UNIT-IV FINITE WORDLENGTH EFFECTS

Introduction: the Basic operations Performed in Convolution is multiplication and addition. The information (on numbers) used in computation are called input and result is o/p data. These i/p are stored in a register. The registers are basic storage device in a digital system. The maximum size of the binary information that can be stored in a register is called register word length. For storing information in registers they have to be quantized and coded in binary. These quantization and coding process depends upon the register word length. This will introduce error in data known as quantization error. The following error occurs due to the quantization of binary data (on information).

*) The error occurs due to the quantization of the input by A-D converter otherwise known as input quantization error

*) Due to quantization filter co-efficient

*) Due to the quantization of product

*) limit cycle oscillation

*) Round off noise.

Input quantization error: the conversion of a CT i/p sig into digital value produces an error

product quantization error: Product quantization error arise at the output of a multiplier. Multiplication of a b bit data with a 'b' bit co-efficient results a product having 2b bits. Since a 'b' bit registe is used, the multiplier o/p must be rounded (or) truncated to 'b' bit which

Co-efficient quantization error: the frequency re... the resulting filter may differ from the desired and sometimes the filter may fail to meet the desired filter specifications. If the poles of the desired filter are close to the unit circle, then those of the filter with quantized coefficients may lie just outside the unit circle, leading to instability.

Number representation:

We can represent a number N to any desired accuracy by a finite series

$$N = \sum_{i=-B}^{A} d_i r^i$$

where

A is the no. of integer digits

B is the no. of fractional digits

and r is the radix.

If $r=10$, the representation is known as decimal representation having numbers from 0 to 9. In this representation the number

$$30.285 = \sum_{i=-3}^{1} d_i 10^i$$

$$= 3\times10^1 + 0\times10^0 + 2\times10^{-1} + 8\times10^{-2} + 5\times10^{-3}$$

If $r=2$, the representation is known as binary with two numbers 0 to 1. For example, the binary number

$$110.010_2 = 1\times2^2 + 1\times2^1 + 0\times2^0 + 0\times2^{-1} + 1\times2^{-2} + 0\times2^{-3}$$

$$= 4 + 2 + 0 + 0 + 0.25 + 0$$

$$= 6.25_{10}$$

Convert the decimal number 20.275 to binary form ...tion

Integer part

$$2\,|\,30$$
$$2\,|\,15 - 0$$
$$2\,|\,7 - 1 \quad \text{Remainder}$$
$$2\,|\,3 - 1$$
$$1 - 1$$

11110_2

Fractional part — Integer part — Binary number

$0.275 \times 2 = 0.550 \rightarrow 0$

$0.55 \times 2 = 1.10 \rightarrow 1$

$0.10 \times 2 = 0.2 \rightarrow 0$

$0.2 \times 2 = 0.4 \rightarrow 0$

$0.4 \times 2 = 0.8 \rightarrow 0$

$0.8 \times 2 = 1.6 \rightarrow 1$

$0.6 \times 2 = 1.2 \rightarrow 1$

$0.2 \times 2 = 0.4 \rightarrow 0$

$\therefore (30.275)_{10} = (11110.01000110\ldots)_2$

Two methods to represent binary number

1. Floating point ~~representation~~.
2. Fixed point ~~representation~~

1. Fixed point: The bits alloted integer and fractional part are fixed. since it is fixed, it is impossible to represent too small and too large values.

2. Floating point: The binary part can be shifted to a desired position. so that the bits in the integer and fractional part of the number can be varied.

