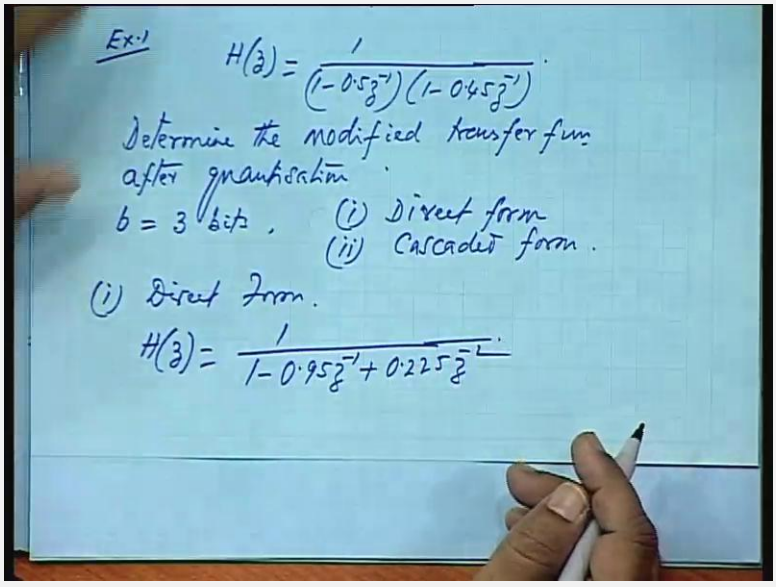
Digital Signal Processing
Prof. T.K. Basu
Department of Electrical Engineering
Indian Institute of Technology, Kharagpur

Lecture - 28
Effects of Quantization (Contd.)

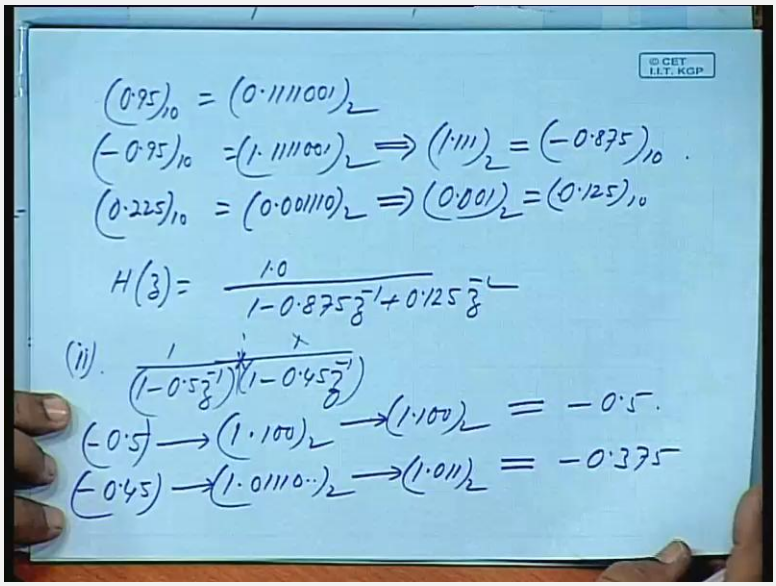
Will be taking up a few numerical problems on the Effects of Quantization and rounding off, and what are the other types of problems that we encounter while quantizing in a digital system.

(Refer Slide Time: 00:57)



I will start with a simple example first, consider a transfer function H z simple second order system say it is like this. So, determine the modified transfer function after quantization these parameters 0.5 and 0.45 take b equal to 3 bits and take first a direct form and then a cascade form. In a cascaded form these are the two factors and the direct form you have to take the product get the quadratic form. So, let us see first we consider the direct form, what will be H z like 1 minus 0.95 z to the power minus 1 plus 0.45 into 0.5, so 0.225 z to the power minus 2 agreed, so let us quantized these two quantities.

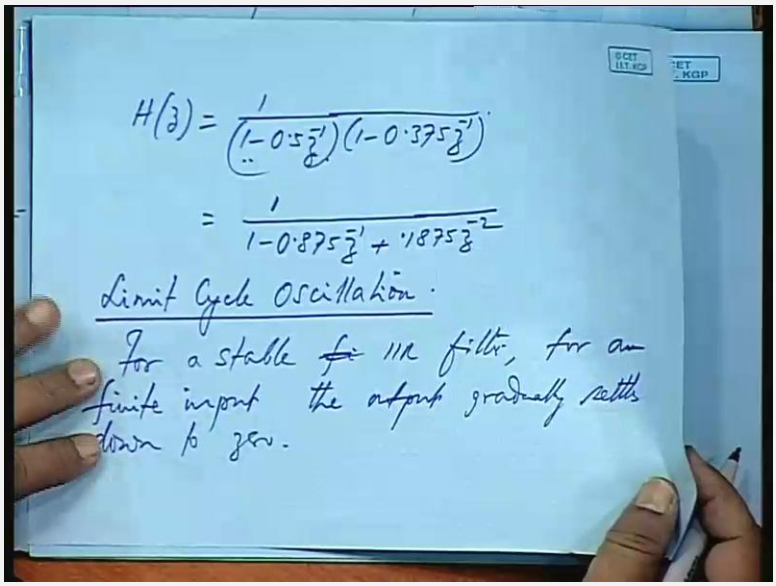
(Refer Slide Time: 03:01)



So, 0.95 10 is 0.1111001 check whether this is, so minus 0.95 will be 1.1111001, now if we truncate it this would be 1.111 and that gives me it is value is minus 875 agreed 0.225 similarly will be 0.001110, after truncation that will be 001 that is equal to 0.125. So, H z therefore, be approximated as 1 minus 0.875 z to the power minus 1 plus 0.125 z to the power minus 2.

So, in the direct form this function has been modified to this, then in the cascade form it is 1 by 1 minus 0.5 z to the power minus 1 in to 1 minus 0.45 z to the power minus 1. So, I can take this as h 1, this as h 2, so I have to approximate these two quantities 0.5 if I take minus 0.5 it will be 1.100. So, that will be in 3 bit approximation remaining as it is, so that is equal to minus 0.5 minus 0.45 will be 1.01110 see, where it ends, how it goes. So, approximation is 1.011, so that gives me minus 375 if I make 3 bit approximation just truncation.

