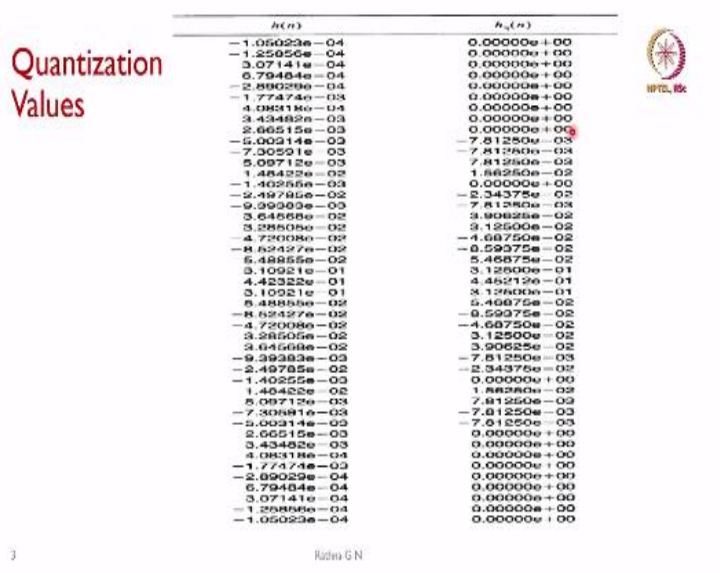
Real – Time Digital Signal Processing
Prof. Rathna G N
Department of Electrical Engineering
Indian Institute of Science - Bengaluru

Lecture – 14
IIR Filters - 3

Come back to the course real time digital signal processing. So, we are continuing with the IIR filter. Today we will see that quantization and then how it is going to effect our design basically.

As a recap, we have been discussing from past 2 lectures IIR filters you know how it is going to have effect on the quantization and then how to design it, why we have chosen IIR filter and then compared to FIR filter except linear phase we are going to have few number of filter coefficient which is going to be designed. So, that is the advantage of it and then we know that most of the earlier work goes on using analog filters.

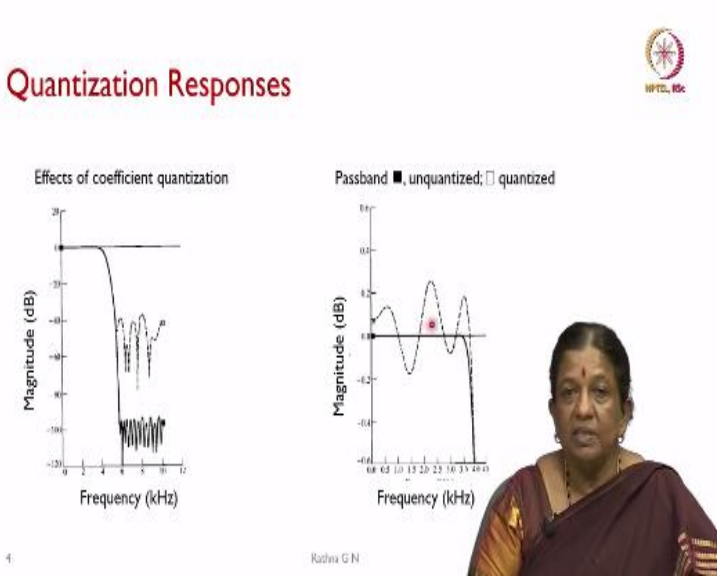
(Refer Slide Time: 01:16)



So, we will see what are the quantization values what we are going to get it is, here it is going to show me $h(n)$. So, you will be seeing the FIR filter continuous signal the values which are represented is -1.10^{-4} to you will be seeing that some more values are present. So, when I do the quantization of it, what we call it as $h_q(n)$ quantized thing. So, you will be seeing a threshold what we will be putting these are the values which are not going to be represented.

So, you will be seeing up to some of the very low values which are going to become 0 then rest of it are going to be quantized to certain values. So, you will be seeing in the end also. So, some of the coefficients are made 0 although they have the values. So, this is the quantization what it is going to happen so, in the FIR filter.

(Refer Slide Time: 02:18)