1. Fixed point representation: To represents the positive number only one format is available. That is sign magnitude format (SMF).

$$N_p = \sum_{i=0}^{B} d_i\, r^{-i}$$

$$\boxed{\ldots d_1\, d_2\, d_3 \cdots d_n}$$

To represent the negative
these are 3 - formats are available

a) sign magnitude
b) 1's complement
c) 2's complement.

a) sign magnitude format (-ve numbers)

$$N_n = \sum_{i=0}^{B} d_i r^{-i}, \quad d_0 = 1 \text{ (sign bit)}$$

$(+0.125)_{10} = (0.001)_2$ → +ve number

$(-0.125)_{10} = (1.001)_2$ → -ve number

$0.125 \times 2 = 0.250 \rightarrow 0$
$0.250 \times 2 = 0.5 \rightarrow 0$
$0.5 \times 2 = 1.0 \rightarrow 1$
0.001

b) 1's complement : (-ve number)

Complement of each digit can be obtained by subtracting the digit from 1.

$$N_{1's} = 1 \times 2^0 + \sum_{i=1}^{B} (1-d_i) r^{-i}, \quad r = 2$$

$(+0.125)_{10} = (0.001)_2$ complementing each bit
smf ↓
$N_{1's} = (1.110)_2 = (-0.125)_{10}$

c) 2's complement :

taking 1's complement of it's +ve num and adding 1 to LSB.

$$N_{2's} = 1 \times 2^0 + \sum_{i=1}^{B} (1-d_i) 2^{-i} + 1 \times 2^{-B}$$

$(+0.125)_{10} = (0.001)_2$ → $N_{1's} = (1.111)_2$

e number fixed point representation of
wing numbers 7/8, −7/8, 5/8, −3/8

Solution

	Sign-magnitude	1's Complement	2's Complement
(0.875) $\frac{7}{8}$ ⇒	0.111		
(0.625) $\frac{5}{8}$ ⇒	0.101		

−7/8, → 3/8 ⇒ (−ve) numbers

(−0.875) − 7/8

(−0.375) − 3/8

0.875 × 2 = 1.75
0.75 × 2 = 0.150
0.150 × 2 = 0.300
0.300 × 2 = 0.600
0.600 × 2 = 1.200

Floating point representation:-

It is used to represent very small and very large numbers and it's representation is given by

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

where,

M → mandisa (0.5 ≤ m ≤ 1) binary fractional value

E → Exponential (may be +ve / −ve integer)

Both mandisa and Exponential use one bit as a sign bit (MSB)

sign bit = 1 for −ve
0 for +ve

Let us consider 8-bit floating point representation

UNIT-IV Finite Impulse Response Filter

Introduction: (2)

In many digital processing applications FIR filters are preferred over their IIR counter parts. The following are the main advantages of the FIR filter over IIR filter.

1. FIR filters are always stable.
2. FIR filters with exactly linear phase can easily be designed.
3. FIR filters can be realized in both recursive and non-recursive structures.
4. FIR filters are free of limit cycle oscillations when implemented on a finite word length digital system.
5. Excellent design methods are available for various kinds of FIR filters.

The disadvantages of FIR filter are,

1. The implementation of narrow transition band FIR filters are very costly, as it requires considerably more arithmetic operations and hardware components such as multipliers, adders and delay elements.
2. memory requirement and execution time are very high

H(z) can be expressed as a sum of terms, with one term containing the even co-efficients and the other containing the odd indexed co-efficients.

$$H(z) = (h[0] + h[2]z^{-2} + h[4]z^{-4} + h[6]z^{-6} + h[8]z^{-8}) + (h[1]z^{-1} + h[3]z^{-3} + h[5]z^{-5} + h[7]z^{-7})$$

$$H(z) = (h[0] + h[2]z^{-2} + h[4]z^{-4} + h[6]z^{-6} + h[8]z^{-8}) + z^{-1}(h[1] + h[3]z^{-2} + h[5]z^{-4} + h[7]z^{-6})$$

By using the notation.

$$E_0(z) = h[0] + h[2]z^{-1} + h[4]z^{-2} + h[6]z^{-3} + h[8]z^{-4}$$

$$E_1(z) = h[1] + h[3]z^{-1} + h[5]z^{-2} + h[7]z^{-3}$$

we can express H(z) as

$$H(z) = E_0(z^2) + z^{-1}E_1(z^2)$$

The above decompostion is more commonly known as the 2-branch polyphase decomposi[tion]

