Fold Here

Finite-Wordlength Effect in Digital filters.

Fold Here

The basic fundamental operations involved in DSP algorithms like convolution, correlation DFT, FFT etc. are multiplication and addition.

These operations are performed using input sequence x(n) impulse sequence h(n) and the co-efficients of the difference equation governing the system.

Fold Here

The input and output data are stored in register in a digital system or different memory location in the form of binary values.

Fold Here

The maximum size of the binary information that can be stored in a memory location or register is known as "Word-length".

* When a register stores an 8-bit data then its wordlength is 8-bit.

Fold Here

* Quantization & Coding depends on register wordlength.

Fold Here

* Quantization & Coding will introduce error in input data, because analog data has infinite precision but digital equivalent has finite precision.

Fold Here

There are some error occurred due to the process of finite precision representation of binary numbers in digital system. These errors are commonly called as " finite word length effects or error".

* Finite wordlength effects in digital filters —

(1) Errors due to quantization of I/p data by A/D converter.

Fold Here

(2) Error due to quantization of filter co-efficients.

(3) Errors due to rounding of products in multiplication.

(4) Error due to overflow in addition.

* Representation of Numbers —

(a) fixed point representation

Fold Here

(b) Floating " " .

(a) fixed point Representation — Bits allowed for integer & fractional part and so position of binary point is fixed.

Drawback → Due to the fixed integer & fractional part: too large and too small values cannot be represented.

Types — (i) Sign magnitude form

$(2.75)_{10} = (010.1100)_2$

$(-2.75)_{10} = (110\ 1100)_2$

(ii) One's Complement form (for negative)

(iii) Two's complement form (" ")

(b) <u>Floating Point Representation</u> :-

$$N_f = M \times 2^E$$

M = Mantissa → it will be in binary fraction format.

$$0.5 \leq M < 1$$

E = Exponent → either a positive or negative integer.

(i) $+7_{10} = +111_2 = 0.1110 \times 2^3 = 0.111 \times 2^{+11_2}$

$= \underbrace{0\,1110}_{M}\ \underbrace{011}_{E}$

(ii) $-7_{10} = -111_2 = 1.111 \times 2^{+3} = 1.1110 \times 2^{+11_2}$

$= \underbrace{1\,1110}_{M}\ \underbrace{011}_{E}$

Quantization :

In fixed point or floating point arithmetic the size of the result of the operation may exceed the size of binary used in the number system. In this case, the low order bits has to be eliminated in order to store the result.

Method → * Truncation

* Rounding.

* Truncation → 8 bits to 4 bits

$0.01110011 \rightarrow 0.0111$

* Rounding → $0.10101\underline{0}$ rounded to 4 bits

0.1010 or 0.1011.

Rounding error → $e_r = N_r - N$.

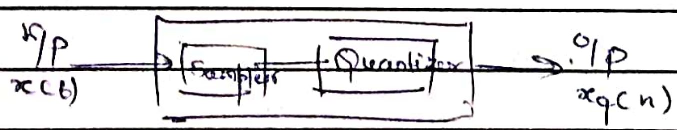
N_r → Quantized or rounded number

N → Unquantized number

Range :

$$-\frac{2^{-b}}{2} \leq e_r \leq \frac{2^{-b}}{2}$$

b = no. of rounded bits.



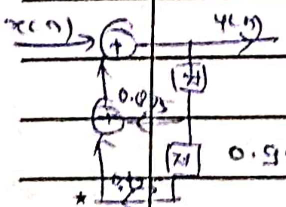
Block diagram of A/D Converter

Quantization in filter - Coefficient -

The filter co-efficients are quantized to word' sizes of the register used to store them either by truncation or by rounding.

* Location of poles & zeros of digital filters directly depends on the value of filter coefficients. The quantization of the filter co-efficients will modify the value of poles and zeros and so the location of poles and zeros will be shifted from the desired location. This will create deviation in the frequency response of the system.

Ex:- $H(z) = \dfrac{1}{(1-0.5z^{-1})(1-0.45z^{-1})}$ 3-bit- coeffecient -



$= \dfrac{z^2}{(z-0.5)(z-0.45)} = \dfrac{1}{(1-0.95z^{-1}+0.225z^{-2})}$

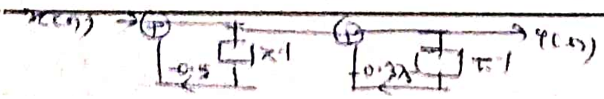
$0.95_{10} \xrightarrow{D \to B} 0.1111_2 \xrightarrow{\text{trun. to 3 bits}} 0.111_2 \xrightarrow{B \to D} 0.875_{10}$

$0.225_{10} \longrightarrow 0.0011_2 \longrightarrow 0.001_2 \xrightarrow{B \to D} 0.125$

- for cascade :- $H_1(z) = \dfrac{1}{1-0.5z^{-1}}$ $H_2(z) = \dfrac{1}{1-0.45z^{-1}}$

$0.5 \xrightarrow{D \to B} 0.1000_2 \longrightarrow 0.100_2 \xrightarrow{B \to D} 0.5_{10}$

$0.45 \longrightarrow 0.0111_2 \longrightarrow 0.011_2 \longrightarrow 0.375$



Fold Here

Quantization in filter - Coefficient -

The filter co-efficients are quantized to word size of the register used to store them either by truncation or by rounding.

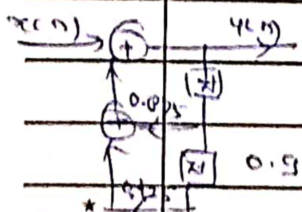
Fold Here

- Location of poles & zeros of digital filters directly depends on the value of filter coefficients. The quantization of the filter co-efficients will modify the value of poles and zeros and so the location of poles and zeros will be shifted from the desired location. This will create deviation in the frequency response of the system.

Fold Here

Fold Here

Ex:- $H(z) = \dfrac{1}{(1-0.5z^{-1})(1-0.45z^{-1})}$ 3-bit-coefficient -



$= \dfrac{z^2}{(z-0.5)(z-0.45)} = \dfrac{1}{(1-0.95z^{-1}+0.225z^{-2})}$

$0.95_{10} \xrightarrow{D\to B} 0.1111_2 \xrightarrow{\text{trun. to 3 bits}} 0.111_2 \xrightarrow{B\to D} 0.875_{10}$

Fold Here

Fold Here

$0.225_{10} \longrightarrow 0.0011_2 \longrightarrow 0.001_2 \xrightarrow{B\to D.} 0.125$

- for cascade :- $H_1(z) = \dfrac{1}{1-0.5z^{-1}}$ $H_2(z) = \dfrac{1}{1-0.45z^{-1}}$

$0.5 \xrightarrow{D\to B} 0.1000_2 \longrightarrow 0.100_2 \xrightarrow{B\to D} 0.5_{10}$.

$0.45 \longrightarrow 0.0111_2 \longrightarrow 0.011_2 \longrightarrow 0.375$