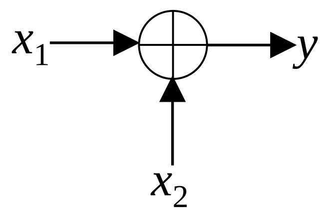
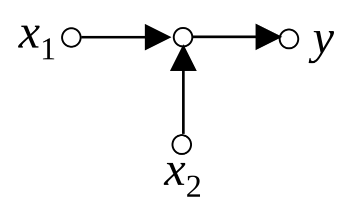
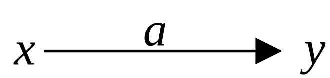
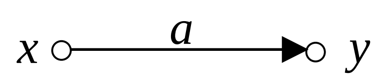
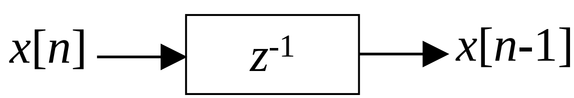
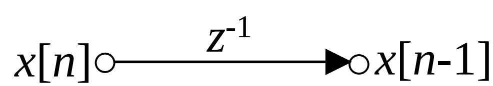
Structures For Discrete-Time Systems

Carrson C. Fung
Institute of Electronics
National Yang Ming Chiao Tung University

Background

- Two branches in the study of filter design
  - Design of the frequency response (actual values of $h[n]$) – discussed in the next chapter
  - Design of the structure
    - Structure discusses the realization/implementation of digital filters/systems
    - Different structures have different advantages and disadvantages
      - Number of delay elements
      - Robustness toward quantization noise

Block Diagram and Signal Flow Graph

	Block diagram	Signal flow graph
Adder		
Scalar (mult. by a constant)		
Unit delay		

- *Nodes* and *branches* are keys in a signal flow graph

  Source node: No entering branches

  Sink node: Only entering branches

Basic Structures for IIR Systems: Direct Form I (Block Diagram)

$$H(z)=\frac{\sum_{k=0}^{M} b_k z^{-k}}{1-\sum_{k=1}^{N} a_k z^{-k}} \Leftrightarrow y[n]-\sum_{k=1}^{N} a_k y[n-k]=\sum_{k=0}^{M} b_k x[n-k]$$

$$H(z)=H_2(z)H_1(z)=\frac{Y(z)}{X(z)} \quad =H_2(z)V(z)\frac{H_1(z)}{V(z)}$$

$$Y(z)=H_2(z)V(z)\leftrightarrow y[n]-\sum_{k=1}^{N} a_k y(n-k)=v[n]$$

$$X(z)=\frac{V(z)}{H_1(z)}$$

$$V(z)=H_1(z)X(z)\leftrightarrow v[n]=\sum_{k=0}^{M} b_k x[n-k]$$

$$V(z)=\sum_{k=0}^{M} b_k X(k)z^{-k}$$

$$Y(z)-\sum_{k=1}^{N} a_k Y(z)z^{-k}=V(z)$$

$$\Rightarrow Y(z)=V(z)+\sum_{k=1}^{N} a_k Y(z)z^{-k}$$

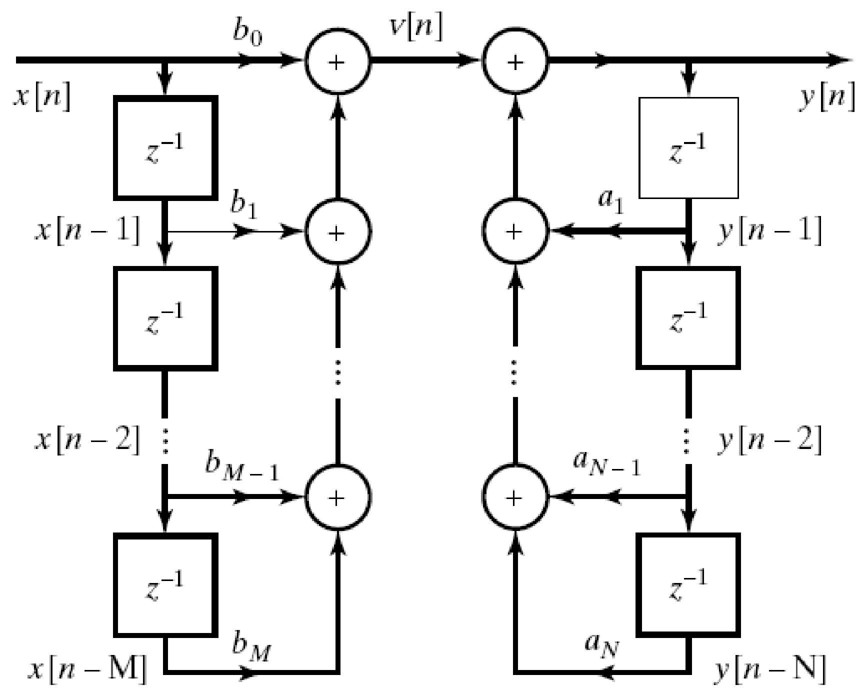


Figure 6.3 Block diagram representation for a general *N*th-order difference equation.

Structures for IIR Systems: Direct Form I (Signal Flow Graph)

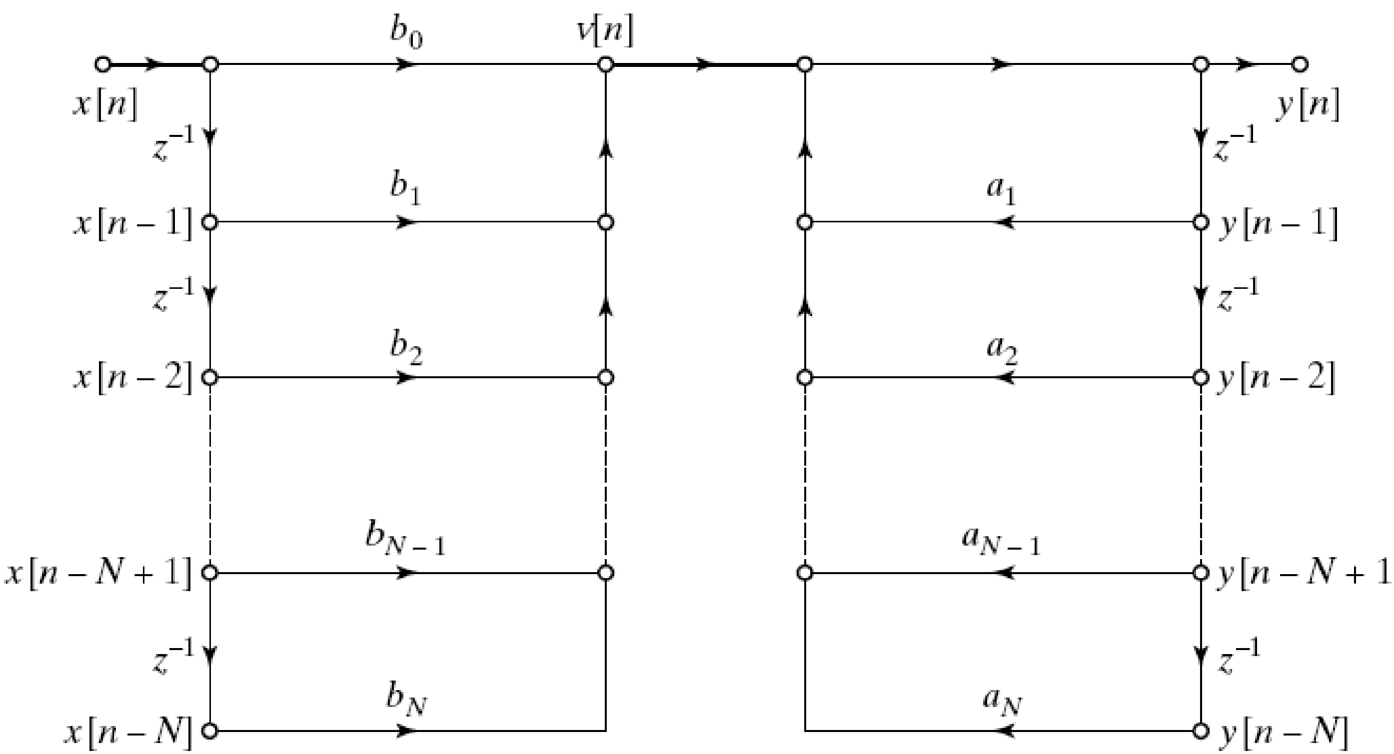


Figure 6.14 Signal flow graph of direct form I structure for an Nth-order system.

