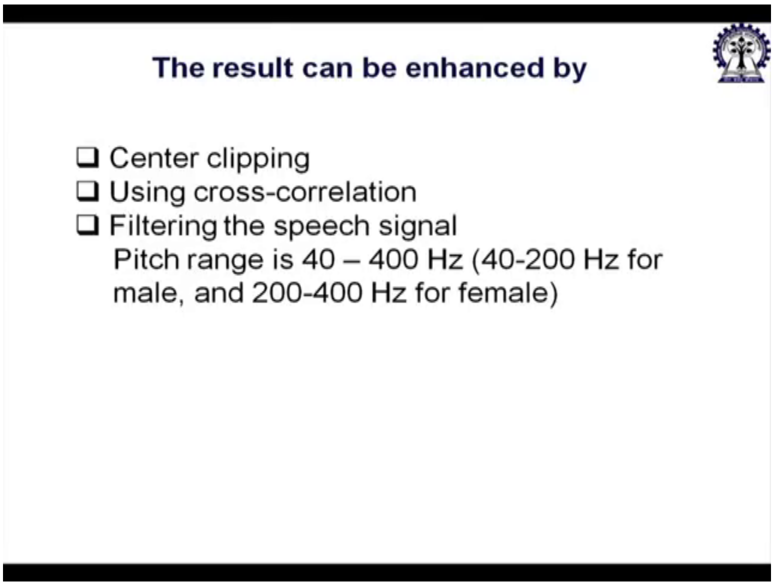
Digital Speech Processing
Prof. S. K. Das Mandal
Centre for Educational Technology
Indian Institute of Technology, Kharagpur

Lecture - 36
Frequency Domain fundamental Frequency Detection Algorithms

So, we have said the cross correlation is one of the method for announce the result of autocorrelation technique. The other method is called central clipping.

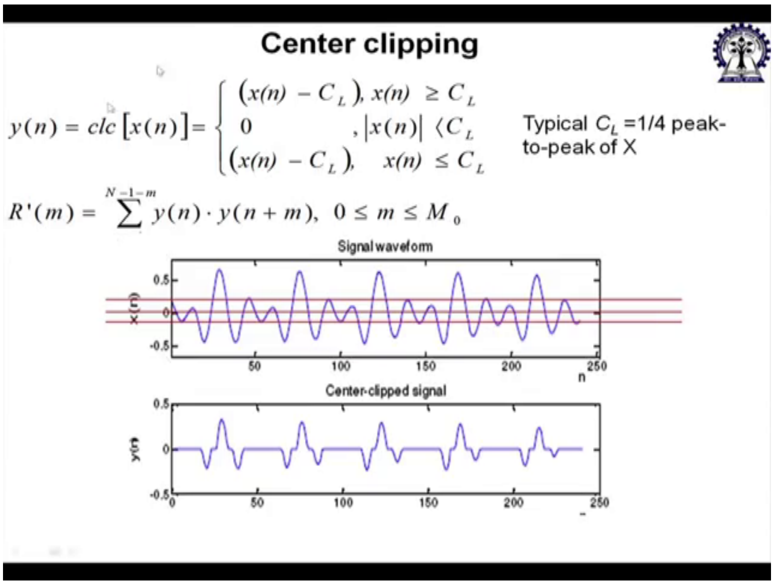
(Refer Slide Time: 00:26)



The result can be enhanced by

❑ Center clipping
❑ Using cross-correlation
❑ Filtering the speech signal
Pitch range is 40 – 400 Hz (40-200 Hz for male, and 200-400 Hz for female)

(Refer Slide Time: 00:30)



