Georgia Institute of Technology

CSIP Center for Signal & Image Processing

ECE4270 Fundamentals of DSP Lecture 20

Fixed-Point Arithmetic in FIR and IIR Filters (part I)

School of ECE

Center for Signal and Information Processing

Georgia Institute of Technology

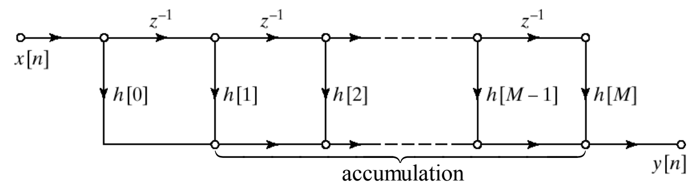
Overview of Lecture

- Two's-complement arithmetic
  - Overflow
- Issues in FIR implementation:
  - Quantizing filter coefficients
  - Roundoff noise and Scaling
- Introduction to IIR Filter Structures and Quantization
- Coefficient Quantization effects in IIR filters

Overflow

- Overflow can occur in two ways:
  - Accumulator overflows due to additions
  - Taking 16 bits out of the wrong part of a product
- Overflow can be prevented by:
  - More precision - use larger accumulator
  - Saturation arithmetic - clip accumulator at largest value
  - Shift products to the right to discard low-order bits. This results in loss of precision.

FIR Digital Filter

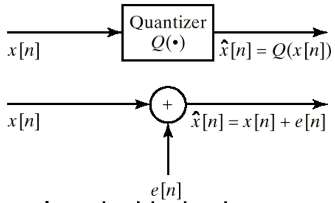


$$y[n] = \sum_{k=0}^{M} h[k]x[n-k]$$

$$= (((h[0]x[n] + \underline{h[1]x[n-1]}) + \ldots) + h[M]x[n-M])$$

To implement this filter, we must do multiply followed by accumulation (MAC).

Linear Noise Model



- Error is uncorrelated with the input.
- Error is uniformly distributed over the interval

$$-(\Delta/2) < e[n] \le (\Delta/2).$$

- Error is stationary white noise, (i.e. flat spectrum)

$$P_e(\omega) = \sigma_e^2 = \frac{\Delta^2}{12}, \quad |\omega| \le \pi$$

Roundoff Noise in FIR Filters

- Using the MAC instruction with quantized coefficients and quantized input, we can compute

$$y[n] = \sum_{k=0}^{M} h[k]x[n-k]$$ (Assume unquantized input and coefficients)

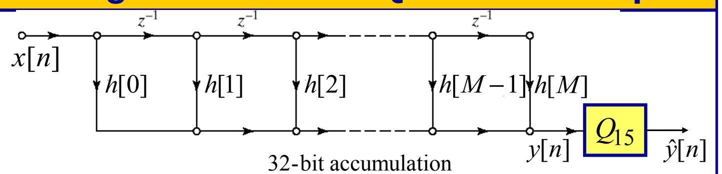
- This sum of products can be computed with 32-bit precision; i.e., with no quantization of the partial sums.
- The result is usually quantized to 15 bits + sign.

$$\hat{y}[n] = Q_{15}\{y[n]\} = y[n] + e[n]$$

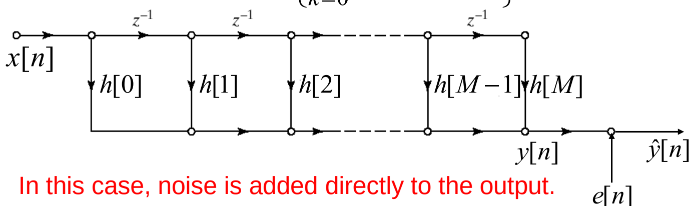
- The resulting noise power in the output is therefore

$$\sigma_e^2 = \Delta^2/12 = 2^{-30}/12$$

FIR Digital Filter with Quantized Output

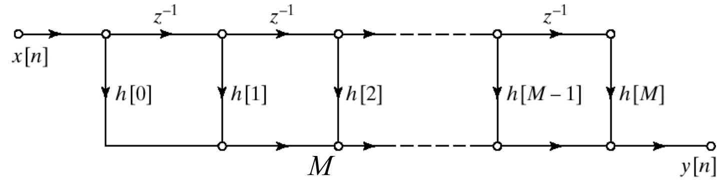


$$\hat{y}[n] = Q_{15}\{y[n]\} = Q_{15}\left\{\sum_{k=0}^{M} h[k]x[n-k]\right\} = y[n] + e[n]$$



In this case, noise is added directly to the output.

