Digital Signal Processing
Lecture 7 - Structures for Discrete-Time Systems

Electrical Engineering and Computer Science
University of Tennessee, Knoxville

October 6, 2015
Discrete-time systems

- Special properties: linearity, TI, stability, causality
- LTI systems: the unit sample response $h[n]$ uniquely characterizes an LTI system

$$y[n] = \sum_{k=-\infty}^{\infty} x[k] h[n-k] = x[n]*h[n]$$

- Frequency response: $H(e^{j\omega})$, complex exponentials are eigenvalues of LTI systems, i.e., if $x[n] = e^{j\omega n}$,

$$y[n] = H(e^{j\omega}) x[n] = \left( \sum_{k=-\infty}^{\infty} h[k] e^{-j\omega k} \right) e^{j\omega n}$$

- Fourier transform: Generalization of frequency response (a periodic continuous function of $\omega$)

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x[n] e^{-j\omega n}, x[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} X(e^{j\omega}) e^{j\omega n} d\omega$$

- The z-transform as a generalization to the Fourier transform, $X(z) = \sum_{n=-\infty}^{\infty} x[n] z^{-n}$, and the system function $H(z)$

- Sampling, Aliasing, Reconstruction
Transform-domain Analysis

- FIR vs. IIR
- FIR filters with generalized linear phase (special pattern for zeros)
- Minimum phase systems (special pole-zero properties)
- All-pass systems (special pole/zero properties)
- $H = H_{\text{min}} H_{\text{ap}}$
- Geometric interpretation of the pole-zero plot
Different system representations

- Using LCDE with initial rest condition

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

- Using system function with ROC \(|z| > R_+\)

\[ H(z) = \frac{\sum_{k=0}^{M} b_k z^{-k}}{1 - \sum_{k=1}^{N} a_k z^{-k}} \]
Block diagram vs. Signal flow graph

- Block diagram symbols: adder, multiplier, unit delay (memory)

- Signal flow graph: directed branches (branch gain, delay branch), nodes (source node, sink node)

- A comparison: nodes in the flow graph represent both branching points and adders, whereas in the block diagram a special symbol is used for adders
Different flow graph representations require different amounts of computational resources.
**Direct form I and II**

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

- **Direct form I: implementing zeros first**

  \[ H(z) = H_2(z) H_1(z) = (\frac{1}{1 - \sum_{k=1}^{N} a_k z^{-k}})(\sum_{k=0}^{M} b_k z^{-k}) \]  
  \[ V(z) = H_1(z) X(z) = (\sum_{k=0}^{M} b_k z^{-k}) X(z) \]  
  \[ Y(z) = H_2(z) V(z) = (\frac{1}{1 - \sum_{k=1}^{N} a_k z^{-k}}) V(z) \]  
  \[ y[n] = \sum_{k=0}^{M} b_k x[n-k] + \sum_{k=1}^{N} a_k y[n-k] \]

- **Direct form II: implementing poles first**

  \[ H(z) = H_1(z) H_2(z) = (\sum_{k=0}^{M} b_k z^{-k})(\frac{1}{1 - \sum_{k=1}^{N} a_k z^{-k}}) \]  
  \[ W(z) = H_2(z) X(z) = (\frac{1}{1 - \sum_{k=1}^{N} a_k z^{-k}}) X(z) \]  
  \[ Y(z) = H_1(z) W(z) = (\sum_{k=0}^{M} b_k z^{-k}) W(z) \]  
  \[ y[n] = \sum_{k=1}^{N} a_k y[n-k] + \sum_{k=0}^{M} b_k x[n-k] \]
Comparison
Comparison (cont’)
Comparison (cont')

Diagram:

\[ x[n] \rightarrow z^{-1} \rightarrow b_0 \rightarrow v[n] \rightarrow y[n] \]

\[ x[n-1] \rightarrow z^{-1} \rightarrow b_1 \rightarrow a_1 \rightarrow y[n-1] \]

\[ x[n-2] \rightarrow z^{-1} \rightarrow b_2 \rightarrow a_2 \rightarrow y[n-2] \]

\[ x[n-N+1] \rightarrow z^{-1} \rightarrow b_{N-1} \rightarrow a_{N-1} \rightarrow y[n-N+1] \]

\[ x[n-N] \rightarrow z^{-1} \rightarrow b_N \rightarrow a_N \rightarrow y[n-N] \]

\[ w[n] \rightarrow b_0 \rightarrow y[n] \]

\[ w[n] \rightarrow a_1 \rightarrow z^{-1} \rightarrow b_1 \]

\[ w[n] \rightarrow a_2 \rightarrow z^{-1} \rightarrow b_2 \]

\[ w[n] \rightarrow a_{N-1} \rightarrow z^{-1} \rightarrow b_{N-1} \]

\[ w[n] \rightarrow a_N \rightarrow z^{-1} \rightarrow b_N \]
Exercises

Ex1: \( H(z) = \frac{1+2z^{-1}}{1-1.5z^{-1}+0.9z^{-2}} \)

Ex2: \( H(z) = \frac{1+2z^{-1}+z^{-2}}{1-0.75z^{-1}+0.125z^{-2}} \)
Solution
A digital filter structure is said to be *canonic* if the number of delays is equal to the order of the difference equation. Otherwise, it is a *noncanonic* structure. That is, minimum number of delays required is $\max(N, M)$. 
The transposition theorem: For single-input, single-output systems, the resulting flow graph has the same system function as the original graph if the input and output nodes are interchanged.

