Digital Speech Processing
Prof. S. K. Das Mandal
Centre for Educational Technology
Indian Institute of Technology, Kharagpur

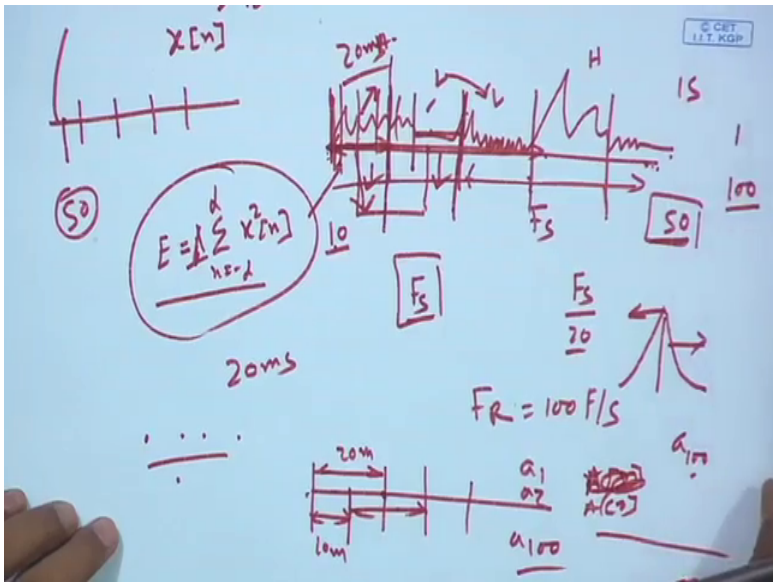
Lecture – 19
Time Domain Method In Speech Processing

So last class we have finished that speech perception that part.

(Refer Slide Time: 00:25)

So, let us start, so it is time, domain methods in speech processing. So, here we want to discuss about some methods which are in time doing methods, now we are not analyzing the frequency of the speech, in time domain methods for speech processing, and that is used many cases those kind of methods are used. Now, first of all that one things is that once the speech is digitized and taken to the computer, so these are we are saying that digital signal processing the speech signal is now in digital domain.

(Refer Slide Time: 01:00)



So, we always denote the speech signal with x of n, so there is a digital signal we know the sampling frequency of x n that is F s. So, now speech if you see the speech is a time varying signal; that means, along the time if this is the time axis, then speech property has been changed this may be voiced, that may be unvoiced, then there may be a voiced, there may be noise, then there may be a voiced again, then there will be noise like that, so speech is not a stationary signal.

So, what I say that if I take this portion of the speech the property is different, if I take this portion of the piece property is different, so speech is changing along the time. Now what happened suppose I want to find out the energy of this whole signal? So, I can easily find out energy E is nothing, but a summation of n equal to minus infinity to infinity x square of n whole signal whatever the signal is there I can take that whole signal at a time x square n that is the energy.

Now if I take that energy, then if I want to know that this portion is voiced, so energy of this signal is high this portion is unvoiced energy is low this portion is noise energy is low compared to this; this is high, this portion again voice it is high. So, if I want to know that information if I take the whole signals at a time and find out the energy. So, it gives me the average energy or if I make it average, then 1 by I can make the average also, so if I make the average, so I can get either total energy or average energy of the whole signal. So, that information is whole signal energy is but that does not give me any

kind of parameter by which I can process or I can use this; those, parameter to some purpose in the speech processing.

So, what we have done instead of doing that whole signal at a time we try to analyze the signal for a particular window, for a particular signal for a particular segment. Now this segment I take this segment and find out the parameters and move to the next segment next segment; next segment, then I can get a segment. So, if this is my F s sampling frequency, so rate of sample in 1 second I get F s number of sample. Now if I window the signal let us 20 millisecond window. So, 1 second signal how much frame I will get 50 frames, I will get if I slice the whole signal in a 20 millisecond window, so I get the 50 slice, so I have a 1 second signal I slice it every 20 millisecond, so I get 50 slide.