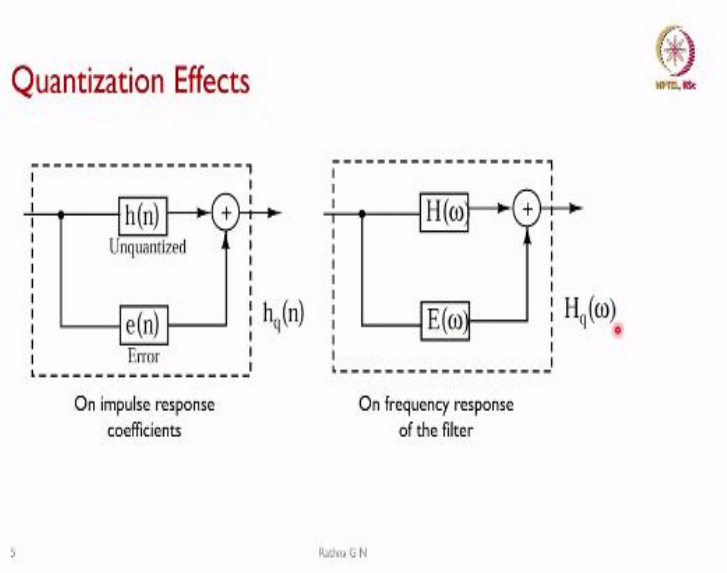


So, how the response is going to be we will see it in a while. So, you will be seeing that the black representation shows that this is the frequency response with respect to magnitude. So, you will be seeing based on the quantization which is shown in those dotted lines. So, you will be seeing that my cutoff frequency or stopband just gets moved up. So, from minus whatever it is around greater than -80 db. So, which will be dropped down to around -40 or 30 it is in the range -30 to 40 db. So, this is the coefficient quantization.

So, one has to consider when we are doing using in the hardware. So, our design has to account for this quantization effects. So, coming to this thing, this is the passband region what it has been taken. So, you will be seeing how it is going to be represented this is in terms of magnitude and this is in frequency in kilohertz. So, this is black one shows our unquantized and then the dotted lines it shows that it is the quantized.

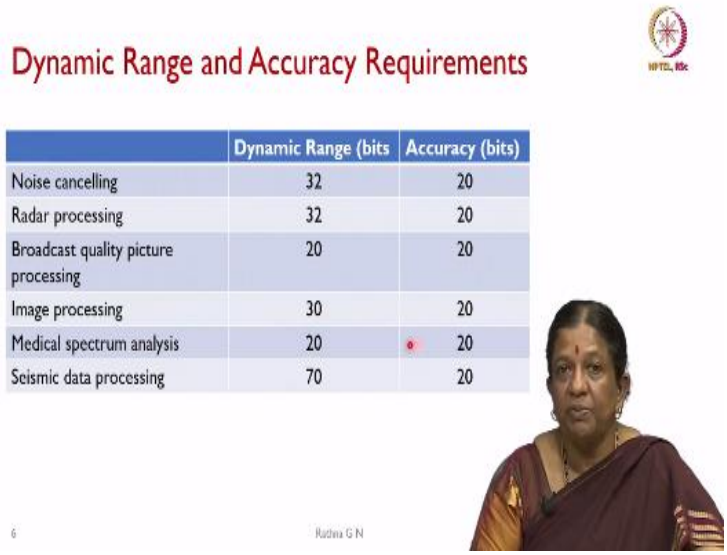
So, here I am supposed to get a flat response. So, where what is the thing is happening, so, there is a ripple in my passband in the representation. So, these are the effects one has to consider.

(Refer Slide Time: 03:54)



So, what is the thing is going to happen with this quantization. So, we will see in effect of it. So, whatever we have $h(n)$ if it is unquantized, then output is going to be a normal. So, if there is an error, because of the things of which is going to be added with respect to that so, output will be my impulse response coefficients what I will be getting it, so, in terms of frequency domain, we know that $h(n)$ is going to be represented as $H(\omega)$ and then our error function is represented as $E(\omega)$. So, output will be quantized basically $H_q(\omega)$ what will be the response of the filter.

(Refer Slide Time: 04:47)

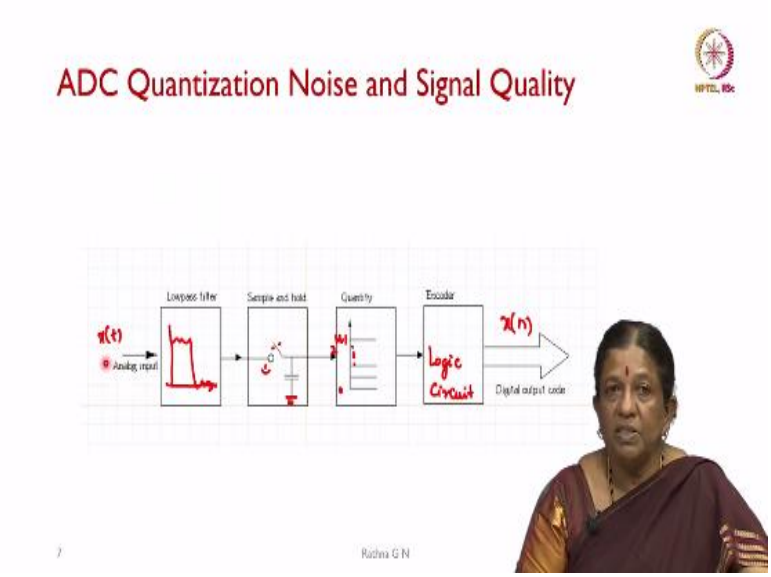


So, how it is going to affect we will see in this block diagram that is basically dynamic range and accuracy requirement of different applications one needs it. So, we know that in noise cancelling, so,

dynamic range in number of bits what will be telling 32 bits what we need it. So, whereas accuracy it is 20 bits is enough whereas, in the radar processing it is 32 and then same thing 20 bits, whereas in case of broadcast quality picture processing, so, both what we need is 20 bits.

