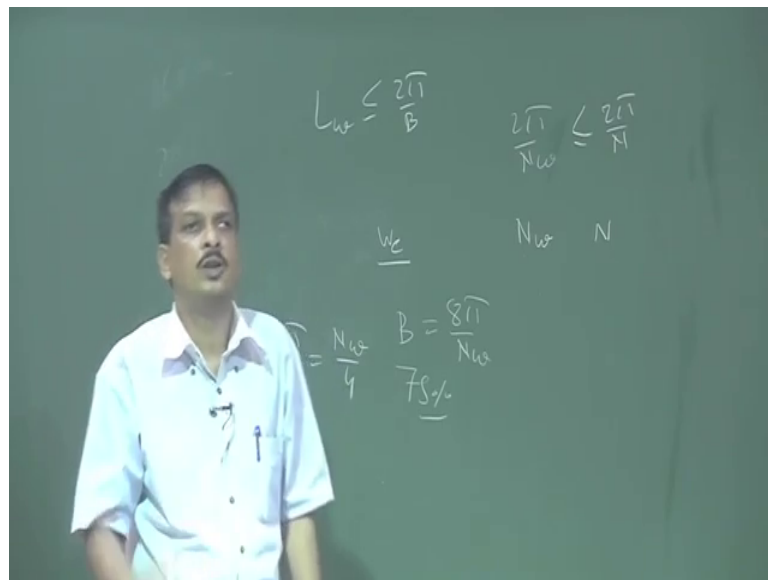


Digital Speech Processing
Prof. S. K. Das Mandal
Centre for Educational Technology
Indian Institute of Technology, Kharagpur

Lecture – 28
Short – Time Fourier Transform Synthesis

So, from the last lecture it is a time frequency sampling that constraint that we said that L_w the decimation in time should be less than equal to 2π by the bandwidth of the analysis window and we said that N_w or you can say 2π by N_w must be less than equal to 2π by N ok.

(Refer Slide Time: 00:25)

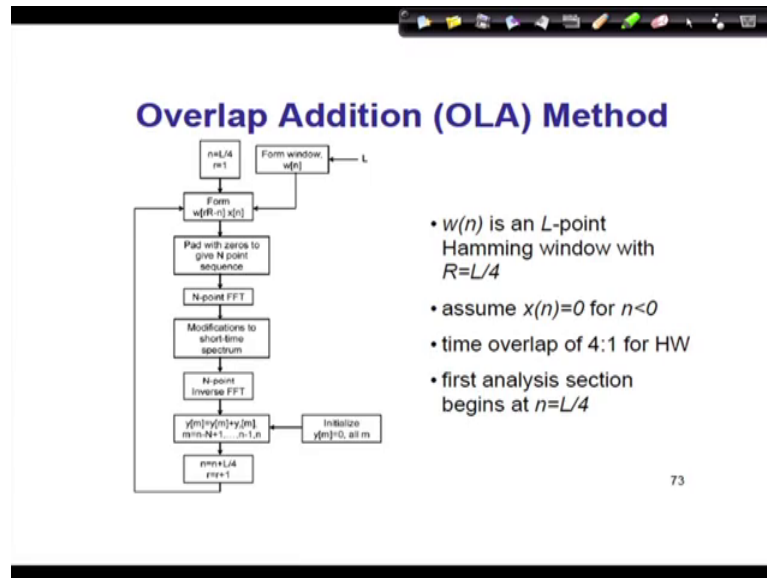


So, when it is said that basically that complete recovery of the signal is possible if I choose the window whose length is N_w , which should be less than the number of depth I have taken N , which define the number of channel and so if I say that I have a 320, 16 kilo hertz speed signal and if I take 20 millisecond window. So, length of the signal is nothing, but the 320 sample, if I take the DFT length, if I take the DFT length N is equal to 512; that means, this is N_w this is N . So, N_w , 2π by N_w must be less than equal to 2π by N and I said that N_w must be 2π by b .

So, how much decimation in time is possible, how much decimation in time, how much decimation in time is allowable is defined by the ω_c which is the bandwidth of the analysis window. So, if it is hamming window then bandwidth b is equal to 8π by N_w .

So, L w should be 2π by B is equal to N w by 4 so, 75 percent overlap. So, L w is equal to 75 percent of overlap upto 75 percent so this way we can find out the time and frequency sampling. So, let us not go, sorry I have done mistake this is should, be this should be 2π by N w must be greater than or equal to 2π by N . This is given in the slide so there may be mistaken while I am writing in the book board ok.

(Refer Slide Time: 03:30)



So, time frequency sampling is there, now suppose I want to write down that overlap add method in algorithm. So, if you see if you go through this slide you can make the algorithm. So, this is given, by this way you can write a program for your synthesis, now I come to the application of this.

(Refer Slide Time: 03:52)

Summary

- **OLA Method (DFT of order N)**
 1. No time aliasing if window length N_w so that:
 $2\pi/N \leq 2\pi/N_w$
 2. No frequency-domain aliasing occurs if decimation factor L is small enough so that filter bandwidth
 $\omega_c = (2\pi/L)$
 3. If zeros are allowed in $W(\omega)$ then condition 2 can be relaxed. In this case we can under-sample in frequency and still recover the sequence.

