 1958.
[13] H. L. VanTrees, *Detection, Estimation, and Linear Modulation Theory*. New York: Wiley, 1968.
[14] C. W. Helstrom, *Statistical Theory of Signal Detection*. New York: Pergamon, 1960.
[15] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*. New York: Wiley, 1965.
[16] T. H. Glisson, C. I. Black, and A. Sage, "On digital replica correlation algorithms with applications to active sonar," *IEEE Trans. Audio Electroacoust.*, vol. AU-17, pp. 190-197, Sept. 1969.
[17] W. B. Davenport, Jr., and W. L. Root, *An Introduction to the Theory of Random Signals and Noise*. New York: McGraw-Hill, 1958, p. 175 and prob. 2.

Analysis of Quantization Errors in the Direct Form for Finite Impulse Response Digital Filters

DAVID S. K. CHAN and LAWRENCE R. RABINER

Abstract—An analysis of the three possible types of quantization effects in the direct form realization of finite impulse response (FIR) digital filters is presented. These quantization effects include roundoff noise, A-D noise, and filter frequency response errors due to coefficient quantization. Since the analysis of roundoff noise and A-D noise for the direct form is straightforward, this paper concentrates on an analysis of the effects of quantized coefficients on the resulting filter frequency response. Based on this analysis, statistical bounds on the error incurred in the frequency response of a filter due to coefficient quantization are developed and verified by extensive experimental data. Using these bounds, a procedure for applying known techniques for FIR filter design to the design of filters with finite word length coefficients is presented. On the whole, the direct form is shown to be a very attractive structure for realizing FIR filters.

Manuscript received January 15, 1973; revised February 9, 1973.
The authors are with Bell Laboratories, Murray Hill, N.J. 07974.

I. Introduction

Important developments in research in recent years have made it possible to readily design finite impulse response (FIR) digital filters with arbitrary frequency or time response characteristics [1]-[6]. However, little is as yet known concerning practical aspects in the implementation of these filters. In particular, although many structures [7], [8] have been proposed for realizing FIR filters, differences in the effects of quantization on these different structures are still not

well understood. In two recent papers [9], [10] the problem of roundoff noise in cascade realizations of FIR filters has been treated in depth. The subject of this paper is the implementation of FIR filters in the direct form.

An analysis of all possible types of quantization effects in the direct form is presented, together with a procedure for obtaining filters whose coefficients are represented by a given number of bits and which satisfy given specifications on their frequency response. Although the method is not optimum, it is simple to apply and yields useful results.

The direct form is usually avoided in the implementation of infinite impulse response (IIR) digital filters mainly because Kaiser [2] has shown that the sensitivity of filter response to the accuracy of representation of the denominator coefficients in the IIR direct form increases very rapidly with increases in filter order compared to either the cascade or the parallel form. However, in this paper it is shown that the same is not true for FIR digital filters. In fact, the direct form is shown to be a very attractive structure for the realization of FIR filters.

II. Quantization Effects in the FIR Direct Form

There are three basic ways in which quantization affects the performance of an FIR digital filter.

1) Quantization of the results of arithmetic operations within the filter causes errors in the filter output, referred to as roundoff noise.

2) Quantization of input samples to the filter causes inaccuracies known as A-D noise.

3) Quantization of the filter coefficients alters the frequency response of the filter.

Overflows can also occur within the filter. However, if proper scaling procedures are used, such overflows can be eliminated. In addition, limit cycles can occur in recursive realizations of FIR filters (such as the frequency-sampling structure [7]). However, limit cycles cannot occur in the direct form, which is a nonrecursive structure. In the following, each of the three types of quantization effects listed above is analyzed for the direct form.

A. Roundoff Noise

To analyze roundoff noise, the most important case of fixed-point arithmetic with rounding is assumed. Similar analyses follow for all other cases. Furthermore, the usual model used for the analysis of roundoff errors [9] is employed, viz., to each rounding point in the filter is associated a zero-mean white noise source of variance $Q^2/12$ (Q = quantization step size) whose samples are uniformly distributed random variables on the interval $(-Q/2, Q/2)$, and all noise sources are assumed to be uncorrelated with each other and with the input signal. Using this model, the mean and variance of roundoff noise at the output of a filter can be easily computed.

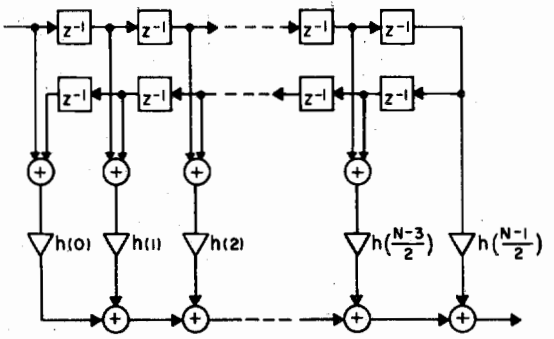


Fig. 1. Block diagram of linear phase direct form FIR filter.

The transfer function for an FIR filter can be written in the form

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

where $\{h(n)\}$ is the impulse response of the filter. The direct form is simply defined to be a straightforward implementation of (1). If the filter has linear phase,[1] then $\{h(n)\}$ satisfies [9]

$$h(n) = h(N-1-n), \qquad 0 \leqslant n \leqslant N-1. \tag{2}$$

In this case, a slightly different version of the direct form, requiring fewer multipliers, can be derived. For simplicity only, the case of odd N is considered. By (1) and (2),

$$\begin{aligned} H(z) &= \sum_{n=0}^{(N-3)/2} h(n) z^{-n} \\ &\quad + \sum_{n=(N+1)/2}^{N-1} h(n) z^{-n} + h\left(\frac{N-1}{2}\right) z^{-(N-1)/2} \\ &= \sum_{n=0}^{(N-3)/2} h(n)\left[z^{-n} + z^{-(N-1-n)}\right] \\ &\quad + h\left(\frac{N-1}{2}\right) z^{-(N-1)/2}. \end{aligned} \tag{3}$$

Equation (3) is the defining equation for the linear phase direct form. It is seen that, given the filter order, the direct form for a linear phase filter requires approximately half as many multipliers as that required for an arbitrary phase filter. A block diagram of the linear phase direct form is shown in Fig. 1.

The statistics of the roundoff noise at the output of a direct form FIR filter depend, as expected, on the location of points in the filter where rounding is performed. Two possibilities are considered. First of all,

[1] The results presented in this paper can be extended in a straightforward manner to the case where the filter impulse response is antisymmetric, i.e., $h(n) = -h(N-1-n)$, $0 \leqslant n \leqslant N-1$. For convenience we shall neglect this case here.

if all multiplication products are represented exactly and rounding is performed only after they are summed, i.e., at the filter output, then only one noise source is present in the filter, and it superimposes noise directly onto the output signal.[2] Thus, independent of filter order and type of direct form (i.e., linear phase or arbitrary phase), the output roundoff noise is uniformly distributed between $-Q/2$ and $Q/2$, with zero mean and variance $Q^2/12$, where Q is the quantization step size.

On the other hand, if all multiplication products are rounded before they are summed, then the arbitrary phase direct form and the linear phase direct form yield different results for a given N. Specifically, denoting by $\{e_i(n)\}$ the noise sequence produced by the ith noise source, and by $\{E(n)\}$ the noise sequence at the filter output, for the arbitrary phase direct form

$$E(n) = \sum_{n=0}^{N-1} e_i(n), \tag{4}$$

but for the linear phase direct form

$$E(n) = \sum_{n=0}^{(N-1)/2} e_i(n). \tag{5}$$

The mean of the output noise is zero in either case, but the variance for the arbitrary phase direct form is

$$\sigma_A^2 = N\frac{Q^2}{12}, \tag{6}$$

whereas for the linear phase direct form it is

$$\sigma_L^2 = \left(\frac{N+1}{2}\right)\frac{Q^2}{12}, \tag{7}$$

Thus for all the possible variations of the FIR direct form mentioned, the output roundoff noise variance always satisfies

$$\sigma_{RO}^2 \geqslant \frac{Q^2}{12}. \tag{8}$$

Note at this point that if the noise sources were statistically independent rather than merely uncorrelated, then by the central limit theorem of probability theory, in both (4) and (5) $E(n)$ would be essentially Gaussian distributed for all n and sufficiently large N since it would be a sum of independent, identically distributed random variables. In fact, the convergence as N is made large would be very rapid since the $e_i(n)$'s are uniformly distributed. Thus, for all values of N of practical interest, the output noise would be essentially Gaussian. Even though an independence assumption on the noise sources may not be valid, the output noise distribution can still be expected to resemble a Gaussian distribution.