So, we see in the image processing it is 30 and 20 so, medical spectrum analysis if you are doing it both can have 20. So, when you are going for the seismic data processing, you will be seeing that dynamic range what we need it 70 bits to 20 bits. So, most of the cases whatever we are designing 6713 DSK board what we are using it which is 16 bit, so, almost it is nearer to the thing what we can have.

(Refer Slide Time: 05:55)

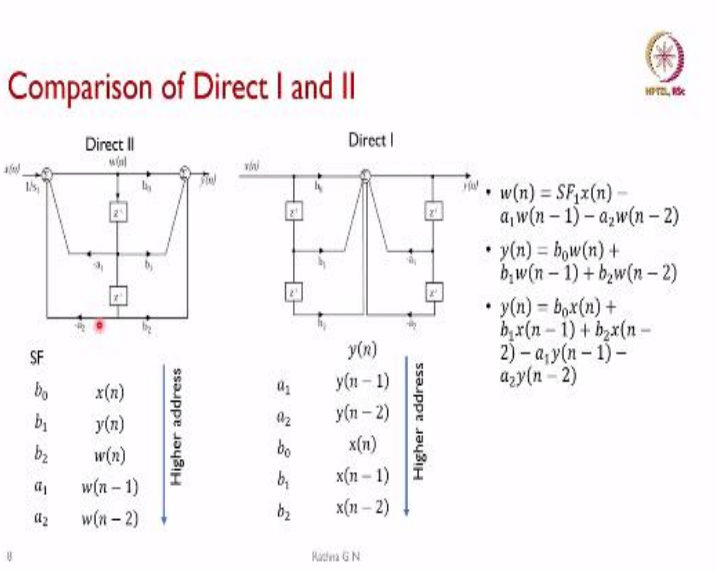


So, coming to quantization, so, all of us know that what we have to do is because most of the signal in nature is analog in nature. So, we represent it analog input as $x(t)$. So, if I want to do what is it band limit signal so, I have to pass it through the lowpass filter. So, we call this as a anti aliasing filter also. So, that will be limiting the stopband frequencies which are going to creep in, most of the stopband frequencies we say higher, the frequencies are the noise, so, we are limiting it, and then we will be feeding it into our sample and hold circuit.

So, when it is closed, you will be seeing that more than whatever the value of it. So, we will be getting the output, in sample and hold circuit is going to work then we are going to do the quantization. So, we know that depends on number of bits. So, the representation will be 0 to 2^{n-1}, n is the number of

bits ADC circuit what it is going to have. So, we have to do the encoding of this number, we use a logic circuit to do the thing and $x(n)$ will be our digital output coming out of our ADC.

(Refer Slide Time: 07:23)

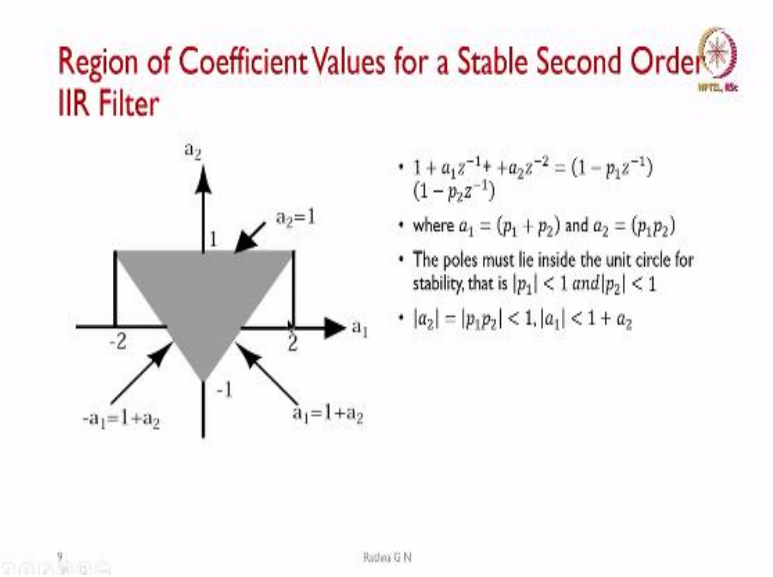


- So, we have seen the different structures here, we are going to see that this is my direct form 2 what we derived it in the last class. So, what are the coefficients for it is b_0, b_1, b_2 and a_1 and a_2 and then $x(n)$, $y(n)$, $w(n)$ and then $w(n-1)$, $w(n-2)$. So, these are the parameters what how to store it. So, when I see the direct form 1 structure, so, we will be seeing that this is a 0s are in the feed forward whereas poles are in the feed backward stages.