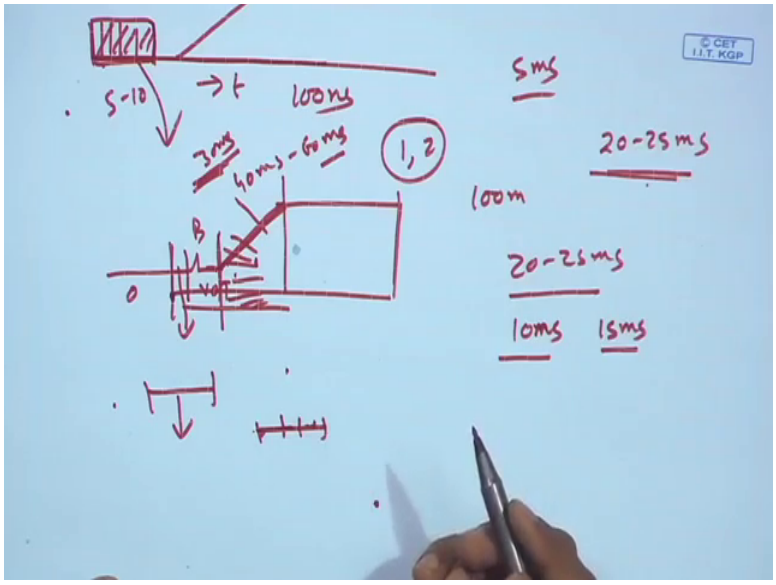
So, then I can say this is F s by 50, so it is F s by 20, so instead of F s sampling rate now for every window I get some parameter suppose parameters are sampled at 20 millisecond per second in 50 some 50 window per second. So, instead of F s number of sample, now I get if each window representing 1 sample then I get 50 samples. Now the problem is that, if I do that things then what will happen since this boundary, in this boundary effect will come in the processing, so there may be a this boundary I just this boundary when you fall in a very large signal. So, that this sample is included this window and this window energy will be less.

So what I want to know, I want to remove this boundary effect also, so then what we will do I may take 20 millisecond window, but I slide the window let us 50 percent of all them, so effectively 10 milliseconds it. So, I can say frame rate in here is in 1 second then I get 100 frames. So, if it is frame rate is F R, then I can say that if it is 10 millisecond sliding of the window, I take the window 20 millisecond, but if I slide it 10 millisecond, then I get 100 frame millisecond, so frame rate is 100 frame per second here sampling frequency I get F s number of sample per second.

Now, if I say for every 10 millisecond I take a point then I get 100 point in 1 second signal, so I can say it is down sample the signal. So, if I want to process the signal taking every sample is very difficult because what F s if it is 8 kilo hertz 8 k sample. So, instead of eight k sample let us I divide that whole 1 second signal in a 100 frame, and 100 frame, a for every 10 millisecond I get a parameter, now slide the window. So, effectively what I am doing I am windowing the signal with overlap methods.

So, this is window length, let us let 20 millisecond and this is frame shaped, 10 millisecond. So, 0 to 20 millisecond I take the signal I process it find the parameter, I find out the parameter vector lets vector a lets it is x 1 or let us A vector 1, then I shift the window with 10 millisecond and again take that 20 millisecond I find out A two vector then, I again shift a 10 millisecond and take 20 millisecond A 3. So, in that case I will get A 100, so if I write like this way A 1 A 2 A 100 data I will get, why I take this one, why I take 20 millisecond, 10 millisecond, why I take this kind of statement, or what is that what is the logic behind it.

(Refer Slide Time: 07:54)



Now, if you see time speech is an time variant signal. So, if I say that that time along the time speech signal change its color. So, if I want to increase the time resolution, I should analyze the signal for every sample, but that analysis does not provide me any information. So, what I want I want a small window. Let us 5 to 10 millisecond if I shift it, so, for 5 to 10 millisecond window very short signal. So, due to small amount of data now if you see the pitch let us I have take a five millisecond window, now it may found that within five millisecond maybe 1 or 2 fundamental period is there if it is male voice, then may 1.5 fundamental period is there.

So, if I analyze it; it is no use now if I take it; lets 100 millisecond, then what will happen I may lose the transitory portion of a consonant to vowel transition because speech is changing, so let us there is a ka bhast then vot you know that, that conclusion bast vot