[2] We are assuming throughout this discussion that all the filter coefficients are nontrivial, e.g., an impulse response consisting of ± 1's would have no roundoff noise. Although such cases do occur in practice, they are trivial to analyze, and hence will not be considered here.

B. A-D Noise

Next, A-D noise is considered. Again, the quantization process is modeled as a white noise source whose samples are uniformly distributed on $(-Q/2, Q/2)$. Since this quantization noise is added to the input signal, its effect on the output signal of a filter is independent of the type of structure used to realize the filter. In particular, if $\{\epsilon(n)\}$ denotes the A-D noise sequence on input to an FIR filter as defined by (1), and $\{\mathcal{E}(n)\}$ denotes the output noise sequence, then

$$\mathcal{E}(n) = \sum_{k=0}^{N-1} h(k)\,\epsilon(n-k). \tag{9}$$

Thus the output A-D noise has zero mean and variance given by

$$\sigma_{AD}^2 = \frac{Q^2}{12}\sum_{n=0}^{N-1} h^2(n), \tag{10}$$

or using Parseval's theorem for discrete signals,

$$\sigma_{AD}^2 = \frac{Q^2}{12}\cdot\frac{1}{2\pi}\int_0^{2\pi} |H(e^{j\omega})|^2\, d\omega. \tag{11}$$

It is shown next that when the gain of a filter is determined by sum scaling, peak scaling, or L_p-norm scaling where $p \geqslant 2$, viz., the practically important scaling methods [9], then

$$\sigma_{AD}^2 \leqslant \frac{Q^2}{12}. \tag{12}$$

Consider first sum scaling, which requires

$$\sum_{n=0}^{N-1} |h(n)| = 1 \tag{13}$$

where $\{h(n)\}$ is the scaled filter impulse response. Clearly, (13) implies $|h(n)| \leqslant 1$ for all n; hence $h^2(n) \leqslant |h(n)|$ for all n; thus

$$\sum_{n=0}^{N-1} h^2(n) \leqslant \sum_{n=0}^{N-1} |h(n)| = 1. \tag{14}$$

Next (see [9]), for any $p \geqslant 2$,

$$\left[\frac{1}{2\pi}\int_0^{2\pi} |H(e^{j\omega})|^2\, d\omega\right]^{1/2}$$

$$\leqslant \left[\frac{1}{2\pi}\int_0^{2\pi} |H(e^{j\omega})|^p\, d\omega\right]^{1/p} \leqslant \max_{\omega} |H(e^{j\omega})|. \tag{15}$$

Hence if either $[1/2\pi \int_0^{2\pi} |H(e^{j\omega})|^p\, d\omega]^{1/p} = 1$, which defines L_p-norm scaling, for some $p \geqslant 2$, or

max $|H(e^{j\omega})| = 1$, which defines peak scaling, then

$$\frac{1}{2\pi}\int_0^{2\pi} |H(e^{j\omega})|^2\, d\omega \leq 1. \tag{16}$$

Equations (10), (11), (14), and (16) show that $\sigma_{AD}^2 \leq Q^2/12$ for all these scaling methods; hence (12) is established.

Thus it is shown that for all practical filters realized in direct form, the A-D noise at the filter output has a variance that is smaller than or equal to that of the roundoff noise. In fact, for the often more economical scheme of rounding before summation, A-D noise is negligible compared to roundoff noise for filter orders greater than about 10.

It is instructive to point out that the output A-D noise is also essentially Gaussian for sufficiently large N if different samples of the quantization error sequence were statistically independent of each other because, although the random variables that make up the sum in (9) are not identically distributed, they are all uniformly distributed with zero probability density outside a finite interval; hence they satisfy the Lindeberg condition [11] of the central limit theorem.

C. Filter Response Errors

The effect of coefficient quantization on the frequency response of an FIR filter in direct form is now analyzed. The linear phase direct form is considered first. Again, N is assumed to be odd, and rounding is used as the quantization process. From (3) the frequency response of a linear phase FIR filter is given by

$$\begin{aligned} H(e^{j\omega}) &= \sum_{n=0}^{(N-3)/2} h(n)\left[e^{-j\omega n} + e^{-j\omega(N-1-n)}\right] \\ &\quad + h\left(\frac{N-1}{2}\right) e^{-j\omega(N-1)/2} \\ &= \left[\sum_{n=0}^{(N-3)/2} 2h(n)\cos\left[\left(\frac{N-1}{2} - n\right)\omega\right] \right. \\ &\quad \left. + h\left(\frac{N-1}{2}\right)\right] e^{-j\omega(N-1)/2}. \end{aligned} \tag{17}$$

The factor $e^{-j\omega(N-1)/2}$ in (17) simply represents a pure delay of an integer number of samples that is unaffected by quantization of the coefficients $\{h(n)\}$. Hence this factor need not be considered, and only the change in the real function $\bar{H}(e^{j\omega}) = H(e^{j\omega}) \times e^{j\omega(N-1)/2}$ due to coefficient quantization is studied.

Following the approach of Knowles and Olcayto [12], let $\{h^*(n)\}$ be the sequence that results when $\{h(n)\}$ is rounded to a quantization step size of Q. Then $h^*(n) = h(n) + e(n)$ and $h^*(N-1-n) = h^*(n)$ for $0 \leq n \leq (N-1)/2$, where $e(n)$ for each n is a number that satisfies $|e(n)| \leq Q/2$. Furthermore, let $H^*(z)$ be the z-transform of $\{h^*(n)\}$ and let $\bar{H}^*(e^{j\omega}) = H^*(e^{j\omega})\, e^{j\omega(N-1)/2}$. Finally, define an error function by

$$E_L(e^{j\omega}) = \bar{H}^*(e^{j\omega}) - \bar{H}(e^{j\omega}). \tag{18}$$

Thus

$$\begin{aligned} E_L(e^{j\omega}) &= \sum_{n=0}^{(N-3)/2} 2e(n)\cos\left[\left(\frac{N-1}{2} - n\right)\omega\right] \\ &\quad + e\left(\frac{N-1}{2}\right). \end{aligned} \tag{19}$$

From (19) it is seen that $E_L(e^{j\omega})$ is simply the frequency response of a linear phase filter that has $\{e(n)\}$ as the first half of its impulse response (the other half by symmetry). The function $E_L(e^{j\omega})$ has the physical significance that a filter which has quantized coefficients can be represented as the parallel connection of the "infinite-precision" version of that filter with a filter whose frequency response is $E_L(e^{j\omega})\, e^{-j\omega(N-1)/2}$.

Since $|e(n)| \leq Q/2$, a bound on $E_L(e^{j\omega})$ that is independent of the $e(n)$'s can be derived as follows:

$$\begin{aligned} |E_L(e^{j\omega})| &\leq \sum_{n=0}^{(N-3)/2} 2|e(n)| \left|\cos\left[\left(\frac{N-1}{2} - n\right)\omega\right]\right| \\ &\quad + \left|e\left(\frac{N-1}{2}\right)\right| \leq \frac{Q}{2}\left[1 + 2\sum_{n=1}^{(N-1)/2} |\cos n\omega|\right]. \end{aligned} \tag{20}$$

Or, independent of ω,

$$|E_L(e^{j\omega})| \leq N\frac{Q}{2}. \tag{21}$$

Now let us consider the arbitrary phase direct form. From (1) the frequency response of an arbitrary phase FIR filter is given by

$$H(e^{j\omega}) = \sum_{n=0}^{N-1} h(n) e^{-j\omega n}. \tag{22}$$

If $H^*(e^{j\omega}) = \sum_{n=0}^{N-1} h^*(n)\, e^{-j\omega n}$ is the frequency response of the filter with quantized coefficients, where $h^*(n) = h(n) + e(n)$ and $|e(n)| \leq Q/2$ for $0 \leq n \leq N-1$, then define an error function by

$$\begin{aligned} E_A(e^{j\omega}) &= H^*(e^{j\omega}) - H(e^{j\omega}) \\ &= \sum_{n=0}^{N-1} e(n)\, e^{-j\omega n}. \end{aligned} \tag{23}$$

Again, the error function is the frequency response of an FIR filter that has impulse response $\{e(n)\}$. However, note that $E_A(e^{j\omega})$ is a complex function. The same bound as (21) also applies to $E_A(e^{j\omega})$ since

$$|E_A(e^{j\omega})| \leq \sum_{n=0}^{N-1} |e(n)||e^{-j\omega n}| \leq N\frac{Q}{2}. \tag{24}$$

Furthermore, the error in the magnitude of the frequency response due to coefficient quantization also satisfies the bound in (24) because of the well-known inequality for complex numbers:

$$\|H^*(e^{j\omega})| - |H(e^{j\omega})\| \leq |H^*(e^{j\omega}) - H(e^{j\omega})|. \quad (25)$$

Unfortunately, all of the bounds derived thus far are overly pessimistic, and thus are of little practical usefulness. In the next section, far more useful statistical bounds are derived.