Structures for IIR Systems: Direct Form II (Block Diagram)

- Interchange 1st and 2nd segments and merge the delay lines (z^{-1})
- Number of delay = max(N,M)

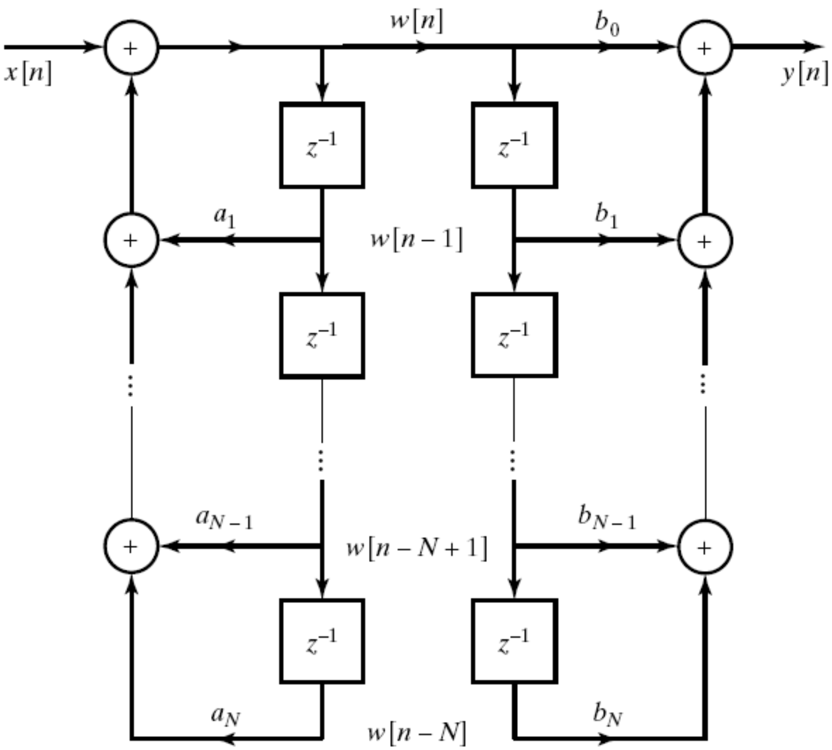


Figure 6.4 Rearrangement of block diagram of Figure 6.3. We assume for convenience that $N = M$. If $N \neq M$, some of the coefficients will be zero.

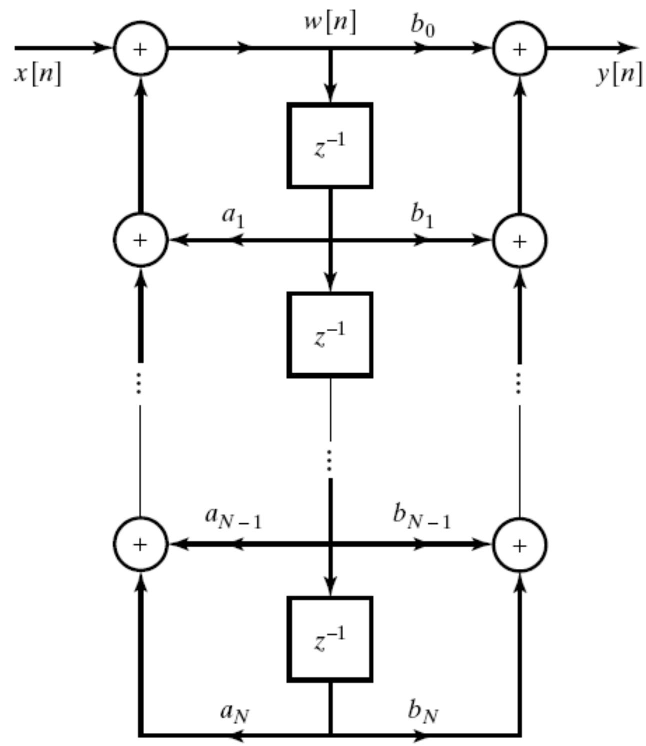


Figure 6.5 Combination of delays in Figure 6.4.

Structures for IIR Systems: Direct Form II (Signal Flow Graph)

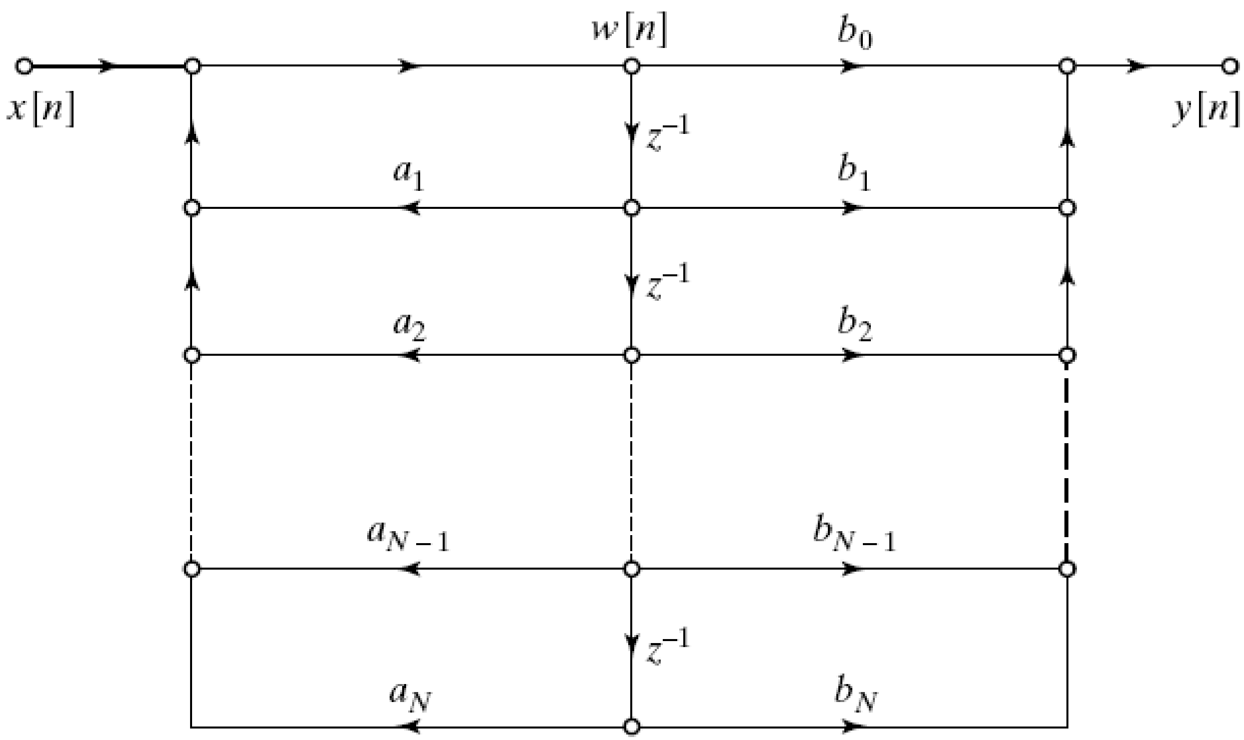


Figure 6.15 Signal flow graph of direct form II structure for an Nth-order system.

Cascade Form

Serial connection of 1^{st} order and 2^{nd} order factors

$$H(z)=\prod_{k=1}^{N_s}\frac{b_{ok}+b_{1k}z^{-1}+b_{2k}z^{-2}}{1-a_{1k}z^{-1}-a_{2k}z^{-2}}$$

Remark: Each factor is a Direct Form II $(2N_s = N)$

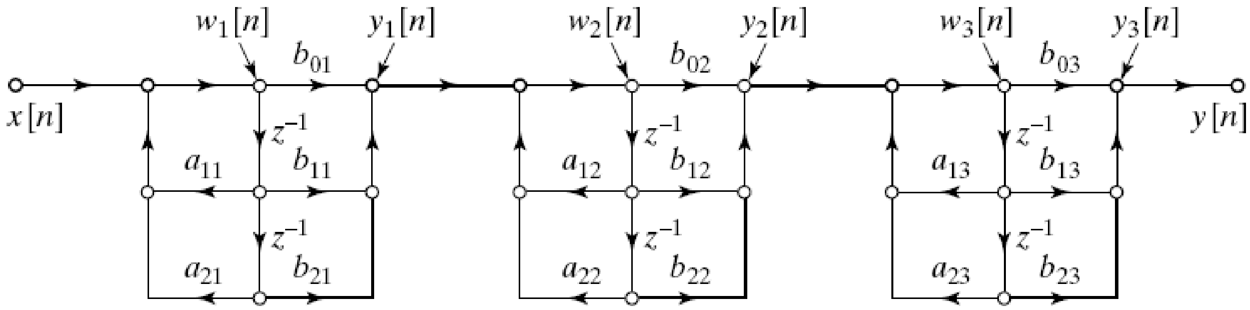


Figure 6.18 Cascade structure for a sixth-order system with a direct form II realization of each second-order subsystem.