Center clipping

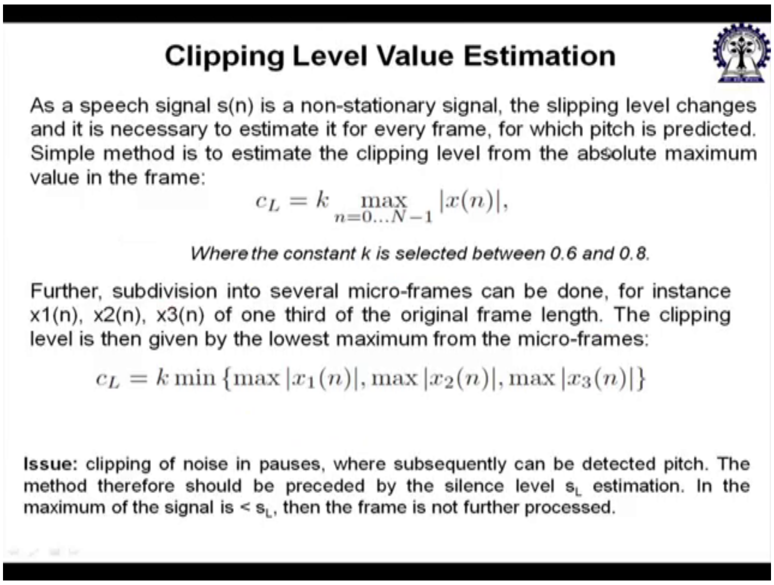
$$y(n) = clc\left[x(n)\right] = \begin{cases} (x(n) - C_L), & x(n) \geq C_L \\ 0 & , |x(n)| \langle C_L \\ (x(n) - C_L), & x(n) \leq C_L \end{cases}$$

Typical C_L =1/4 peak-to-peak of X

$$R'(m) = \sum^{N-1-m} y(n) \cdot y(n+m), \ 0 \leq m \leq M_0$$

If I have this speech signal like this, I want the only the central part. Because if you see that f 0 I want f 0 is or you can say the filtering also. Clipping and that f 0 is up to 5 100 is sufficient. So, I do the central clipping in the signal and all the high amplitude variation can be avoided. Because high amplitude variation give me the auto correlation function is long variation. So, instead of taking the high amplitude variation signal, I can clip the signal in central part and that signal can be used to find out the auto correlation and calculating the f 0. So, this, the algorithm is very simple. So, clipping how do you find out the C L generally it is C L one 4th of the peak. So, how do you do that? Suppose I have taken a 20 millisecond window signal.

(Refer Slide Time: 01:25)



Clipping Level Value Estimation

As a speech signal s(n) is a non-stationary signal, the slipping level changes and it is necessary to estimate it for every frame, for which pitch is predicted. Simple method is to estimate the clipping level from the absolute maximum value in the frame:

$$c_L = k \max_{n=0\ldots N-1} |x(n)|,$$

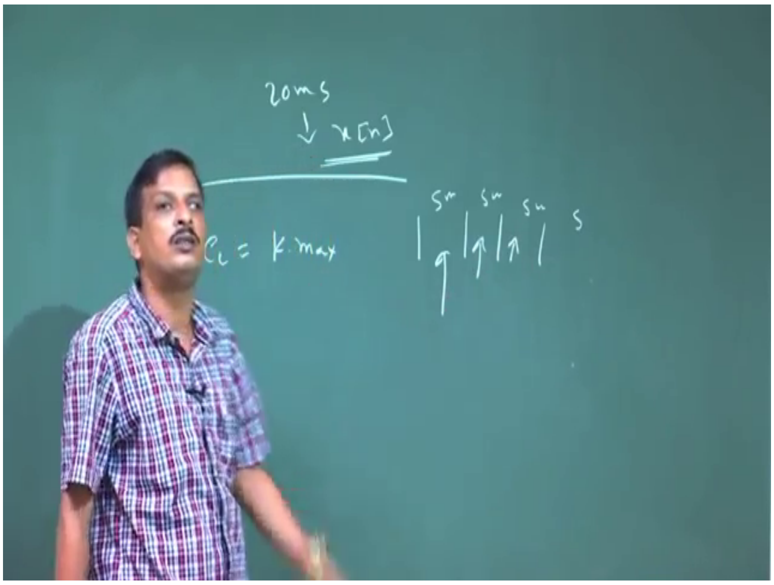
Where the constant k is selected between 0.6 and 0.8.