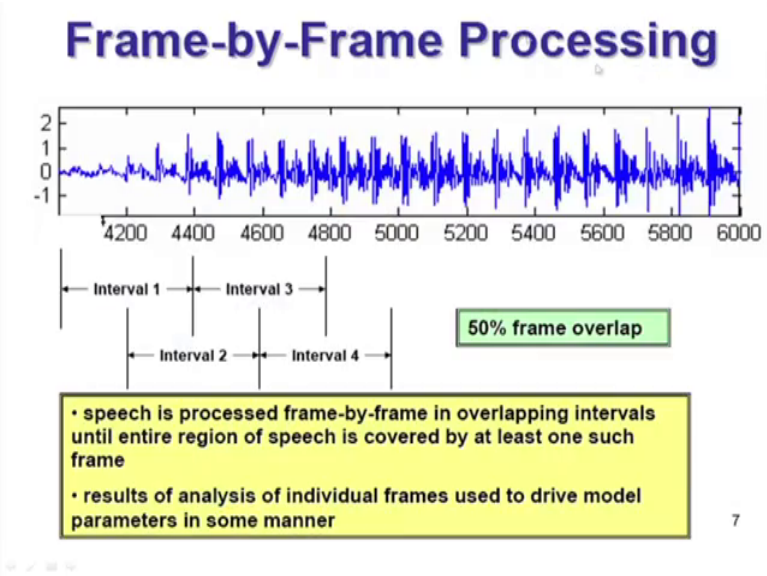
transition. Now if I take a 100 millisecond my window may fall around here, so what is the information which is very important to know this information I lost this information, because this information is taken an average with the vowel and some part of the consonant.

So, I lose the transitory part, so instead of 100, millisecond. So, if at small amount signal processing problem large amount signal I lose the time resolution, so what I want; I want a optimum length of the window, so that it does not affect so much. So, I can say let us 20 to 25 millisecond of window, if I take then one transitory period, roughly 40 millisecond to 60 millisecond maybe even faster speech it may be a 30 millisecond. So, at least I can ensure that my window length is lower than the transitory part ok.

So, then I can say yes that details signal information we along that time I can find out, if I take the window length 20 to 25 millisecond, so that is why if you see in whole all processing of the speech, we use 20 to 25 millisecond as a window length, and shifted the frame by 10 millisecond, so that I get 100 frame per second you can shift at it 15 millisecond also, but what will happen for every 15 millisecond I get a vector in here for every 10 millisecond I get a vector for analysis or I get a parameter vector.

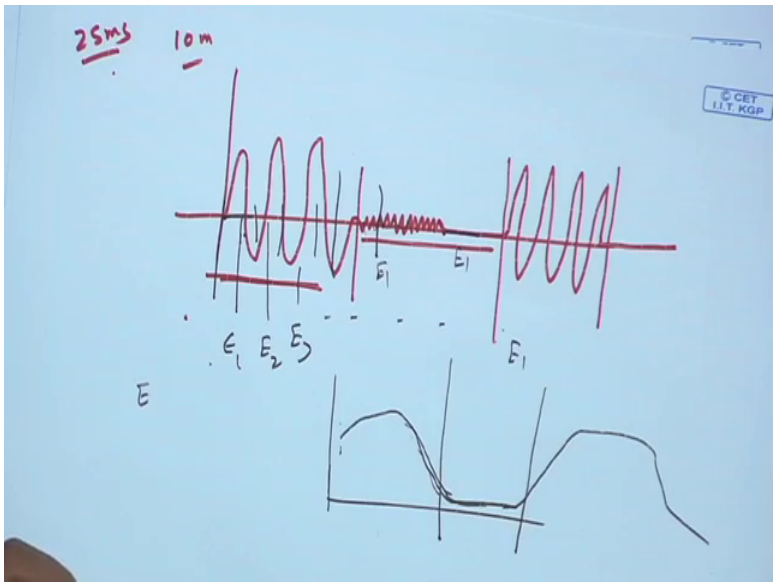
What kind of parameter I can analyze for every 10 millisecond I get a if it is fundamental frequency, then also for 10 millisecond I get one pitch parameter. So, it may contain average of three or four pitch period inside the 10 millisecond maybe 3 or 4 to 5 pitch period is there if it is female voice. So, now, we are framing the speech signal and shifting the window by 10 milliseconds.

(Refer Slide Time: 11:29)



So, this is the example if it is it is here, so this is the interval 1 and then, shifting of the frame by 10 millisecond or less 50 percent, if it is 20 millisecond window 50 percent of overlap means 10 milliseconds Shifting.

(Refer Slide Time: 11:52)



So, I get 100 frame per second, I can take 25 millisecond as a window and shift at the frame 10 millisecond that also I get 100 frames per second ok.

Now, this is the processing part I have discussed, now what kind of time domain parameters are there, so I can say short time energy. So, for every window I can find out the short term energy of the speech signal. So, suppose let us application side you can

come, suppose I have a speech signal there is a voicing, then there is a let us noisy part then, there is a silence part again there is a voicing part and then again there is in silence part.

