Fixed-point Analysis of Digital Filters

For Advanced VLSI Design
and
VLSI Signal Processing

台灣大學電機系 吳安宇

Modified: 11-02-2002
Quantization in implementing system

A. Ideal system

B. Nonlinear model

C. Linear model
Effect of coefficient quantization in IIR system

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

\[ \hat{H}(z) = \frac{\sum_{k=0}^{M} \hat{b}_k z^{-k}}{1 - \sum_{k=1}^{N} \hat{a}_k z^{-k}} \]

\[ \hat{a}_k = a_k + \Delta a_k \]

\[ \hat{b}_k = b_k + \Delta b_k \]
Effect of coefficient quantization in FIR system

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

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

\[ = \sum_{n=0}^{M} (h[n] + \Delta h[n]) z^{-n} \]

\[ = \sum_{n=0}^{M} h[n] z^{-n} + \sum_{n=0}^{M} \Delta h[n] z^{-n} \]

\[ = H(z) + \Delta H(z) \quad \text{(consider as a parallel FIR structure)} \]
Sensitivity

\[ A(z) = 1 - \sum_{k=1}^{N} a_k z^{-k} = \prod_{j=1}^{N} (1 - z_j z^{-1}) \]

New pole of \( \hat{H}(z) \) becomes

\[ \hat{z}_i = z_i + \Delta z_i, \quad i = 1, 2, \ldots, N \]

\[ \Delta z_i = \sum_{k=1}^{N} \frac{\partial z_i}{\partial a_k} \Delta a_k \]

\[
\left( \frac{\partial A(z)}{\partial z_i} \right)_{z=z_i} \frac{\partial z_i}{\partial a_k} = \left( \frac{\partial A(z)}{\partial a_k} \right)_{z=z_i} \left( z_i^{N-k} \prod_{j=1, j \neq i}^{N} (z_j - z_i) \right)
\]
Example: 2nd-order IIR Filter (I)

\[ H(z) = \frac{1}{1 + k_1 z^{-1} + k_2 z^{-2}} \]

\[ = \frac{1}{1 - 2r \cos \theta z^{-1} + r^2 z^{-2}} \]

\[ \rightarrow \begin{cases} 
  k_1 = 2 \cos \theta \cdot R \\
  k_2 = R^2 
\end{cases} \]
Example:

2nd-order IIR Filter (II)

4-bit coefficient (k1,k2)

7-bit coefficient (k1,k2)
Three scaling methods: To avoid overflow

- $S_1X_{\text{max}} < \frac{1}{\sum_{m=-\infty}^{\infty} |h[m]|}$
- $S_2X_{\text{max}} < \frac{1}{\max |H(e^{j\omega})|}$
- $(S_3X_{\text{max}})^2 \leq \frac{1}{\sum_{n=-\infty}^{\infty} |h[n]|^2}$

2’s complement

$X^S_{\text{max}}$ : Scaling factor
$X_{\text{max}}$ : Maximum value of input signal
Scaling method

■ Method I

Transform IIR to FIR system

\[
\frac{1}{1-az^{-1}} = 1 + az^{-1} + a^2 z^{-2} + \ldots
\]

Extreme Case:

\[x[n] = \{1,1,1\ldots\}\]

\[\rightarrow 1 + a + a^2 + a^3 + \ldots = \frac{1}{1-a}\]

\[a = 0.9 \rightarrow \text{放大10倍, 退4bits}\]
Scaling method

- method II

検測 single tone => 不安全

- method III

結合 method I 與 method II, 以能量來檢測
Rounding error model

\[ x[n] \rightarrow \text{Round (word length} = B \text{ bit)} \rightarrow y[n] \]

- Range: \[ -2^{-B} < e[n] < 2^{-B} \]
- Variance: \[ \sigma^2_e = \frac{2^{-2B}}{3} \]
Direct form I

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

Assume \( x[n] \) is uniform in \([-1, 1]\)

- **Variance of output signal**

\[ \sigma_y^2 = \frac{1}{3} \left[ \frac{1}{2\pi i} \int_{c} H(z) H(z^{-1}) z^{-1} dz \right] \]

- **Variance of rounding output error**

\[ \sigma_f^2 = (M + N + 1) \frac{2^{-2B}}{3} \left[ \frac{1}{2\pi i} \int_{c} \frac{z^{-1} dz}{A(z) A(z^{-1})} \right] \]

A(z) : denominator part of H(z)
Direct form II

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

- Variance of output signal

\[
\sigma_y^2 = \left(\frac{1}{3}\right) \left[ \frac{1}{2\pi j} \int_{c} H(z)H(z^{-1})z^{-1}dz \right]
\]

- Variance of rounding output error

\[
\sigma_f^2 = N \left(\frac{2^{-2B}}{3}\right) \left[ \frac{1}{2\pi j} \int_{c} H(z)H(z^{-1})z^{-1}dz \right] (e_a[n])
\]

\[
+ (M + 1) \left(\frac{2^{-2B}}{3}\right) (e_b[n])
\]
Calculation of SQNR

SQNR: Signal-to-Quantization-Noise Ratio

\[
x(t) \xrightarrow{\text{Q(.)}} \sum_{i=1}^{N} y_i^2(n) \xrightarrow{\text{Q(.)}} 10 \log_{10} \left( \frac{\sum_{i=1}^{N} y_i^2(n)}{\sum_{i=1}^{N} e_i^2(n)} \right) = 10 \log_{10} \left( \frac{\sum_{i=1}^{N} y_i^2(n)}{\sum_{i=1}^{N} (y(n) - \hat{y}(n))^2} \right)
\]

In general, SQNR ≥ 30dB for practical implementation
Conclusions

- Fixed-point analysis is required in converting floating-point based DSP algorithm into fixed-point based implementation (e.g. VLSI circuits & fixed-point Programmable DSP processors)
- Usually it is done by doing extensive simulations
- Closed-form analytical results help to see the effectiveness of each design parameters (Wordlength, Scaling factors, etc.)
- Each DSP algorithm has its own numerical property in fixed-point implementations.