5 FINITE WORD LENGTH EFFECTS

5.4 For a two-port adaptor we have:

$$b_1 = a_2 + \alpha(a_2 - a_1)$$
$$b_2 = a_1 + \alpha(a_2 - a_1)$$
$$\alpha = \frac{R_1 - R_2}{R_1 + R_2}$$

The pseudo-power entering into the adaptor is:

$$p = \frac{1}{R_1}(a_1{}^2 - b_1{}^2) + \frac{1}{R_2}(a_2{}^2 - b_2{}^2)$$

Simple, but long and tedious simplification, yields $p = 0$.

5.5 a) In safe scaling, we shall have $S \leq 1$.

$$S = \sum_{n=-\infty}^{\infty} h(n) = a_0 + a_1 + a_2 + a_1 + a_0$$

The new scaled values are: a_0/S, a_1/S, and a_2/S.

b) In L_2-norm scaling , we shall have $S \leq 1$, where

$$S = \sum_{n=-\infty}^{\infty} h(n)^2 = a_0{}^2 + a_1{}^2 + a_2{}^2 + a_1{}^2 + a_0{}^2$$

Let $c = \sqrt{S}$. The new values are: a_0/c, a_1/c, and a_2/c.

5.7 Rank the different scaling criteria in terms of severity. Most severe is: Safe scaling, then L_∞, and finally, L_2.

5.8 a) $$\left\| \, |F(e^{j\omega T})|^2 \right\|_p = \sqrt[p]{\frac{1}{2\pi}\int_0^{\pi} \left| \, |F(e^{j\omega T})|^2 \right|^p d\omega T} =$$

$$= \sqrt[p]{\frac{1}{2\pi}\int_0^{\pi} |F(e^{j\omega T})|^{2p} d\omega T} =$$

$$= \left(\sqrt[2p]{\frac{1}{2\pi}\int_0^{\pi} |F(e^{j\omega T})|^{2p} d\omega T} \right)^2 = \| F(e^{j\omega T}) \|_{2p}^2$$

b) We have:

$$\| F(e^{j\omega T})\, G(e^{j\omega T}) \|_p = \sqrt[p]{\frac{1}{2\pi}\int_{-\pi}^{\pi} |F(e^{j\omega T})\, G(e^{j\omega T})|^p d\omega T} \leq$$

$$\leq \sqrt[p]{\frac{1}{2\pi}\int_{-\pi}^{\pi} |F(e^{j\omega T})|_{\max}{}^{p}\ |G(e^{j\omega T})|^{p}\, d\omega T} \quad \leq$$

$$\leq |F(e^{j\omega T})|_{\max} \sqrt[p]{\frac{1}{2\pi}\int_{-\pi}^{\pi} |G(e^{j\omega T})|^{p}\, d\omega T} \quad \leq$$

$$\leq \left\|F(e^{j\omega T})\right\|_{\infty} \left\|G(e^{j\omega T})\right\|_{p} \quad , \quad p > 1$$

Hence, $\left\| F(e^{j\omega T})\, G(e^{j\omega T}) \right\|_{p} \leq \left\|F(e^{j\omega T})\right\|_{\infty} \left\|G(e^{j\omega T})\right\|_{p}$, p > 1

5.9 The input $x(n)$ is assumed to be scaled. Hence, the inputs to the multipliers with coefficients a is already scaled. Now, scale the next critical node, i.e., the inputs to the multipliers. In this case, the output node.

a) L_2-norm scaling. Determine the impulse response, by first determining the transfer function.

$$H(z) = \frac{a(z-1)}{z-b} = \frac{a}{1-b\,z^{-1}} - \frac{a\,z^{-1}}{1-b\,z^{-1}} = a \sum_{n=0}^{\infty} (b\,z^{-1})^{n} - \frac{a}{b} \sum_{n=1}^{\infty} (b\,z^{-1})^{n}$$

or

$$h(n) = \begin{cases} 0 & \text{for n} < 0 \\ a & \text{for n} = 0 \\ a(1-1/b)\,b^{n} & \text{for n} > 0 \end{cases}$$