(Refer Slide Time: 06:15)



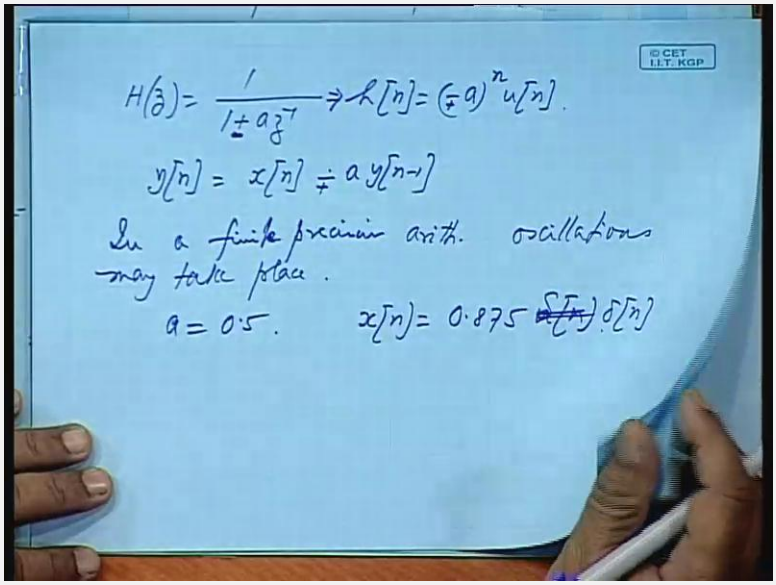
So, this will become H z becomes 1 by 1 minus 0.5 z inverse in to 1 minus 0.375 z inverse. Now, if I multiply what kind of denominator do I get 1 minus 0.875 z inverse plus 0.375 half of that 1875 z to the power minus 2. Now compare the two with the original one ((Refer Time: 07:04)) original one was 1 minus 0.95 z inverse plus 0.225 z to the power minus 2.

If I approximate this in the direct form it is 1 minus 0.875 z inverse plus 0.125 z to the power minus 2. And in this case, it is 1 minus 0.875 z inverse plus 0.1875 z to the power minus 2, so this is this coefficient is closer to this one as compared to this one, the other one is distorted by the same amount. Now, one thing is clear, if you factorize it then some of the factors may remain as it is, so in the cascaded form we have discussed earlier also. In cascaded form in each block the particular root will be sensitive to only one coefficient or the coefficient associated with only that block when you are rounding off or truncate, it will not be affecting the other poles or 0s.

So, this is the advantage of cascaded form you try to restrict the sensitivity of the roots the dependence on the parameter truncation, within that particular block. Next, will take up the issue of limit cycle oscillation, limit cycle oscillations are because of the truncation. We have seen in case of say in case of parameter truncation or rounding off there will be some errors, we represent specially when you are multiplying a quantity by a constant, the end product will be quantized along with an error.

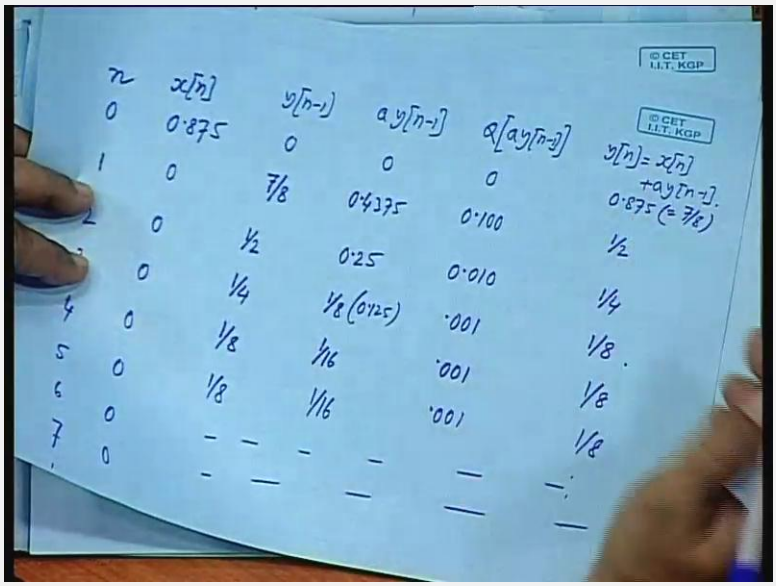
Now, if you are having a stable filter and if you are giving a finite input, after sometimes say in an IIR filter, after sometime the output will gradually come down to 0 for example, if you test it with an impulse. So, impulse will give you a response which will gradually come down to 0 if it is a stable system, but if you are having a truncation, then they will be a of kind of a feedback, it is something like a feedback, so there will be a sustained oscillation I will give an example, it will be clear. So, for a stable filter say IIR filter for a finite input say an impulse for a finite input, which is gradually digging to 0 it may be or an impulse, the output gradually settles down to 0 this is for an ideal filter.

(Refer Slide Time: 11:08)



Now, if you have truncation let us see what happens H z 1 plus a z inverse, what will be h n corresponding h n is minus a to the power n u n. And we can write the difference equation as x n minus a times y n minus 1 had it been minus I have got plus here, say if I have minus here, then it will be minus plus minus plus. So, in a finite precision arithmetic oscillations may take place. Let us take some values of a say let a b 0.5 and let us excited with a finite, so let us take delta n with a finite input 0.875 delta n that is an impulse input, what will be the response.

(Refer Slide Time: 12:59)



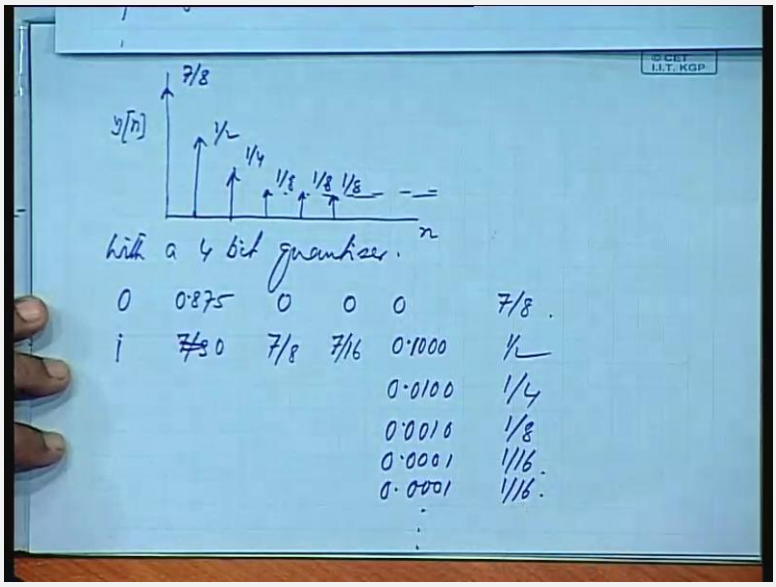
Let us tabulate the response all the components, it will be clear this is n 0, this is x n which is 0.875, then y n minus 1 which is 0 a times y n minus 1 that is also 0, we have started with initial value 0. Then quantized value of this quantity a y n minus 1 that is also 0, then y n which is x n plus a times y n minus 1. Let us see what will be the final output, in the first case it is 0.875; one may write this as 7 by 8, next n is equal to 1 this is 0 it is a delta input, so next moment it is 0, 2, 3, 4, 5, 6, 7, and so on, so these are all 0's.