III. Statistical Analysis of Filter Response Errors

Because of the inherently hard-to-predict nature of quantization errors, a statistical analysis of the effect of coefficient quantization on filter response is appropriate, even though for a given filter the quantization process is performed only once, after which the filter response is exactly determined. The aim of such an analysis is to provide the filter designer with some means of predicting, without knowing the values of the coefficients of a filter, how much accuracy for the coefficients is required to obtain a desired filter response. In the following, such an analysis is presented for the case of coefficient rounding.

The statistical model [12] to be used is a very reasonable one that assumes that errors due to the quantization of different coefficients are statistically independent, and that each error is uniformly distributed between $-Q/2$ and $Q/2$, and thus has zero mean and variance $Q^2/12$, where Q is the quantization step size.

The linear phase direct form is considered first. From (19) and the statistical model it is clear that for each ω, the error in the frequency response due to coefficient rounding has zero mean and a variance given by

$$\begin{aligned}\overline{E_L(e^{j\omega})^2} &= \sum_{n=0}^{(N-3)/2} 4\,\overline{e(n)^2}\cos^2\left[\left(\frac{N-1}{2}-n\right)\omega\right] \\ &\quad + \overline{e\left(\frac{N-1}{2}\right)^2} \\ &= \frac{Q^2}{12}\left[1 + 4\sum_{n=1}^{(N-1)/2}\cos^2\omega n\right]. \end{aligned} \quad (26)$$

Thus, defining a weighting function by

$$W_N(\omega) = \left[\frac{1}{2N-1}\left(1 + 4\sum_{n=1}^{(N-1)/2}\cos^2\omega n\right)\right]^{1/2} \quad (27)$$

the standard deviation of the error is given by

$$\begin{aligned}\sigma_{EL}(\omega) &= \overline{E_L(e^{j\omega})^2}^{1/2} \\ &= \sqrt{\frac{2N-1}{3}}\cdot\frac{Q}{2}W_N(\omega). \end{aligned} \quad (28)$$

$W_N(\omega)$ may be expressed in closed form by summing the series in (27), thus giving

$$W_N(\omega) = \left[\frac{1}{2N-1}\left(N + 2\cos\left(\frac{N+1}{2}\right)\omega\right)\frac{\sin\left(\frac{N-1}{2}\right)\omega}{\sin\omega}\right]^{1/2}, \quad (29)$$

or following some arithmetic manipulation,

$$W_N(\omega) = \left[\frac{1}{2} + \frac{1}{2N-1}\left(\frac{-1}{2} + \frac{\sin N\omega}{\sin\omega}\right)\right]^{1/2}. \quad (30)$$

Equation (30) shows $W_N(0) = W_N(\pi) = 1$, and $0 < W_N(\omega) \leq 1$ for all N. Thus

$$\sigma_{EL}(\omega) \leq \frac{Q}{2}\sqrt{\frac{2N-1}{3}}. \quad (31)$$

The behavior of $W_N(\omega)$ in the limit of large N is readily seen from (30). In the range $0 < \omega < \pi$,

$$\lim_{N\to\infty} W_N(\omega) = \frac{1}{\sqrt{2}}, \quad (32)$$

whereas for $\omega = 0, \pm\pi, \pm 2\pi, \cdots$,

$$W_N(\omega) = 1, \quad \text{all } N. \quad (33)$$

The convergence of (32) is not uniform on $(0, \pi)$. However, on $[\epsilon, \pi - \epsilon]$ for any $\epsilon > 0$ the convergence is uniform. Fig. 2 shows $W_N(2\pi f)$ versus f for $N = 7$ and $N = 67$. The significance of these plots and (32) is that they show $\sigma_{EL}(\omega)$ to be well described by its bound in (32), viz., $\sigma_{EL}(\omega)$ is given to better than a factor of 2 by its bound for all ω.

From (19) it is seen that for any ω, $E_L(e^{j\omega})$ is a sum of independent random variables whose probability density functions vanish outside some finite interval, and hence satisfy the Lindeberg condition [11] of the central limit theorem. Thus $E_L(e^{j\omega})$ is essentially Gaussian for sufficiently large N. Furthermore, the convergence to Gaussian distribution is expected to be quite rapid since the individual terms in the summation are all uniformly distributed random variables. Because of this tendency of the errors to be Gaussian, their mean and variance alone constitute an excellent description of their statistical behavior.

A bound similar to (31) is derived next for the arbitrary phase direct form. The complex error incurred in this case in the frequency response due to coefficient quantization is given in (23). Note that $E_A(e^{j\omega})$ is a two-dimensional random variable. Clearly, its mean is zero for all ω. Furthermore, its probability density function is symmetric about the origin in the complex plane. Thus the second moment of the magnitude of $E_A(e^{j\omega})$ about the origin is a good measure of its deviation from the mean. Now

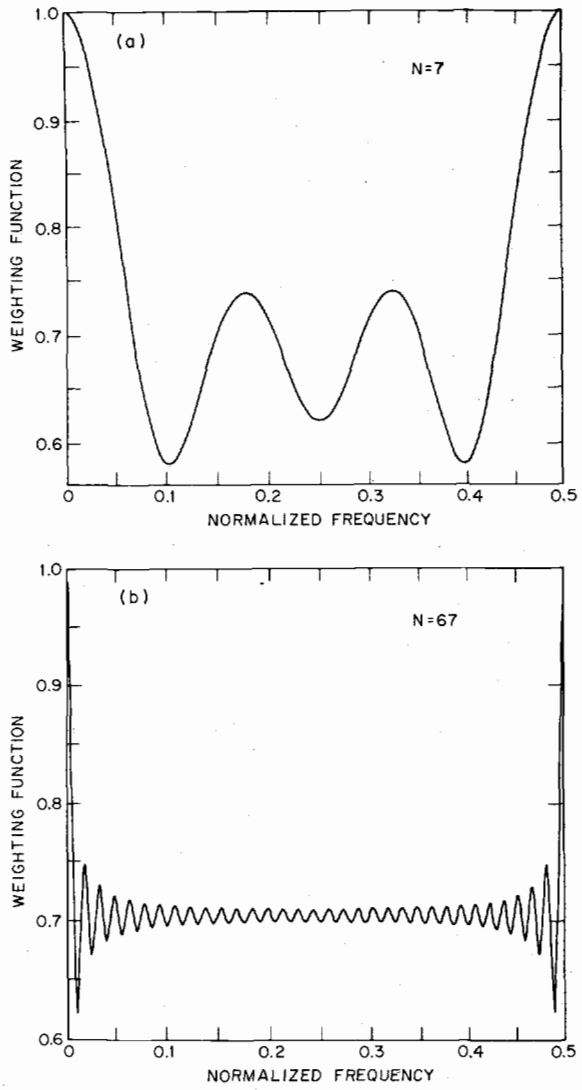


Fig. 2. Weighting function defined in (27) for $N = 7$ and $N = 67$.

$$|E_A(e^{j\omega})|^2 = \left(\sum_{n=0}^{N-1} e(n)\,e^{-j\omega n}\right)\left(\sum_{n=0}^{N-1} e(n)\,e^{j\omega n}\right)$$

$$= \sum_{n=0}^{N-1}\sum_{k=0}^{N-1} e(n)\,e(k)\,e^{-j\omega(n-k)}. \tag{34}$$

Therefore

$$\overline{|E_A(e^{j\omega})|^2} = \sum_{n=0}^{N-1} \overline{e(n)^2} = N\frac{Q^2}{12} \tag{35}$$

using the independence assumption. Thus a measure of the deviation of $|E_A(e^{j\omega}|)$ from zero can be defined as

$$\sigma_{EA} = \frac{Q}{2}\sqrt{\frac{N}{3}}. \tag{36}$$

Since, as shown in (25), the error incurred in the magnitude of the frequency response is bounded by $|E_A(e^{j\omega})|$, σ_{EA} is a useful bound on its expected amount of deviation from zero.