Further, subdivision into several micro-frames can be done, for instance x1(n), x2(n), x3(n) of one third of the original frame length. The clipping level is then given by the lowest maximum from the micro-frames:

$$c_L = k \min \{\max |x_1(n)|, \max |x_2(n)|, \max |x_3(n)|\}$$

Issue: clipping of noise in pauses, where subsequently can be detected pitch. The method therefore should be preceded by the silence level s_L estimation. In the maximum of the signal is < s_L, then the frame is not further processed.

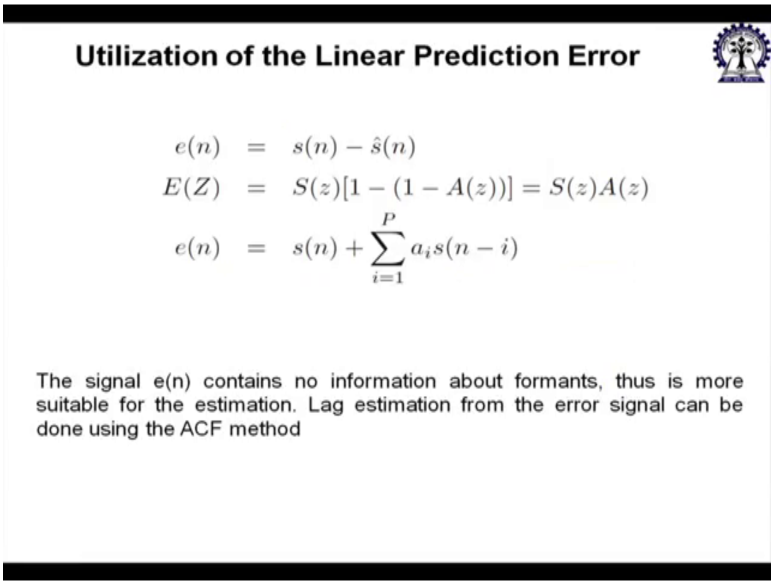
(Refer Slide Time: 01:28)



So, within that 20 millisecond I can find out the maximum peak. So, let us 20 millisecond signal is represented by x n.

So, if I find out the maximum amplitude within the 20 millisecond, then I can further then I can say the k is k into that max amplitude will give me the C L. K value is between 0.6 to 0.8. Now if I want that I have a signal which is lot of variation in amplitude.

(Refer Slide Time: 02:35)



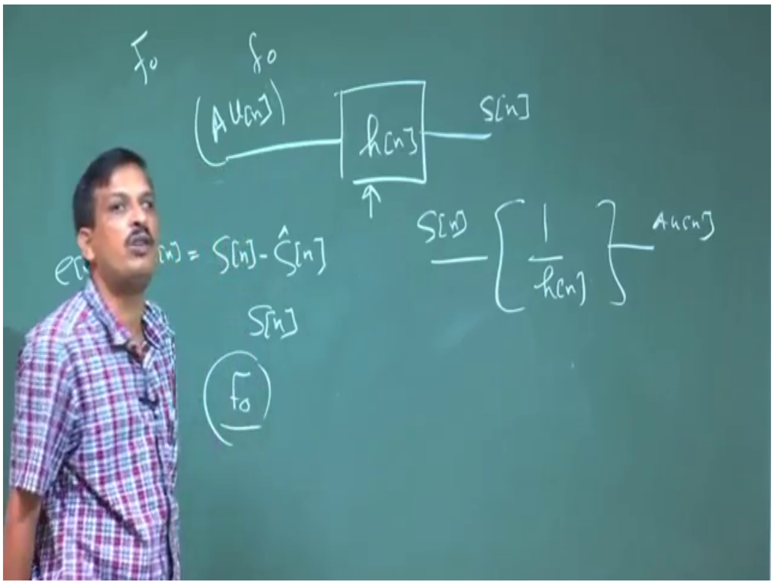
Utilization of the Linear Prediction Error