27

So, summary in summary OLA method there is no time aliasing if window length N_w is. So, that $2\pi/N$ is less than equal to $2\pi/N_w$ and no frequency domain aliasing occur if that ω_c is equal to $2\pi/L$ so if 0 are allowed in ω then condition 2 can be relaxed ok.

(Refer Slide Time: 04:15)

Summary

- **FBS Method**
 1. No frequency-domain aliasing occurs if the decimation factor L meets the Nyquist criterion, i.e., $L \leq N_w (2\pi/\omega_c)$ where ω_c is the $w[n]$ bandwidth.
 2. Not time-domain aliasing occurs if $2\pi/N \leq 2\pi/N_w$
 $\Rightarrow N_w \leq N$.
 3. If zeros in $w[n]$ are allowed then condition 2 can be relaxed. In this case we can under-sample in time and still recover the sequence.

28

Now, in summary FBS method $2\pi/N$ this one, N_w is greater than equal to N , if 0 in w N are allowed then the condition 2 can be relaxed ok.

(Refer Slide Time: 04:31).

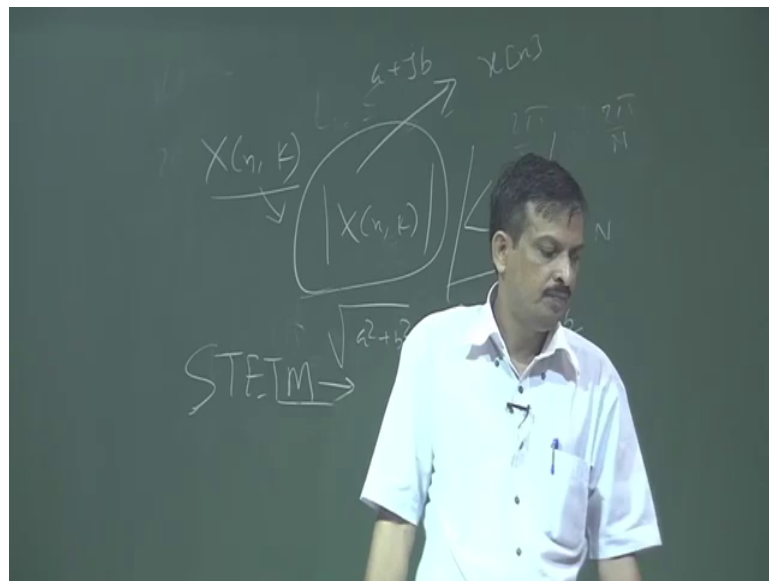
Short-Time Fourier Transform Magnitude (STFTM)

- STFTM discards (possibly) phase information, which has numerous uses in application areas:
 - Time-scale modification
 - Speech Enhancement
- In all these applications phase information estimation of speech is difficult (e.g., presence of noise in the signal)
- Furthermore, a number of techniques have been developed to obtain phase estimate from a STFT magnitude.

29

So, now I come to the short time Fourier transform magnitude. So, before that we should explain it.

(Refer Slide Time: 04:50)



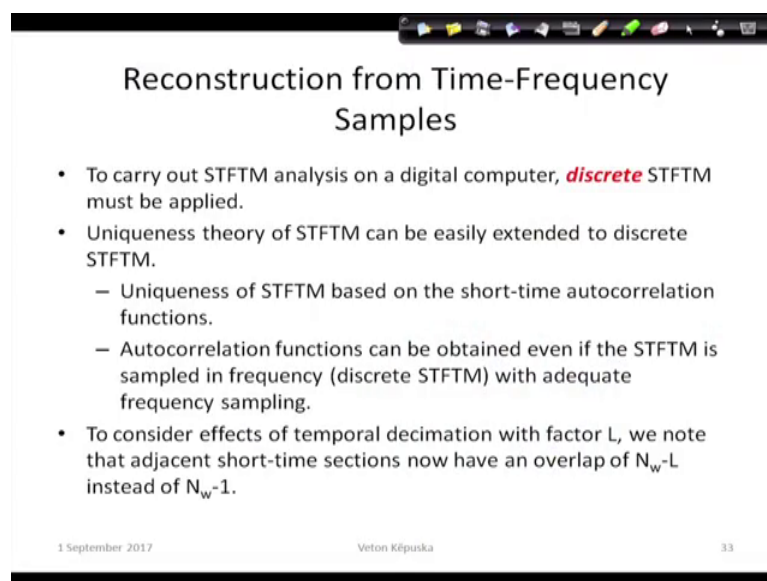
So, what I said that x of N k is nothing, but a complex number Fourier transform so every Fourier transform component has 2 parts, one is called magnitude and another is called angle phase. So, if it is all component let us one of the component written is $a + jb$ in complex number then magnitude is nothing, but root over square a square plus b square and θ is nothing, but a $\tan^{-1} b/a$.