$H(z) = E_0(z^3) + z^{-1}E_1(z^3) + z^{-2}E_2(z^3)$

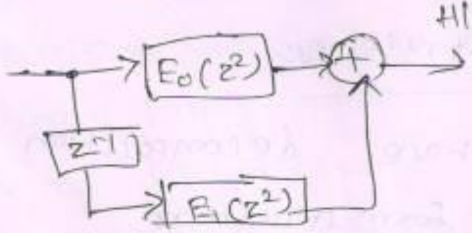


Fig. 2-branch polyphase

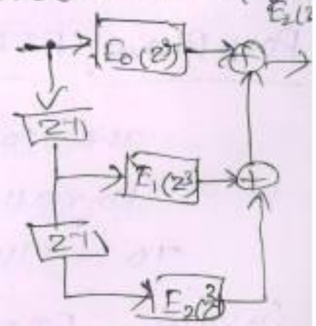


Fig. 3-branch polyphase

phase FIR Filter:

The Transfer function of a FIR Causal filter is given by

$$H(Z) = \sum_{n=0}^{N-1} h(n) z^{-n} \rightarrow ①$$

where, $h(n)$ is the impulse response of the filter. The Fourier Transform of $h(n)$ is

$$H(e^{jw}) = \sum_{n=0}^{N-1} h(n) e^{-jwn} \rightarrow ②$$

Which is periodic in frequency with period 2π

$$H(e^{jw}) = \pm |H(e^{jw})| e^{j\theta(\omega)} \rightarrow ③$$

where $H(e^{jw})$ is magnitude response and $\theta(\omega)$ is phase response.

We define the phase delay & group delay of a filter as.

$$\tau_p = \frac{-\theta(\omega)}{\omega} \text{ and } \tau_g = \frac{-d\theta(\omega)}{d\omega} \rightarrow ④$$

For FIR filters with linear phase we can ~~write~~ define

$$\theta(\omega) = -\alpha\omega \; ; \; -\pi \le \omega \le \pi \rightarrow ⑤$$

where α is a constant phase delay in samples.

Sub equ ⑤ in ④ we have $\tau_p = \tau_g = \alpha$, which means that α is independent of frequency we can write,

$$\sum_{n=0}^{N-1} h(n) e^{-jwn} = \pm |H(e^{jw})| e^{j\theta(\omega)} \rightarrow ⑥$$

which gives us.

$\ldots |H(e^{jw})| \cos\theta(\omega) \rightarrow ⑦$

By taking ratio of Equ (8) to Equ (9) we

$$\frac{-\sum_{n=0}^{N-1} h(n)\sin\omega n}{\sum_{n=0}^{N-1} h(n)\cos\omega n} = \frac{\pm|H(e^{j\omega})|\sin\theta(\omega)}{\pm|H(e^{j\omega})|\cos\theta(\omega)} = \frac{\sin\theta}{\cos\theta}$$

$$\frac{\sum_{n=0}^{N-1} h(n)\sin\omega n}{\sum_{n=0}^{N-1} h(n)\cos\omega n} = \frac{\sin\alpha\omega}{\cos\alpha\omega} \longrightarrow (9)$$