IV. Implications of Analysis for Design

Considerable insight into the problem of implementing FIR filters in direct form can be gained from the analysis of the previous section if the results are phrased in terms of the usual filter design parameters. In this section, useful relationships between different parameters are derived for the linear phase direct form.

Most filter design problems of interest involve finding a filter with a frequency response that approximates some ideal response to within a specified amount of error. In general, the specified bound on the error is a function of frequency. However, for the important case of band-select filters, a single bound on the error is usually specified for each frequency band of interest, where bounds for different frequency bands may be different. For simplicity, the discussions of this paper are restricted to band-select filters. However, this is only a matter of convenience, and the ideas developed can be easily generalized.

Let $L(\omega)$ be some real, ideal band-select function that is desired to be approximated by the frequency response of a linear phase FIR filter. The usual design specifications consist of a set of disjoint frequency bands $\Omega_k \subset [0, \pi]$, $k = 1, \cdots, M$ and a set of error bounds $\delta_k > 0$, $k = 1, \cdots, M$ such that for each k, $L(\omega)$ is to be approximated to within an error of δ_k for all $\omega \in \Omega_k$. The frequency bands Ω_k are separated by transition bands where the filter frequency response is unconstrained. Quantizing the coefficients of a filter designed to meet these specifications can increase the approximation error for $\omega \in \cup_{k=1}^{M} \Omega_k$ beyond their specified bounds. A statistical bound on the effect of such error increases is first developed.

Define, as in Section 11-C, $\bar{H}(e^{j\omega})$ and $\bar{H}^*(e^{j\omega})$ to be the frequency response, less the pure delay factor, of a linear phase FIR filter in direct form that has unquantized and quantized coefficients, respectively. Suppose $\bar{H}(e^{j\omega})$)has been designed to approximate $L(\omega)$ so that

$$\max_{\omega\in\Omega_k} |\bar{H}(e^{j\omega}) - L(\omega)| = \delta_k, \qquad k = 1, \cdots, M. \tag{37}$$

Clearly, for all ω

$$|\bar{H}^*(e^{j\omega}) - L(\omega)| \leq |\bar{H}^*(e^{j\omega}) - \bar{H}(e^{j\omega})| + |\bar{H}(e^{j\omega}) - L(\omega)|. \tag{38}$$

Thus for all $\omega \in \Omega_k$,

$$|\bar{H}^*(e^{j\omega}) - L(\omega)| \leq \max_{\omega\in\Omega_k} |E_L(e^{j\omega})| + \delta_k \tag{39}$$

where $E_L(e^{j\omega})$ is as defined in (18). With high probability $E_L(e^{j\omega})$ ought to be bounded by two to three

times its standard deviation; thus for all ω

$$|E_L(e^{j\omega})| \lessapprox 2\sigma_e \tag{40}$$

where $\lessapprox$ denotes "is with high probability less than or equal to" and σ_e is defined as

$$\sigma_e = \max_{\omega} \sigma_{EL}(\omega) = \frac{Q}{2}\sqrt{\frac{2N-1}{3}} \tag{41}$$

with $\sigma_{EL}(\omega)$ defined in (28). Consequently,

$$|\bar{H}^*(e^{j\omega}) - L(\omega)| \lessapprox \underbrace{2^{-t}\sqrt{\frac{2N-1}{3}} + \delta_k}_{\epsilon_k}, \omega \in \Omega_k \tag{42}$$

where t is the number of bits to which the coefficients of $\bar{H}^*(e^{j\omega})$ are rounded, so that $Q = 2^{-t}$ (sign bit not included).

Equation (42) says that, with high probability, $L(\omega)$ can be approximated on Ω_k by the frequency response of an N-point linear phase filter, with t-bit coefficients, to an error bounded by ϵ_k if the frequency response of the filter with infinite-precision coefficients $\bar{H}(e^{j\omega})$ can be designed so that for all $\omega \in \Omega_k$

$$|\bar{H}(e^{j\omega}) - L(\omega)| \leq \epsilon_k - 2^{-t}\sqrt{\frac{2N-1}{3}}. \tag{43}$$

Equations (42) and (43) are investigated further by rephrasing them in terms of in-band "rejection"[3] in decibels. Define the in-band rejection on Ω_k of the quantized and unquantized filter, respectively, as

$$DL_k^* = -20\log_{10}\left(\max_{\omega\in\Omega_k} |\bar{H}^*(e^{j\omega}) - L(\omega)|\right) \tag{44}$$

and

$$DL_k = -20\log_{10}\left(\max_{\omega\in\Omega_k} |\bar{H}(e^{j\omega}) - L(\omega)|\right). \tag{45}$$

Then from (37) and (42),

$$DL_k^* \gtrapprox -20\log_{10}\left(10^{-(DL_k/20)} + 2^{-t}\sqrt{\frac{2N-1}{3}}\right). \tag{46}$$

The lower bound in (46) is plotted in Fig. 3 as a function of DL_k for $N = 129$ and several values of t. It is seen that for each t, there is a maximum value of in-band rejection that can be statistically guaranteed (i.e., with high probability) as a lower bound to the attainable in-band rejection of an N-point filter if its coefficients are rounded to t-bits, no matter how high

[3] The use of the term "rejection" is not meaningful in frequency regions where the filter is designed to pass the signal, i.e., the filter passbands. It is used here for consistency in describing the characteristics of the filter in *any* frequency band. It should also be noted that the rejection in any band is not referenced to a peak filter gain. Thus if all the filter coefficients are multiplied by a constant, the rejection in each band changes even though the filter is substantially the same.

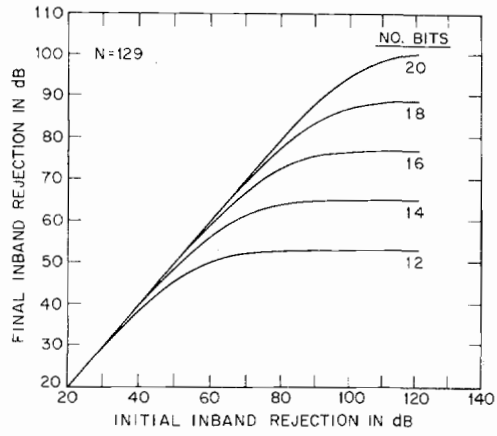
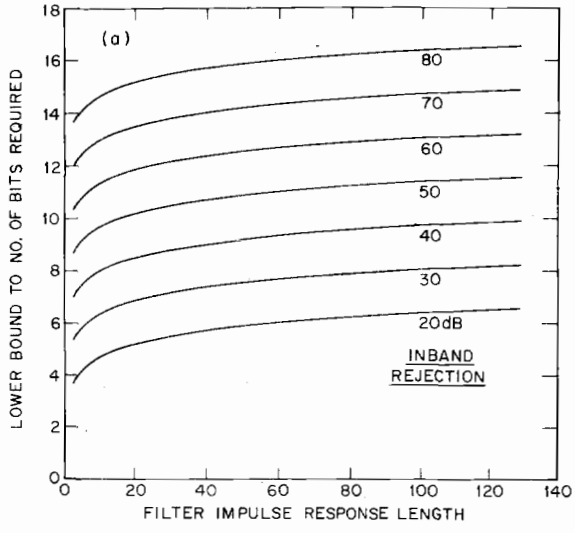
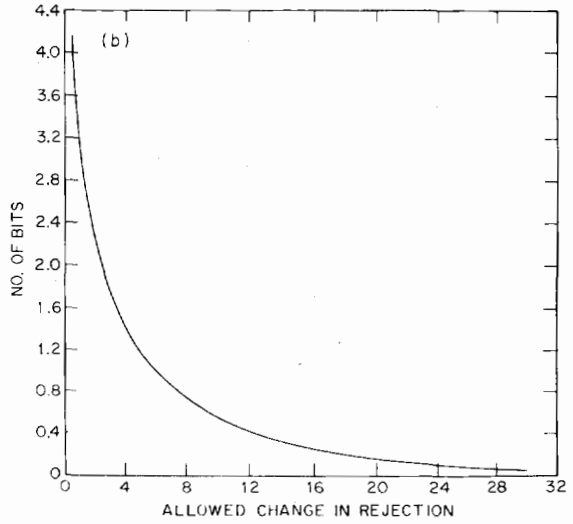


Fig. 3. Lower bounds on in-band rejection obtained after coefficient quantization as a function of initial rejection.