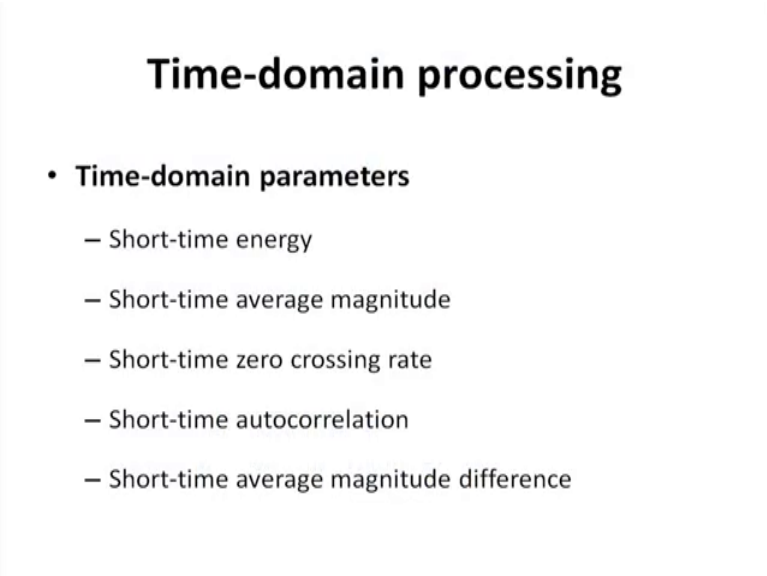
Now, suppose some anyhow I want to find out which part is voicing part and which part is unvoiced or silenced part this I want to find out. So, I want to distinguish between this I want to mark this point, this point and this point that I want to mark, now even in assumption if you say that if I find out the short term energy of the whole speech signal.

So, for a let us I take this path this window find out the energy, then 50 percent overlap taken window find out the energy E 1 E 2 then take an energy e three. So, that way if I find out the energy now if you see since it is a voicing signal the energy will be very high since it is as noisy signal the energy value will very low it is a silence energy will be more low.

So, from that short term energy this E value can give me a plot like this that this is maybe energy is very high. So, this is the high; high, then coming down low then low then again it will go high, because of this kind of transition, because window may take this portion and this portion. So, it is average out the energy, so that is why you get this kind of transitory part also get, so from that curve I can find out this up to this point may be the voicing up to this point may be the silence part.

So, that I can easily understand, so short term energy is one kind of parameter, then short term average magnitude.

(Refer Slide Time: 14:21)



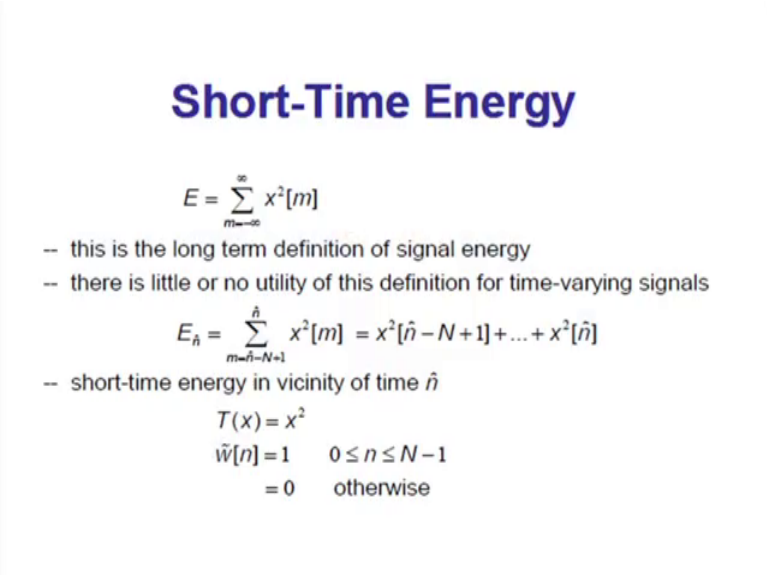
Short time energy, or short time average energy, or certain average magnitude, now what is the problem in energy, when I calculate energy the signal x n has to be square up. So, every sample value let us see if it is a 16 bit speech sample. So, let us the value is come around 28000 for a 1 sample, so it is 32000 plus and 32000 minus if with 8 bit maximum possible, now if it is 28000 if you square it the number will be integer number will be very high.

So, instead of squaring the signal, what I can do I can find out the sort of average energy short term magnitude energy average magnitude means, if it is x n is my signal. So, magnitude is mod of x n is the magnitude, and then I can add it take the average they take the sum value and divide it by the number of sample I get the average magnitude, So, that I can use as a parameter for this; this instead of E 1, so that kind of mathematical little sun coming. So, that I can use it, then short time 0 crossing, if you see a let us come in here I if you see I do not know either there is a specter time signal is there or not I do not have any signal.