- reverse direction of all branches
- interchange input and output

Implement zeros first, then poles as compared to the direct II form
Example
Comparison - Direct form II vs. Transposed direct form II
Cascade form

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

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

\[ N_s = \left\lfloor (N + 1)/2 \right\rfloor \]
Ex: \( H(z) = \frac{1+2z^{-1}+z^{-2}}{1-0.75z^{-1}+0.125z^{-2}} = \frac{(1+z^{-1})(1+z^{-1})}{(1-0.5z^{-1})(1-0.25z^{-1})} \)
Why cascading?

- Use of computation resource
  - Direct form II structure: $2N + 1$ constant multipliers
    \[
    H(z) = \frac{\sum_{k=0}^{M} b_k z^{-k}}{1 - \sum_{k=1}^{N} a_k z^{-k}}
    \]
  - Cascade form structure: $5N/2$ constant multipliers (assume $M = N$ and $N$ is even)
    \[
    H(z) = \prod_{k=1}^{N_s} \frac{b_{0k} + b_{1k} z^{-1} + b_{2k} z^{-2}}{1 - a_{1k} z^{-1} - a_{2k} z^{-2}}
    \]
- Precision
$H(z) = \sum_{k=0}^{N_p} C_k z^{-k} + \sum_{k=1}^{N_s} \frac{e_{0k} + e_{1k} z^{-1}}{1 - a_{1k} z^{-1} - a_{2k} z^{-2}}$
Ex: \( H(z) = \frac{1+2z^{-1}+z^{-2}}{1-0.75z^{-1}+0.125z^{-2}} = 8 + \frac{-7+8z^{-1}}{1-0.75z^{-1}+0.125z^{-2}} \)
Solution
Feedback in IIR systems

- Closed path: **necessary** to generate infinite long impulse responses (not sufficient)
- The computability of a flow graph is that all loops must contain at least one unit delay element
FIR - Direct and transposed direct form

- tapped delay line structure (transversal filter structure)
- discrete convolution

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

\[ H(z) = \sum_{n=0}^{M} h[n]z^{-n} = \prod_{k=1}^{M_s} (b_{0k} + b_{1k}z^{-1} + b_{2k}z^{-2}) \]

\[ M_s = \lfloor (M + 1)/2 \rfloor \]
Linear phase FIR systems

\[ h[M - n] = h[n], \quad n = 0, 1, \ldots, M \]

When \( M \) is even
Linear phase FIR systems (cont’)

\[ h[M - n] = h[n], \quad n = 0, 1, \ldots, M \]

When \( M \) is odd
Sources of errors

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

- Coefficient quantization problem: \( a \rightarrow \hat{a} \)
- Input quantization error: \( x[n] \rightarrow \hat{x}[n] = x[n] + e[n] \)
- Product quantization error: 
  \[ v[n] = ay[n - 1] \rightarrow \hat{v}[n] = v[n] + e_a[n] \]
- Limit cycles: caused by the nonlinearity by the quantization of arithmetic operations. When the input is absent or constant input or sinusoidal input signals are present, the output is in the form of oscillation
Quantization problem in Implementation

(a) Narrowband passband implementation

(b) Narrowband stopband implementation

(c) Narrowband implementation with quantized coefficients
Number representations

- The two's complement format

\[ x = X_m(-b_0 + \sum_{i=1}^{\infty} b_i 2^{-i}) \]

- \( X_m \): an arbitrary scale factor
- \( b_0 \): the sign bit. \( 0 \leq x \leq X_m \) if \( b_0 = 0 \); \( -X_m \leq x < 0 \) if \( b_0 = 1 \)

- Fix-point binary numbers

\[ \hat{x} = Q_B[x] = X_m(-b_0 + \sum_{i=1}^{B} b_i 2^{-i}) = X_m \hat{x}_B \]

- Quantizing a number to \( B + 1 \) bits. Quantization error:

\[ e = Q_b[x] - x \]
Quantization error

- Rounding: $-\Delta/2 < e \leq \Delta/2$
- Truncating: $-\Delta < e \leq 0$
Quantization error (cont’)
- When \( x > X_m \)
- Saturation overflow (Clipping)
Effect of coefficient quantization of an IIR digital filter implemented in direct form (5th-order IIR elliptic lowpass filter)
Coefficient quantization - IIR (cont’)

- Effect of coefficient quantization of an IIR digital filter implemented in cascade form (5th-order IIR elliptic lowpass filter)
Effect of coefficient quantization of an FIR digital filter implemented in direct form (39th-order FIR equiripple lowpass filter)
Pole sensitivity of second-order structures (Product quantization)

- The direct form structure exhibits high pole sensitivity with poles closer to the real axis and low pole sensitivity with poles closer to $z = \pm j$
The coupled form structure is more suitable for implementing any type of second-order transfer function.