So, if I draw this one with respect to frequency then I call magnitude spectra which already I have discussed in beginning class, magnitude spectra if I tell frequency versus theta then I call phase spectra. So, STFTM, STFTM short term Fourier transform magnitude, short term Fourier transform magnitude. So, only magnitude part will take we discard this phase part do not use this phase part and this is many cases it is used for time scale modification and speech enhancement, now it was found that I will not go in details mathematics of STFTM synthesis. So, STFTM analysis is that I do the same analysis and calculate the magnitude part.

So, analysis part is ok, so synthesis part there is a detail mathematics available, I am not going that details you can go through the books and you can, if you want you can really go through the books and find out of the details. So, it is said that from the magnitude part if I maintain certain constraints then the from magnitude part also signal recovery is possible. So, it is said that from the magnitude part itself I can recover the signal. So, signal recovery is possible if I only take the magnitude part. So, that if I take the magnitude part then analysis is called STFTM short term Fourier transform in magnitude and synthesis is STFTM synthesis. So, I am not going details of the mathematics ok.

So, there will be a constant on analysis window and the signal if I there is some constant is allowed then it is, it can be shown that from the STFTM also I can recover the signal ok.

(Refer Slide Time: 08:00)

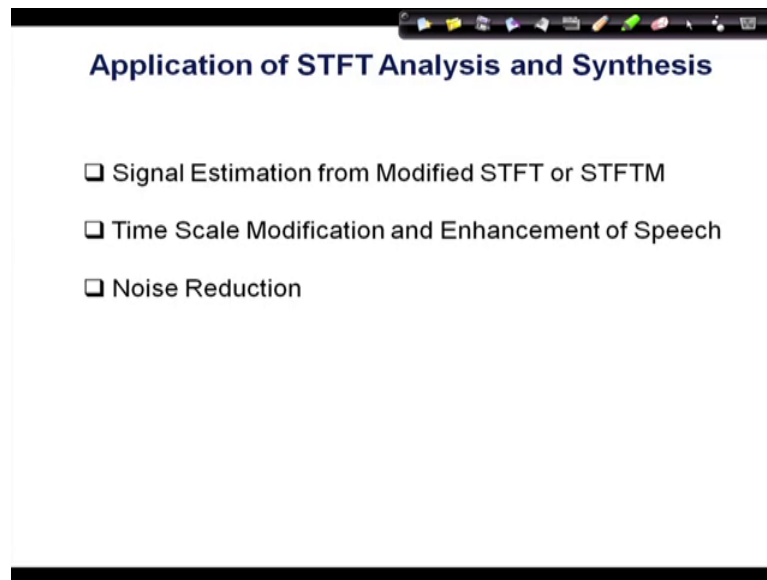


Reconstruction from Time-Frequency Samples

- To carry out STFTM analysis on a digital computer, **discrete** STFTM must be applied.
- Uniqueness theory of STFTM can be easily extended to discrete STFTM.
 - Uniqueness of STFTM based on the short-time autocorrelation functions.
 - Autocorrelation functions can be obtained even if the STFTM is sampled in frequency (discrete STFTM) with adequate frequency sampling.
- To consider effects of temporal decimation with factor L , we note that adjacent short-time sections now have an overlap of $N_w - L$ instead of $N_w - 1$.

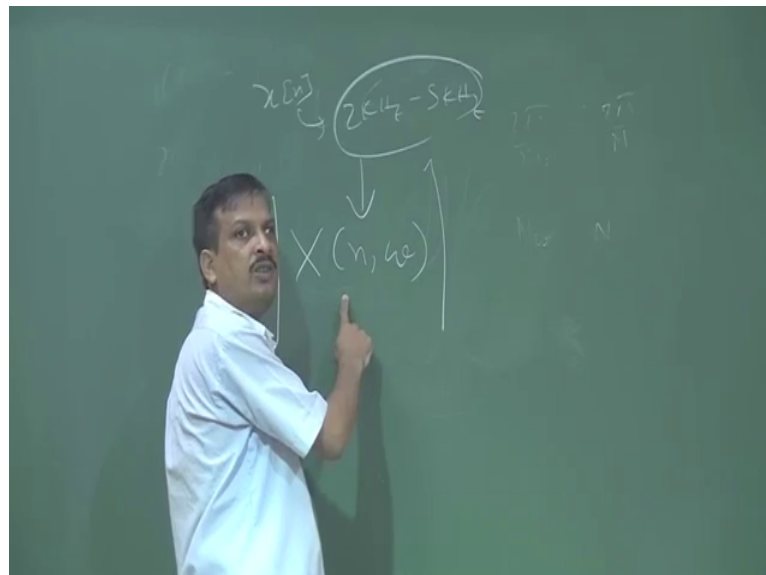
1 September 2017 Veton Këpuska 33

(Refer Slide Time: 08:04)



So, I am not reading the slides slide you can read better than me. So, now, come to the applications. So, why we do STFT, STFT analysis, synthesis all kinds of things why we will do it that what is the application of STFT signal estimation and modification STFT or STFTM both.