Now, if you see any signal any speech signal, if you see that this portion lot of time signal is cross the 0 line, this portion may be the number of zero, so this is the zero line zero amplitude line, so zero line crossing will number within a particular window will be less compared to this. So, short times zero crossing can be used to find out whether it is a voice signal or it is a noisy signal. So, short term energy is short term zero crossing can be used, so the number of zero crossing can be a one parameter to extract the number of zero crossing can be a time domain parameter.

So, this is a time domain parameter I have to extract, the for per frame extract the number of zero crossing then short time autocorrelation I come, then short time average magnitude difference, those are the parameter I can use from the speech signal and those every parameter has its own purpose also, now I just come to that one by one short time energy mathematics.

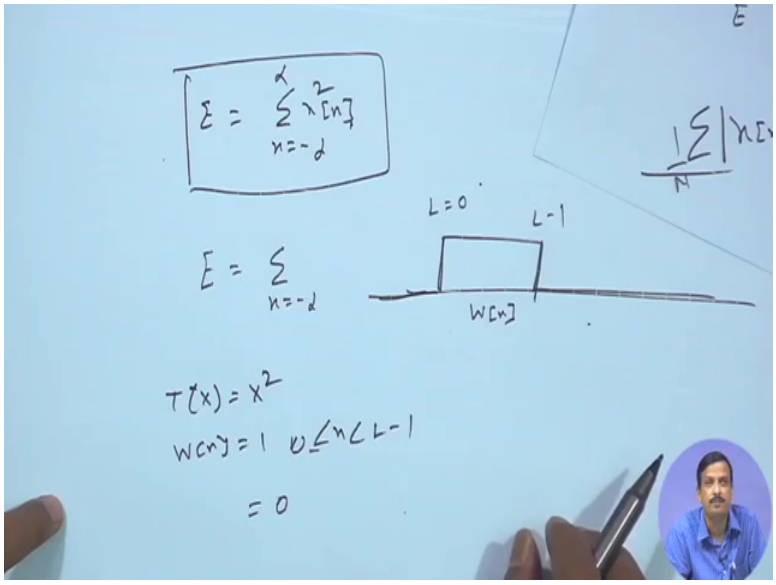
(Refer Slide Time: 17:15)



So, energy E of a signal which is n is equal to minus infinity to infinity x of n square is the energy. So, this is the whole signal energy long term definition.

Now, if I take the long term definition lets this is the my signal and if I take the whole energy at a time or this is my signal, and I take the whole energy at a time no use what I do that with that parameter, because I want to find out where voice signal un voice signal silence, so if I take the whole signal at a time find out the energy I get a value which is energy value of that speech signal, but I can I infer any information from that energy value. So, there is no use if I use whole signal at a time.

(Refer Slide Time: 18:06)

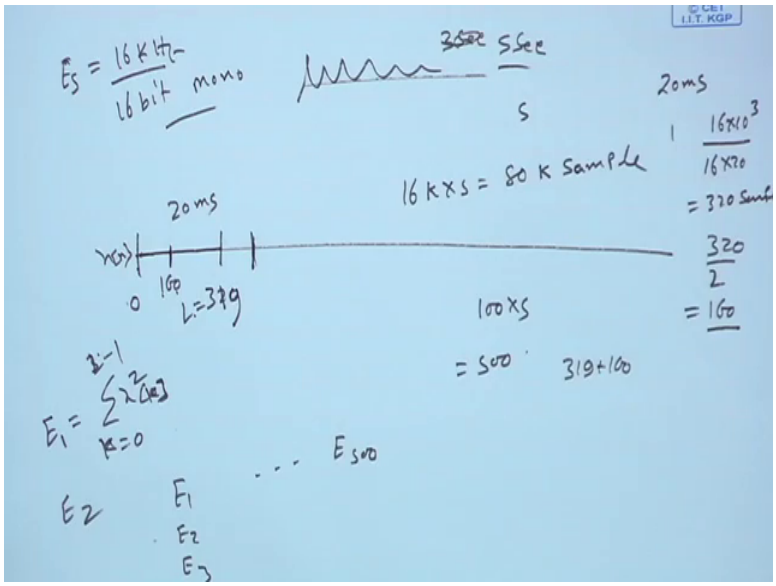


So, what I will do instead of taking whole signal at a time, I will multiply the signal this signal E for a particular window.