So, you will be seeing that I need so many coefficients to, coefficients are 5 in this case also, except that we need $y(n-1)$, $y(n-2)$output also in the case of it. So, that is $3+2=5$, but we need output to be fed back into the system. Whereas, as we can see here, I have one adder as I have discussed in the last class also, the quantization effect may nullify sometimes here, whereas, in this case it may little bit override.

So, that is the representation we have $y(n)$ as we know about it, it is $b_0x(n) + b_1x(n-1) + b_2x(n-2) - a_1y(n-1) - a_2y(n-2)$. So, in both the cases it is the same, but the delay elements here what we need is 4 delays whereas in this case it becomes only 2 delays in that case. So, that is the reason why most of the applications use direct form 2 structure compared to direct form 1 structure.

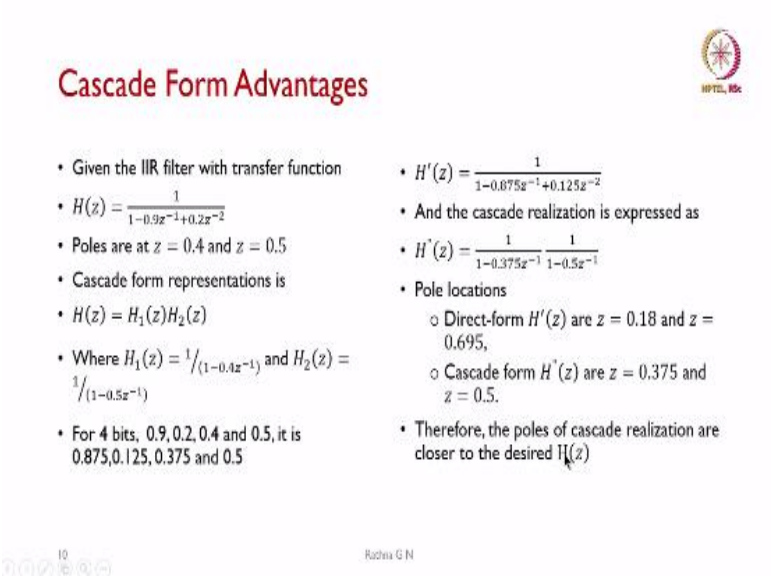
(Refer Slide Time: 09:33)



So, now we have to see that when we have taken the direct form 2 so how the coefficients are going to be represented whether they are going to cause any instability. So, we will be seeing that it is the coefficient values for a stable second order IIR filter. So, what should be the region of it? So, we say this is our coefficients a_1 and then a_2 what we are representing most of the time it is the poles which are going to have unstability for the filter. So, we will be considering a_1 and then a_2 coefficients what we have plotted.

So, we call this is the triangle which is going to be represented. So, we know that 1 is maximum, and then -1 in this, and then we will be having 2 for a_1 and -2. So, the rest of the thing represented $-a_1$ will this is equal to $1 + a_2$ location and then this is $a_1 = 1 + a_2$. So, how we arrived at this is shown with the poles here. So, we consider the feed backward section and then analyze it. So, what we have is $1 + a_1z^{-1} + a_2z^{-2} = (1 - p_1z^{-1})(1 - p_2z^{-1})$. So, we represent $a_1 = (p_1 + p_2)$ and $a_2 = (p_1p_2)$. So, we see that poles must lie inside the unit circle for stability. So that means that magnitude of p_1 should be less than 1 and magnitude of p_2 should be less than 1 in this case. So, magnitude of p_2 also has to be 1, then what happens magnitude of a_2 what it is given by magnitude of p_1 into p_2, so which is less than 1. So, a_1 as you can see it is $1 + a_2$. So, it will be magnitude of it can go up to 2 that is what what we get it from this triangle, this is how it has been plotted.

(Refer Slide Time: 12:00)