$$
\begin{aligned}
e(n) &= s(n) - \hat{s}(n) \\
E(Z) &= S(z)[1-(1-A(z))] = S(z)A(z) \\
e(n) &= s(n) + \sum_{i=1}^{P} a_i s(n-i)
\end{aligned}
$$

The signal e(n) contains no information about formants, thus is more suitable for the estimation. Lag estimation from the error signal can be done using the ACF method

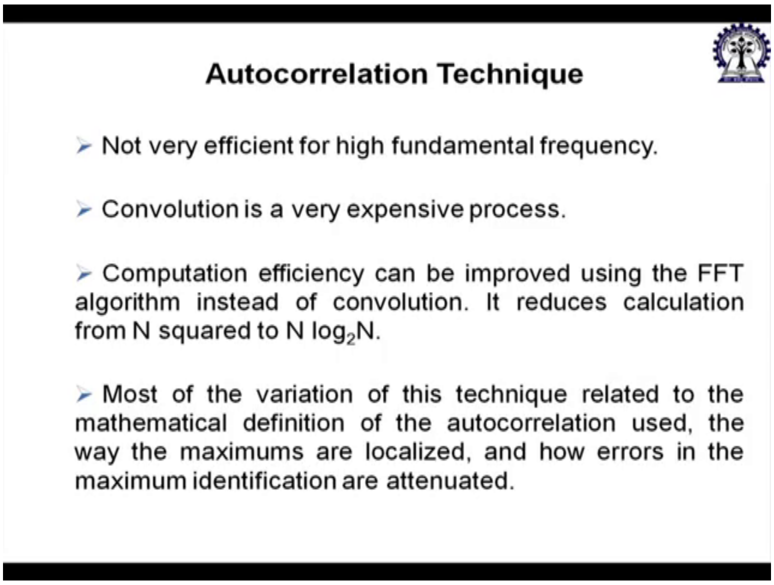
So, instead of taking the single max I can divided this 20 millisecond signal in maybe another 4 segment. Maybe 5 millisecond, 5 millisecond, 5 millisecond, 5 millisecond and from each segment find out the maxima. And take the mean of that maxima to find out the C L, value that is the formula. So, this is called central clipping to improve the autocorrelation base f 0 extraction. Another procedure, instead of taking the signal if you remember in LPC analysis, that if this is my vocal tract transfer function h n, and if I apply A u n then I get speech signal S n.

(Refer Slide Time: 02:47)



Now, if you remember that linear prediction, what I want to predict signal S n for I and if I must have substrate the predicted signal. So, the remaining is nothing but a A u n. Which is nothing but e n a r signal is nothing but excitation signal, inverse filtering. I can say that I have designed 1 by h n and I pass the signal S n, I get A u n which is the error LPC prediction error. Now if you remember the f 0 is part of the excitation not the vocal tract. So, f 0 information is here not in here. So, I can extracted the f 0 from en signal. So, instead of using S n, speech signal whose had a lot of radiation which is introduced by h n. So, instead of take the S n, I can pass the S n to a universe filter to get the A u n which is the error signal of the LPC analysis, from the LPC error signal I can find out the f 0 value using autocorrelation is ok.