We get:

$$S = \sum_{n=-\infty}^{\infty} h(n)^2 = \frac{2\,a^2}{1+b} \quad \text{which shall be} = 1 \ \Rightarrow a = \sqrt{\frac{1+b}{2}}$$

b) L_∞-norm scaling. Generally, it is difficult to find the maximal value of the magnitude function by analytical methods. However, in this simple case, (highpass filter) we have max for $z = -1$. Hence,

$$|H(e^{j\omega T})| = \frac{2\,a}{1+b} \quad \text{should be set to } 1 \Rightarrow a = \frac{1+b}{2}$$

5.10 According to Jackson's lower bounds, Eqs.(5.20)–(5.21), the variance of round-off noise at the output is bounded from below by a measure of the sensitivity. Hence, the sensitivity is low. Therefore, this filter can satisfy stringent requirements with short coefficient word length.

5.11 The input is assumed to be scaled. The first critical node is after the first adder. The impulse response to this node is:

$0.4, -0.4, 0, \ldots \Rightarrow S = 0.8$

⇒ Increase the coefficients by a factor $1/S$ to 0.5 and –0.5, respectively. The impulse response to the output node is (with the two first coefficients scaled):

$(0.5 \cdot 0.75), (0.5\,(-0.75)) + (-0.5)\,0.75, (-0.5)(-0.75), 0, 0 \ldots =$

$= 0.375, -0.75, 0.375, 0, 0 \Rightarrow S = 1.5$

⇒ Decrease the coefficients by multiplying with a factor $1/S$. We obtain the new values 0.5 and –0.5, respectively. The new scaled impulse response becomes:

$(0.5 \cdot 0.5),\ (0.5\,(-0.5)) + (-0.5)\,0.5, (-0.5)(-0.5), 0\ldots =$

$= 0.25, -0.5, 0.25, 0\ldots$

Using L_∞-norm scaling we get the transfer function to the first critical noise node: $H(z) = 0.4(1 - z^{-1})$. The maximal value of the magnitude function is gotten for $z = -1$. Hence, $|H|_{max} = 0.8$. Thus, increase the first two coefficients to 0.5. The transfer function to the output node is:

$$H(z) = 0.5(1 - z^{-1})\, 0.75(1 - z^{-1})$$

The maximal value of the magnitude function is occur for $z = -1$. Hence, $|H|_{max} = 1.5$. Thus, decrease the last two coefficients to 0.5. In this case, the two scaling criteria leads to the same coefficients.

5.12 a) The node $v(n)$ must be scaled since it is input (after the delay element) to the multiplier with non-integer coefficient. Insert a scale coefficient, c, in front of the filter. The impulse response from the input to the node is:

$$h_v(n) = \begin{cases} 0 & , n < 0 \\ c\, b^n & , n \geq 0 \end{cases}$$

We get: $$S_v = \sum_{n=0}^{\infty} h_v(n)^2 = \sum_{n=0}^{\infty} c^2\, b^{2n} = \frac{c^2}{1 - b^2} = 25.2525c^2$$

Now, let $S_v = 1 \Rightarrow c = \dfrac{1}{\sqrt{25.2525}} = 0.198997$

The impulse response from the input to the output is:

$$h(n) = a_0\, h_v(n) + a_1 h_v(n-1) = \begin{cases} 0 & , n < 0 \\ a_0\, c & , n = 0 \\ a_0\, c\, b^n + a_1\, c\, b^{n-1} & , n \geq 1 \end{cases}$$

$$S_y = \sum_{n=0}^{\infty} h(n)^2 = (a_0c)^2 + \sum_{n=1}^{\infty} (a_0c)^2 [b^n - b^{n-1}]^2 =$$

$$= (a_0c)^2[1 + \frac{(b-1)^2}{b^2} \sum_{n=1}^{\infty} b^{2n}] = (a_0c)^2[1 + \frac{(b-1)^2}{b^2} \frac{b^{\,2}}{1-b^2})] =$$

$$= (a_0c)^2[1 + \frac{1-b}{1+b}] = a_0^2\ 0.02$$

Let $S_y = 1 \Rightarrow a_0 = \frac{1}{\sqrt{0.02}} = 7.07107$

b) All of the coefficients are non-integers. Hence, there are two noise sources to each "adder". The contribution from c and b is:

$$\sigma_{y1}^2 = 2\ \sigma_0^2 \sum_{n=0}^{\infty} h_{noise}(n)^2 = 2\ \sigma_0^2 \sum_{n=0}^{\infty} (\frac{h(n)}{c})^2 =$$

$$\sigma_{y1}^2 = \frac{2\ \sigma_0^2}{c^2} \sum_{n=0}^{\infty} h(n)^2 = \frac{2\ \sigma_0^2}{c^2}$$