(a)



(b)

Fig. 4. (a) Lower bound to number of bits required to realize a filter of given minimum in-band rejection. (b) Additional number of bits required in excess of lower bound in (a) if change in rejection due to quantization is constrained.

a value of in-band rejection the unquantized filter has. Thus if the significance level chosen in (40) is used for design purposes, a minimum number of bits is required to implement any filter of a given order and a given minimum in-band rejection for the most tightly specified band if design specifications are to be met with as much probability as there exists for (40) to be satisfied. To find this lower bound for t, let DL^*m_k equal the right-hand side of (46), so that $DL_k^* \gtrsim DL^*m_k$, and solve for t to give

$$t \geqslant t_{\min} = -\log_2 \left[\sqrt{\frac{3}{2N-1}} \left(10^{-(DL^*m_k/20)} - 10^{-(DL_k/20)} \right) \right]$$

$$= \frac{DL^*m_k + 10 \log_{10} [(2N-1)/3]}{20 \log_{10} 2} - \log_2 \left(1 - 10^{(DL^*m_k - DL_k)/20} \right). \quad (47)$$

Defining t_∞ to be the limit of $t_{\min}$ as $DL_k \to \infty$, viz.,

$$t_\infty = \frac{DL^*m_k + 10 \log_{10} [(2N-1)/3]}{20 \log_{10} 2} \cong \frac{DL^*m_k}{6} + \frac{5}{3} \log_{10} [(2N-1)/3], \quad (48)$$

(47) can be written as

$$t \geqslant t_{\min} = t_\infty - \log_2 (1 - 10^{-(\Delta_k/20)}) \quad (49)$$

where $\Delta_k = DL_k - DL^*m_k$ is the maximum allowable change in in-band rejection on Ω_k due to coefficient quantization.

Thus t_∞ is a lower bound (statistical) for t given N and DL^*m_k (desired minimum rejection on Ω_k of quantized filter), but $t_{\min}$ is a tighter, more practical bound that is a function of Δ_k. Equations (48) and (49) show that, for fixed N and Δ_k, t is bounded to increase linearly with DL^*m_k at a rate of 1 bit per 6 dB of increase in the in-band rejection desired for the quantized filter. Fig. 4(a) and (b) show plots of t_∞ and $(t_{\min} - t_\infty)$, respectively. It is of interest to note that, from Fig. 4(b), approximately 1 bit above the number given by t_∞ is required to realize any filter if $(DL_k - DL_k^*)$ for that filter is to be bounded by 6 dB, i.e., if the in-band approximation error on Ω_k is not to increase by more than double due to quantization.

A useful formula needed in the next section is derived from (46) as follows:

$$DL_k - DL_k^* \lesssim 20 \log_{10} \left(1 + 10^{(DL_k/20)} \, 2^{-t} \sqrt{\frac{2N-1}{3}} \right) = 20 \log_{10} \left(1 + 2^{((DL_k/6.02)-t)} \sqrt{\frac{2N-1}{3}} \right) \quad (50)$$

taking $20 \log_{10} 2 = 6.02$.

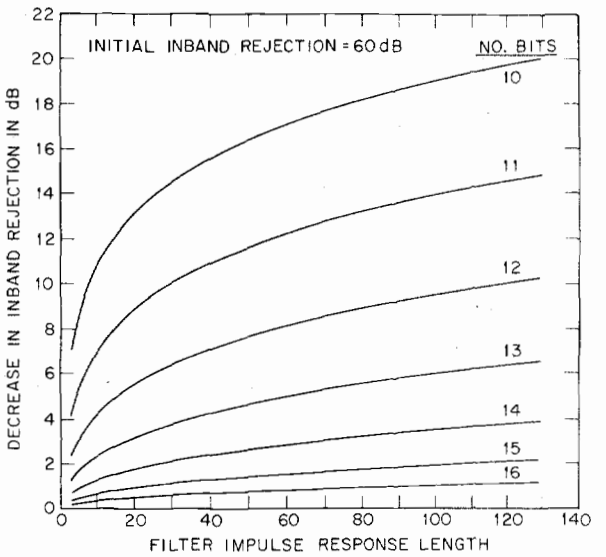


Fig. 5. Bounds on the change of in-band rejection due to coefficient quantization.

Equation (50) gives a statistical bound on the loss of in-band rejection due to coefficient quantization. This bound is plotted in Fig. 5 as a function of N for $DL_k = 60$ dB and $t = 10$ to 16 bits.

V. Experimental Verification and Discussion

To verify the validity of (40), the values of $DL_k - DL_k^*$ for Ω_k equal to both the passband and stopband were found for a large number of low-pass extra-ripple [5] filters with equal number of ripples in their passband and stopband and values of N ranging from 15 to 127. Fig. 6(a) shows a plot of these values for filters with passband and stopband rejection of 40 and 60 dB, respectively, and coefficients quantized to 11 bits, together with the bound in (50) (solid curves) for these values of in-band rejection. Also, the bound of (50) derived using σ_e rather than $2\sigma_e$ in (40) [equivalent to replacing t by $(t+1)$ in (50)] is plotted for both passband and stopband as dashed curves in Fig. 6(a). Similar results, except with 16-bit coefficients, are shown in Fig. 6(b). Clearly, (40) is well supported by these results.

It is important to point out that, contrary to claims by Herrmann and Schuessler [8], [13], the analysis presented in previous sections shows that the passband of a linear phase direct form filter is in no way favored over the stopband to have lower sensitivity to coefficient accuracy. It is only a difference in in-band rejection that makes one band more sensitive than another. For instance, in the example given by Herrmann and Schuessler [13], the absolute errors due to quantization added to the passband and stopband responses of the filter are actually comparable. However, since the ripples in the passband are 100 times larger than those in the stopband, percentagewise the errors in the passband are not seen.

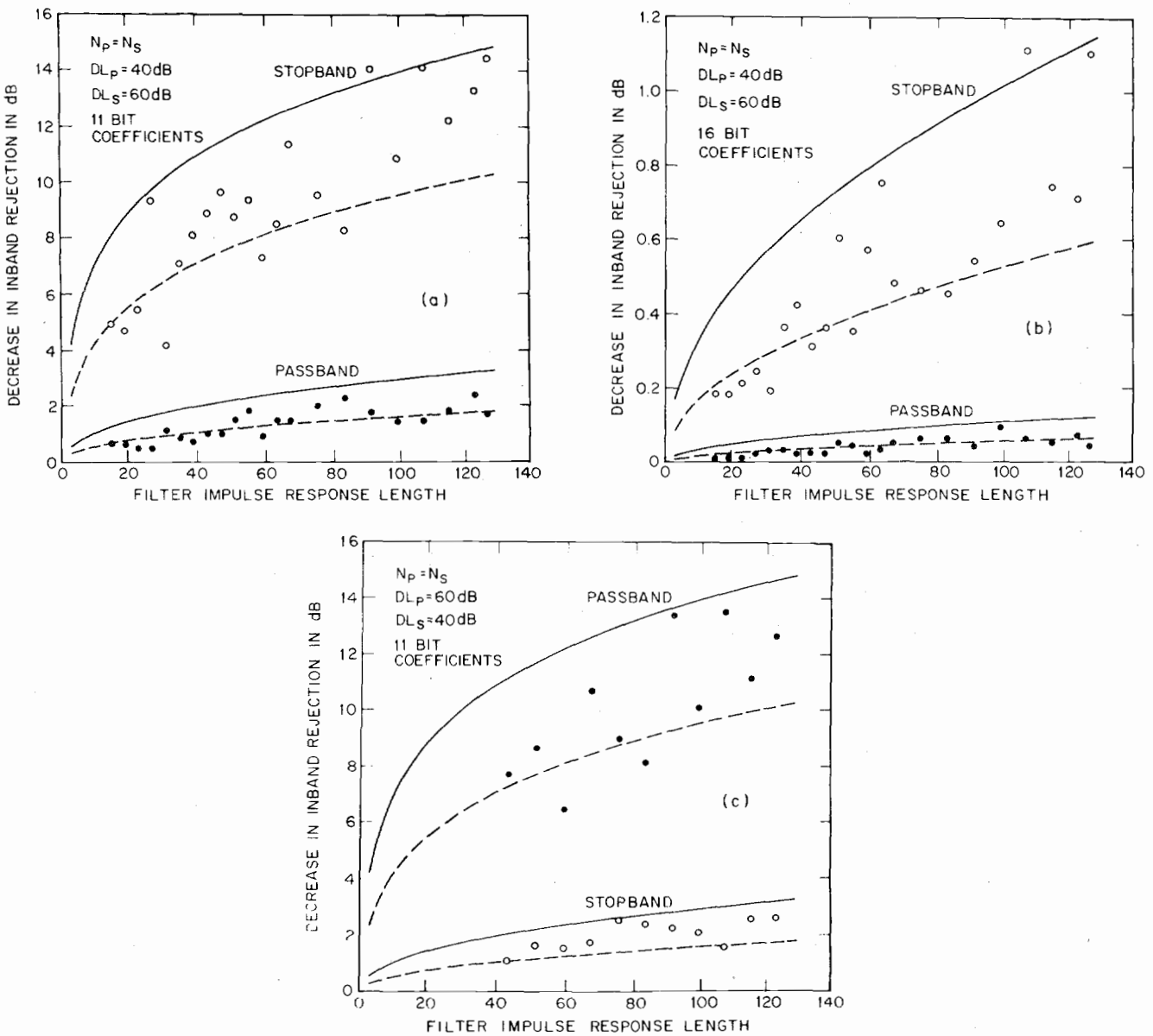


Fig. 6. Comparisons of experimental data to predicted statistical bounds.