(Refer Slide Time: 04:42)

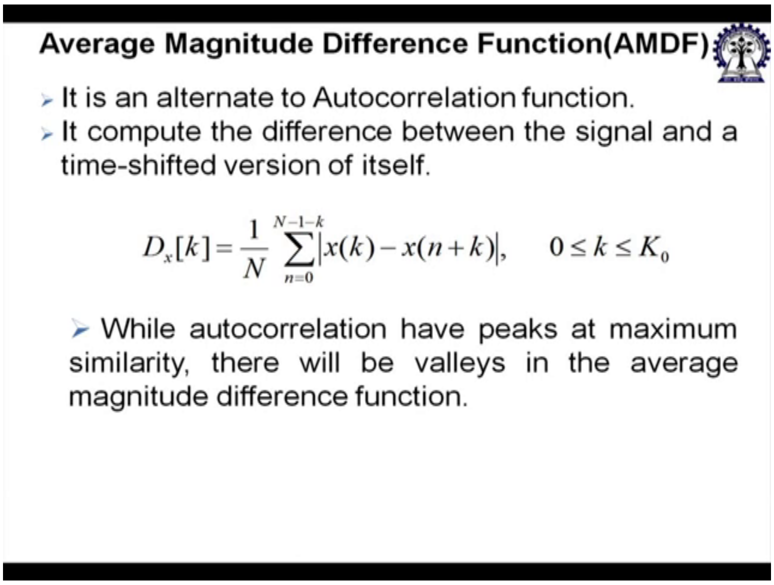


Autocorrelation Technique

- Not very efficient for high fundamental frequency.
- Convolution is a very expensive process.
- Computation efficiency can be improved using the FFT algorithm instead of convolution. It reduces calculation from N squared to N $\log_2 N$.
- Most of the variation of this technique related to the mathematical definition of the autocorrelation used, the way the maximums are localized, and how errors in the maximum identification are attenuated.

Next that not very efficient or high fundamental convolution is very expensive process that some close and cause of autocorrelation technique. So, computation efficiency can be improved if I am if I use the FFT algorithm to implement that correlation calculation of the correlation function ok.

(Refer Slide Time: 05:07)



Average Magnitude Difference Function(AMDF)

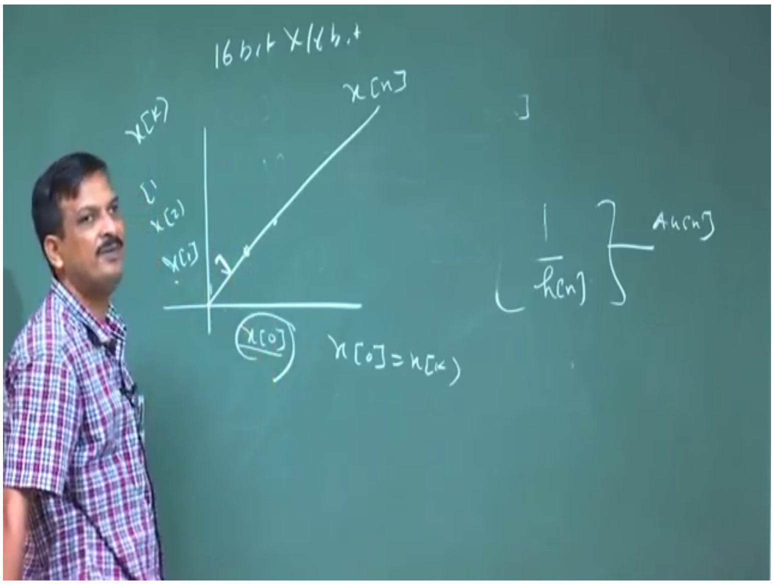
- It is an alternate to Autocorrelation function.
- It compute the difference between the signal and a time-shifted version of itself.

$$D_x[k] = \frac{1}{N}\sum_{n=0}^{N-1-k}\left|x(k) - x(n+k)\right|, \qquad 0 \le k \le K_0$$

- While autocorrelation have peaks at maximum similarity, there will be valleys in the average magnitude difference function.

So, this you can read it. Now instead of autocorrelation since, autocorrelation is required the same sample is multiplied by the sample. So, if it is a if you see, if it is a 16 bit signal, 16 bit multiplied by another 16 bit sample. See worst case it requires lot of memory, and also multiplication is computer scenario complex.