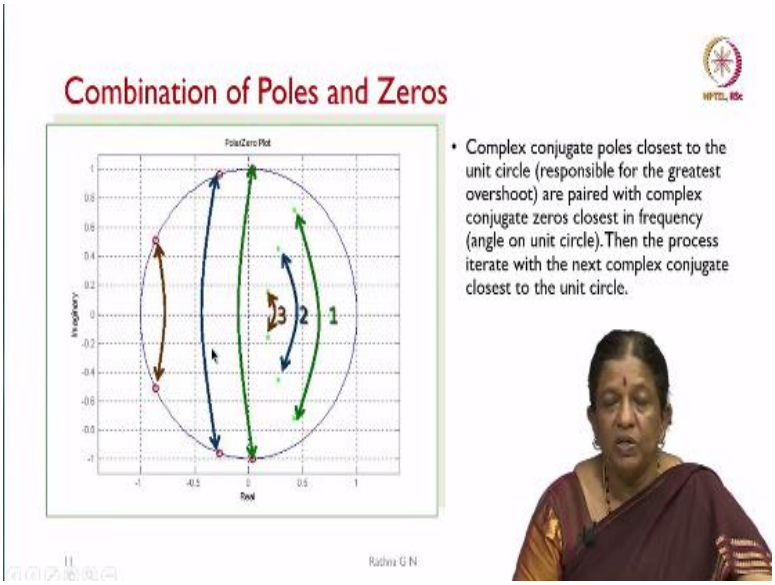
So, coming to the advantages of cascade section, we will see with an example here. So, impulse response of the filter in Zee domain is given as $H(z) = \frac{1}{1-0.9z^{-1}+0.2z^{-2}}$. So, if we consider the pole, then what happens $z = 0.4$ and $z = 0.5$. So, in the cascade form representation, we represent this as $H_1(z)H_2(z)$. So, what is it we will be having $H_1(z)$ is nothing but so ${}^{1}/_{(1-0.4z^{-1})}$, and then $H_2(z) = {}^{1}/_{(1-0.5z^{-1})}$.

So actually, this has been converted into $(1-0.4z^{-1})(1-0.5z^{-1})$ will realize this equation, that is how the poles are there. So, if we say that I am going to represent it with 4 bits, you have to call back our number system, then what happens so we have to represent all these values we will see in a while, that is 0.9, 0.2, 0.4 and 0.5. So, what is the nearest number I can represent with all these values? So, this is the direct form what I have taken, this is the second order section what we have taken the thing. So, 0.9 is represented as 0.875.

And then 0.2 is represented as 0.125. So, in the case of 0.4, we will represent it as 0.375, and then 0.5 directly is represented as 0.5 in our fixed form. So, coming with the thing, so, this is what we have done the second cascade section. So, what happens to the pole locations in that direct form, what we call it as because this is the quantized values what we are going to represented. $H'(z)$ is our $z = 0.18$ and $z = 0.695$. In the cascade form, what happens to our $H'(z)$ so, here it is going to be $z = 0.375$ and then 0.5.

So, you will be seeing that instead of point what we are supposed to get is 0.4 and 0.5. So, in the direct form, pole location is at 0.1 and 0.695. So, you will be seeing it that very far apart whereas in the cascade form, you are seeing that nearer to whatever the value is what we have represented. So that is the reason why we use the cascade form compared to the direct form, so, you will be seeing that the poles are nearer, the poles cascade realizations are closer to the desired are impulse response in the zee domain what it shows.

(Refer Slide Time: 15:14)



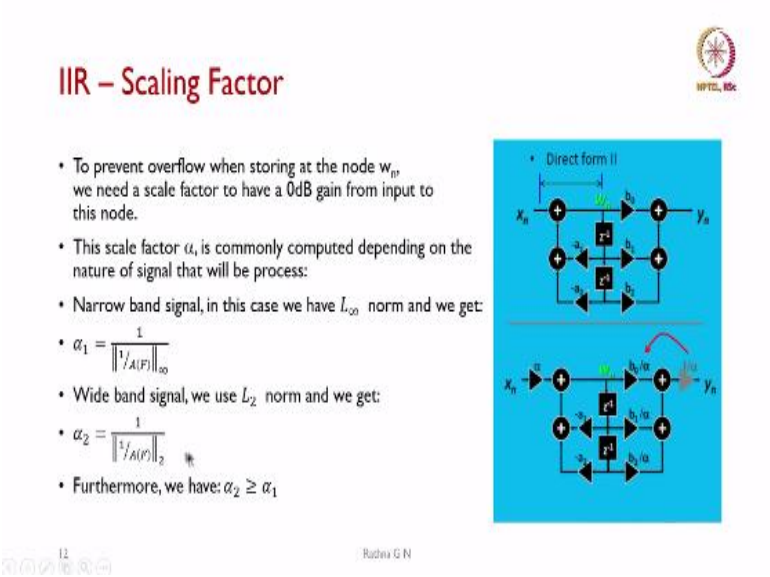
So, now, how because we are going to have many cascade sections, it is only second order 1 section what we have taken the thing if we have multiple sections, how we are going to combine because we will be arriving at poles and 0s. So, how we are going to combine them is shown by this diagram that is first what we do the thing complex conjugate poles what we are considering here. So, they have been named as first conjugate pole here and then the second set and this the third set. So, we see the closest to the unit circles.

So, that is which is closer to unit circle is named as 1 responsible for the greater overshoot are paired with complex conjugate 0s closer in frequency that is angle on unit circle. So, we paired this one with this one. So, that is the reason why we have colored them the first one which is shown with green is paired with this gray. The next one is the blue. So, that is what it says then the process iterate with

the next complex conjugate closest to the unit circle. So, this one with next 0s and then 1 is nearer to the 0 is connected with the nearer to the pole of 0s.