Let us compute this one y n minus 1 is how much, this quantity then a times this it will be half of this 7 by 16. So, that is 0.4375. And if I quantize it by a 3 bit quantizer, then this will be I will write the quantized value or rounded off value 0.100, if we round it off it will be 0.100. So, that will be equal to half if you round it off, so you can add if it is on the higher side as you round off a decimal quantity, so it will be 0.100 it will go to the other side.

And then this one is half 0 half, then half of that is 0.25 or 1 by 4 this will be 0.010, this will be 1 by 4, next this will be 1 by 4 this will be 1 by 8. Either, you write in a decimal form does not matter or fraction form and what will be this approximated as in a 3 bit quantizer. So, it is 1 by 8 now see the fun 1 by 16, if you take rounding off may be 001, then 1 by 8, again 1 by 8 1 by 16, 001, 1 by 8 and it continues 1 by 8. So, it gives me a steady output of 1 by 8, what was this looks like the same thing continues.

(Refer Slide Time: 17:14)

(Refer Slide Time: 18:35)



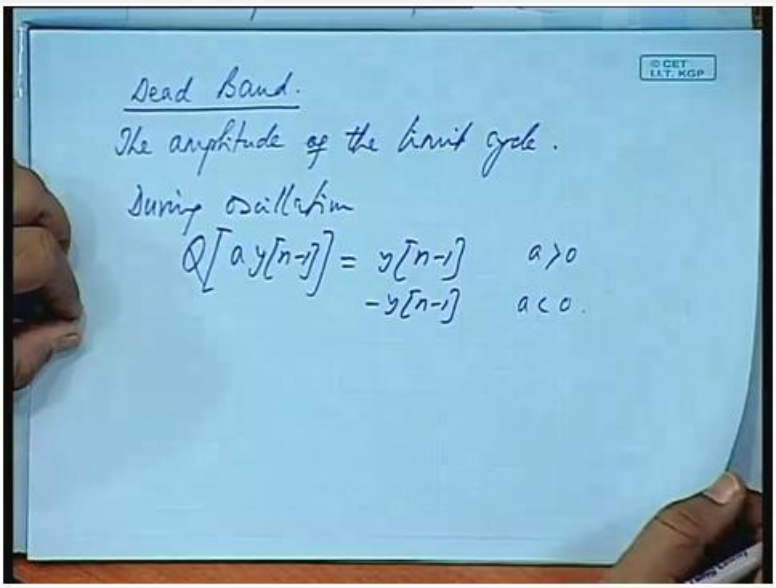
So, if you look at the output function n y n, you started off with 7 by 8 this is not to the scale. Next you show, it was half, next it was 1 4 th. Then 1 by 8, 1 by 8, 1 by 8 it is not decreasing progressively afterwards, it is remain in steady here. Now, if you have a 4 bit quantizer, you verify for yourself you will find will be following a similar kind of a of a table, but this value will be reduced and 0.875, then 0, 0, 0 has same columns I am writing and then last 1 is 7 by 8. Next it will be 1 this one will be 0, this is 7 by 8, it will follow this, then 7 by 16 or 0.4375 7 by 16 and then that will be 0.1000 half.

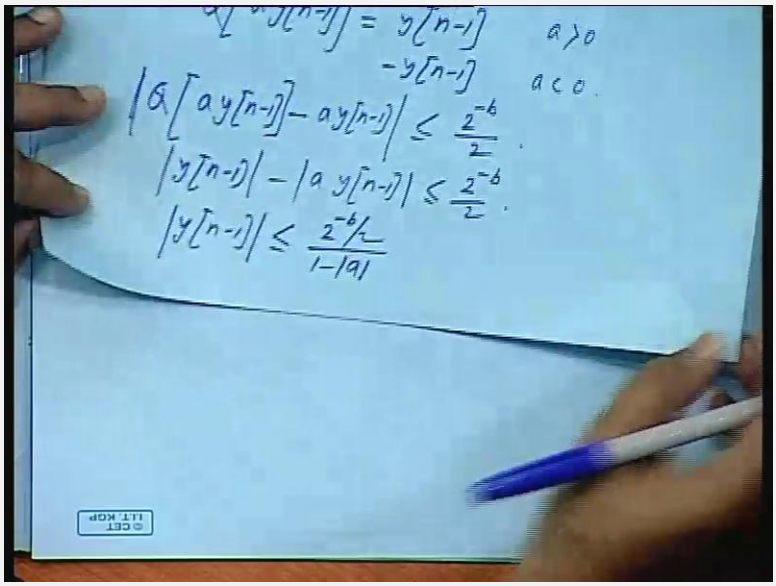
Then follow the same sequence this will become 0.0100 1 by 4 is that next 0.0010 1 by 8 next it will be 0.0001 1 by 16 and after that it will be always 0001 it will be oscillating with a magnitude 1 by 16. So, the oscillation will continue, but with reduced magnitude, so per bit you get the magnitude of their oscillation halved per additional 1 bit. If we change the sign of a it will be alternating between plus a 1 by 8 and minus 1 by 8, you can see for yourself in the earlier table ((Refer Time: 20:20)).

You just put minus 1 by minus a in to y n minus 1 and then see the effect, it will be oscillation with 1 by 8 and minus 1 by 8. For a speech signal for example, when we have a silence period there should not be any output, suppose the signal is discritised that is we take a digitized signal and then we filter with a digital filter. And then again we convert it back to a continuous signal by d to a converter, then in that silence period you

are having a study oscillation. So, it will generate a speech with a noisy period with a noisy sound, when the otherwise period should have been with a 0 signal, that is a silence period will also be containing some noise, so it is not desirable.