So, I cut the signal particular window here, if I cut the signal I take the signal of particular window. So, what I am doing it, so there is a infinite length signal I place a window over the signal, so this is l equal to 0 to l minus 1 which is W n. So, I can say short time energy in vicinity of n cap n, so T x is equal to x square value and omega n is equal to 1, if it is within l equal to lets n equal to 0 to l minus 1 else it is zero. So, I cut the signal from the long signal I cut a window and take the energy ok.

So, if I say that find out the short term energy, first I recorded record your let us record your name in a let us F s is equal to 16 kilo hertz and 16 bit mono record your record your name in computer.

(Refer Slide Time: 19:31)

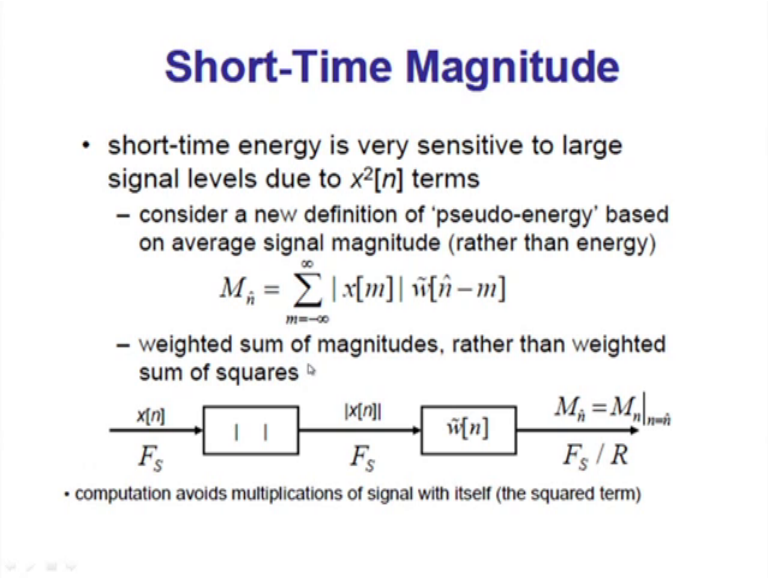


So, put the microphone connect to the computer record your name, once you let your name consists of three second signal all let us 5 second signal long name 5 second signal it is consist of 5 second speech signal so; that means, 16 kilo hertz sample 16 k sample per second, so I can get 16 k into 5 80 k sample ok.

So, I have a 80 k sample this side, I take a window of 20 millisecond. So, I take the sample number 0, so if it is F s is 16 kilohertz then how many sample, will be there in 20 millisecond, in 1 second there will be a 16 k sample, so in 1 millisecond 16 sample. So,1 20 millisecond I can get 3 2 0 sample, so there is a 3 2 0 sample in 20 millisecond I have a take the first frame 0 to 320, and since I am multiplying a window function with amplitude 1 so; that means, from take every sample so, let us take this is x n, so I take x 0 then I take sum x equal let us k equal to 0 to 320 minus 1. So, n l is equal to 320 l minus 1 x of k x square of k equal to E 1.

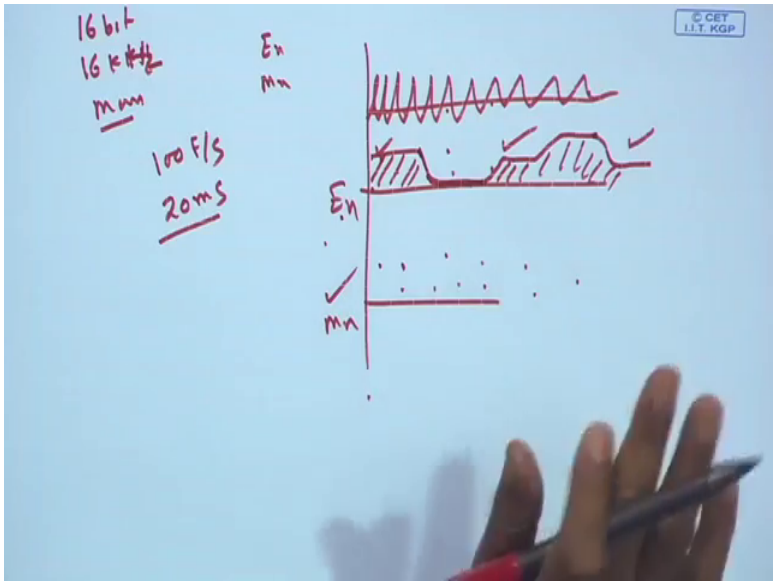
So, first frame power I get, then I shifted this frame by 10 millisecond; that means, 10 millisecond means 320 divided by 2 so; that means, 1 6 0 160 sample. So, I take that another window from 160 samples, to here so this is 0 to 319 samples, so this is 160 sample to another. So, here will be 319 plus 160. So, I get take that that sample and find out E 2, then that way I can so, if it is 10 milliseconds shifting, so I get 100 frame per second, so in five second I will get 500 frame. So, I can get E 1 E 2 E 3 dot dot dot dot E 500. So, I get 500 short term energy data point.