So, otherwise one has to do trial and error method or you have to see how the pairing is going to happen as we have seen in the last class. So, the quality factor how it is going to affect our pairing also is one of the important aspects. So, this is one of the methods what one can use, so, that we may get a good pairing of poles and 0s in our cascade section.

(Refer Slide Time: 17:14)

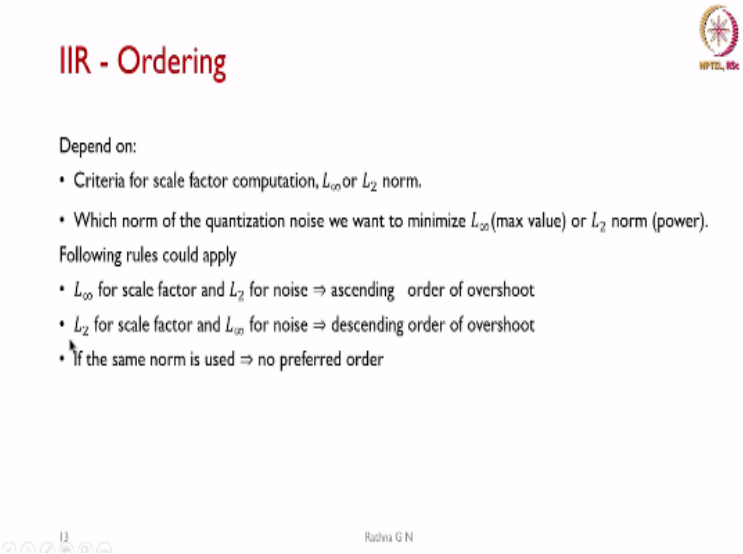


So, coming to the thing as we know that direct form 1 we said that there is no over flow it may or may not have it, but definitely we have in the cascade section, 2 adders which are coming in so, because it is the same what magnitude what it is going to be added or same sign magnitude magnitudes are going to be added so, I may have a overflow or an underflow. So, to avoid this overflow and under flow, we have to have a scale factor.

So, you will be a triggered with why did we choose a scaling thing that is barrel shifter what we call it either shift right or shift left which is required for scaling my input also in this case. So that is what what it is shown here to prevent our overflow in storing at node w n, we need a scale factor to have a 0 dB gain from input to this node. This scale factors what we call it as α_k is commonly computed depending on the nature of signal that will be processed.

So, for narrowband signal, in this case we have we use the L_∞ norm, then we get $\alpha_1 = \frac{1}{\left\| 1/A(F) \right\|_\infty}$ what we are taking it so, if we want to process the wideband signal, we usually use L_2 norm in that case, our $\alpha_2 = \frac{1}{\left\| 1/A(F) \right\|_2}$ input signal. So furthermore, we have $\alpha_2 \geq \alpha_1$ always this is at a higher distinct value compared to α_1 we are going to have the scaling this is the maximum scaling what we can provide.

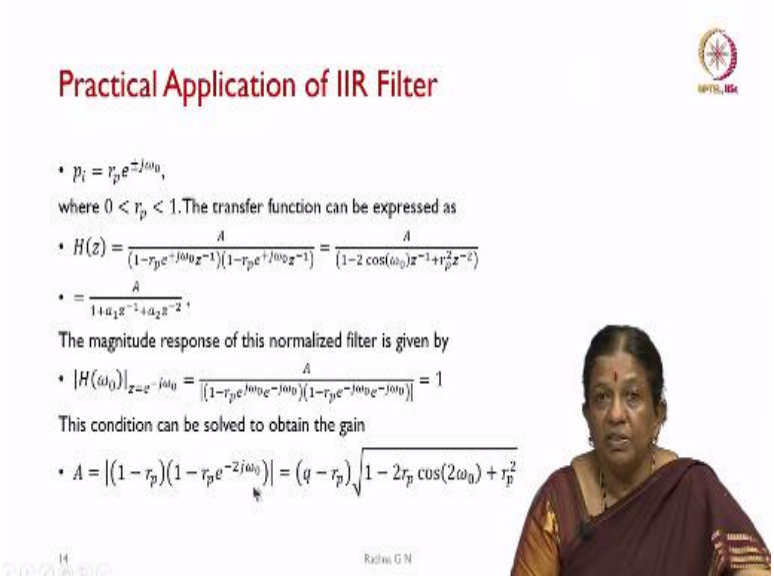
(Refer Slide Time: 19:31)



So, how we are going to order that is what we said criteria for scale factor computation. We are going to have it as L_∞ or L_2 norm one way one of the thing what we will be selecting it, so which norm of the quantization noise we want to minimize whether we want to L_∞ that is the maximum value or L_2 norm that is the power which is going to be considered. So, these are the rules one has to apply. So, L_∞ for scale factor, and L_2 for noise, that is we will be ordering the second order section ascending order of overshoot.