(Refer Slide Time: 08:25)

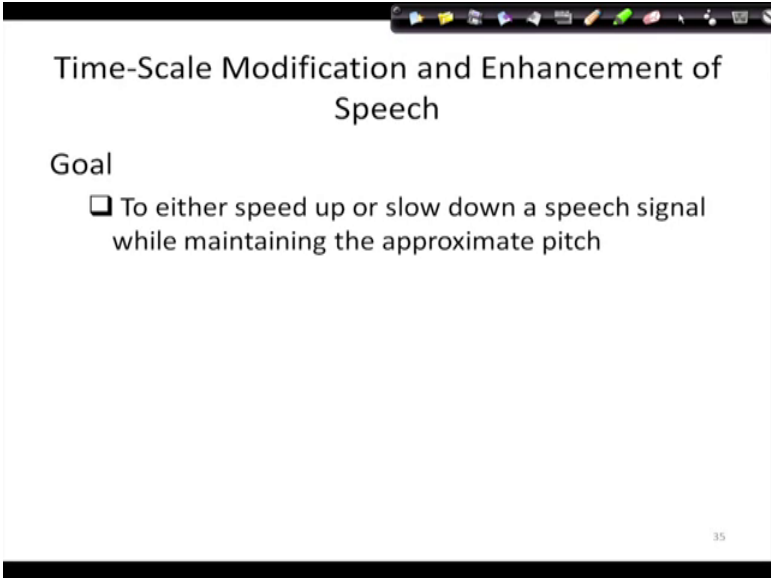


Suppose I have a signal $x[n]$ that may contain lets 2 kilo hertz to 5 kilo hertz component, I want to reduce the amplitude of this portion. So, I analyze $x[n]$ in frequency given and

reduce whatever modification I want, I modify the spectrum and then take the synthesis again I will get back the x_N which is modified signal ok.

So, I can do it in both terms, I can calculate x_N and do the modification in x_N or I can calculate x_N and take the magnitude part and modify in the magnitude part then also recovery is possible. So, STFT or STFTM is used to signal estimation. So, I want to do the power spectral magnitude of this estimate that spectral density of the signal, I can do the estimation or I can modify the some spectra and that I, then another one is called time scale modification I can do the time scale modification of the speech signal and noise reduction. So, time scale modification goal, what is time scale modification?

(Refer Slide Time: 09:49)



Time-Scale Modification and Enhancement of Speech

Goal

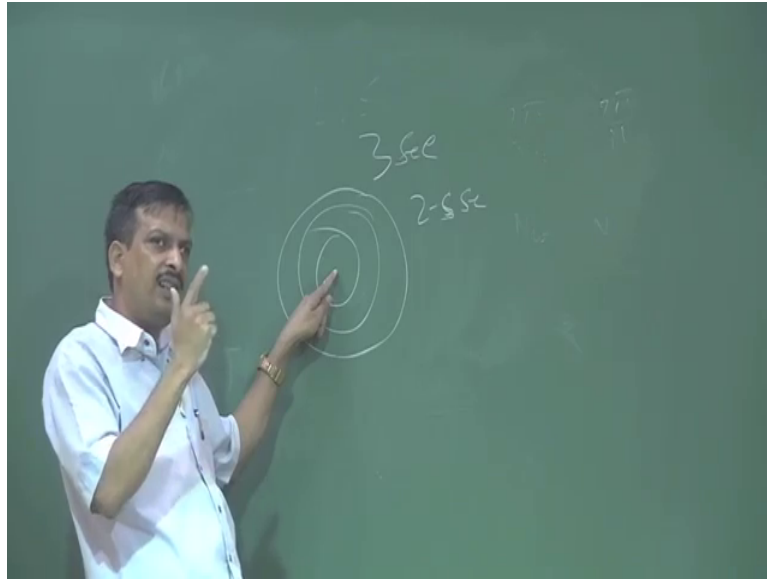
- To either speed up or slow down a speech signal while maintaining the approximate pitch

35

I can give you example, I want to either speed up or slow down the speech signal while maintaining the approximate speech, if you remember earlier that there is a you can tape recorder or you can say the gramophone that like gramophone recorder, gramophone that disk gramophone disk if you see the top of the gramophone disks the r p m was mentioned. R p m was mentioned it is 60 r p m record, it is 40 r p m record, it is 30 r p m record that is mentioned during the recording time the recording speed is mentioned there. So, that plain time if I play the same speed then only I get the exact recorded signal, if I increase the speed what will happen?

So, suppose there is a 3 second signal is there.

(Refer Slide Time: 11:02)



Now, I increase the speed so suppose if I increase the r p m. So, if there is a 5 track is 3 second now increase the r p m time will be reduced, speed up. So, time will be reduced. So, instead of 3 seconds song I can get 2.5 second song or lets 2 second song if I modify more then what will happening, once I increase the speed the spectral characteristics of re sampling that is nothing, but re sampling spectral characteristics is also changing.

So, the fundamental frequency or speech is also changing. So, that is why when you play a 60 r p m record in 40 r p m what will happen, if it is 60 r p m was there you quickly play it. So, fundamental frequency may shifted to female voice you can do that. So, male voice become female voice, female voice become male voice, this is happens in cassette also if you see the cassette if your tape recorder is very low very old then if the motto speed it change then the quality of the sound will change ok.