If there are N_s second-order sections, there are $N_s!$ pairings of the poles with zeros and $N_s!$ orderings of the resulting 2^{nd}-order sections

$\Rightarrow$ $(N_s!)^2$ different pairings and orderings

- $\prod_{k=1}^{N_s}\frac{b_{ok}+b_{1k}z^{-1}+b_{2k}z^{-2}}{1-a_{1k}z^{-1}-a_{2k}z^{-2}}$ needs 5 constant multipliers for each section
- $b_0\prod_{k=1}^{N_s}\frac{1+\tilde{b}_{1k}z^{-1}+\tilde{b}_{2k}z^{-2}}{1-a_{1k}z^{-1}-a_{2k}z^{-2}}$ needs 4 constant multipliers for each section, where $\tilde{b}_{1k}=\frac{b_{ik}}{b_{0k}}$

Parallel Form

Parallel connection of 1st and 2nd order factors $\left(N_s = \left\lfloor (N+1)/2 \right\rfloor\right)$

$$H(z) = \sum_{k=0}^{N_p} C_k z^{-k} + \sum_{k=1}^{N_s} \frac{e_{ok} + e_{1k} z^{-1}}{1 - a_{1k} z^{-1} - a_{2k} z^{-2}}$$

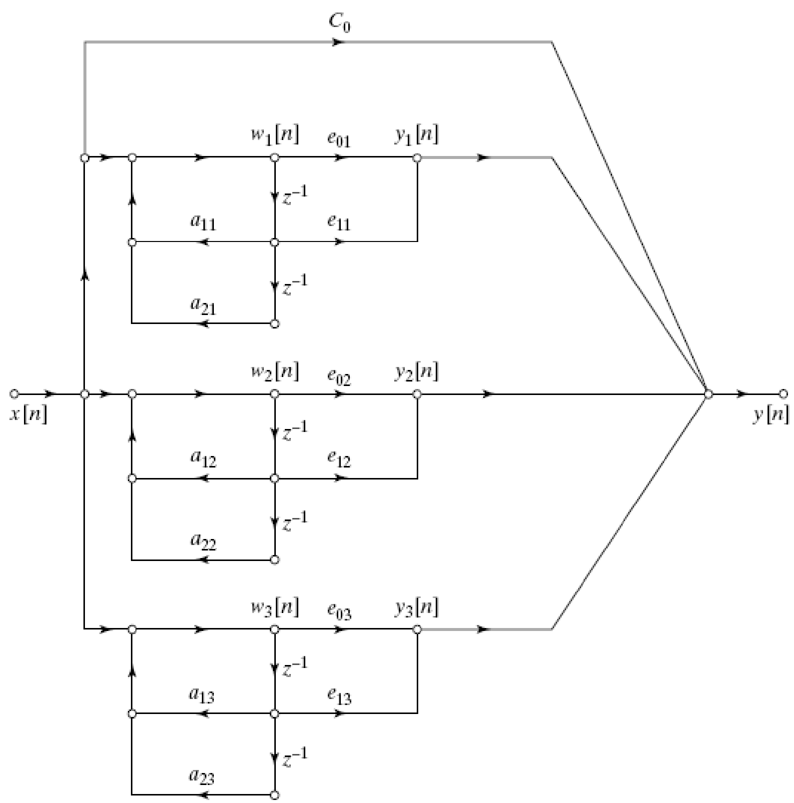
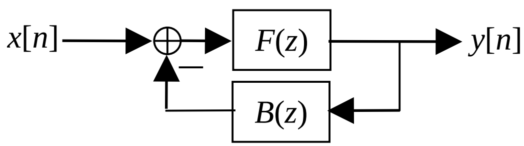


Figure 6.20 Parallel-form structure for sixth-order system ($M = N = 6$) with the real and complex poles grouped in pairs.

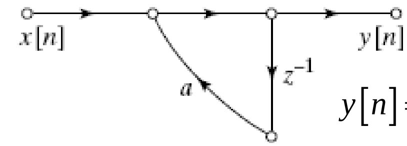
Feedback In IIR Systems

- Basic formula of a feedback system with negative feedback

$$H(z) = \frac{F(z)}{1 + F(z)B(z)}$$

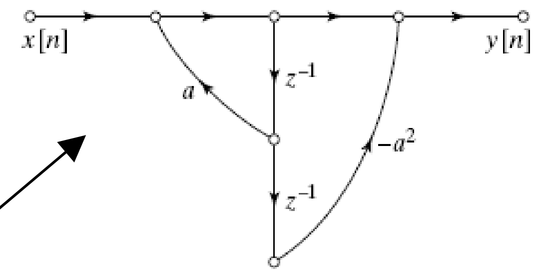


- If a system has non-zero poles, a corresponding block diagram or signal flow graph will have feedback loops
  - But neither poles in the system function nor loops in the network are sufficient for the impulse response to be infinitely long because zero/pole cancellation may occur
- A delay element is necessary in the feedback loop, otherwise, it is noncomputable.



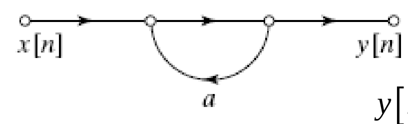
$$y[n] = ay[n-1] + x[n]$$

(a)



(b)

E.g. $H(z) = \frac{1 + (a_0 + a_0^*)z^{-1} + |a_0 a_0^*|^2 z^{-2}}{1 + a_0 z^{-1}} = \frac{(1 + a_0 z^{-1})(1 + a_0^* z^{-1})}{1 + a_0 z^{-1}} = (1 + a_0^* z^{-1})$



$$y[n] = ay[n] + x[n]$$

(c)

Figure 6.23 (a) System with feedback loop. (b) FIR system with feedback loop. (c) Noncomputable system.

Transpose Forms

- *Transposition* of a flow graph is reversing the *directions* of *all* branches in the network while keeping the branch transmittances (as they were) and reversing the roles of the input and output (so that source nodes become sink nodes and vice versa).
- Flow Graph Reversal Theorem
  - For single-input, single-output systems, the transposed flow graph has the same system function as the original graph if the input nodes and output nodes are interchanged.

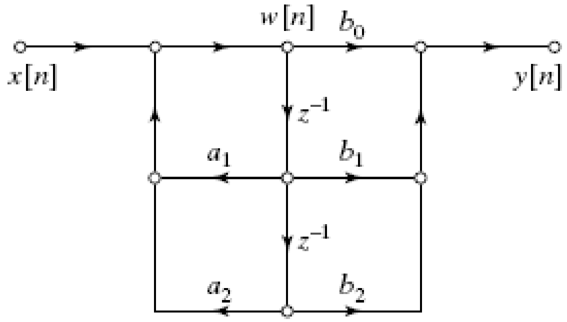


Figure 6.27 Direct form II structure for Example 6.8.

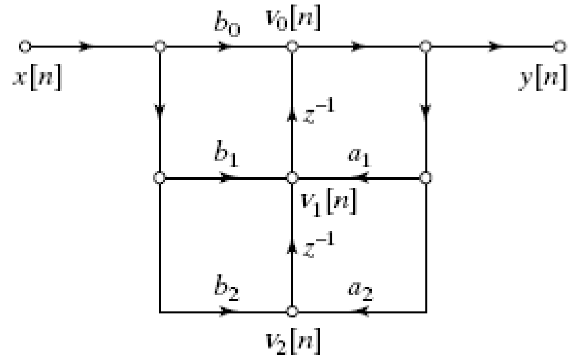


Figure 6.28 Transposed direct form II structure for Example 6.8.

Basic Structures for FIR Systems: Direct Form

Transversal filter or tapped delay line

$$y[n]=\sum_{k=0}^{M} b_k x[n-k] \quad (M \text{ is the order})$$

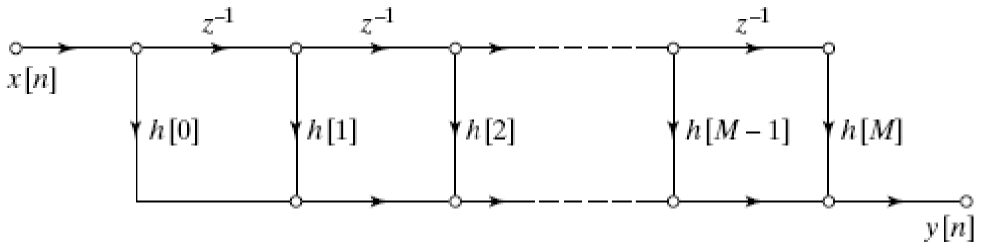


Figure 6.31 Direct-form realization of an FIR system.