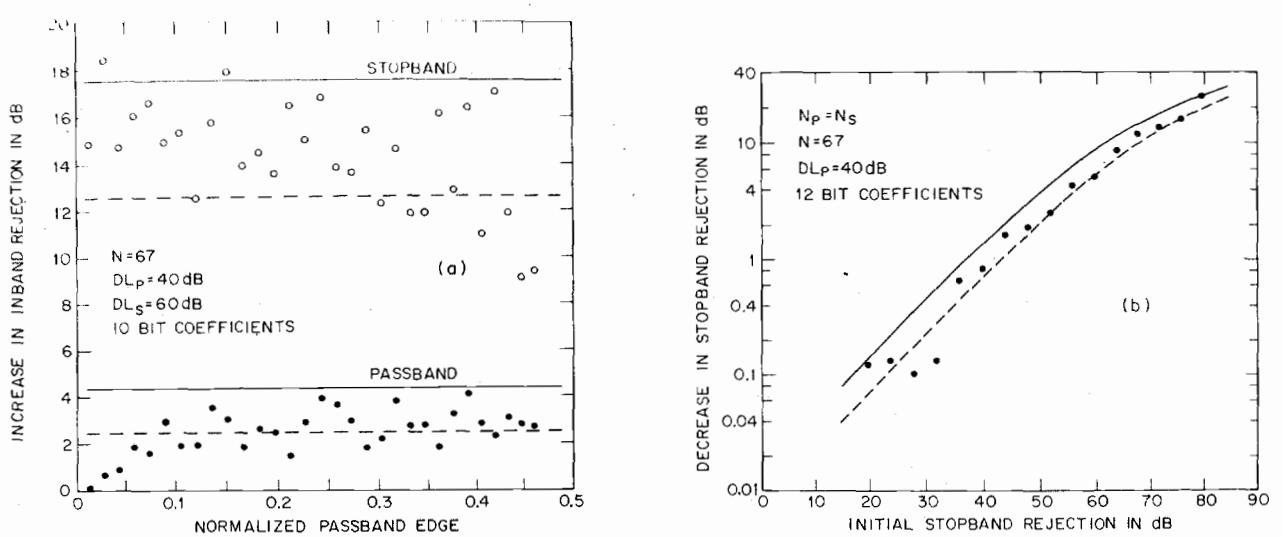


Fig. 7. Comparisons of experimental data to predicted statistical bounds.

To illustrate this point, the results of quantizing filters with values of passband and stopband rejection exactly reversed from those of the filters of Fig. 6(a), (b), viz., with 60 and 40 dB, respectively, are shown in Fig. 6(c). As can be seen, the roles of the passband and stopband are simply exactly interchanged. In many cases of interest, filters often do require greater rejection in the stopband than in the passband; thus in such cases it is true that the passband is less sensitive.

It is also a consequence of the presented analysis that the sensitivity of the response of a filter in any frequency band to coefficient quantization should not depend strongly on the width of the band. In Fig. 7(a), the results of coefficient quantization for low-pass extraripple 67-point filters of passband and stopband rejection of 40 and 60 dB, respectively, and all possible bandwidths (except the two extremes) are shown, together with the statistical bounds. It is seen that, indeed, the data points do not have any strong average dependence on bandwidth. Finally, Fig. 7(b) shows the statistical bound (50) and measured data points for the stopband of several 67-point filters, with 40-dB passband rejection, as a function of stopband rejection. All the results shown in Figs. 6 and 7 clearly demonstrate the validity and usefulness of (40).

VI. Application to Design of Filters with Finite Word Length Coefficients

In Section V, the usefulness of the bound (40) for filter design has been discussed. Specifically, it enables a set of specifications for a filter with coefficients of a given word length to be converted to a set of specifications for the design of an "infinite word length" filter in such a way that when the coefficients of a filter designed to meet the new specifications are quantized, the filter meets the original specifications. In general, this process involves converting a given set of minimum in-band (i.e., referring to both passbands and stopbands) rejections $\{DL^*m_k\}$ into a new set $\{DL_k\}$. Clearly, in any given problem, a filter with the minimum N that will satisfy the given specifications is desired, since both complexity and sensitivity increase with N. If N can be determined *a priori*, the rule for conversion implied by the previous analysis is immediate from (46) as

$$DL_k = -20 \log_{10} \left(10^{-(DL^*m_k/20)} - 2^{-t} \sqrt{\frac{2N-1}{3}} \right)$$
$$= DL^*m_k - 20 \log_{10} \left[1 - 2^{((DL^*m_k/6.02)-t)} \cdot \sqrt{\frac{2N-1}{3}} \right] \quad (51)$$

taking $\log_{10} 2 = 0.301$. However, in general, for a given set of Ω_k, $k = 1, \cdots, M$, the minimum N is a function of DL_k, $k = 1, \cdots, M$. Thus, $\{DL_k\}$ is the solution of a set of M simultaneous nonlinear equations when a solution exists. Unfortunately, the exact functional relationship among N, $\{DL_k\}$, and $\{\Omega_k\}$ is not known for any of the available design methods. However, a useful approximate relation has been found [14] for the case of low-pass optimal filters [4]-[6].

Let DL_p, DL_s, F_p, and F_s denote the passband and stopband rejection, and passband and stopband cutoff frequency (with sampling frequency normalized to unity), respectively, of any low-pass filter. Using data from over 1500 extraripple filters with equal number of ripples in their passband and stopband and various values of N, DL_p, and DL_s, viz., $3 \leq N \leq 127$ and $6 \leq DL_p \leq DL_s \leq 100$ dB, it has been shown that for fixed DL_p and DL_s, the product $(N-1)(F_s - F_p) \triangleq D_N(DL_p, DL_s)$ is essentially a constant independent of N for sufficiently large N. Denoting this limiting value of the product by $D_\infty(DL_p, DL_s)$, it is found to be given to an excellent approximation by

$$D_\infty(DL_p, DL_s) = [-6.64 \times 10^{-7} (DL_p)^2 + 1.78 \times 10^{-4} DL_p + 0.0238] DL_s + [-6.65 \times 10^{-6} (DL_p)^2 + 0.0297 DL_p - 0.4278]. \quad (52)$$

A correction function is then added to improve the approximation for small values of N, with the result that $D_N(DL_p, DL_s)$ can be well approximated for all N by

$$\hat{D}(DL_p, DL_s, F_s - F_p) = D_\infty(DL_p, DL_s) - f(DL_p, DL_s)(F_s - F_p)^2 \quad (53)$$

where

$$f(DL_p, DL_s) = 0.0256(DL_s - DL_p) + 11.012. \quad (54)$$

Thus, since $D_N(DL_p, DL_s) = (N-1)(F_s - F_p)$, the minimum N required for a filter with cutoff frequencies F_p and F_s and in-band rejections at least DL_p and DL_s can be found as

$$N_{\min} = \frac{D_{N_{\min}}(DL_p, DL_s)}{F_s - F_p} + 1 \approx \frac{\hat{D}(DL_p, DL_s, F_s - F_p)}{F_s - F_p} + 1. \quad (55)$$

The error in $N_{\min}$ incurred by substituting the approximate value of $N_{\min}$ given above in place of its actual value was found to be less than 1 for all N, $3 \leq N \leq 127$, and $20 \leq DL_p \leq DL_s \leq 100$ dB.