So, I want some time I want time scale modification, suppose you experience with the movie I if you know that fast you have done that acting and then the voice is synchronized with your acting. So, suppose when you dub it that time you found that recorded speech is slower than the movement of the lips during the acting. So, what I want, I want therefore, recorder speech little bit of faster. So, I can play it fast that much changing the sample re sampling it or what I can do I can cut some portions. So, that is called time scale modification. So, in time scale modification cut and paste.

(Refer Slide Time: 13:20)

Time-Scale Modification and Enhancement of Speech

Goal

- ❑ To either speed up or slow down a speech signal while maintaining the approximate pitch

Applications

- Change voice mail playback
- Court stenographers-play proceedings quicker
- Sound effects
- Etc...

35

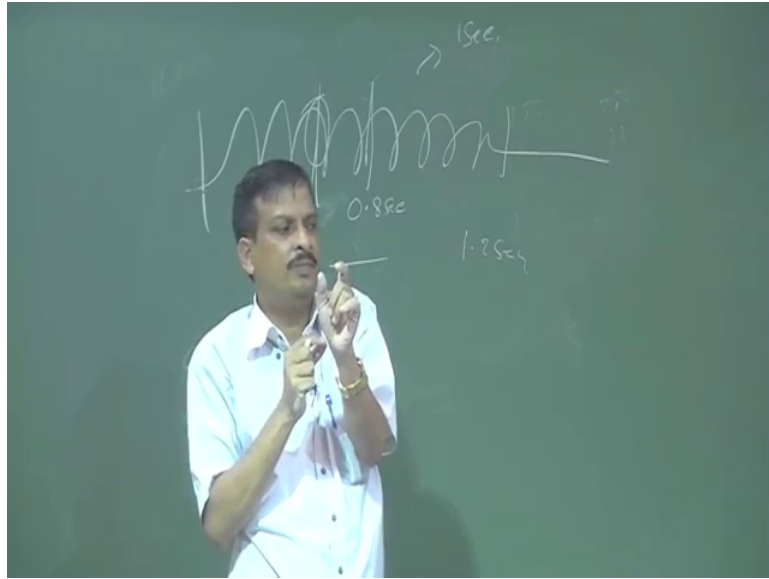
(Refer Slide Time: 20:01)

Time-Scale Modification

- **Methods:**
 - **Cut & Paste (Fairbanks method):**
 - Discard or duplicate frames, in order to speed up or slow down the articulation respectively.
 - Problem:
 - Pitch period mismatch at adjacent frames causes distortion.
 - **Pitch-synchronous OLA (Scott & Gerber)**
 - Select frame size & location synchronous to pitch periods. Problem of pitch period mismatch is avoided.
 - Problem:
 - Pitch synchronization is not always easy.
 - **STFTM Synthesis**
 - To avoid pitch synchronization problems use only the magnitude of STFT (i.e., STFTM)
 - 1. Compute $|X(nL, \omega)|$ at an appropriate frame interval – decimation rate L (e.g., $L=128$ at $F_s=10000$ Hz, and N is several T_0 long)
 - 2. Modify decimation rate with new rate M (e.g., $M=L/2$) for a speed-up of factor of $1/2$: $|Y(nM, \omega)| = |X(nL, \omega)|$
 - 3. Apply the Least-Squared Error iterative estimation algorithm until $|Y(nM, \omega)|$ converged.
 - Problem:
 - Occasional reverberant characteristic of synthesized signal are perceived due to lack of STFT phase control.

36

(Refer Slide Time: 13:28)

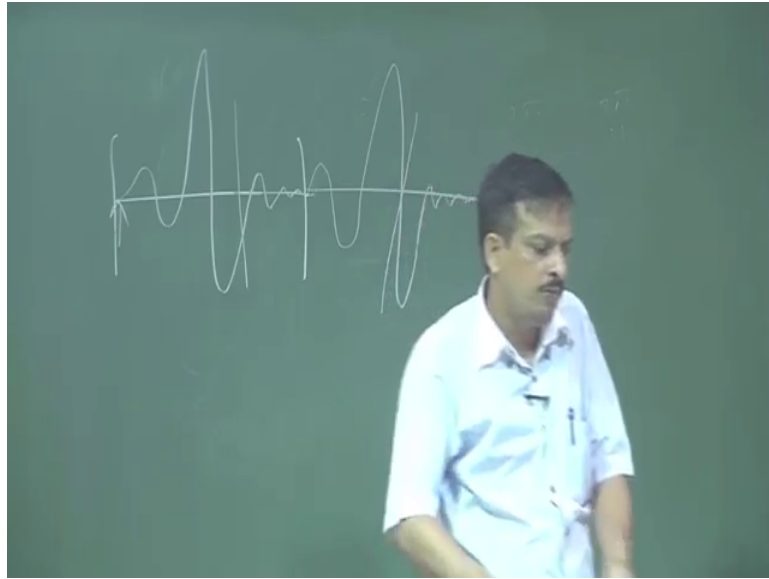


Suppose I have this is my voicing segment let's say something and there is a voicing segment this voicing segment represents let's say, this to this represents is let's say one second if I want to match with the lip movement in the video I have to make it let's say 50 instead of 1 second I have to make it 0.8 second.