(Refer Slide Time: 21:20)

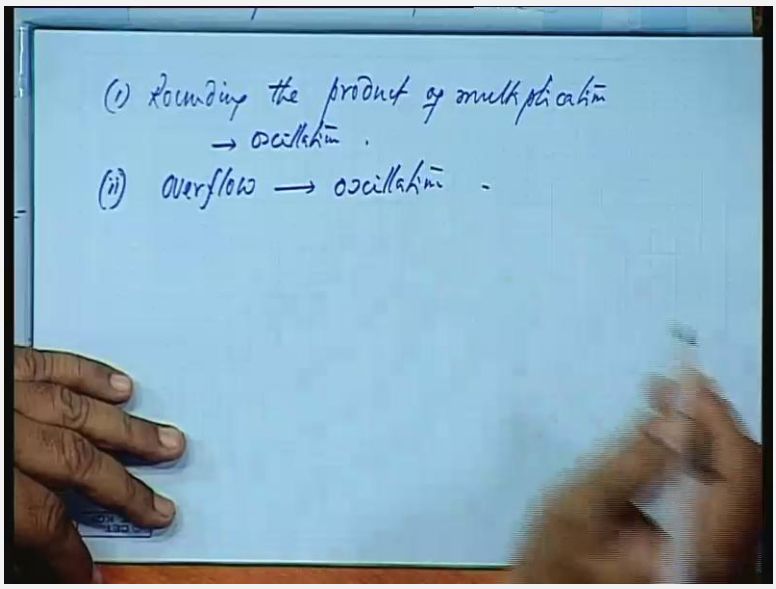


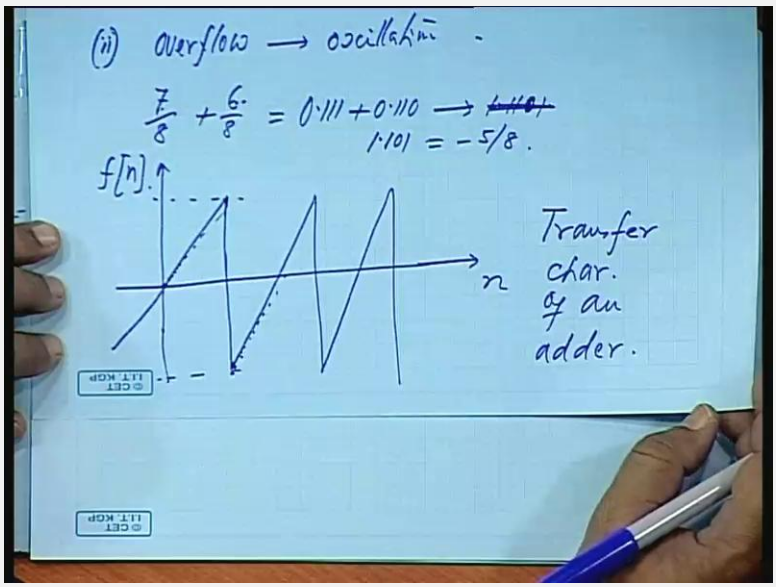


Next, we take up dead band, the amplitude of the limit cycle of the amplitude is limited to a particular band, say in the last example it was either 1 by 8 or 1 by 16, so we call it dead band. So, during oscillation the quantized value or say a in to y n minus 1 in the previous example because, ((Refer Time: 22:16)) for a delta input x n does not come in to picture, after n is equal to 0. So, you are quantizing only this one, when the input has vanished, then also it is oscillating; that means, it is only this quantity a in to y n minus 1.

And this quantized value is y n minus 1 or minus y n minus 1 is it not when a is greater than 0 or a is less than 0. So, quantized value of a y n minus 1 minus a y n minus 1 should be less than 2 to the power minus b by 2 or y n minus 1 magnitude minus a times y n minus 1 magnitude should be less than 2 to the power minus b by 2 or y n minus 1 magnitude should be less than equal to 2 to the power minus b by 2 divided by 1 minus magnitude a. So, this is the dead band for a first order IIR filter for a first order filter, this is the dead band.

(Refer Slide Time: 24:33)





So, rounding of the product of multiplication that causes oscillation, now there are other types of oscillations due to overflow, this is because of the rounding off there can be also

overflow. So, when you are having an overflow, you will have to control the magnitude will see that later on when the sum off…so that is an oscillation due to overflow see 7 by 8 plus 3 by 4 or 6 by 8 this is 0.111 plus 0.110. So, that gives me 1.1101, 0.5, 0.2, 4 by 8, 6 by 8, 7 by 8, so this is 1 is this all right.

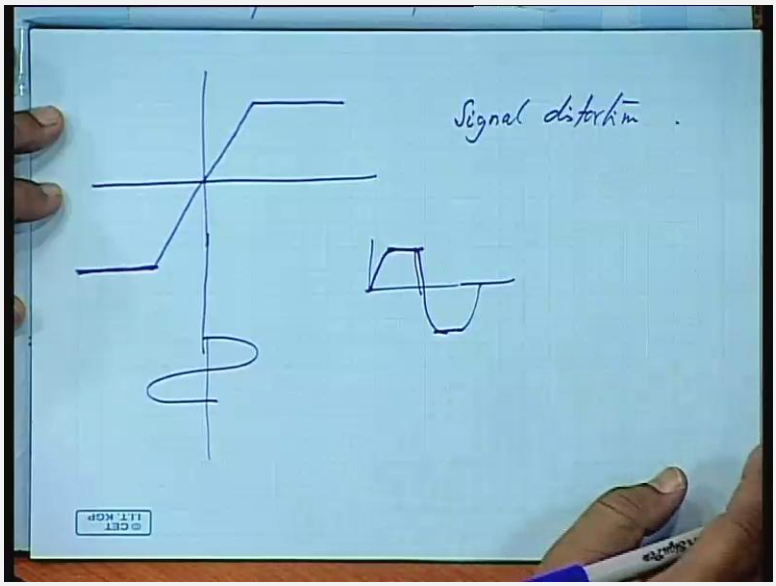
Student: ((Refer Time: 26:32))

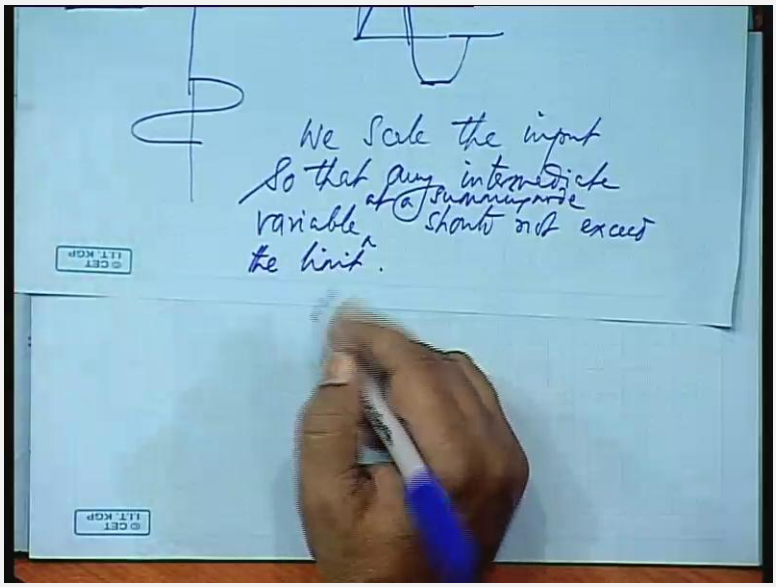
1 point.