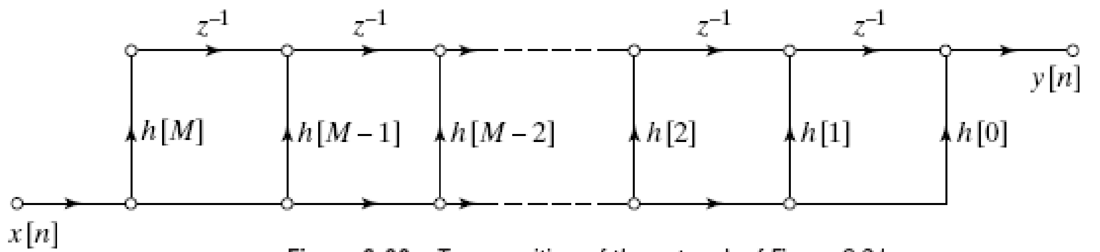


Figure 6.32 Transposition of the network of Figure 6.31.

Basic Structures for FIR Systems: Cascade Form

Serial connection of 1^{st} and 2^{nd} order factors

$$H(z)=\prod_{k=1}^{M_s}\left(b_{ok}+b_{1k}z^{-1}+b_{2k}z^{-2}\right)$$ (Each factor is a Direct Form)

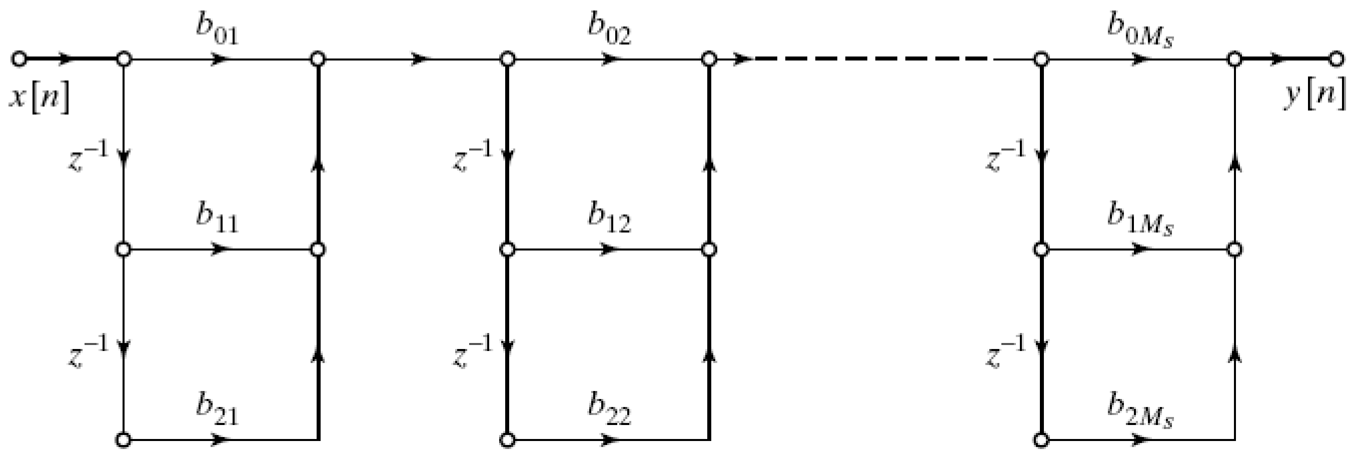


Figure 6.33 Cascade-form realization of an FIR system.

Linear Phase FIR Filters

Take advantage of the symmetry property of the impulse response

$$h[M-n]=h[n]$$

$$h[M-n]=-h[n]$$

$$y[n]=\sum_{k=0}^{M}h[k]x[n-k] \quad (M \text{ is the order of the filter})$$

$$\sum_{p=0}^{\frac{M}{2}-1}h[M-p]x[n-M+p]$$

$$=\sum_{p=0}^{\frac{M}{2}-1}h[k]x[n-M+p]$$

Type I or III: M is even

$$y[n]=\sum_{k=0}^{M/2-1}h[k]x[n-k]+h\left[\frac{M}{2}\right]x\left[n-\frac{M}{2}\right]+\sum_{k=\frac{M}{2}+1}^{M}h[k]x[n-k]$$

$$k=M-p \Rightarrow p=M-k$$

$$k=\frac{M}{2}+1 \Rightarrow p=M-\frac{M}{2}-1=\frac{M}{2}-1$$

$$k=M \Rightarrow p=0$$

$$=\sum_{k=0}^{\frac{M}{2}-1}h[k]\left(x[n-k]\pm x[n-M+k]\right)+h\left[\frac{M}{2}\right]x\left[n-\frac{M}{2}\right]$$

(for Type-I: $+x[n-M+k]$, for Type-III: $-x[n-M+k]$)

Type-II or IV: M is odd

$$y[n]=\sum_{k=0}^{\frac{M-1}{2}}h[k]\left(x[n-k]\pm x[n-M+k]\right)$$

(for Type-II: $+x[n-M+k]$, for Type-IV: $-x[n-M+k]$)

Linear Phase FIR Filters

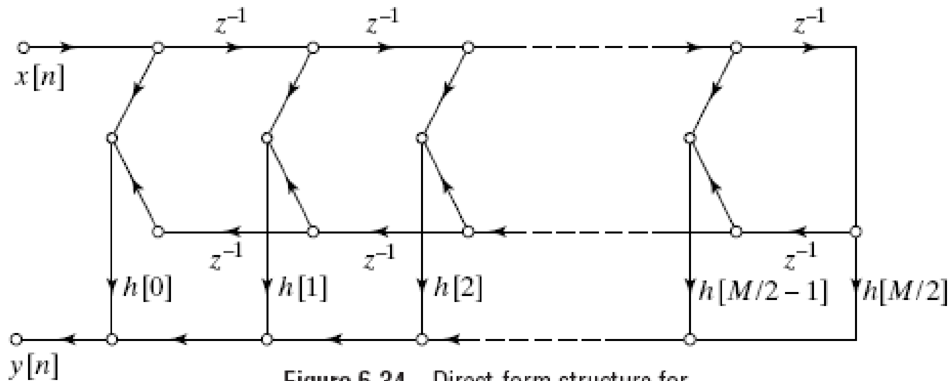


Figure 6.34 Direct-form structure for an FIR linear-phase system when M is an even integer.

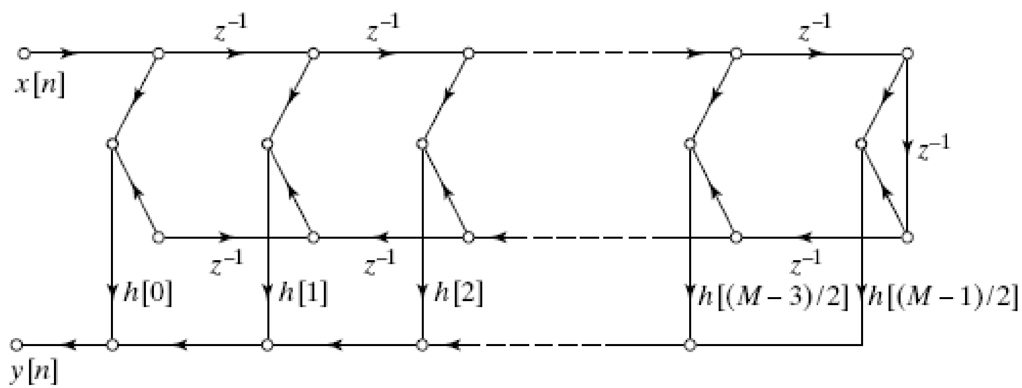


Figure 6.35 Direct-form structure for an FIR linear-phase system when M is an odd integer.

Linear Phase FIR Filters

- Instead of M multipliers, we can reduce the number of computations by taking advantage of the linear phase property (note that we do not include multiplication by -1 as this can be implemented by a flipping bits, e.g. flipping signed bit for one's complement)
  - Type-I: only need $M/2+1$ multipliers
  - Type-II: only $(M+1)/2$ multipliers
  - Type-III: only $M/2$ multipliers
  - Type-IV: only $(M+1)/2$ multipliers
- Linear-phase FIR filters can also be implemented as a cascade of 1st-order, 2nd-order, and 4th-order real-coefficient systems. (The 4th-order system is formed by grouping the conjugate and the reciprocal zeros together.)

Linear Phase FIR Filters

- Linear-phase FIR filters can also be implemented as a cascade of 1st-order, 2nd-order, and 4th-order real-coefficient systems. (The 4th-order system is formed by grouping the conjugate and the conjugate reciprocal zeros together.)

$$H(z) = h[0]\left(1+z^{-1}\right)\left(1+az^{-1}+z^{-2}\right)\left(1+bz^{-1}+z^{-2}\right)\left(1+cz^{-1}+dz^{-2}+cz^{-3}+z^{-4}\right)$$

$$a = z_2 + \frac{1}{z_2},\quad b = 2\,\mathrm{Re}\{z_3\},\quad c = -2\,\mathrm{Re}\left\{z_1 + \frac{1}{z_1}\right\},\quad d = 2 + \left|z_1 + \frac{1}{z_1}\right|^2$$

- This representation suggests a cascade structure consisting of linear-phase elements. The order of the system is M=9, and number of different coefficient multipliers is (M+1)/2 = 5 (for Type-II and IV). This is the same number as seen in Figure 6.34 when Direct Form I is used.