So, how can I do it, either I can cut some portion of the signal and add them, if I cut then I can paste them. So, if I cut in an arbitrary position what will happen there may be a pitch mismatch a half period is cut this side and half period is cut this side. So, there is a pitch mismatch also if I want to expand it one second I want to make it 1.2 seconds. So, what I will do I will cut some signal from here and paste it here and so at that boundary there may be a pitch mismatch in adjacent phase or this boundary. So, there may be a pitch mismatch. So, that is the method, but it is a good not good method, but if you cut it precisely it is a good method. I can show you in the next class if I take I can show you in a usual cool edit I can cut exactly one period of voice and test it I can show you how it should be cut ok.

So, if we cut in an arbitrary position there will be a problem then they said how do we solve it, if you see any speech signal, any speech signal if you see if you remember the opening closer and the opening of the vocal fold if you see in if I want to see the vowel it will see look like this vowel so this is a complete here, again it will be look like this.

(Refer Slide Time: 15:29)

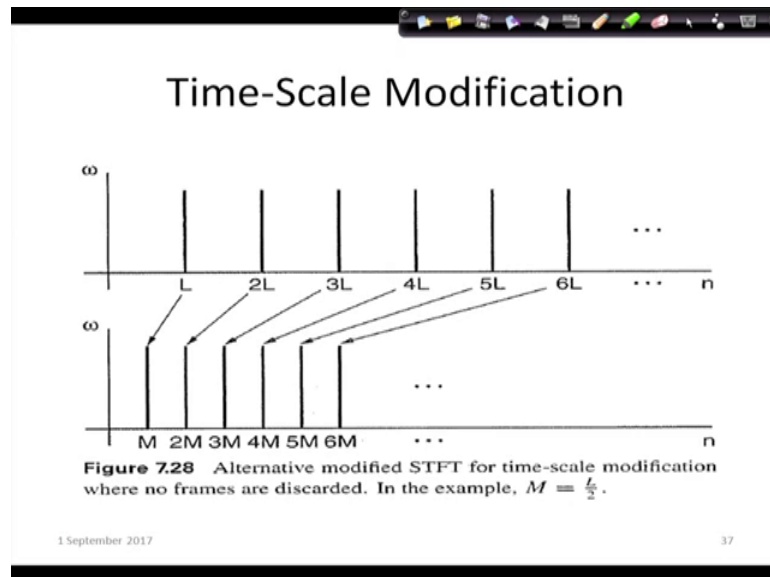


Now, suppose I cut in here and cut in here and cut it out so what will happen this is half speech this is also half speech so there will be pitch mismatch. So, to avoid the pitch mismatch next method is called piece OLA, pitch synchronous overlap add method details I will discuss these things during the pitch synthesis. Pitch p s OLA t d, p s ola e s p s ola all we have discussed. So, pitch synchronous overlap add method then they said if I able to find out the hip (Refer Time: 16:25) point the open closing and opening point of the vocal fold from the signal, then I can synchronize the pitch. So, I can exactly cut one period then no problem if I cut one period and paste it one period here there is no problem. So, that is called pitch synchronous.

Now, to get that exact pitch period from the voice signal automatically it is a very difficult. So, estimation of pitch synchronization or you can make the pitch synchronization is very difficult. So, that is why this process is also very difficult other is called STFTM s t STFTM synthesis to avoid pitch synchronize problem use only the magnitude spectra of the frame, from the magnitude spectra it is possible to synthesize the pitch signal. So, compute x of $N L \omega$ at appropriate frame interval or decimation rate and appropriate window length and appropriate you can say the length of the DFT, then modify decimation rate with the new rate m equal to L by 2 to speed up factor by half and then take the inverse transform and estimate the same.

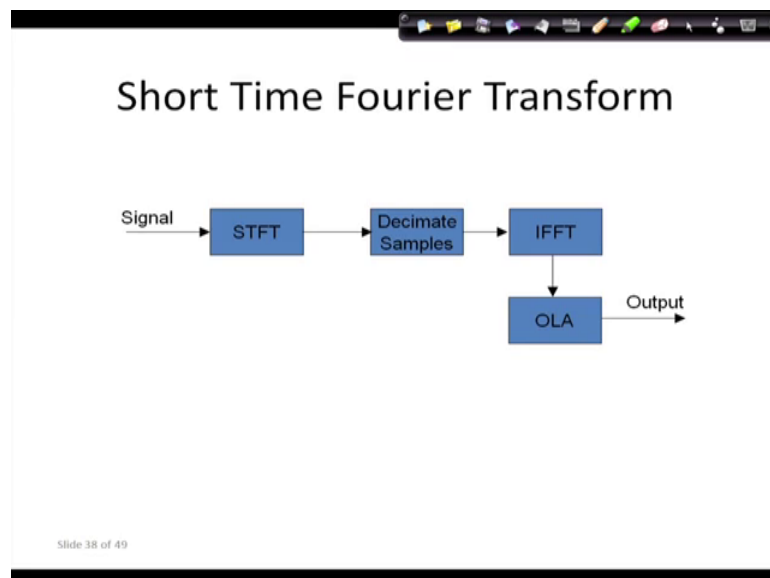
So, STFTM I can use to time scale modification of the pitch signal if you see it here.