Absolute Scaling an FIR Digital Filter



$$y[n] = \sum_{k=0}^{M} h[k]x[n-k]$$

$$|y[n]| = \left|\sum_{k=0}^{M} h[k]x[n-k]\right| \le \sum_{k=0}^{M} |h[k]||x[n-k]|$$

$$|y[n]| \le \max\{x[n]\} \sum_{k=0}^{M} |h[k]| < 1 \quad \Rightarrow \quad \text{no overflow}$$

Sinusoidal Scaling for an FIR Filter

- Assume that the input is a sinusoid

$$x[n] = \cos(\omega_0 n)$$

- Then the output is

$$y[n] = \left|H\left(e^{j\omega_0}\right)\right|\cos\left(\omega_0 n + \angle H\left(e^{j\omega_0}\right)\right)$$

- Therefore, if we want $|y[n]| < 1$, then we must guarantee that

$$\left|H\left(e^{j\omega}\right)\right| < 1 \quad \text{for all } \omega$$

- This scaling is appropriate for most narrowband input signals.

Conclusions on FIR

- Coefficient quantization can modify the zero locations and therefore the frequency response.
  – This is usually not severe for linear phase filters
- Using the MAC instruction, we can avoid roundoff (quantization) until the very end of the computation
- Scaling must be used to avoid overflow

IIR Filter Structures and Quantization

- IIR filters are more complicated with regard to the effects of quantization.
  – Many different “equivalent” structures
  – Coefficient sensitivity may be high
  – Feedback is inherent in IIR structures
  – Possibility of instability
  – Roundoff noise is shaped by filter
- Some analysis is possible, but given the power of software tools such as MATLAB, an empirical approach is often most effective.

Fixed-Point Implementation Issues

- We need to represent coefficient and signal values by integers in a fixed range.
- Quantization errors in coefficients imply shifts of poles and zeros (even instability).
- For a given word-length, the quantization error is fixed in size. Therefore, signal values should be maintained as large as possible to maximize SNR.
- If signal values get too large, additions can overflow (or clip), thereby creating large errors.
- Thus, fixed-point implementations require careful attention to scaling the signal values.

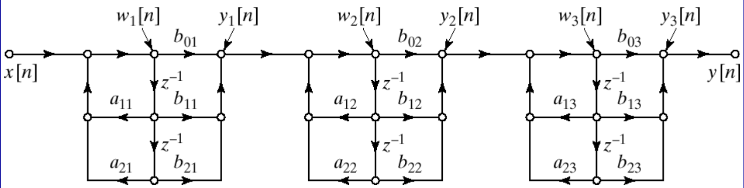
Cascade Form

$$H(z) = \prod_{k=1}^{N_s} \left(\frac{b_{0k} + b_{1k} z^{-1} + b_{2k} z^{-2}}{1 - a_{1k} z^{-1} - a_{2k} z^{-2}} \right)$$

$$w_k[n] = a_{1k} w_k[n-1] + a_{2k} w_k[n-2] + y_{k-1}[n]$$

$$y_k[n] = b_{0k} w_k[n] + b_{1k} w_k[n-1] + b_{2k} w_k[n-2]$$

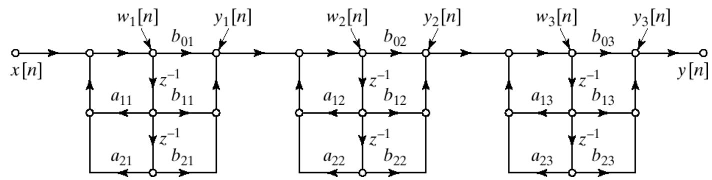
$$y_0[n] = x[n], \quad y[n] = y_{N_s}[n]$$