So, when we have L_2 for a scale factor, and L_∞ for noise, then we will be having descending order of overshoot, what we have to look into the thing based on this, we will be organizing our second order section in ascending order or descending order depending on one of these. So, if the same norm is used, so, there is no preferred order, so, you can go in whatever order you want to have it.

(Refer Slide Time: 20:44)

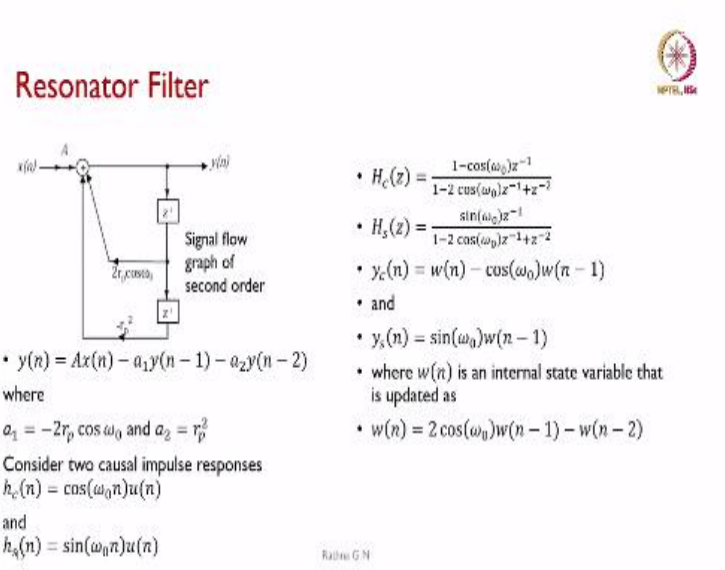


So, what is the practical application of IIR filter, so, here we have considered p_i is our row of pole positions, which is given by r_p. So, r is the radius of the circle into $e^{\pm j\omega_0}$, omega naught is the angular frequency what we are considering. So, we assume that r_p is in the range 0 and then 1 so, the transfer function can be expressed in that case as we have $H(z) = \frac{A}{(1-r_p e^{\pm j\omega_0} z^{-1})(1-r_p e^{\pm j\omega_0} z^{-1})}$. So, when you this is my conjugate poles what I have taken the thing.

So, we will be this is the second order sections what we are representing this the first one and this is a second one what we are multiplying it, so, then I expand the thing it becomes $\frac{A}{(1-2\cos(\omega_0)z^{-1}+r_p^2 z^{-2})}$. So, we represent these values as a_1 and then a_2. So, what happens in the frequency domain so, if we are considering the normalized filter the magnitude response is given by that is we will be substituting $z = e^{-j\omega_0}$ in this equation.

Then it becomes $|H(\omega_0)|_{z=e^{-j\omega_0}} = \frac{A}{|(1-r_p e^{j\omega_0} e^{-j\omega_0})|}$ and then the second term what you will be getting it so, this condition can be solved to obtain the gain that is what is my A is equal to so, the magnitude of this so, some quantized values $q - r_p$ what we will be assuming it into $\sqrt{1-2r_p\cos(2\omega_0)+r_p^2}$.

(Refer Slide Time: 23:07)



So, how we can represent this ok, so, this is my $Ax(n)$ is the input and then what I have is z^{-1} is the delay. So, the equation has in from the previous case what we are taking $2r_p \cos\omega_0$ here. So, that is what $2r_p \cos\omega_0$ is represented as a coefficient and this is my $-r_p^2$ is the other coefficient. So, which are fed into my adder and $y(n)$ is the output. So, we say that this is the signal flow graph of our second order section.

So, what will be $y(n)$ if I consider this is the filter, as the name says it is a resonator filter, I will comment on it in a while this is $Ax(n) - a_1 y_s(n-1) - a_2 y_s(n-2)$. So, where we have $a_1 = -2r_p \cos\omega_0$ and $a_2 = r_p^2$. So, consider 2 causal impulse responses, that is, $h_c(n) = \cos(\omega_0 n)u(n)$. So, the second one I can consider it as a sine function $h_s(n) = \sin(\omega_0 n)u(n)$.