(Refer Slide Time: 18:06)



So, this is decimation in time in L this is $L, 2L, 3L, 4L, 5L, 6L$ now once I synthesize time I make the decimation time is L by 2. So, it is half $m, 2m, 3m, 4m, 5m, 6m$ so I can say squeeze the number of samples. So, I can speed up the second if I want to slow down the signal so instead of half I can say $2L$ so it is double. So, this way I can do the time scale modification of the speech signal which is the application of STFTM.

(Refer Slide Time: 18:54)



So, this is the block diagram ok.

(Refer Slide Time: 18:56)

Noise Reduction

- A number of techniques developed to remove/reduce additive noise:
- Noise corrupted signal is given by:

$$y[n]=x[n]+b[n]$$
 - STFT Synthesis:
 - Subtract Noise spectrum $\hat{S}_b(\omega)$

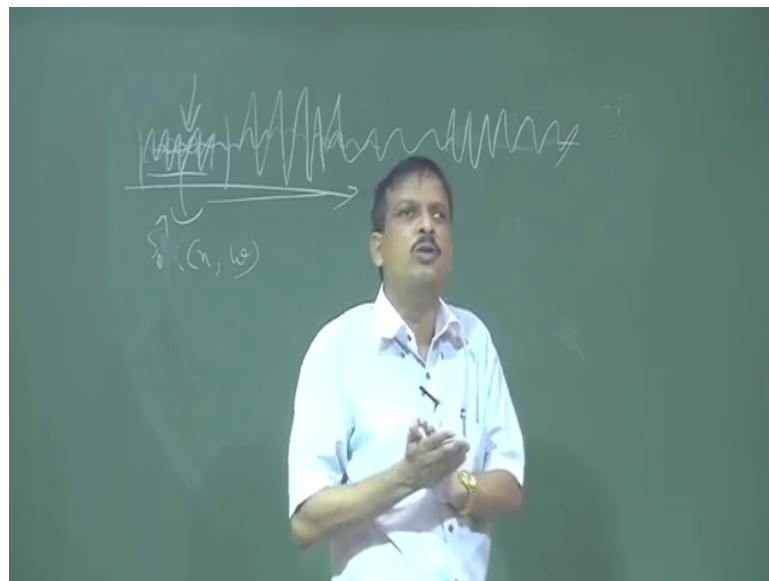
$$\hat{X}(nL,\omega)=\left[|Y(nL,\omega)|^2-\alpha\hat{S}_b(\omega)\right]^{\frac{1}{2}}e^{j\angle Y(nL,\omega)}$$

$$\text{if } |Y(nL,\omega)|^2-\alpha\hat{S}_b(\omega) < 0 \Rightarrow |Y(nL,\omega)|^2-\alpha\hat{S}_b(\omega)=0$$
 - Original phase spectrum $\angle Y(nL,\omega)$ is retained because phase of the noise can not be reliably estimated in general.
 - Factor α is a control of the degree of noise reduction.

39

Next one is the noise reduction, you know that in cool edit there is a also a button called noise reduction.

(Refer Slide Time: 19:08)



So, suppose I have recorded a signal, speech signal this is silence then I start recording this portion I have not speech speak anything then I will start speaking. So, the noise in the silence region also is here the noise also there noise is spread. So, much during the speech and this speech is going on, I want to remove the noise if I consider the noise is stationary signal mean; that means, noise is not changing over the signal then I can estimate the noise power from this silence zone and subtract it from the spectral information of this speech and again re synthesis I can remove the noise. So, what I will

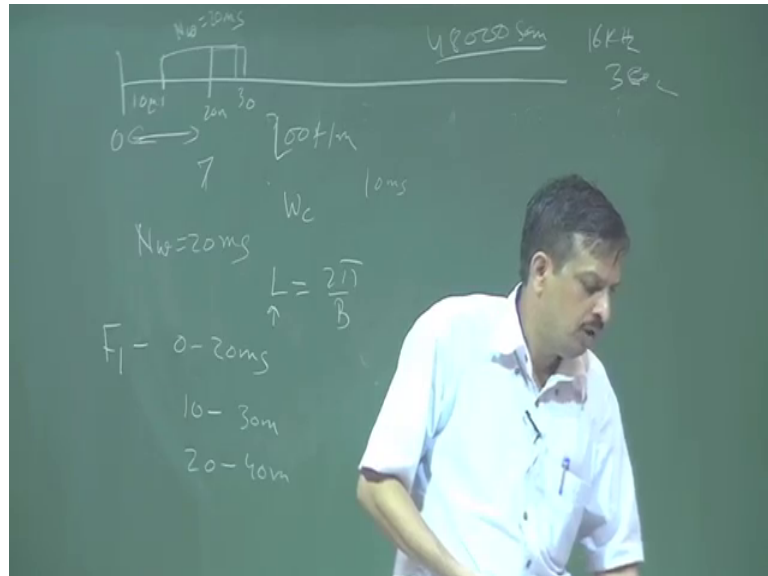
do, I first estimate the noise power by analyzing this portion take a single window and this portion and I can estimate that $n_k n_\omega$ then I apply that noise estimation. So, noise estimation is let's s_b , s_b is the noise estimation. So, I can subtract the noise estimation from the signal spectral estimation and I re-synthesize it I can get back the signal the noise is removed ok.