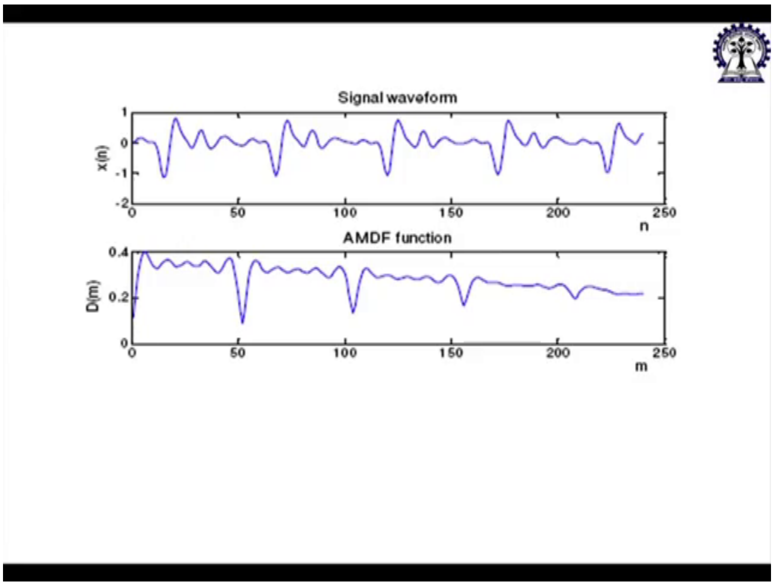
(Refer Slide Time: 05:21)



So, for hardware implementation or when that computation facility is not that enough, you can use average magnitude difference function, which is called AMDF, average magnitude difference function, average magnitude difference function. So, what is the what are the theory what are the methods in here. So, suppose you have a signal x n, which may have a period which has a period which a periodic signal, some fundamental frequency is there. Now if I want to plot x 0 versus x 1, then x 0 versus x 2, x 0 versus x 3 are that way. So, x 0 versus x 0 will become here. X 0 versus x 1 some come here. So, I can get a scatter plot, if x this sample is exactly match with another sample let us x k, then I get a point in here. X 0 is equal to x k. Then I can get a point in here, then I can get a point in here.

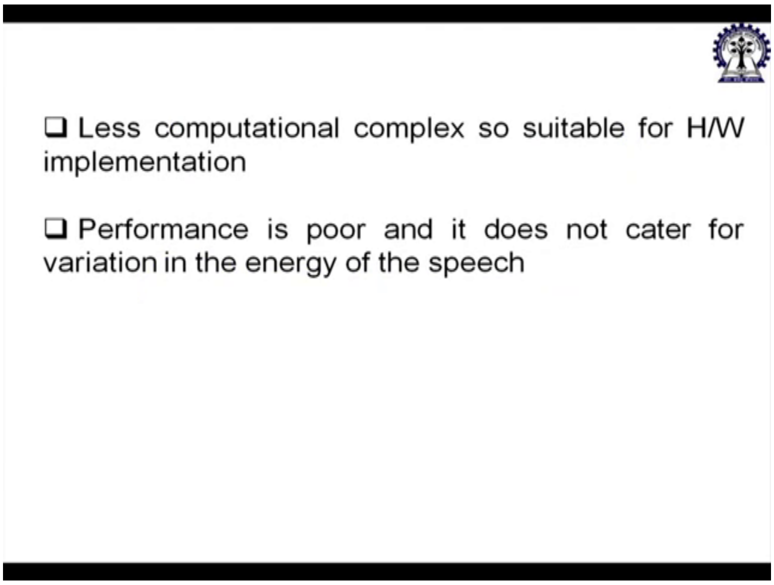
So, along the 45 degree line all x 0 which is matched with x k. Since which is not a you can say it is a not exactly periodic signal then what is it? So, instead of calculating the exactly 45 degree line what I want I want to calculate the deviation from the 45 degree line. And find out the average value of the deviation for each delay, then plot that average deviation value against the delay. So, what is the function 1 by n n equal to 0 to n minus k x k minus x sub n plus k. So, for all the sample and d x k. So, k equal to 0. So, it is 0 minus. So, x 0 minus x 0 x 0 minus x 1 x 0 minus x 2 x 0 minus x 3. So, all for all deviation I am finding out the deviation and add them and find out the average value of the deviation (Refer Time: 08:13) next I calculate k equal to 1. So, instead of x 0, I put x 1, I put x 2, I put x 3. So, all deviation average deviation will be there. So, deviation will be minima when it matched with the period. So, if I plot the deviation deviation plotted will look like this.