Student: 101.

101 that is equal to minus 5 by 8, so you can see addition of two quantities is giving me a different answer it is like this. So, it is an adder if you keep on adding quantities, the total sum goes up, suddenly it becomes negative, whenever it exceeds 1 this is 7 plus 6 is 13, 13 by 8 if it is up to 8 by 8 it will go up to this point. The moment it becomes 9 by 8 the overflow takes place, so it becomes negative quantity, so like that and in the adder if you keep on adding again some quantities it will go on increasing again it will drop, so this is a kind of response you get. So, this is the dead band it will be oscillating between these two quantities, this is the transfer characteristics of an arc adder, so the total output varies between minus 1 and plus 1, so how do we modify this.

(Refer Slide Time: 28:43)





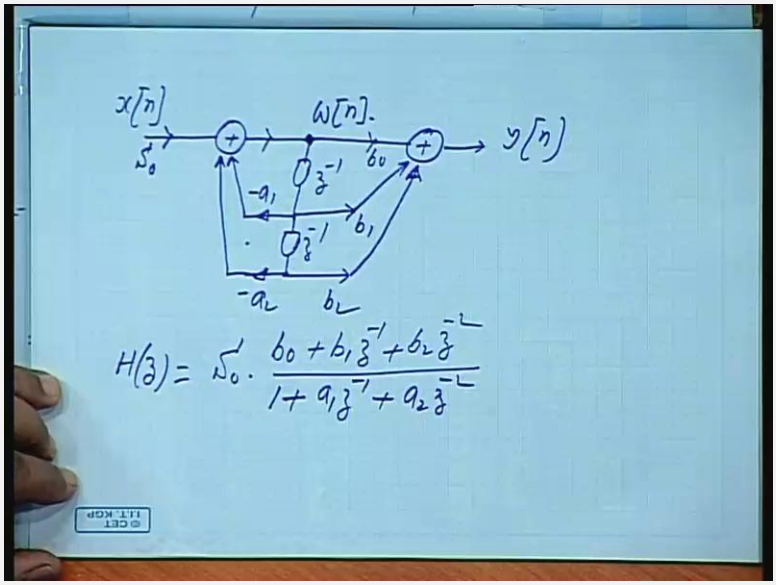
So, we try to modify the characteristics by changing in to this form, that is you clip it, it get saturated at one. So, two quantities when they are added 7 by 8 plus 6 by 8 is equal to 13 by 8 will chop it of will store it as 1, otherwise it will be stored as a negative quantity. So, similarly on the negative side, if it is going up to minus 1 then beyond minus 1 will take it as minus 1, so when you are chopping it there is a lot of error is it not, so that error again will cause problem.

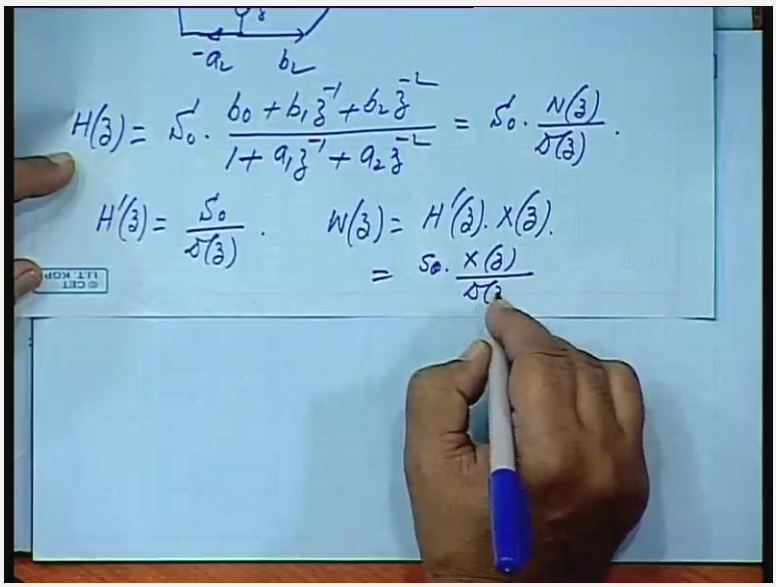
So, there will be signal distortion you have seen non-linearity's of this kind you know a for example, in a regular in an analog signal, if you feed a sinusoid what will be the corresponding output it is clipped here is it not. So, there is a distortion, so what is a

remedy we try to reduce the magnitude, so magnitude you multiply by a scaling factor, so that it oscillates within this limit, the variation takes place only within this. So, whenever you are designing a digital filter, you have to see for a given input at any intermediate stage, the overall output should be restricted within this range.

So, we should have a scaling factor a multiplier to control the inputs such that at any intermediate point, the output should not exceed plus 1 or minus 1. So, we scale the input, so that any intermediate variable at any summing note should not exceed the limit plus minus 1 that is there should not be any overflow.

(Refer Slide Time: 32:00)

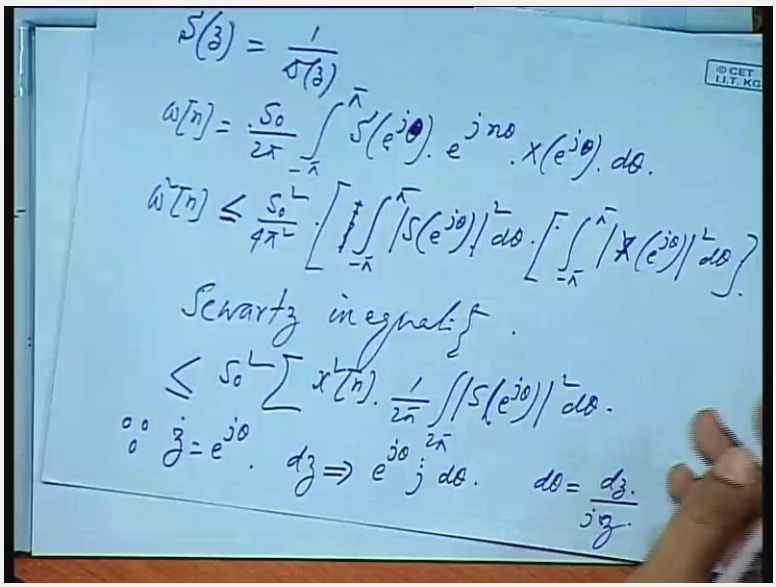




So, x n for example, if we have a second order filter a 1 again z inverse a 2, we multiply by a factor s naught, this is a scaling factor s naught, here we have b 1 and b 2 this is y n this intermediate variable let us call it as w n. So, H z is this is a constant S naught in to b 0, this is b 0 plus b 1 z inverse plus b 2 z to the power minus 2 by 1 plus a 1 z inverse plus a 2 z to the power minus 2 any question I should have put minus a 1 here minus a 2.