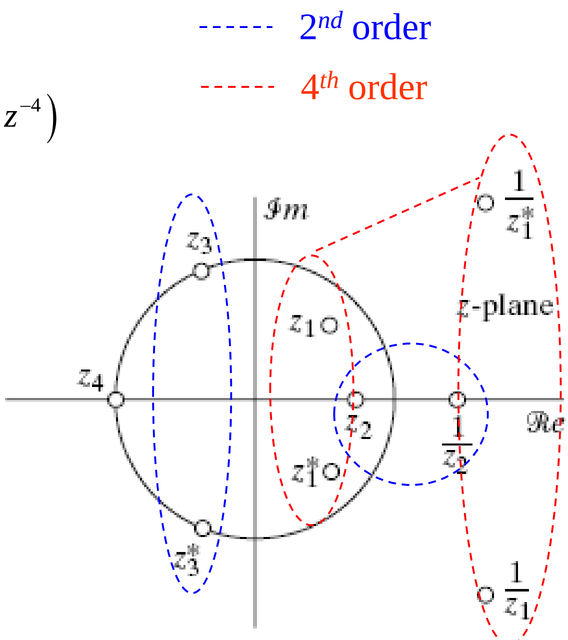


Figure 6.36 Symmetry of zeros for a linear-phase FIR filter.

Effects of Coefficient Quantization (IIR Systems)

- Effects depend on the filter structure
- Want to demonstrate the sensitivity of the filter frequency response characteristics to quantization of the filter coefficients is minimized by realizing a filter having a large number of poles and zeros as an interconnection of 2^{nd}-order filter sections.
- For Direct Form:

$$H(z) = \frac{\sum_{k=0}^{M} b_k z^{-k}}{\underbrace{1 - \sum_{k=1}^{N} a_k z^{-k}}_{A(z)}} \rightarrow \hat{H}(z) = \frac{\sum_{k=0}^{M} \hat{b}_k z^{-k}}{1 - \sum_{k=1}^{N} \hat{a}_k z^{-k}}$$

Note: $\hat{a}_k = a_k + \Delta a_k; \quad \hat{b}_k = b_k + \Delta b_k$

  - Note that quantization error in a given coefficient affects *all* the poles of the system function.
  - Suppose that the poles are all first order and they are located at $z = z_i$, *for* $i = 1, 2, \ldots, N$.

Effect On Pole Locations

- This will affect the frequency response and stability
- Consider the denominator of an unquantized and quantized rational system function

$$\begin{cases} A(z) = 1 - \sum_{k=1}^{N} a_k z^{-k} = \prod_{j=1}^{N} \left(1 - z_j z^{-1}\right) \\ \hat{A}(z) = 1 - \sum_{k=1}^{N} \hat{a}_k z^{-k} = \prod_{j=1}^{N} \left(1 - \hat{z}_j z^{-1}\right) \end{cases}$$

where $\hat{z}_i = z_i + \Delta z_i, \quad i = 1, \ldots, N$

- Δz_i is the error or perturbation of the i^{th} pole resulting from the quantization of the filter coefficients. The perturbation error Δz_i can be expressed in terms of the errors in the coefficient as

$$\Delta z_i = \sum_{k=1}^{N} \left(\frac{\partial z_i}{\partial a_k} \right) \Delta a_k, \qquad i = 1, 2, \ldots, N,$$

where $\partial z_i / \partial a_k$ represents the incremental change in the pole z_i due to a change in the coefficient a_k. In other words, $\partial z_i / \partial a_k$ is the sensitivity of the pole location to quantization of a_k. The total error Δz_i is expressed as a sum of the incremental errors due to changes in each of the coefficients $\{a_k\}$.

Remark: This formula is approximately true when Δa_k and Δz_k are small.

Effects of Pole Locations

The partial derivative $\partial z_i/\partial a_k$, $k = 1, 2, \ldots, N$, can be obtained by differentiating $A(z)$ with respect to each of the $\{a_k\}$.

$$\left(\frac{\partial A(z)}{\partial a_k}\right)_{z=z_i} = \left(\frac{\partial A(z)}{\partial z}\right)_{z=z_i} \left(\frac{\partial z_i}{\partial a_k}\right)$$

Then

$$\frac{\partial z_i}{\partial a_k} = \frac{\left(\partial A(z)/\partial a_k\right)_{z=z_i}}{\left(\partial A(z)/\partial z\right)_{z=z_i}}$$

The numerator is

$$\left(\frac{\partial A(z)}{\partial a_k}\right)_{z=z_i} = -z^{-k}\Big|_{z=z_i} = -z_i^{-k}$$

The denominator is

$$\left(\frac{\partial A(z)}{\partial z}\right)_{z=z_i} = \left\{\frac{\partial}{\partial z}\left[\prod_{j=1}^{N}\left(1-z_j z^{-1}\right)\right]\right\}_{z=z_i} = \left\{\sum_{k=1}^{N}\frac{z_k}{z^2}\prod_{\substack{j=1\\ j\neq i}}^{N}\left(1-z_j z^{-1}\right)\right\}_{z=z_i}$$

$$= \frac{1}{z_i^N}\prod_{\substack{j=1\\ j\neq i}}^{N}\left(z_i - z_j\right)$$

Effects On Pole Locations

Therefore, $\partial z_i / \partial a_k$ is equal to

$$\frac{\partial z_i}{\partial a_k} = \frac{-z_i^{-N-k}}{\prod_{\substack{j=1 \\ j \neq i}}^{N} \left(z_i - z_j \right)}$$

Therefore, the total perturbation is

$$\Delta z_i = -\sum_{k=1}^{N} \frac{z_i^{-N-k}}{\prod_{\substack{j=1 \\ j \neq i}}^{N} \left(z_i - z_j \right)} \Delta a_k$$

- That is, if $(z_i - z_j)$ is small, then $\partial z_i / \partial a_k$ is large, which contributes to large errors and hence a large perturbation error Δz_i results. For example, in narrowband filter, the lengths $|z_i - z_j|$ are small for poles in the vicinity of z_i. This problem can be alleviated if we implement high-order filter with either single-pole (and single-zero) filter sections. To avoid complex-valued arithmetic, the complex poles/zeros are combined with their conjugates to form 2nd-order sections. These conjugates pairs are usually sufficiently far apart, thus, $\{\Delta z_i\}$ is minimized.
- *Remark:* The preceding analysis can be derived for the sensitivity of the zeros to the quantization errors in b_k's.

Mitigation Effects of Quantization Noise: Parallel and Cascade Forms

- Mainly consists of 1^{st} and 2^{nd} order sections to avoid complex-valued multiplications
- Errors in a particular pole pairs (section) are independent of the other poles (sections). This is also true for zeros in cascade form. ➔ In general, both the *cascade form* and the *parallel form* are less sensitive to coefficient quantization (because zeros are often widely distributed the unit circle).

Example: Bandpass IIR Elliptic Filter

$$0.99 \le \left|H(e^{j\omega})\right| \le 1.01 \qquad 0.3\pi \le \omega \le 0.4\pi$$

$$\left|H(e^{j\omega})\right| \le 0.01 \quad \text{(-40dB)} \qquad \omega \le 0.29\pi$$

$$\left|H(e^{j\omega})\right| \le 0.01 \quad \text{(-40dB)} \qquad 0.41\pi \le \omega \le \pi$$

- Coefficients in Table 6.1 were computed with 32-bit floating point (defined as "unquantized"). The table gives the coefficients of the six 2nd order sections, i.e.

$$H(z) = \prod_{k=1}^{N_s} \frac{b_{0k} + b_{1k} z^{-1} + b_{2k} z^{-2}}{1 - a_{1k} z^{-1} - a_{2k} z^{-2}}, \quad \text{where } N_s = \left\lfloor \frac{N+1}{2} \right\rfloor,$$

  and N is the total number of poles
- To show the effects of quantization, Table 6.2 shows the coefficients in Table 6.1 quantized using 16-bit accuracy using fixed-point representation. The fixed-point coefficients are shown as a decimal integer times a power of 2 scale factor. The binary representation would be obtained by converting the decimal integer to a binary number. The scale factor would be represented only implicitly in the data shifts that would be necessary to line up the binary points of products prior to addition to other products. Note that the binary points of the coefficients are not all in the same location.