$\cos A \sin B = \sin(A-B)$

$$\sum_{n=0}^{N-1} h(n)\,\cancel{\sin\omega n}\cdot\cos\alpha\omega\,\sin\omega n = \sum_{n=0}^{N-1} h(n)\cos\omega n \sin\alpha\omega$$

$$\sum_{n=0}^{N-1} h(n)\sin(\alpha-n)\omega = 0 \longrightarrow (10)$$

Equation (10) will be zero when impulse response for symmetric

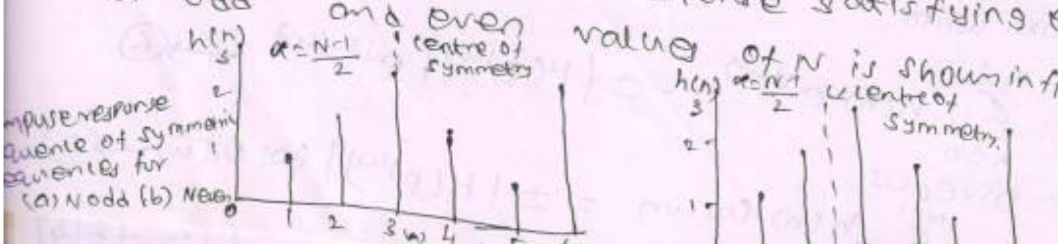
$$h(n) = h(N-1-n) \longrightarrow (11)$$

and

$$\alpha = \frac{N-1}{2} \longrightarrow (12)$$

$\therefore$ FIR filters will have constant phase and group delays when the impulse response is symmetrical about $\alpha = \frac{N-1}{2}$

The impulse response satisfying equ (12) for odd and even value of N is shown in fig.



3-4

when $N=7$ the centre of symmetry of the sequence is at third sample and when $N=6$, the filter delay is $2\frac{1}{2}$ samples.

If only constant group delay is required and not the phase delay we can write.

$$\theta(\omega) = \beta - \alpha\omega \longrightarrow (13)$$

Now we have

$$H(e^{j\omega}) = \pm|H(e^{j\omega})|e^{j\overset{\theta(\omega)}{(\beta-\alpha\omega)}} \longrightarrow (14)$$

$$\sum_{n=0}^{N-1} h(n)e^{-j\omega n} = \pm|H(e^{j\omega})|e^{j\overset{\theta(\omega)}{(\beta-\alpha\omega)}} \longrightarrow (15)$$

which gives us

$$\sum_{n=0}^{N-1} h(n)\cos\omega n = \pm|H(e^{j\omega})|\cos(\beta-\alpha\omega) \longrightarrow (16)$$

and

$$-\sum_{n=0}^{N-1} h(n)\sin\omega n = \pm|H(e^{j\omega})|\sin(\beta-\alpha\omega) \longrightarrow (17)$$

By taking ratio of eqn (17) to eqn (16) we get

$$\frac{-\sum_{n=0}^{N-1} h(n)\sin\omega n}{\sum_{n=0}^{N-1} h(n)\cos\omega n} = \frac{\pm|H(e^{j\omega})|\sin(\beta-\alpha\omega)}{\pm|H(e^{j\omega})|\cos(\beta-\alpha\omega)} = \frac{\sin(\beta-\alpha\omega)}{\cos(\beta-\alpha\omega)}$$

$\omega n - \beta + \alpha\omega$

$\beta - \omega n - \alpha\omega$

$$\sum_{n=0}^{N-1} h(n)\cos(\beta-\alpha\omega)\sin\omega n = \sum_{n=0}^{N-1} \sin(\beta-\alpha\omega)\cos\omega n$$

from which we obtain

$$\sum_{n=0}^{N-1} h(n)\sin[\beta-\alpha\omega-\omega n]$$

$$\sum_{n=0}^{N-1} h(n)\sin[\beta-(\alpha-n)\omega] = 0 \longrightarrow (18)$$

$\because \sin(90^\circ-\theta) = \cos\theta$

If $\beta = \pi/2$

$$\sum_{n=0}^{N-1} h(n)\cos(\alpha-n)\omega = 0 \longrightarrow (19)$$

Fourier Transform on both sides

$$Y(e^{j\omega}) = 0.25\, X(e^{j\omega}) + e^{-j\omega} X(e^{j\omega}) + 0.25\, e^{-2j\omega} X(e^{j\omega})$$