(Refer Slide Time: 08:40)



So, minimum deviation will be happen at the period or at the f 0. So, this is f 0, this is twice f 0 this is thrice f 0. So, this can this technique is used to find out the fundamental frequency, when the computation power is less, but what is the drawback. Less computational complexities ok.

(Refer Slide Time: 09:06)



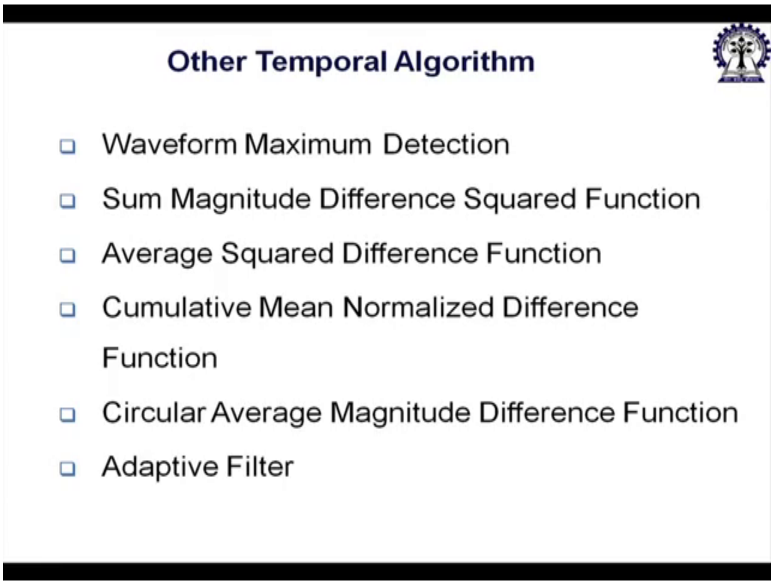
□ Less computational complex so suitable for H/W implementation

□ Performance is poor and it does not cater for variation in the energy of the speech

In implement hardware, the performance is poor and it does not cater for variation in energy of the speech signal. If you see the deviation depends on the variation of the energy. So, suppose this time I speak some speech which is less energy and next time is high energy, the deviation will change. All I can if I say suppose that a long vocalic region, that they the amplitude has energy had changing. Then there will be a variation will f 0 will a call here also. So, that kind of energy the amplitude problem is there.

So, autocorrelation is highly used and this, this is used when I can say I require a average (Refer Time: 09:55) estimation of f 0 is sufficient, but computational complexity will be very less in that time I use AMDF average magnitude difference.