Since the filter is scaled.

$$\sigma_{y1}^2 = \frac{2\ \sigma_0^2}{c^2} = 100.5026\ \frac{Q^2}{12}$$

The contribution from σ_0 and σ_1 is: $\sigma_{y2}^2 = 2\ \frac{Q^2}{12}$

$$\sigma_y^2 = \sigma_{y1}^2 + \sigma_{y2}^2 = 102.5026\ \frac{Q^2}{12} \text{ where } Q = 2^{-7}$$

5.13 a) We have for a white stochastic sequence $x(n)$ (which is a wide-band signal): $E\{x(n)\} = 0$ and the autocorrelation function is: $r(nT) = \sigma_x^2\ \delta(n)$. Hence, we have:

$$S_x(e^{j\omega T}) = \sum_{n=-\infty}^{\infty} r(nT)\ e^{-j\omega nT} = \sigma_x^2$$

and

$$\|\ S_x\ \|_{\infty} = max\{\,|\,S_x(e^{j\omega T})\,|\,\} = \sigma_x^2$$

b) For a narrow-band signal, for example, a sinusoidal we have: $\|\ S_x\ \|_1 = A$ where A is the amplitude.

5.17 a) The critical overflow node that shall be scaled in direct form I is the output node. The transfer function to this node is:

$$H_{Critical}(z) = H_{AP}(z) = \frac{z^2 + a\ z + b}{b\ z^2 + a\ z + 1}$$

The L_p-norm of the magnitude function is:

$$\| H_{Critical}(e^{j\omega T}) \|_p = \sqrt[p]{\frac{1}{2\pi}\int_{-\pi}^{\pi} |H_{AP}(e^{j\omega T})|^p \, d\omega T} =$$

$$= \sqrt[p]{\frac{1}{2\pi}\int_{-\pi}^{\pi} 1^p \, d\omega T} = 1$$

Hence, all of the outputs are properly scaled and their scaling is independent of the L_p-norm. Of course, this is only valid for direct form I. For example, the direct form II structure is not automatically scaled.

b) The ordering of the sections does not effect the signal range since the transfer function from the input of the allpass filter to each of the critical nodes are allpass functions. Also this fact is a special case that is valid only for direct form I.

5.18 a) The variance of the quantization noise for System #1 is:

$\sigma_1^2 = Q_1^2/12$ where $Q_1 = 2^{-17}$

The variance of the quantization noise for System #2 is:

$\sigma_2^2 = Q_2^2/12$ where $Q_2 = 2^{-13}$

We assume that the average values are = 0. The power spectrum after the A/D converter is:

$S_2 = \sigma_2^2$

since the noise is white. A lowpass filter with cutoff angle $\omega_c T$ is placed after the converter in order to remove noise that should be aliased into the passband after the decimation. The variance of the noise after the lowpass filter is:

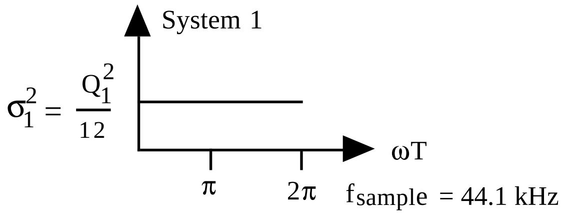
$$\sigma_3^2 = r(0) = \frac{1}{2\pi}\int_0^{2\pi} S_2 \, e^{j0\omega T} \, d\omega T = \frac{1}{2\pi}\int_0^{2\pi} S_2 \, d\omega T =$$