$$= X(e^{j\omega})\left[0.25 + e^{-j\omega} + 0.25\, e^{-2j\omega}\right]$$

$$\therefore H(e^{j\omega}) = \frac{Y(e^{j\omega})}{X(e^{j\omega})} = 0.25 + e^{-j\omega} + 0.25\, e^{-2j\omega}$$

$$= e^{-j\omega}\left(0.25\, e^{j\omega} + 1 + 0.25\, e^{-j\omega}\right)$$

$$= e^{-j\omega}\left(0.25\cos\omega + j0.25\cancel{\sin\omega} + 1 + 0.25\cos\omega - j0.25\cancel{\sin\omega}\right)$$

$$= e^{-j\omega}\left(1 + 0.5\cos\omega\right)$$

$$H(e^{j\omega}) = e^{-j\omega}\, \bar{H}(e^{j\omega}) \rightarrow ①$$

Comparing equation ① and ③ then we get

$$\theta(\omega) = -\omega,$$

The phase delay $\tau_p = \frac{-\theta(\omega)}{\omega} = \frac{\omega}{\omega} = 1$

The group delay $\tau_g = \frac{d\theta(\omega)}{d\omega} = \frac{-d}{d\omega}(-\omega) = 1$

The Fourier series method of designing FIR filters

The frequency response $H(e^{j\omega})$ of a system is periodic in 2π.

Any periodic fn can be expressed as a linear combination of complex exponentials

⇒ Desired frequency response of an FIR filters can be represented by Fourier series.

where, $h_d(n) \leftarrow$ Fourier co-efficients of Impulse response sequence

$$h_d(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} H_d(e^{j\omega})\, e^{j\omega n}\, d\omega \rightarrow (2)$$

The z-transform of the sequence is given

$$H(z) = \sum_{n=-\infty}^{\infty} h_d(n)\, z^{-n} \rightarrow (3)$$

eqn (3) represents a non-causal digital of infinite duration. transfer function.

To get an ~~filter~~ FIR filter can be truncated by assigning

$$h(n) = \begin{cases} h_d(n) & \text{for } |n| \le \frac{N-1}{2} \\ 0 & \text{otherwise} \end{cases} \rightarrow (4)$$

Then, taking z-transform of the above equation

$$H(z) = \sum_{n=-\left(\frac{N-1}{2}\right)}^{\frac{N-1}{2}} h(n)\, z^{-n} \rightarrow (5)$$

$$= h\left(\frac{N-1}{2}\right) z^{-(N-1)/2} + \dots + h(1)z^{-1} + h(0) + h(-1)z + \dots + h\left[-\left(\frac{N-1}{2}\right)\right] z^{(N-1)/2}$$

$$H(z) = h(0) + \sum_{n=1}^{\frac{N-1}{2}} \left[h(n)\, z^{-n} + h(-n)\, z^{n} \right] \rightarrow (6)$$

For a symmetrical impulse response having at $n=0$

$$h(-n) = h(n) \rightarrow (7)$$

Therefore equation (6) can be written as

$$H(z) = h(0) + \sum_{n=1}^{N-1/2} h(n) \left[z^{-n} + z^{+n} \right] \rightarrow (8)$$

The eqn (8) transfer fn is not physically re[alizable]

*) By multiplying eqn (8) by $z^{-(N-1)/2}$ we can m[ake] ... $(N-1) \leftarrow$ delay in samples

Design an ideal lowpass filter with a frequency response

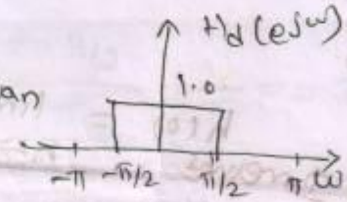
$$H_d(e^{j\omega}) = 1 \quad \text{for } -\pi/2 \le \omega \le \pi/2$$
$$= 0 \quad \text{for } \pi/2 \le |\omega| \le \pi$$

Find the values of $h(n)$ for $N=11$. Find $H(z)$. Plot the magnitude response.