It has also been shown [14], [15] that given DL_p, DL_s, and N, for all optimal filters, $(F_s - F_p)$ is fairly independent of F_p or F_s if they are not too close to 0 or half the sampling frequency. Thus $D_N(DL_p, DL_s)$ should be fairly independent of the filter cutoff frequencies, so that the approximate equation for $N_{\min}$ given in (54) found for the case of extraripple filters

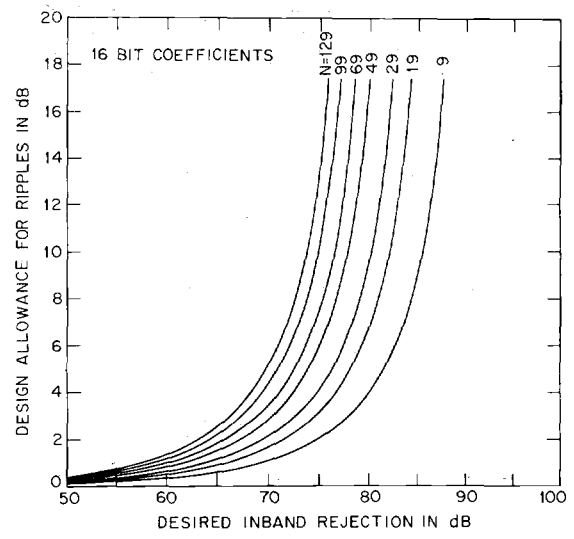


Fig. 8. Design curves for conversion of specifications on in-band rejection.

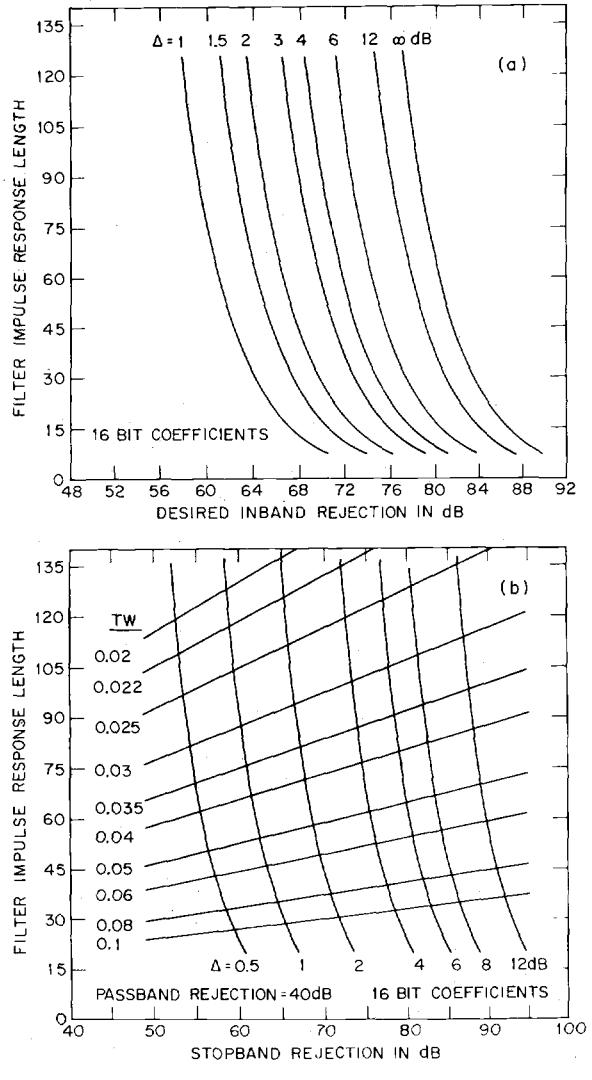


Fig. 9. (a) Bounds on N for design procedure to converge. (b) Design curves for conversion of stopband rejection of filter with 40-dB passband rejection.

with an equal number of passband and stopband ripples should still be fairly good for other values of F_p or F_s.

Although an expression explicit in DL_p and DL_s has been found for $N_{\min}$, substituting it into (51) still leaves that set of equations too complex to be solved in closed form. However, letting DL_p^* and DL_s^* be the desired minimum passband and stopband rejection for a t-bit filter, also denoting the right-hand side of (51) by $g_t(DL^*m_k, N)$, and writing $N_{\min}$ as $N_{\min}(DL_p, DL_s, F_s - F_p)$, the following simple iterative procedure can be used to find the specifications DL_p and DL_s for a t-bit filter.

Step 1:

Set $N_0 \leftarrow N_{\min}(DL_p^*, DL_s^*, F_s - F_p)$
$DL_p \leftarrow g_t(DL_p^*, N_0)$
$DL_s \leftarrow g_t(DL_s^*, N_0)$.

Step 2:

Set $N_1 \leftarrow N_{\min}(DL_p, DL_s, F_s - F_p)$
$\overline{DL}_p \leftarrow g_t(DL_p^*, N_1)$
$\overline{DL}_s \leftarrow g_t(DL_s^*, N_1)$.

Step 3:

If $|\overline{DL}_p - DL_p| > \epsilon$ or $|\overline{DL}_s - DL_s| > \epsilon$, then set $DL_p \leftarrow \overline{DL}_p$, $DL_s \leftarrow \overline{DL}_s$, and return to Step 2; otherwise stop.

The number ϵ is some tolerance level to be chosen.

The function $g_t(x, N) - x$ for $t = 16$ bits is plotted for several values of N in Fig. 8. It is seen that the function is rather insensitive to N for a given x if the combination of x and N is not too large. Thus the algorithm described above will converge rapidly in this region. On the other hand, for a given x, there is a maximum value for N, however large, above which the algorithm cannot converge. This value of N can be found from (51) as

$$N_{\max} = \frac{1}{2} + \frac{3}{2}\left(1 - 10^{-(\Delta/20)}\right)^2 \cdot 2^{2(t-(DL^*m_k/6.02))} \tag{56}$$

where Δ is the maximum allowable change in in-band rejection due to quantization. Equation (56) is plotted in Fig. 9(a) for several values of Δ and $t = 16$ bits. The curve for $\Delta = \infty$ is the upper bound on N required for convergence of the algorithm described. Clearly, the smaller N is compared to this bound, the faster the convergence will be. The curves for other values of Δ are bounds on N that must be satisfied if coefficient quantization is required to incur no more than a change of Δ dB in in-band rejection. Note that both the function $g_t(DL^*m_k, N) - DL^*m_k$ and $N_{\max}$ depend on t and DL^*m_k only through $(t - (DL^*m_k/6.02))$. Thus the labels for the horizontal axes of Figs. 8 and 9(a) simply need to be shifted if plots of

the functions for values of t other than 16 bits are desired.

To further illustrate the procedure described, the important case of a low-pass filter whose stopband rejection is significantly higher than its passband rejection is considered. As an example, $DL_p^* = 40$ dB and $t = 16$ bits is chosen. As can be concluded from Fig. 8, 40 dB of rejection is hardly affected by quantization to 16 bits for at least all $N \leqslant 129$. Thus the passband need not be considered. In Fig. 9(b), (54) is plotted for $DL_p = 40$ dB and various values of $(F_s - F_p)$, the transition bandwidth [denoted by TW in Fig. 9(b)]. Also plotted are curves that describe the functional relationship between N and DL_s for various values of the bound Δ on in-band rejection change due to quantization to 16 bits. Note that as opposed to Fig. 8 or 9(a), the horizontal coordinate of Fig. 9(b) is the rejection for the infinite-bit filter (DL_s) rather than the minimum rejection obtainable after quantization (DL_s^*). For a given DL_s and TW or N, the latter is equal to $DL_s - \Delta$, where Δ is found by observing which curve of constant Δ the given point (DL_s, N) falls on.

Thus to design a filter with $t = 16$ bits, $DL_p^* = 40$ dB, and given F_p, F_s, and DL_s^*, Fig. 9(b) is used as follows. Choose a value of Δ that is desired not to be exceeded and find the bound on N, N_Δ from Fig. 9(a). Find the point with $DL_s = DL_s^*$ on the curve $TW = (F_s - F_p)$ in Fig. 9(b). Move to the right on that curve until either N exceeds N_Δ, in which case the method will not converge to a solution, or else a value of Δ is attained such that $DL_s - \Delta = DL_s^*$. The value of DL_s at this point is the desired value. This method is reasonably accurate as long as F_p and F_s are not too close to 0 or half the sampling frequency. Since the TW curves are only approximate, it may be necessary to try one or more adjacent values of the N obtained if a satisfactory filter is not found.

To conclude this section, two examples are given to enlighten the procedures described.

Example 1: Find a low-pass transfer function with 14-bit coefficients in the direct form and $DL_p^* = 40 \pm 0.5$ dB, $DL_s^* \geqslant 56.5$ dB, $F_p = 0.246$, and $F_s = 0.272$.