(Refer Slide Time: 22:39)



Now, if I told you that if I draw that this diagram, so what I what is the diagram here is x n this is F s if it is energy then this function is square.

(Refer Slide Time: 22:56)



So, if it is short term energy then, I can find out let us this is my signal x n whose sampling frequency is F s I operator is x of so, I can say square; square, is the operator signal square I take the signal square then, I pass the signal, so all signal is every sample is square rate is still F s. Now once I pass through the window W n after the window my

rate is F s by R F s by R, R is the 15 10 millimeters if it is 10 millisecond, so if it is for every 10 millisecond I get one sample you can say that one data.

So, I can say here I get E n which is nothing, but it for every 10 if it is 10 milliseconds shifting, so 100 frame 100 per second earlier it is F s let us see it is 1 F s per F s number of sample per second instead of F s number of sample I get 100 number of samples per second. So, I get E n value E n is equal to on E 2 dot 100 for 1 second I here F s sampling E 1 2 up to F s, F s number of sample or not. So, this is called I can say that short term energy.

Now, if I interested instead of energy I can replace these walks same walks by only magnitude again it is F s again I if I put a window W n I get F s by R does this M n, so instead of squaring the signal, so M n is nothing, but a lets k equal to 0 to l minus 1 that is a within the vicinity of the window length you are taking x of m lets x of k mod into W n minus k where n is the here M n, n is the starting polar window number of window, so if it is first window n value equal to 0 if it is second window n value is equal to 160 shifting is 160 sample if it is third window n value will be starting of the window will be 320 forth window starting of the window will be 1 320 plus 60, I can get. So, for every window I get 1 M n ok.

So, then I get M n, so instead I can get E n I can get M n E n is the short term energy M n is the average magnitude average magnitude, it is sorry it is a sum of magnitude if I want to make it average 1 by put 1 by l here. So, I can normalize it also 1 by l l number of sample up there, so I can one by l I get the average. So, average magnitude. So, using these 2 see that, so this is some plot of a boost to parameter you cannot extract after recording your name extract that E n.

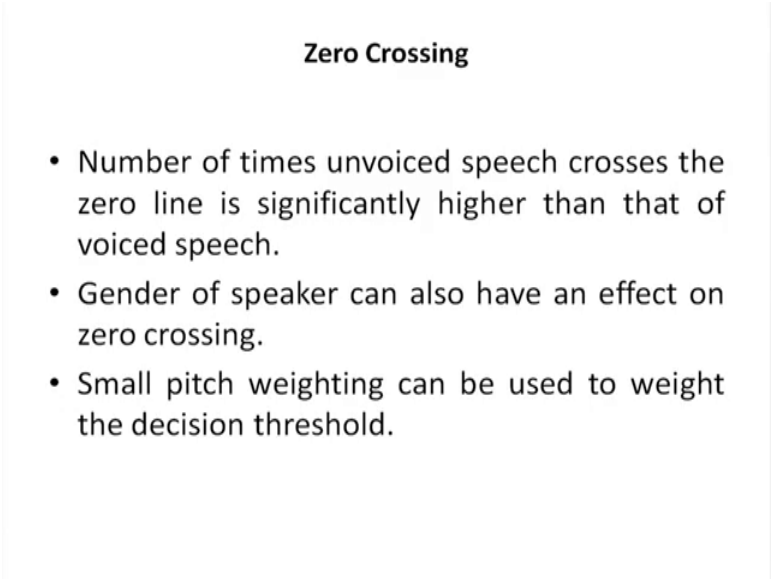
So, I give you the problem, so problem is you record your name in 16 bit 16 kilohertz and 16 bit 16 kilo hertz and mono record your name using any you can use cool edit you can use plot record it, then find out E n and M n value of E n and M n E n in the rate of 100 frame per second and 20 millisecond is the window length take 20 millisecond window length and find out the 100 pairs of (Refer Time: 27:29) shift is 10 millisecond find out the E n and M n.