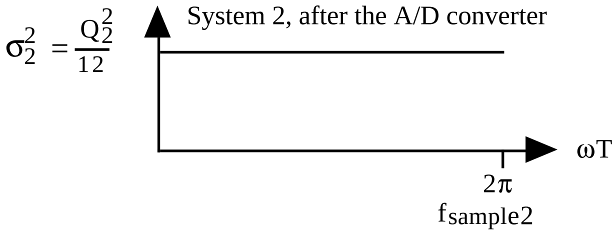
$$= \frac{2}{2\pi}\int_0^{\omega_c T} \sigma_2^2 \, d\omega T = \frac{2}{2\pi} \sigma_2^2 \frac{2\pi f_{sample}}{2 f_{sample2}} = \sigma_2^2 \frac{f_{sample}}{f_{sample2}}$$

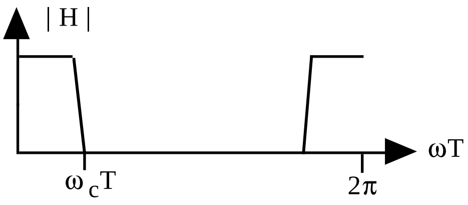
($r(\tau)$ is the autocorrelation function). We assume that the noise has zero mean. We must have:

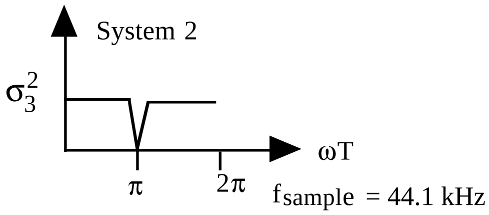
$$\omega_c T \geq \frac{2\pi f_{sample}}{2 f_{sample2}}$$

so that the audio signal will not be attenuated by the lowpass filter. If the quantization noise is to be less then or equal to the noise in System #1, we must have:









$$\sigma_3{}^2 \le \sigma_1{}^2$$

Hence, $\frac{Q_2{}^2}{12}\frac{f_{sample}}{f_{sample2}} \le \frac{Q_1{}^2}{12}$ and

$$f_{sample2} \ge \frac{Q_2{}^2}{Q_1{}^2} f_{sample} = 2^8 f_{sample};$$

$$f_{sample2} \ge 256 \cdot 44.1 \text{ kHz} = 11.264 \text{ MHz}$$

b) Decimation with a factor 1024 yields a sample frequency of:

$$f_{sample2} \ge 1024 \cdot 44.1 \text{ kHz} = 45.1584 \text{ MHz}$$

Decimation can be done in steps of 2 and the decimation filters can work at the lower sample frequency. The filter orders are almost the same for all stages. We need 10 decimation stages. The workload for one stage is only half of that of the preceding stage. The workload of the whole decimation filter, counted from the output, is:

$$N[2^0 + 2^1 + 2^2 + 2^3 + \ \dots + 2^9] = N\, 2^9 \frac{1 - 2^{-10}}{1 - 2^{-1}} \approx N\, 2^{10} = 1024\, N$$

where N is the workload of a single decimation filter working at the lowest sample frequency.

c) System #1: $\frac{f_{sample} - f_c}{f_c} = \frac{44.1 - 20}{20} = \frac{24.1}{20} = 1.205$

System #2: $\frac{f_{sample2} - f_c}{f_c} = \frac{11264 - 20}{20} = \frac{11224}{20} = 561.2$

d) We assume that the acceptable ripple in the passband is about 0.1 dB and the required stopband attenuation is about 50 dB in both cases.

We get:

$N_{System1} =$

$N_{System2} =$

5.19 Let $E(e^{j\omega T})$ be the error function caused by rounding the coefficients. We have:

$$E(e^{j\omega T}) = e^{-j\omega(K+1)T}\,[\delta h_0 + 2 \sum_{n=1}^{K} \delta h_n \, cos(\omega n T)]$$

but $|\,\delta h_n\,| \le Q_c/2$. The deviation can be estimated in may ways, for example, the maximal deviation or the variance of the deviation. Here we use an estimate of the maximal deviation:

$$|E(e^{j\omega T})| \ = \ |e^{-j\omega(K+1)T}\,[\delta h_0 + 2 \sum_{n=1}^{K} \delta h_n \, cos(\omega n T)]| \le$$

$$\le |\,\delta h_0 + 2 \sum_{n=1}^{K} \delta h_n \, cos(\omega n T)| \le |\,\delta h_0| \ + 2 \sum_{n=1}^{K} |\,\delta h_n \, cos(\omega n T)| \le$$

$$|E(e^{j\omega T})| \le |\,\delta h_0| \ + 2 \sum_{n=1}^{K} |\,\delta h_n|$$

Hence, $|E(e^{j\omega T})| \le \frac{Q}{2}(1 + 2\,K) = \frac{N\,Q}{2}$ (Bound #1)