Solution. Given $N=11$ (odd)

Step 1: To find $h_d(n)$

The frequency response of lowpass filter with cutoff frequency $\omega_c = \pi/2$ is given below.



From the freq response we can find $\alpha = 0$. So we can select non-causal filter co-effs symmetrical about $n=0$

i.e $\boxed{h_d(n) = h_d(-n)}$

or zero phase frequency response.

$$h_d(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} H_d(e^{j\omega})\, e^{j\omega n}\, d\omega$$

$$= \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} 1 \cdot e^{j\omega n}\, d\omega$$

$$= \frac{1}{2\pi}\left[\frac{e^{j\omega n}}{jn}\right]_{-\pi/2}^{\pi/2}$$

$$= \frac{1}{\ldots}\left[e^{j\pi/2 n} - e^{-j\pi/2 n}\right]$$

whe

$$\boxed{\therefore h_d(n) = \frac{\sin \pi n/2}{\pi n}} \quad -②$$

Step2: To find h(n):

Here N=11, so truncate $h_d(n)$

T 11 samples.

$$\therefore h(n) = \frac{\sin \pi/2\, n}{\pi n} \quad \text{for } |n| \le 5 \rightarrow$$ bec symm about n=0 so we consider −5 to ...

$= 0$ otherwise.

of

T for n=0 equation h(n) become indeterminate

can so.

$$h(0) = \lim_{n \to 0} \frac{\sin \pi/2\, n}{\pi n}$$

I method

To make $\lim_{\theta \to 0} \frac{\sin \theta}{\theta} = 1$ and $\lim_{\theta \to 0} \frac{\sin n\theta}{\theta} = n$ multiply ½ and divide ½

$$h(0) = \lim_{n \to 0} \frac{\frac{1}{2} \sin \pi/2\, n}{\pi \frac{1}{2} n}$$

$$= \frac{1}{2} \lim_{n \to 0} \frac{\sin \pi/2\, n}{\pi/2\, n}$$

$$= \frac{1}{2}(1)$$

at r

$$\boxed{h(0) = \frac{1}{2}}$$

(or)

II method: substitute n=0 in $h_d(n)$ equat

$\Rightarrow$, $h_d(0) = h(0) = \frac{1}{2\pi}\int_{-\pi/2}^{\pi/2} e^0 \, d\omega$

$= \frac{1}{2\pi}[\omega]_{-\pi/2}^{\pi/2}$

$= \frac{1}{2\pi}\left[\pi/2 - (-\pi/2)\right] \Rightarrow 1/2$

$\boxed{h(0) = 0.5}$

for n=1 sub in $h_d(n)$.

formula. $h(n) = \frac{\sin \pi/2\, n}{\pi n}$

$h(n) = h(-n)$

So.

$h(1) = h(-1) = \frac{\sin \pi/2}{\pi} = \frac{1}{\pi} = 0.3183$

similarly,

$h(2) = h(-2) = \frac{\sin \pi}{2\pi} = 0$

$h(3) = h(-3) = \frac{\sin 3\pi/2}{3\pi} = -\frac{1}{3\pi} = -0.106$

$h(4) = h(-4) = \frac{\sin 4\pi/2}{4\pi} = 0$

$h(5) = h(-5) = \frac{\sin 5\pi/2}{5\pi} = \frac{1}{5\pi} = 0.06366$

step3
To find ~~Transfe~~ fun ~~ium~~.

The Transfer function of the filte is given by.

$$H(z) = h(0) + \sum_{n=1}^{\frac{N-1}{2}} \left[h(n)(z^n + z^{-n})\right]$$

$$= 0.5 + \sum_{n=1}^{5} h(n)\,(z^n + z^{-n})$$

$= 0.5 + h(1)(z^1 + z^{-1}) + h(2)(z^2 + z^{-2}) + h(3)$