So, I can write this as S naught in to some N z by D z, let us take H dashed z is equal to S naught by D z, S naught by D z that is the denominator. So, this will correspond to input x z if I multiply by H dash z what would be the output, if I multiply the input x z by H dash z it is that intermediate output W z. So, W z is equal to H dash z in to x z is that all right. So, x n is multiplied by S naught and then only this block will give me the output here. So, this the intermediate variable I want to check whether this is within that restricted value, this does not cause any overflow, so what should be the relationship between S naught and H dashed z.

(Refer Slide Time: 35:33)



So, s z ((Refer Time: 35:43)) s naught, we have taken this let me write S naught in to x z by D z S naught by D z in to x z. So, S z let us call it 1 by d z as S z then omega n is S naught by 2 pi minus pi to plus pi S e to the power j theta s e to the power j 0 e to the power j n theta in to x e to the power j 0 d theta is this all right? Do you agree w n will it look like this S naught is a constant and this s a let me write this as because, there are scopes for confusion one is S naught, the other one is s z s naught is a constant.

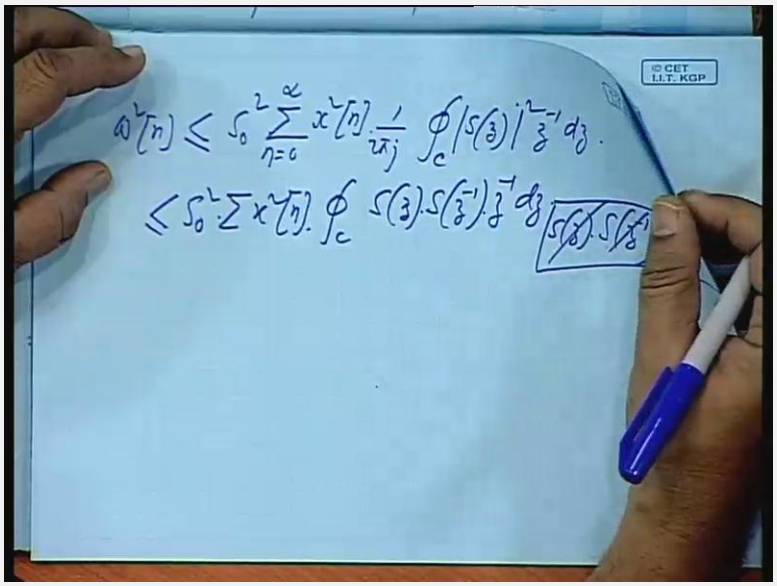
Whenever S is put with an argument that is basically S z, so S z z evaluated at 0 at omega equal to 0 no I does not give you, it should be e to the power j theta j n theta this is also theta not 0. ((Refer Time: 38:12)) So, x z in to S z, so basically S naught by d z is x z S z, so it is S z in to x z, so this quantity I have taken the inverse transform to get the time domain response.

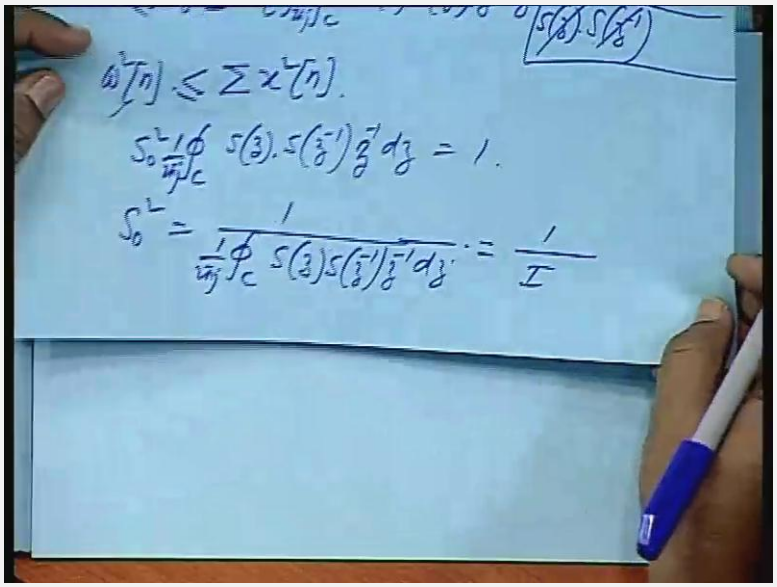
So, omega squared n will be s naught square by 4 pi squared check whether this is all right, integration minus pi to plus pi s e to the power j theta squared d theta, it should be less than this one if I square it, it should be less than square of this, square of this because, this is always less than equal to 1. So, if I square it up it should be multiplied by minus pi to plus pi x e to the power j theta square d theta, do you all agree I can always take the product of the magnitude square, integrated for the two term separately.

And that is equal to this is known as Schwartz in equality, now what is this basically square some of all the frequency component is it not. So, you apply pursuers theorem S naught square this 1 by 2 pi 1 by 2 pi will go there. Actually I should not have taken 4 pi square here I should have put 2 pi inside and then square individually, that would have been better any way. So, what I wanted is summation of this square integration of square of this, square of this is same as summation of the corresponding time domain term square again for this also.

So, product of these two if we take it will become sigma x squared n for this one and for this one it will be 1 by 2 pi integrated over 2 pi s e to the power j theta square d theta, this I have written as it is this one I have written in this form. So, because z is equal to e to the power j theta d z is e to the power j theta in to j in to d theta, so d theta will be d z by j times z, so we can write this integration.