Then plot the 3 signal 1 is your first recorded signal, then plot it E n then plot M n. So, for recorded signal there are lot of samples will be there for E n for 1 second, I get 10

sample values for M n also for 1 second I get 10 sample values. Now you see whether you are able to find out the voice zone using E n and M n or not. So, maybe if it is voiced zone E n value will be high, if it is unvoiced or sibilant yeah maybe if it is noisy unvoiced amplitude may not be that much of holly (Refer Time: 28:17) it may rise somewhere, and then again voice in you it may be raised sometime if it is voice bar like this. So, I do not know either it is a voice bar and whether it is a noise.

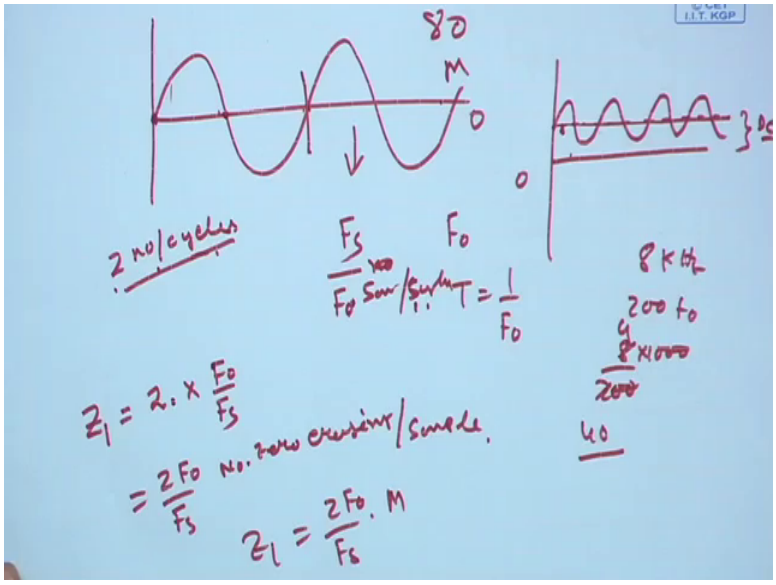
So, for that purpose we use another parameter, but at least using En I can find out where the speech signal has high energy where it is there is no energy that I can find out. So, this can be used as a voice detection, you drop plot it and see how it is behaved like, then you can get a real exposure. Now another parameter is called 0 crossing.

(Refer Slide Time: 28:56).



If you see zero crossing yes is in zero crossing if it is a sin wave.

(Refer Slide Time: 29:10)



So, how many time signal crosses the 0 line, so it is 1 2 if it is period, so if it is I start from 1 then 1 2 then this point is part of the next period. So, I can say if it is a pure sine wave then 2 number of zero crossing per cycle; per cycles, 2 number of zero crossing ok.

Now, what will happen if you say some time you may say that I have recorded the signal and this is my 0 line m 0 amplitude line by my sin wave in here, this is never crosses the 0 line; that means there is a 0 line, but the sine wave is d c shifted d c shifted means the 0 there is a sine wave during the recording there is a d c shift in above that is why 0 line shifted to here, 0 line is this line. So, if it is d c shifted what I can do either I can pass the signal through a low pass filter or I can take the average and subtract it. So, first normalize that d c shift of the signal then find out the 0's ok.

Now, I just come to the calculation then derive that equation then find out the parameters. Now if we see that if it is a pure sine wave 2 number of zero crossing per cycle, now if the sampling frequency of this sine wave is a F s, and the fundamental frequency is a zero, that means length of the period is 1 by f 0, then I can say F s by F 0 number of F s by F 0 sample per cycle, so how many sample will be there in per cycle F s by f 0. So, if it is 8 kilohertz sampling frequency and I have a 200 hertz f 0, then how many samples will be there in one cycle 8 k divided by 200 hertz, so 40 samples per cycle ok.

So, this is number of sample or I can say sample per cycle. Now if you see 2 number of zero crossing per cycle, so number of zero crossing Z 0 is equal to 2 crossing per cycle

into cycle per sample, this is sample per cycle. So, cycle per sample is nothing, but a F 0 by F s, so I can say twice F 0 by F s number of zero crossing per sample, number of zero crossing per sample, so if I want to find out find out the number of zero crossing of 80 sample. So, let us M sample then Z 1 is nothing, but a 2 F 0 by F s into M ok.

So, next class we try to derive the generalized formula, how do I find out the zero crossing for non periodic signal or you can say the for speech signal or the periodicity is periodicity quasi periodic signal, so for there, how do I find out or even non periodic signal how do I find out the number of zero crossing ok.

Thank you.