Now, assume that the coefficients are randomly rounded and the errors are considered as independent random variables that are uniformly distributed in the interval $[-Q/2, Q/2]$. The variance is $Q^2/12$. Let e_0 be the effective value of $E(e^{j\omega T})$ in the passband:

$$e_0{}^2 = \frac{1}{f_c} \int_0^{f_c} |E(e^{j\omega T})|^2 d\omega T = \sum_{n=0}^{N-1} |\,\delta h_0|^2$$

The variance of $e_0{}^2$ is:

$$\sigma^2 = E\{\, e_0{}^2\} = \frac{N\,Q^2}{12} \qquad \text{(Bound \#2)}$$

The required coefficient word length is estimated as follows. Let δ_m be the acceptable deviation in the passband of the stopband. Now, we must have:

$$|E(e^{j\omega T})| < \delta_m - \delta_0$$

where δ_0 is the deviation before quantization of the coefficients. Let the level of acceptance of a coefficient set that does not fit the requirements be 5%. Hence the probability the coefficient set does not meet the specification is (assuming a normal distribution):

$$P(\,|E(e^{j\omega T})| \geq 2\sigma_x) = P(\frac{|E(e^{j\omega T})|}{\sigma_x} \geq 2) \approx 0.05$$

$$2[1 - \Phi(\frac{|E(e^{j\omega T})|}{\sigma_x})] = 0.05 \qquad \Rightarrow \qquad \frac{|E(e^{j\omega T})|}{\sigma_x} \approx 2$$

$$|E(e^{j\omega T})| \approx 2\sigma_x \qquad \Rightarrow \qquad 2\sigma_x \approx \delta_m - \delta_0$$

$$\sqrt{\frac{N\,Q^2}{12}} \approx \frac{\delta_m - \delta_0}{2} \quad \text{and } Q \approx (\delta_m - \delta_0)\sqrt{\frac{3}{N}}$$

The number of bits that is required to represent the coefficient depends on the largest coefficient. Hence.

$Q = 2^{(1-W_c)}\,[max\{\,[h_n\,|\,\} = 2^{(1-W_c)}h_0 \approx 2^{(1-W_c)}\frac{f_s + f_c}{f_{sample}}$ for a lowpass filter. Hence,

$$1\text{–}W_c \approx log_2\{\frac{f_{sample}}{f_s + f_c}\,Q\,\} \approx log_2\{\frac{f_{sample}}{f_s + f_c}\,(\delta_m - \delta_0)\sqrt{\frac{3}{N}}\,\}$$

$$W_c \approx 1 - log_2\{\frac{f_{sample}}{f_s + f_c}\,(\delta_m - \delta_0)\sqrt{\frac{3}{N}}\,\}$$

For most lowpass filter we have: $\frac{N}{3} \leq \frac{f_s - f_c}{f_{sample}}$.

$$W_c \geq 1 - log_2\{\frac{f_{sample}}{(f_s + f_c)}\sqrt{\frac{f_{sample}}{f_s - f_c}}\,(\delta_m - \delta_0)\}$$

In practice, we may select $\delta_m = 2\,Min\{\,\delta_1, \delta_2\} = 2\delta_0$

$$W_c \geq 1 - log_2 \left\{ \frac{f_{sample}}{(f_s + f_c)} \sqrt{\frac{f_{sample}}{f_s - f_c}} \; \frac{2}{\delta_m} \right\}$$

$$W_c \geq 1 - log_2 \left\{ \frac{f_{sample}}{(f_s + f_c)} \sqrt{\frac{f_{sample}}{f_s - f_c}} \right\} + log_2 \left\{ Min\{ \delta_1, \delta_2 \} \right\}$$

See also: Niedringshaus W.P., Steglitz K., and Kodek D.: An Easily Computed Performance Bound for Finite Wordlength Direct-Form FIR Digital Filters, IEEE Trans. on Circuits and Systems, Vol. CAS-29, No. 3, pp. 191-193, March 1982.