(Refer Slide Time: 42:46)





As I rewrite it omega squared n is less than s naught square summation x square n 1 by 2 pi j, this if you remember s z square z to the power minus 1 d z ((Refer Time: 43:25)) I have just replace z by e to the power j theta. So, it becomes like this j in to z, so that becomes z inverse. So, this becomes this limit of 2 pi becomes a close path of this, now this one is applied that residue theorem basically this is s z in to s z to the power minus 1, whenever we take magnitude square it is this.

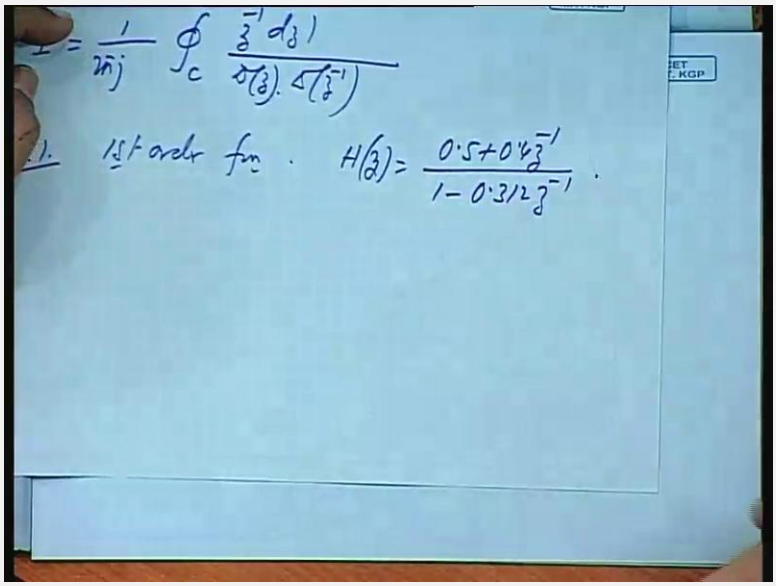
And what is this s z in to s z minus 1 in to z inverse d z, it basically residue within that circle and if s z has a pole inside, then s z minus z inverse will have a pole outside, whose residue will be 0, so will have to take this one. So, omega squared n therefore,

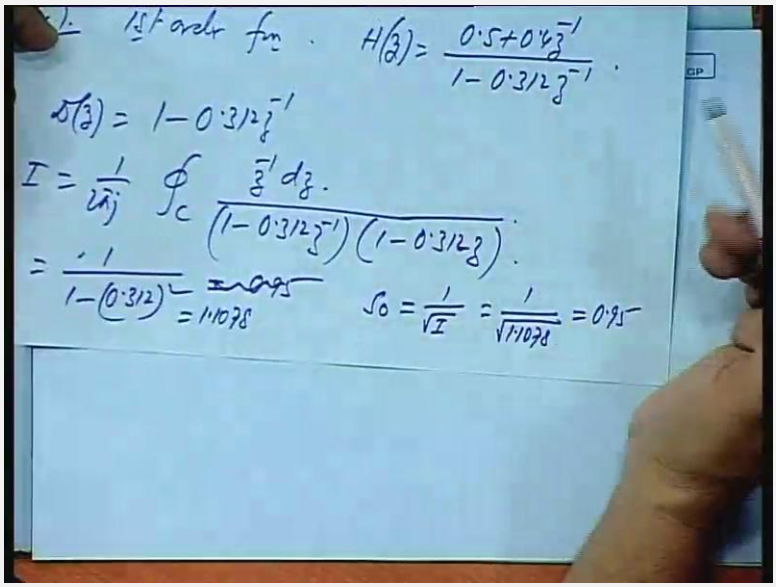
whatever be this omega square n should be less than sigma x square n, when s naught square in to this one s z s z inverse d z is equal to 1.

What does this mean omega square n; that means, at any instant the value of the signal should be less than the energy content in a signal, if it is restricted, if it is less than the total energy of the input signal, that can be met by putting this condition then there would not be any overflow. What does it physically mean, can you conceive of the situation when you are say it is like this, you are try to push something, you are applying a force. If I want at the see the total amount of energy that you have spend, if it is more than the energy of any at any point, if you measure it as a variable may be pressure or any quantity displacement.

If it is always less than the input energy, then it will not have any overflow, it is something like if you have something oscillating, sometimes the energy of a variable that is energy associated with a signal in the intermediate stage may go up. Then there is an overflow, but there would not be any overflow if it is always less than the input energy. So, that condition can be met if this is equal to made 1, so s naught squared is equal to 1 by this quantity.

(Refer Slide Time: 47:43)





So, I call this integration 1 by 2 pi j there was a 2 pi j missing somewhere here, ((Refer Time: 47:52)) we have met a small slip 2 pi j, so 1 by 2 pi j. So, this integral I call it I, so s naught square should be less than this, where I is equal to 1 by 2 pi j z inverse d z what is s z, this 1 by d z is it not? If you remember I have taken S z as 1 by d z. So, it is d z in to d z inverse, so let us consider a second order system a first order system, let us take a first order system H z is equal to 0.5 plus 0.4 z inverse divided by 1 minus 0.312 z inverse.

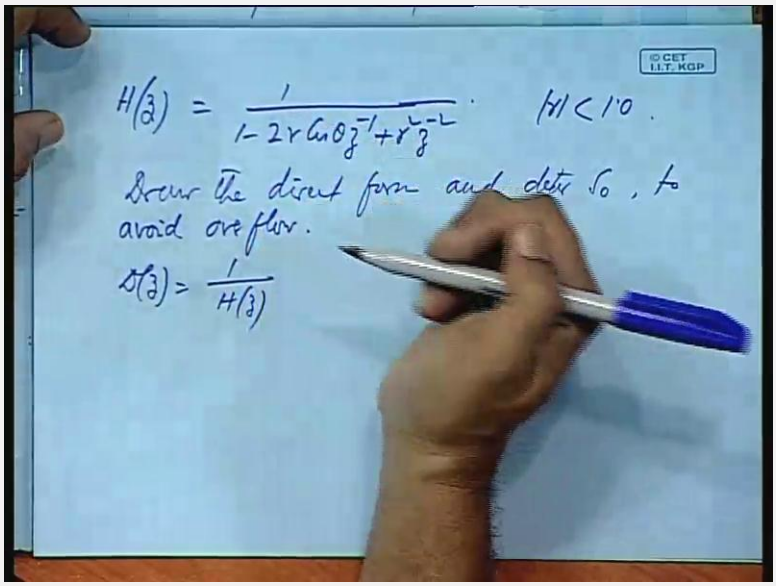
Now, what will be D z 1 minus 0.312 z inverse this one, so what is the integration I 1 by 2 pi j z inverse d z divided by capital D z it is this 1 minus 0.312 z inverse 1 minus 0.312