Example: Bandpass IIR Elliptic Filter

- For example, all the coefficients with scale factor 2^{-15} have their binary points as

$$b_0 \diamond b_1\, b_2\, b_3 \ldots b_B,$$

where B+1 is the total number of bits, b_0 is the sign bit, and b_1 is the highest fractional bit. This is commonly known as Q15 format. Converting coefficient a_{11} in Table 6.2 to binary, we have (b_0 here is implied to be 0)

$$24196 \Rightarrow 0.101111010000100$$

where the decimal point is on the left side of leftmost '1'. Thus, representing this as a decimal value

$$\left(1\times 2^{-1}\right)+\left(1\times 2^{-3}\right)+\left(1\times 2^{-4}\right)+\left(1\times 2^{-5}\right)+\left(1\times 2^{-6}\right)+\left(1\times 2^{-8}\right)+\left(1\times 2^{-13}\right)=0.7384$$

For coefficients in Table 6.1 whose values do not exceed 0.5, we can represent them using the Q16 format, i.e. scale factor is 2^{-16}, where the binary points is located in front of b_0, i.e.

$$\diamond b_0\; b_1 b_2\, b_3 \ldots b_B.$$

For example, the integer part (and the binary equivalent) of coefficient b_{02} in Table 6.2 is

$$18278 \Rightarrow .0100011101100110.$$

Thus, representing this as a decimal value

$$\left(1\times 2^{-2}\right)+\left(1\times 2^{-6}\right)+\left(1\times 2^{-7}\right)+\left(1\times 2^{-8}\right)+\left(1\times 2^{-10}\right)+\left(1\times 2^{-11}\right)+\left(1\times 2^{-14}\right)+\left(1\times 2^{-15}\right)=0.2789$$

Example: Bandpass IIR Elliptic Filter

- But for coefficients whose value exceeds 1 but less than 2, then we must move the decimal point one place to the right, i.e. between b_1 and b_2

$$b_0\, b_1 \lozenge\, b_2\, b_3 \ldots b_B,$$

  For example, a_{16} the integer part (and the binary equivalent) of coefficient b_{02} in Table 6.2 is

$$19220 \Rightarrow 01.00101100010100.$$

  Thus, representing this as a decimal value

$$\left(1\times 2^{0}\right)+\left(1\times 2^{-3}\right)+\left(1\times 2^{-5}\right)+\left(1\times 2^{-6}\right)+\left(1\times 2^{-10}\right)+\left(1\times 2^{-12}\right)=1.1731$$

- Use of different binary locations retains greater accuracy in the coefficients, but it complicates the programming or system architecture

Example: Bandpass IIR Elliptic Filter

- We can see that if Direct Form I structure is used, as in Figure 6.47e, quantization has destroyed the frequency response of the system since it is more sensitive to quantization error than either cascade form (Figure 6.47c) or parallel form (Figure 6.47d).

TABLE 6.1 UNQUANTIZED CASCADE-FORM COEFFICIENTS FOR A 12TH-ORDER ELLIPTIC FILTER

k	a_{1k}	a_{2k}	b_{0k}	b_{1k}	b_{2k}
1	0.738409	−0.850835	0.135843	0.026265	0.135843
2	0.960374	−0.860000	0.278901	−0.444500	0.278901
3	0.629449	−0.931460	0.535773	−0.249249	0.535773
4	1.116458	−0.940429	0.697447	−0.891543	0.697447
5	0.605182	−0.983693	0.773093	−0.425920	0.773093
6	1.173078	−0.986166	0.917937	−1.122226	0.917937

One's complement (flip all bits)

$$b_{12} = -29131 \Rightarrow .1000111000110101$$

$$-1 + 2^{-1} + 2^{-5} + 2^{-6} + 2^{-7} + 2^{-11} + 2^{-12} + 2^{-14} + 2^{-16} = -0.4445037842$$

TABLE 6.2 SIXTEEN-BIT QUANTIZED CASCADE-FORM COEFFICIENTS FOR A 12TH-ORDER ELLIPTIC FILTER

k	a_{1k}	a_{2k}	b_{0k}	b_{1k}	b_{2k}
1	24196×2^{-15}	-27880×2^{-15}	17805×2^{-17}	3443×2^{-17}	17805×2^{-17}
2	31470×2^{-15}	-28180×2^{-15}	18278×2^{-16}	-29131×2^{-16}	18278×2^{-16}
3	20626×2^{-15}	-30522×2^{-15}	17556×2^{-15}	-8167×2^{-15}	17556×2^{-15}
4	18292×2^{-14}	-30816×2^{-15}	22854×2^{-15}	-29214×2^{-15}	22854×2^{-15}
5	19831×2^{-15}	-32234×2^{-15}	25333×2^{-15}	-13957×2^{-15}	25333×2^{-15}
6	19220×2^{-14}	-32315×2^{-15}	15039×2^{-14}	-18387×2^{-14}	15039×2^{-14}

Hidden zero after decimal

One's complement (flip all bits)

$$b_{16} = -18387 \Rightarrow 10.11100000101101$$

$$-2 + 2^{-1} + 2^{-2} + 2^{-3} + 2^{-9} + 2^{-11} + 2^{-12} + 2^{-14} = -1.122253418$$

unquantized (log)

unquantized cascade

Cascade structure (16-bit)

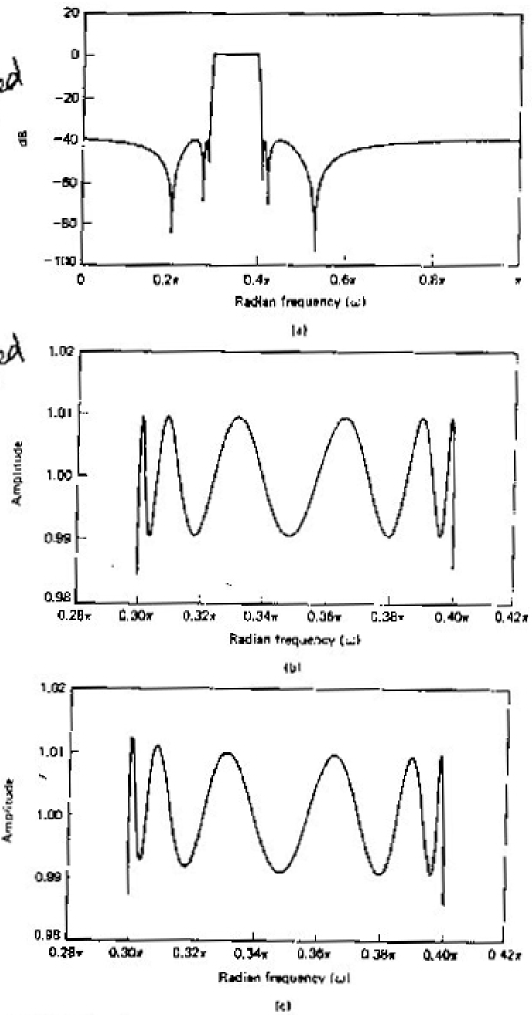


Figure 6.47 IIR coefficient quantization example. (a) Log magnitude for unquantized elliptic bandpass filter. (b) Passband for unquantized cascade case. (c) Passband for cascade structure with 16-bit coefficients

parallel structure (16-bit)

direct form (16-bit) (log)

normalized lattice (16-bit)

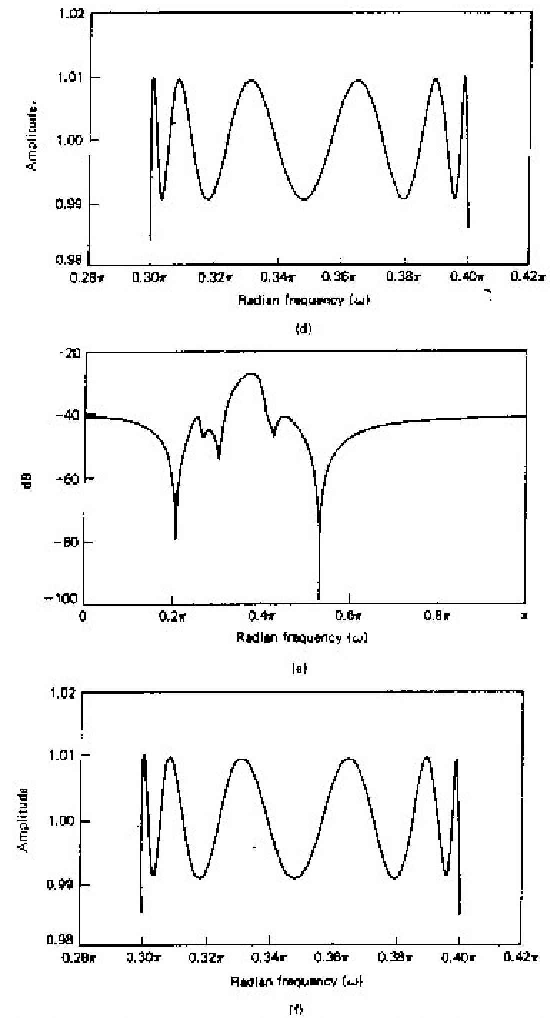


Figure 6.47 *(continued)* (d) Passband for parallel structure with 16-bit coefficients. (e) Log magnitude for direct form with 16-bit coefficients. (f) Passband for normalized lattice with 16-bit coefficients.