First of all, by evaluating $N_{\min}(DL_p^*, DL_s^*, F_s - F_p)$, $N = 95$ is obtained as a first estimate of the minimum N required. Clearly, even if the final N becomes as high as 129, 40-dB rejection is not decreased by more than 0.5 dB by quantizing to 14 bits (see Fig. 8, noting that x-axis labels must be shifted 12 dB to the right to correspond to 14-bit coefficients). Thus one can set $DL_p = 40$ dB without further calculation. To find DL_s, the three-step algorithm is followed. Using $DL_s^* = 56.5$ dB, Step 1 gives $N_0 = 94.7$ and $DL_s = 59.9$ dB. Step 2 then gives $N_1 = 98.4$ and $\overline{DL}_s = 60$ dB. Using $\epsilon = 0.1$ dB in Step 3, the algorithm has already converged. The specifications $N = 99$, $DL_p = 40$ dB, $DL_s = 60$ dB, $F_p = 0.246$, and $F_s = 0.272$ are used to design an optimal infinite-precision transfer function. After quantizing the coefficients of this transfer function in direct form to 14 bits, the passband and stopband rejections are measured to be $DL_p^* = 39.8$ dB and $DL_s^* = 56.7$ dB.

Example 2: Find a low-pass transfer function with 16-bit coefficients in the direct form and $DL_p^* = 40$ dB $\pm$ 0.1 dB, $DL_s^* \geqslant 74$ dB, $F_p = 0.2480$, and $F_s = 0.2955$.

Clearly, $DL_p = 40$ dB can be specified *a priori* with very little possible error. Following the three-step algorithm yields in the end $DL_s = 80$ dB and $N = 67$. An infinite-precision optimal transfer function satisfying these specifications yields, when quantized to 16 bits (direct form), $DL_p^* = 40.0$ dB and $DL_s^* = 75.0$ dB.

In both of the above examples, the final transfer function meets specifications quite tightly. In general, however, much leeway may still be left for possible improvement. In that case, lower values of N can be attempted. However, whether or not any improvement (or how much) is possible cannot, in general, be foreseen.

VIII. Conclusions

A thorough statistical analysis of all three types of quantization effects in the direct form FIR filter has been presented. Resulting from the analysis of filter response errors due to coefficient quantization, a procedure for designing filters with finite word length coefficients has been proposed and tested using on the order of 500 filter examples. The direct form has been shown to be a very attractive structure for realizing FIR filters because it exhibits very low A-D and roundoff noise, and although its response is rather sensitive to the accuracy of its coefficients, this factor is more than adequately compensated for by the minimal number of multipliers it requires. For instance, for a given filter order, the linear phase direct form uses approximately only one-third as many multipliers as required for the cascade form [9], [10]. Thus a linear phase direct structure using 18-bit coefficients is comparable in complexity to a cascade structure using 6-bit coefficients. Clearly, an 18-bit direct form filter can realize a far greater range of transfer functions than can a 6-bit cascade form filter. It remains to be seen exactly how the direct structure compares with other types of structures for realizing FIR filters.

Acknowledgment

The authors would like to acknowledge the helpful comments and criticisms of Dr. J. F. Kaiser on this paper. In particular, Dr. Kaiser noted a simple way to sum the series in (27).

References

[1] L. R. Rabiner, "Techniques for designing finite-duration impulse-response digital filters," *IEEE Trans. Commun. Technol.*, vol. COM-19, pp. 188–195, Apr. 1971.

[2] J. F. Kaiser, "Digital filters," ch. 7 in *Systems Analysis by Digital Computer*, F. F. Kuo and J. F. Kaiser, Ed. New York: Wiley, 1966.

[3] L. R. Rabiner, B. Gold, and C. A. McGonegal, "An approach to the approximation problem for nonrecursive digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-18, pp. 83–106, June 1970.

[4] L. R. Rabiner, "The design of finite impulse response digital filters using linear programming techniques," *Bell Syst. Tech. J.*, vol. 51, pp. 1177–1198, July–Aug. 1972.

[5] T. W. Parks and J. H. McClellan, "Chebyshev approximation for nonrecursive digital filters with linear phase," *IEEE Trans. Circuit Theory*, vol. CT-19, pp. 189–194, Mar. 1972.

[6] E. Hofstetter, A. V. Oppenheim, and J. Siegel, "A new technique for the design of nonrecursive digital filters," in *Proc. 5th Annu. Princeton Conf. Inform. Sci. Syst.*, 1971, pp. 63–72.

[7] L. R. Rabiner and R. W. Schafer, "Recursive and nonrecursive realizations of digital filters designed by frequency sampling techniques," *IEEE Trans. Audio Electroacoust.*, vol. AU-19, pp. 200–207, Sept. 1971.

[8] H. W. Schuessler, "On structures for nonrecursive digital filters," *Arch. Elek. Übertragung*, vol. 26, pp. 255–258, 1972.

[9] D. S. K. Chan and L. R. Rabiner, "Theory of roundoff noise in cascade realizations of finite impulse response digital filters," *Bell Syst. Tech. J.*, vol. 52, pp. 329–245, Mar. 1973.

[10] —, "An algorithm for minimizing roundoff noise in cascade realizations of finite impulse response digital filters," *Bell Syst. Tech. J.*, vol. 52, pp. 347–385, Mar. 1973.

[11] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 2. New York: Wiley, 1968, pp. 491–492.

[12] J. B. Knowles and E. M. Olcayto, "Coefficient accuracy and digital filter response," *IEEE Trans. Circuit Theory*, vol. CT-15, pp. 31–41, Mar. 1968.

[13] O. Herrmann and H. W. Schuessler, "On the accuracy problem in the design of nonrecursive digital filters," *Arch. Elek. Übertragung*, vol. 24, pp. 525–526, 1970.

[14] O. Herrmann, L. R. Rabiner, and D. S. K. Chan, "Practical design rules for optimum finite impulse response lowpass digital filters," *Bell Syst. Tech. J.*, vol. 52, pp. 769–799, July–Aug. 1973.

[15] T. W. Parks, L. R. Rabiner, and J. H. McClellan, "On the transition width of finite impulse response digital filters," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 1–4, Feb. 1973.

Stability Test for Two-Dimensional Recursive Filters

BRIAN D. O. ANDERSON and ELIAHU I. JURY

Abstract—For deciding the stability of a two-dimensional filter, it is of interest to determine whether or not a prescribed polynomial in the variables z_1 and z_2 is nonzero in the region $|z_1| \leqslant 1 \cap |z_2| \leqslant 1$. A new procedure for testing for this property is given, which does not involve the use of bilinear tranformations. Key parts of the test involve the construction of a Schur-Cohn matrix and the checking for positivity on the unit circle of a set of self-inversive polynomials.

I. Introduction

A two-dimensional digital recursive linear filter can be defined by its two-dimensional z-transform

$$G(z_1, z_2) = \frac{\sum_{i=0}^{m} \sum_{j=0}^{n} a_{ij} z_1^i z_2^j}{\sum_{i=0}^{p} \sum_{j=0}^{q} b_{ij} z_1^i z_2^j} \tag{1}$$

where the quantities a_{ij} and b_{ij} are real constants. The filter is stable if and only if (see [1]–[3])

$$B(z_1, z_2) = \sum_{i=0}^{p} \sum_{j=0}^{q} b_{ij} z_1^i z_2^j \neq 0,$$

$$|z_1| \leqslant 1 \cap |z_2| \leqslant 1. \tag{2}$$

The object of this paper is to present a procedure for checking the stability condition; the procedure is believed to be simpler than those of [2] and [3]. In fact, the procedure of [2] is not finite in the sense that the procedure requires the construction of a theoretically infinite number of mappings. The procedure of [3], though finite, requires application of two bilinear transformations to pose the problem in a form solved by Ansell [4]. In essence, Ansell's main contribution is to couple the use of a Hermite test for checking stability [5] with a series of Sturm tests [6] checking positivity.

Our procedure, like that of [3], is finite. We require no bilinear transformation and we replace the Hermite test component of the main part of Ansell's procedure by a Schur-Cohn matrix test [7]–[9]. Then we allow either a series of Sturm tests, or, what turns out to be equivalent, a series of tests for establishing the root distribution of a polynomial.

Manuscript received November 9, 1972. This work was supported in part by the Australian Research Grants Committee.

B. D. O. Anderson is with the Department of Electrical Engineering, University of Newcastle, New South Wales 2308, Australia.

E. I. Jury is with the Department of Electrical Engineering and Computer Science, University of California, Berkeley, Calif. 94720.