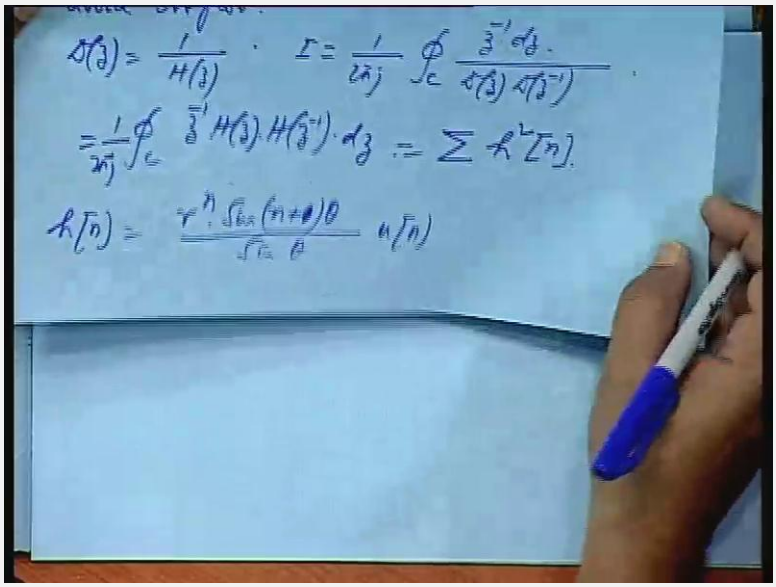
z. Now, this is the residue at this point z is equal to 0.312 and z is equal to 1 by 0.312 for 1 0.312 it will be 0. So, you take sum of residues, so it will become 1 by finally, 1 minus 0.312 squared, you know how to calculate the residue you have to multiply by this factor and then.

Student: ((Refer Time: 50:50))

Which inverse, no, finally it comes to you can check, you just calculate the residues it will be approximately 0.95. So, s naught will be 1 by root I it is 1 by I is not this much I is, so much this becomes 1.1078, so 1 by 1.1078 which is 0.95. So if you have a multiplier s naught equal to 0.95, then there would not be any overflow at any stage or at the output side.

(Refer Slide Time: 51:56)





H z equal to 1 by 1 minus 2 r cos theta z inverse plus r square z to the power minus 2 r less than 1.0, draw the direct form and determine s naught to avoid overflow to avoid overflow. So, D z is 1 by H z here there is nothing in the numerator except one, so I will be 1 by 2 pi j z inverse d z divided by D z D z inverse, which is 1 by 2 pi j integration z inverse D z. So, that will become H z H z inverse in to d z and what will be this H z H z inverse in to z inverse integration, this we have done somewhere earlier you know.

So, what was H n yesterday I give you a problem.

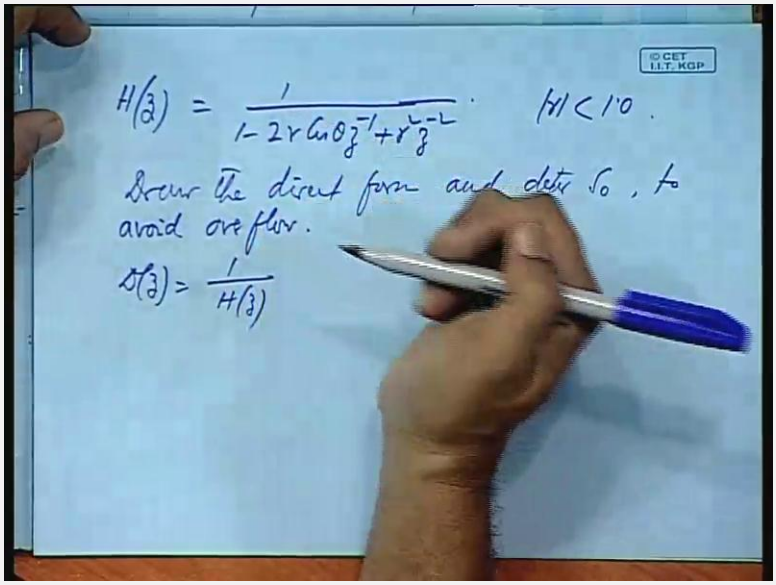
Student: ((Refer Time: 54:13))

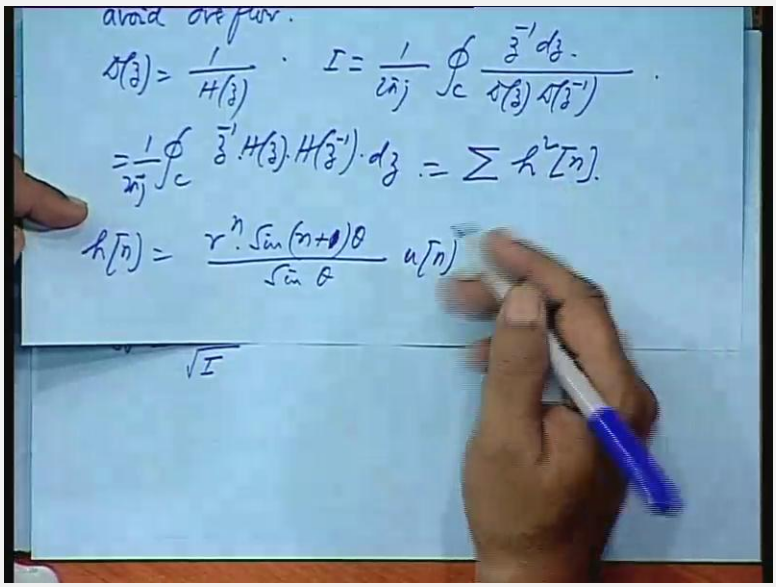
H n corresponding to this function just check up.

Student: R to the power n.

R to the power n sin of n plus 1 theta, thank you very much by sin theta in to u n, so substitute here square of this.

(Refer Slide Time: 54:45)





So, d z or that is I which is equal to sigma h square n will be equal to r to the power 2 n sin of sin square n plus 1 theta divided by sin square theta submitted over n, this your

proof. And check whether this result can be obtained 1 plus r square divided by 1 minus r square in to 2 r square cos square theta plus r to the power 4. So, S naught will be 1 by root I where I is this much, now if you are given a value of r and theta, we can calculate s naught is that all right.

So, I would request you to verify all these relations I have taken for granted certain relationships, you derive them and solve. So, in the next class we shall be discussing about random signals that is the simple properties of random signals as applicable to this subject DSP.

Thank you very much.