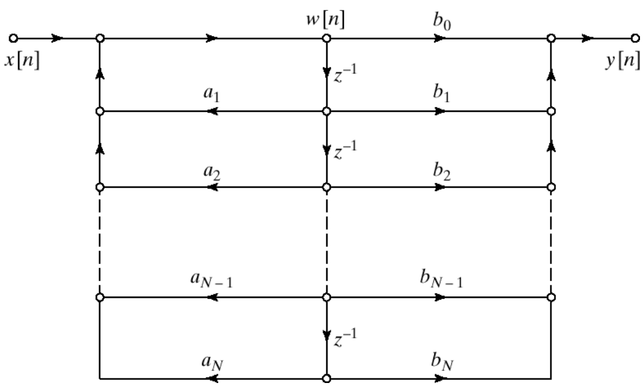
Cascade Implementation

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



Direct Form II (flow graph)



$$w[n] = \sum_{k=1}^{N} a_k w[n-k] + x[n] \qquad y[n] = \sum_{k=0}^{M} b_k w[n-k]$$

Equivalent Direct Form

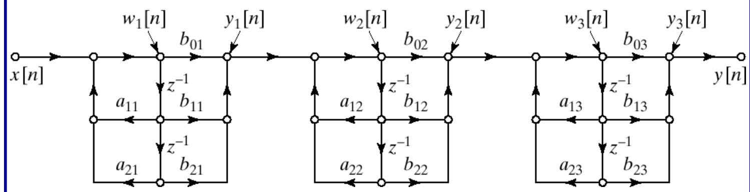
- The cascade form groups zeros and poles in pairs (second-order factors).
- These can be multiplied out to obtain single numerator and denominator polynomials in the direct forms I and II.

$$H(z) = \prod_{k=1}^{N_s} \left(\frac{b_{0k} + b_{1k} z^{-1} + b_{2k} z^{-2}}{1 - a_{1k} z^{-1} - a_{2k} z^{-2}} \right) = \frac{\sum_{k=0}^{M} b_k z^{-k}}{1 - \sum_{k=1}^{N} a_k z^{-k}}$$

16-Bit Quantized Coefficients

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}



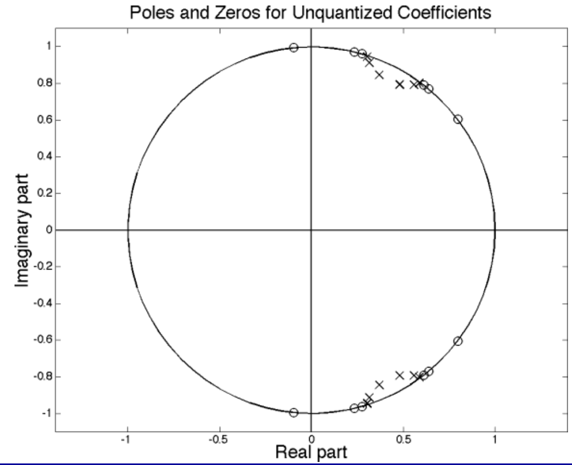
Quantized Numerator Coefficients

b_k	$\hat{b}_k$
0.01004671277613	0.01004028320312
-0.04940368759854	-0.04940795898438
0.15047336191420	0.15048217773438
-0.31987136089868	-0.31988525390625
0.53335872862212	0.53335571289062
-0.71133037924498	-0.71133422851562
0.78462412594880	0.78463745117188
-0.71133037924498	-0.71133422851562
0.53335872862212	0.53335571289062
-0.31987136089868	-0.31988525390625
0.15047336191420	0.15048217773438
-0.04940368759854	-0.04940795898438
0.01004671277613	0.01004028320312

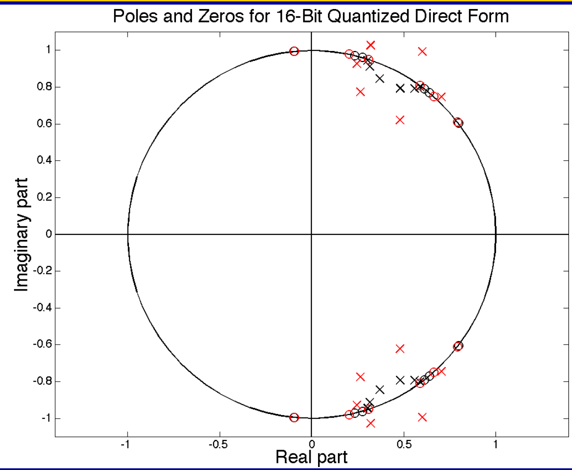
Quantized Denominator Coefficients

a_k	$\hat{a}_k$
1.00000000000000	1.00000000000000
-5.22295000000000	-5.22265625000000
16.76588243738500	16.76562500000000
-36.82056676783901	-36.82031250000000
62.25444673309200	62.25390625000000
-82.41640461036937	-82.41796875000000
88.36682598383915	88.36718750000000
-76.16156361479057	-76.16015625000000
53.16012384772395	53.16015625000000
-29.04773747211416	-29.04687500000000
12.21807179345019	12.21875000000000
-3.51452513378527	-3.51562500000000
0.62178984772852	0.62109375000000