Coefficient Quantization in FIR Systems

$$H(z)=\sum_{n=0}^{M} h[n]z^{-n} \rightarrow \hat{H}(z)=\sum \hat{h}[n]z^{-n}$$

$$= H(z)+\underbrace{\Delta H(z)}_{\sum \Delta h[n]z^{-n}}, \text{ where } \Delta H(z)=\sum_{n=0}^{M}\Delta h[n]z^{-n},$$

and $\{\hat{h}[n]=h[n]+\Delta h[n]\}$ is a new set of coefficients obtained if the coefficients $h[n]$ has been quantized and $\{\Delta h[n]\}$ are the quantization error samples of $\{h[n]\}$

- Effect on the zero locations
  - The sensitivity function of this form is similar to that of the direct form I IIR filter. That is, if the zeros are tightly clustered, their locations will then be highly sensitive to quantization errors. However, for most linear phase FIR systems, the zeros are more or less uniformly spread on the z-plane.

Effect on $H(e^{j\omega})$

- After scaling (all coefficients < 1), each $h[n]$ is represented by $(B+1)$ bits 2's complement number; i.e. $-2^{-(B+1)} < \Delta h[n] \leq 2^{-(B+1)}$

$$\Delta H\left(e^{j\omega}\right) = \sum_{n=0}^{M} \Delta h[n] e^{-j\omega n}$$

$$\left|\Delta H\left(e^{j\omega}\right)\right| = \left|\sum_{n=0}^{M} \Delta h[n] e^{-j\omega n}\right| \leq \sum_{n=0}^{M} \left|\Delta h[n]\right|\left|e^{-j\omega n}\right|$$

$$\underbrace{\leq (M+1)2^{-(B+1)}}_{worst\ \ case!}$$

- This gives us the (pessimistic) error bounds for different quantized systems. For M+1 = 28: 16-bit: 0.000427; 14-bit: 0.001709; 13-bit: 0.003418; 8-bit: 0.109375.
- It's pessimistic because if we look at Figures 6.46c – 6.46f, we will see that the approximation error is always less than what the bound provides. To reach the bound, the quantization errors would have to be of the same sign and equal to the maximum value
- Note that there will no effect on the linear phase property as long as $\hat{h}[n] = \hat{h}[M-n]$

Example: Linear Phase Lowpass Filter

$$0.99 \le \left|H(e^{jw})\right| \le 1.01 \qquad 0 \le \omega \le 0.4\pi$$

$$\left|H(e^{jw})\right| \le 0.001 \quad (-60\text{dB}) \qquad 0.6\pi \le \omega \le \pi$$

TABLE 6.3 UNQUANTIZED AND QUANTIZED COEFFICIENTS FOR AN OPTIMUM FIR LOWPASS FILTER ($M = 27$)

Coefficient	Unquantized	16 bits	14 bits	13 bits	8 bits
$h[0] = h[27]$	1.359657×10^{-3}	45×2^{-15}	11×2^{-13}	6×2^{-12}	0×2^{-7}
$h[1] = h[26]$	-1.616993×10^{-3}	-53×2^{-15}	-13×2^{-13}	-7×2^{-12}	0×2^{-7}
$h[2] = h[25]$	-7.738032×10^{-3}	-254×2^{-15}	-63×2^{-13}	-32×2^{-12}	-1×2^{-7}
$h[3] = h[24]$	-2.686841×10^{-3}	-88×2^{-15}	-22×2^{-13}	-11×2^{-12}	0×2^{-7}
$h[4] = h[23]$	1.255246×10^{-2}	411×2^{-15}	103×2^{-13}	51×2^{-12}	2×2^{-7}
$h[5] = h[22]$	6.591530×10^{-3}	216×2^{-15}	54×2^{-13}	27×2^{-12}	1×2^{-7}
$h[6] = h[21]$	-2.217952×10^{-2}	-727×2^{-15}	-182×2^{-13}	-91×2^{-12}	-3×2^{-7}
$h[7] = h[20]$	-1.524663×10^{-2}	-500×2^{-15}	-125×2^{-13}	-62×2^{-12}	-2×2^{-7}
$h[8] = h[19]$	3.720668×10^{-2}	1219×2^{-15}	305×2^{-13}	152×2^{-12}	5×2^{-7}
$h[9] = h[18]$	3.233332×10^{-2}	1059×2^{-15}	265×2^{-13}	132×2^{-12}	4×2^{-7}
$h[10] = h[17]$	-6.537057×10^{-2}	-2142×2^{-15}	-536×2^{-13}	-268×2^{-12}	-8×2^{-7}
$h[11] = h[16]$	-7.528754×10^{-2}	-2467×2^{-15}	-617×2^{-13}	-308×2^{-12}	-10×2^{-7}
$h[12] = h[15]$	1.560970×10^{-1}	5115×2^{-15}	1279×2^{-13}	639×2^{-12}	20×2^{-7}
$h[13] = h[14]$	4.394094×10^{-1}	14399×2^{-15}	3600×2^{-13}	1800×2^{-12}	56×2^{-7}

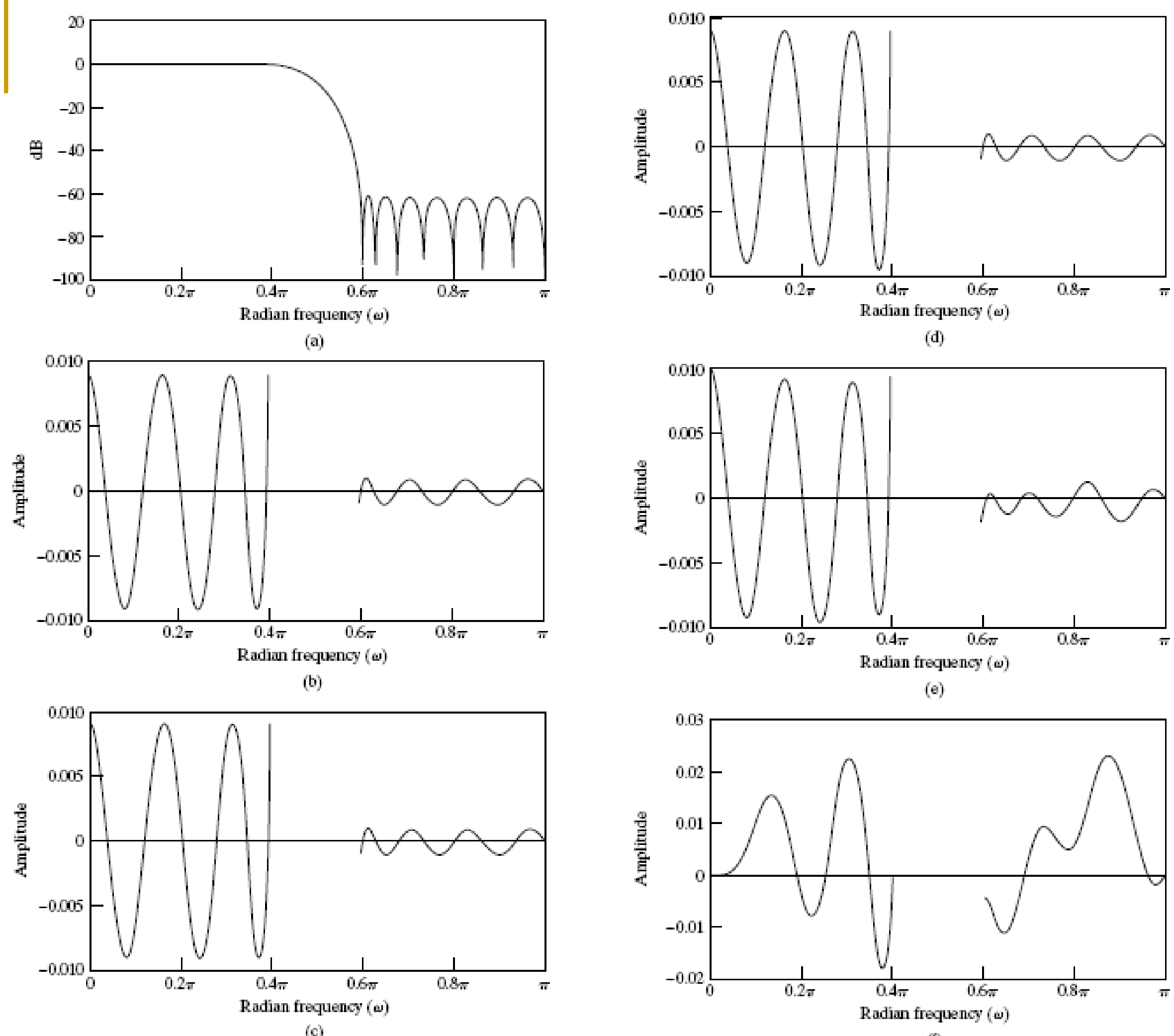


Figure 6.46 FIR quantization example. (a) Log magnitude for unquantized case. (b) Approximation error for unquantized case. (Error not defined in transition band.) (c) Approximation error for 16-bit quantization.