So, this way there is a lot of, lot of algorithms you can develop. So, there is an algorithm there may be a complex procedural I can find out I can estimate the noise in different way I can subtract not full. So, what is the problem is that if the noise this is a noise and speech signal contains a sibilant sound which is also a noise, if I subtract it the sibilant sound may go so this is the problem. So, then what kind of estimation I can say if this is a sibilant sound then do not subtract this.

So, those kind of algorithm I can write and I can remove the noise from the speech signal. So, STFTM can be used or STFT used for noise reduction, time scale modification, speech enhancement all those all things is that de-synthesis is done by o L a method or overlap add method or f b s method. O L A method is recommended you can write an o L a method algorithm to reconstruct the because I have to get back the signal again only deconstruction is only possible if I suitably choose the window and also the shifting of the window.

So, that is why if you see in any parameter extraction we do those frequency analysis of the speech signal what we done.

(Refer Slide Time: 22:10)

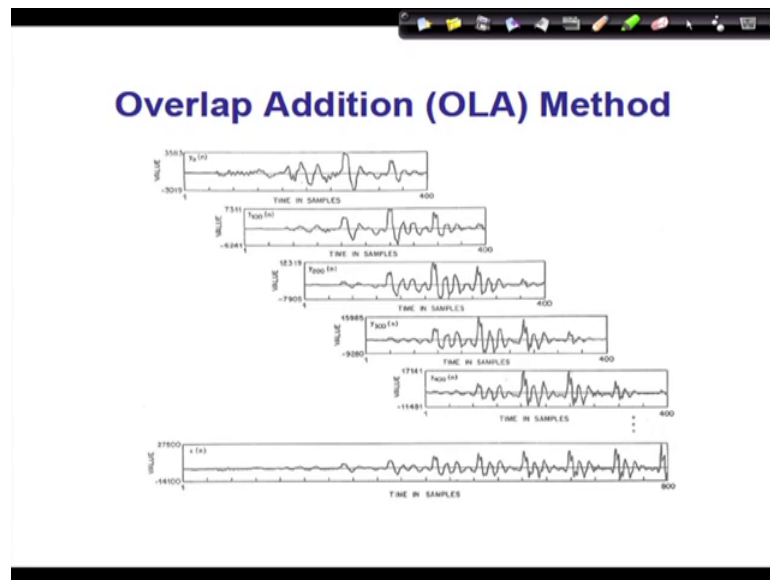


So, suppose I have a long speech signal let's say this is 48000 samples which is 16 kHz sampling rate and 3 second signal. So, I have that number of samples, you have not taken all samples at a time and do the frequency analysis what I will do, we will divide this signal with a frame rate and window length. So, we choose a window length analysis window N_w is equal to let's say 20 milliseconds.

So, N_w is equal to 20 milliseconds, then what we will do which one kind of window you will use either Hamming window, Hanning window or rectangular window what any kind of window we can use depending on the window that defines the ω_c . So, L is equal to $2\pi / B$. So, that defines me what kind of shifting is possible then I shifted the signal that way then take another window from here to here. So, if it is 10 milliseconds shifting. So, this is 10 milliseconds. So, again 10 to so this is 10 milliseconds, this is 20 milliseconds. So, I can first window is 0 to 20 milliseconds. So, first window first frame, first frame 0 to 20 milliseconds, second frame 10 to 30 milliseconds. So, I take 10 to 30 milliseconds third frame 20 to 40 milliseconds.

So, I can now instead of milliseconds I can number of samples I can do. So, I said 100 frames per second; that means, the shifting of the window is 10 milliseconds if I say I want 200 frames per second shifting of the window I can easily understand 5 milliseconds. So, that is why we do the speech processing in frame rate then I can do the STFT I can use theOLA method to get back the signal. So, I analyze the signal this way if you see there is a picture I have shown you, you can go through this slide also this one.

(Refer Slide Time: 24:44)



So, I take frame by frame 70, I can make 75 percent of overlap also I analyze frame this frame this frame and I get back. So, those are the inverse Fourier. So, I get back the frame then I add all those frame I get back the signal ok.

So, this is the overall STFT analysis, the STFTM synthesis is I am not described if you want you can req you can just I want that if you raise in the forum that STFTM is also required or you can go with that book which I have referred that book you can go with the mathematics and if you want, you can I can take one another half an hour class in the end because time will not be permitted to details go to the STFTM. So, this is nothing, but a mathematics all kinds of mathematics deductions will be there. So, you just go through the book if you not understand the deductions then again I come back ok.

Thank you.