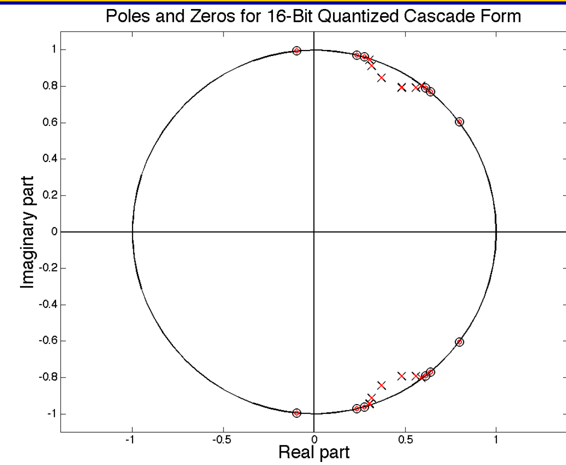
No Quantization



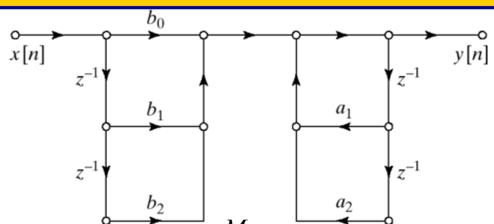
16-Bit Quantized Direct Form



16-Bit Quantized Cascade Form



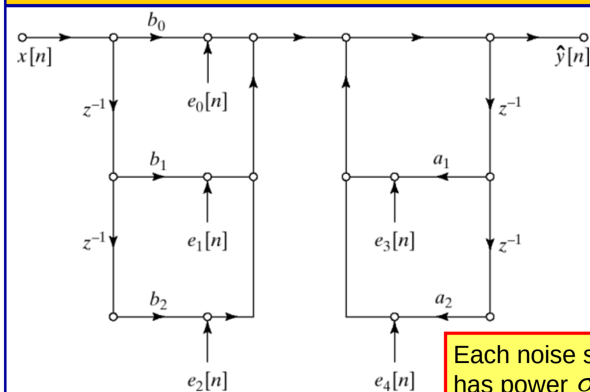
Direct Form I IIR Filter



$$H(z)=\frac{\sum_{k=0}^{M} b_k z^{-k}}{1-\sum_{k=1}^{N} a_k z^{-k}}=\frac{B(z)}{A(z)}$$

$$y[n]=\sum_{k=1}^{N} a_k y[n-k]+\sum_{k=0}^{M} b_k x[n-k]$$

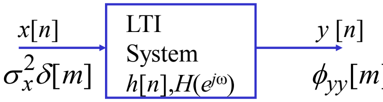
Linear Noise Model



Each noise source has power $\sigma_e^2=2^{-2B}/12$

The noise sources are independent so their powers add.

Linear System with a White Noise Input



$$\phi_{xx}[m] = \sigma_x^2 \delta[m] \qquad \phi_{yy}[m] = \sigma_x^2 c_{hh}[m]$$

$$\Phi_{xx}(e^{j\omega}) = \sigma_x^2 \qquad \Phi_{yy}(e^{j\omega}) = \sigma_x^2 C_{hh}(e^{j\omega})$$

$$\phi_{yy}[m] = \phi_{xx}[m] * c_{hh}[m] = \sigma_x^2 \delta[m] * c_{hh}[m] = \sigma_x^2 c_{hh}[m]$$

$$c_{hh}[m] = \sum_{k=-\infty}^{\infty} h[m+k] h^*[k] = h[-m] * h^*[m]$$

$$\Phi_{yy}(e^{j\omega}) = \Phi_{xx}(e^{j\omega}) C_{hh}(e^{j\omega}) = \sigma_x^2 C_{hh}(e^{j\omega})$$

$$C_{hh}(e^{j\omega}) = H(e^{-j\omega}) H^*(e^{-j\omega}) = \left| H(e^{-j\omega}) \right|^2$$