Figure 6.46 *(continued)* (d) Approximation error for 14-bit quantization. (e) Approximation error for 13-bit quantization. (f) Approximation error for 8-bit quantization.

From Figure 6.46, we see that using 13-bit quantization, the stopband approximation error becomes larger than -60dB. And using 8-bit quantization, the stopband approximation error becomes 10 times as large as what is specified. So, we need at least 14-bit coefficients in order to meet the design specifications using Direct Form I. This is also reflected in the pole-zero diagram in Figure 6.47. The zero location is hardly affected in the 16-bit case. However, the zeros location have moved significantly in the 13-bit case (and lower), thus affected the frequency response in the passband and stopband

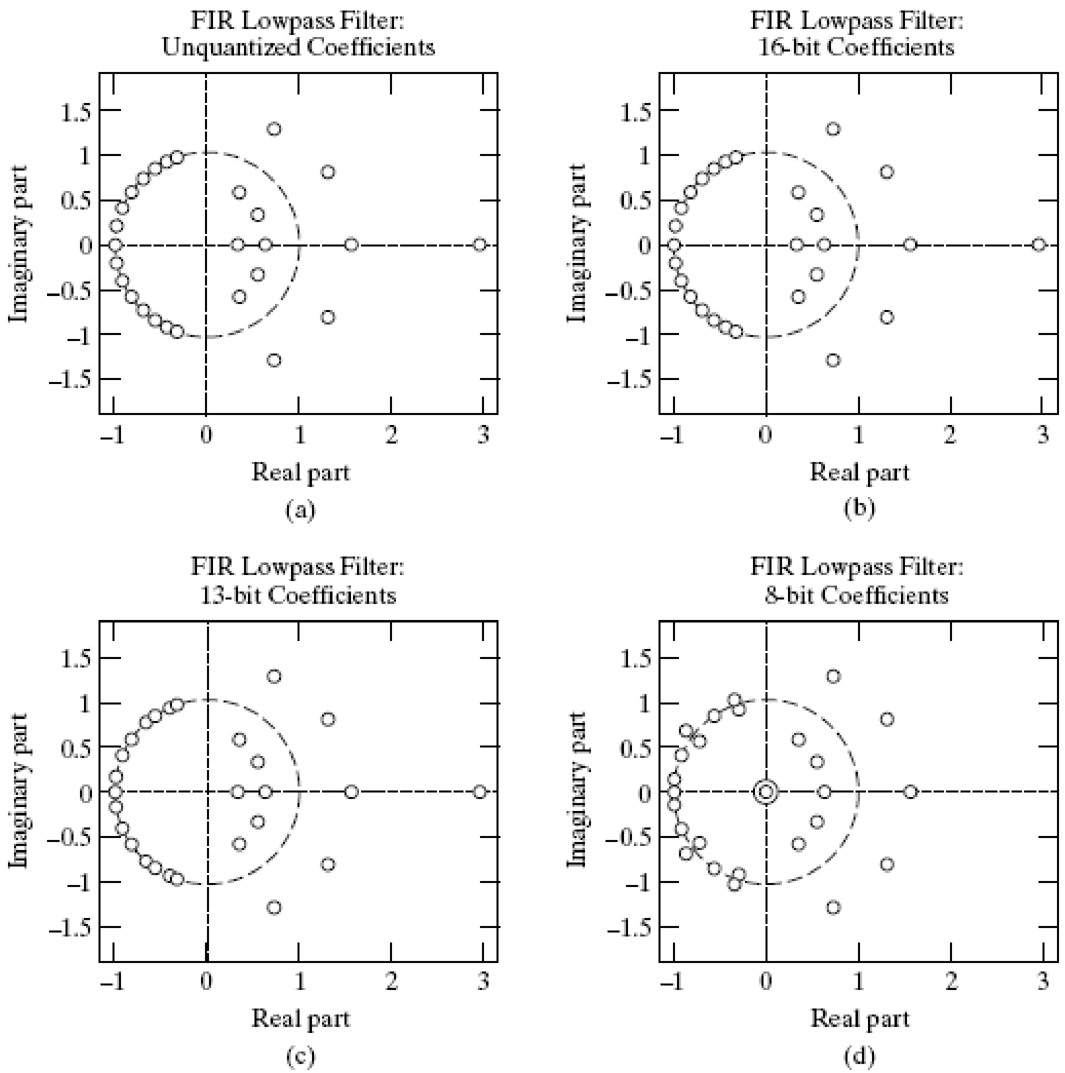


Figure 6.47 Effect of impulse response quantization on zeros of $H(z)$. (a) Unquantized. (b) Sixteen-bit quantization. (c) Thirteen-bit quantization. (d) Eight-bit quantization.

Cascade Form For Linear Phase FIR Filters

- less sensitive than Direct Form I because it isolates the quantization errors from the other sections
- To preserve linear phase, each section has to have linear phase
  - $(1+az^{-1}+z^{-2})$**:** this 2nd order section would contain complex conjugate pair of zeros on the unit circle, so the zeros can move only on the unit circle when a is quantized → prevents zeros from moving away on the unit circle as in Figure 6.47c.
  - Same 2nd order section can also be used for real zeros inside the unit circle and its reciprocal (outside the unit circle) so that the zeros would remain real
  - zeros at ±1 can be realized by first-order systems
  - If a 2nd order section is used to implement a pair of complex-conjugate zeros inside the unit circle instead of a 4th-order system (since there are a total of 4 zeros which we need to take care of due to the fact that zeros for linear-phase FIR systems occur as reciprocal pairs), then to ensure that linear-phase property is preserved, we can factorize 4th order system as

$$1+cz^{-1}+dz^{-2}+cz^{-3}+z^{-4}=\left(1-2r\cos\theta z^{-1}+r^2z^{-2}\right)\frac{1}{r^2}\left(r^2-2r\cos\theta z^{-1}+z^{-2}\right)$$

where the zeros are at: $z=re^{j\theta}$ and $z=\frac{1}{r}e^{-j\theta}$

Cascade Form For Linear Phase FIR Filters

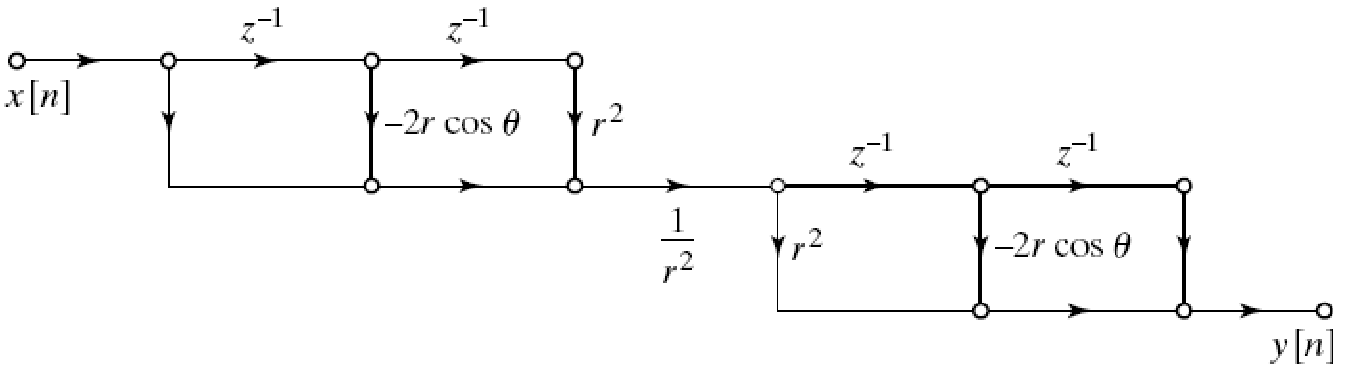


Figure 6.48 Subnetwork to implement fourth-order factors in a linear-phase FIR system such that linearity of the phase is maintained independently of parameter quantization.

so that the system uses the same coefficients, namely, $-2r\cos\theta$ and r^2, to realize both the zeros inside the unit circle and the conjugate reciprocal zeros outside the unit circle. Therefore, linear phase condition is preserved under quantization.