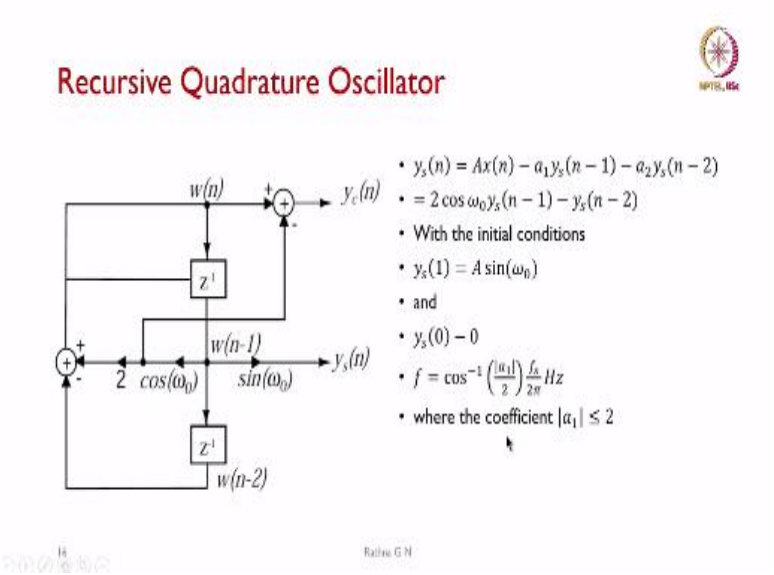
So, then the equation as you will be seeing that the cos function is given by this equation and then sine function is represented this way, then what happens to my $y_c(n) = w(n) - \cos(\omega_0)w(n-1)$ and $y_s(n) = \sin(\omega_0)w(n-1)$. So, we when we say $w(n)$ is an internal state variable that is updated as using this equation $w(n-2)\, cos(\omega_0)w(n-1) - w(n-2)$.

So, coming to the resonator so, all of us know that infinite impulse response why the name comes is once I removed the input, I can remove the $x(n)$, so, the system is going to oscillate on its own. So, what is the advantage of it, so, this resonating response what we will be using it in our generating

sine wave. So, we will see in the lab how we can now are giving only these 3 coordinates as we know that sine function is represented using the table.

We can do the thing or we derived the harness way of representing it by expanding in series or you can use the math function sine or cos function, all of us know that when we use these functions, it is going to take a long time to compute. So, with this IIR filter with the 2 feedback, we know that we can calculate the amplitude with the 2 input because it is LTI system usually we gave initially 0 then automatically by giving $x(n) = 1$ so, it automatically start resonating and then we can generate different kinds of sine frequencies and then we can use it in our sine generation.

(Refer Slide Time: 27:08)



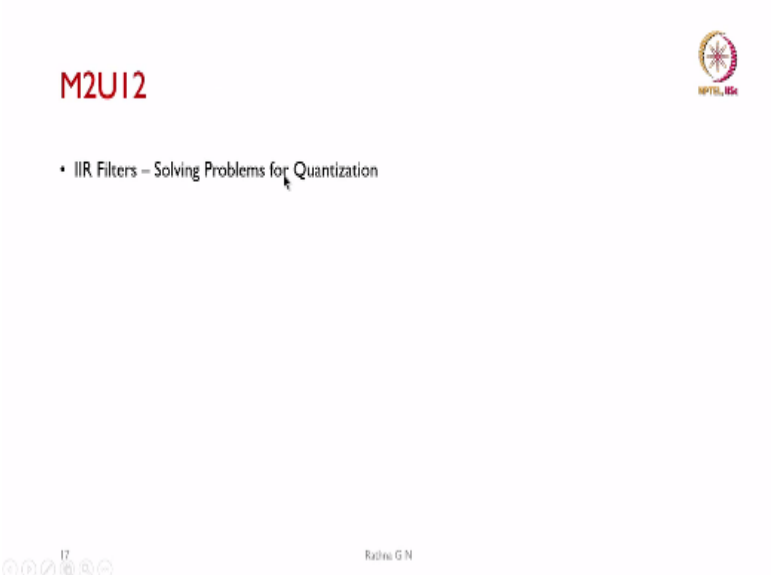
So, the next application is in the recursive quadrature oscillators. So, those who are comfortable with communication you know, where you need the quadrature oscillator. So, how we are going to generate this, so, you will be seeing this as a $w(n)$ and then $w(n-1)$ in this case also. So, you will be seeing that $2\cos\omega_0$ is the feedback signal what it is going and then feed forward what we have is a multiplication with $\sin(\omega_0)$ and you will be seeing 2 delays, this is $w(n-2)$ which you will be subtracting with the thing.

And you will be getting your cos and then here you will be getting the sine. So, how we are going to get it so, you will be seeing that $y_s(n)$ is nothing but $Ax(n)$. So, that is what you are getting the thing

$-a_1 y_s(n-1) - a_2 y_s(n-2)$ which is nothing but if you simplify it, it is going to be $2\cos\omega_0 y_s(n-1) - y_s(n-2)$. So, with the initial conditions, $y_s(1) = A\sin(\omega_0)$ and $y_s(0) - 0$.

Then I can calculate the frequency of it as $\cos^{-1}\left(\frac{|a_1|}{2}\right)\frac{f_s}{2\pi}Hz$, this is the frequency what I will be generating it where the coefficient $|a_1| \leq 2$ which we got it in our triangular representation.

(Refer Slide Time: 28:58)



So, this completes our IIR filter for with respect to structure and then what is the quantization effect. So, we will be solving some of the problems how our centre frequency is going to change with quantization. So, without quantization at with number of bits that is represented so, we will be seeing in our next class. Thank you, happy learning in this class.