(Refer Slide Time: 10:10)

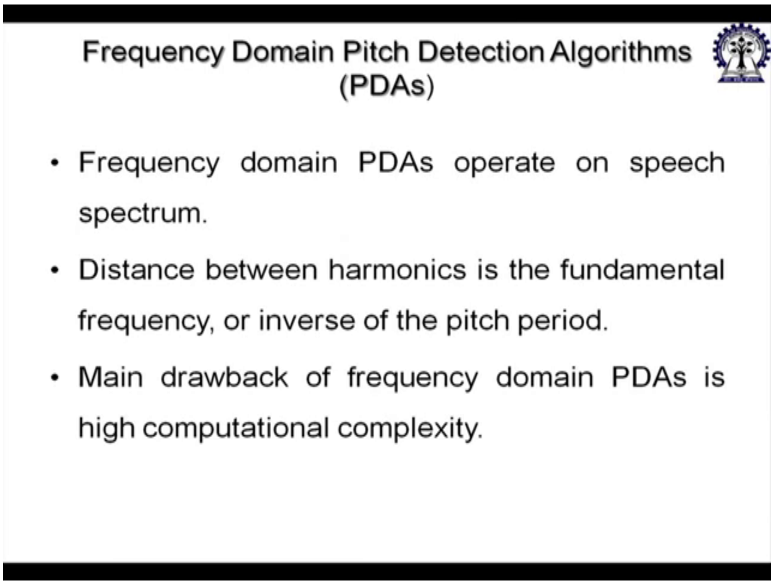


Other Temporal Algorithm

- □ Waveform Maximum Detection
- □ Sum Magnitude Difference Squared Function
- □ Average Squared Difference Function
- □ Cumulative Mean Normalized Difference Function
- □ Circular Average Magnitude Difference Function
- □ Adaptive Filter

There is other algorithm also wavefrom maximum detection some magnitude difference. So, all others methods are also adaptive filtering circular average magnitude difference function cumulative mean normalized different function, so all others methods also available.

(Refer Slide Time: 10:27)



Frequency Domain Pitch Detection Algorithms (PDAs)

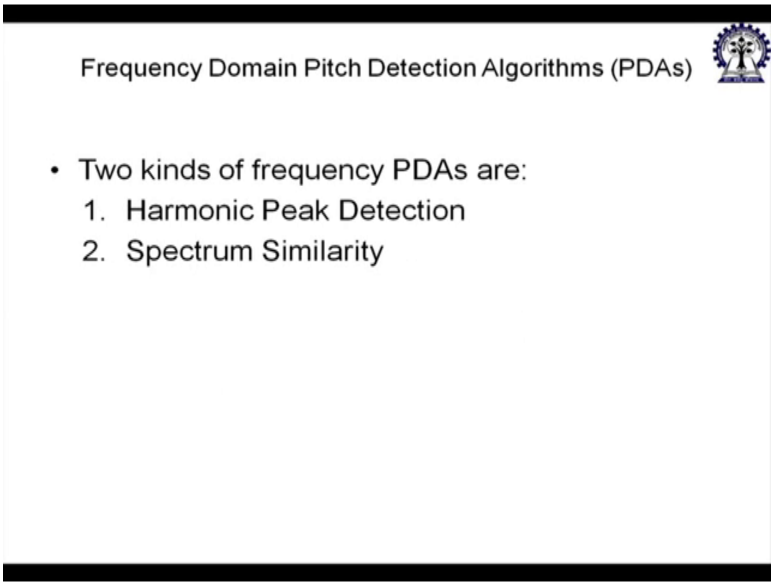
- Frequency domain PDAs operate on speech spectrum.
- Distance between harmonics is the fundamental frequency, or inverse of the pitch period.
- Main drawback of frequency domain PDAs is high computational complexity.

Then frequency domain PDA's pd detection algorithm or f 0 detection algorithm think (Refer Time: 10:32) and pd operates on speech spectrum. So, distance between the harmonics is the fundamental frequency or inverse of the speech period. So, what is the

(Refer Time: 10:45) behind the frequency domain PDA, that if a signal consists f 0 then harmonics will be repeated f 0 twice f 0, 3 f 0, 4 f 0.

So, from there I can calculate the f 0. So, main drawback of frequency domain PDA is the high computational complexity, any signal to trans transform in frequency domain is a computation complex ok.

(Refer Slide Time: 11:17)

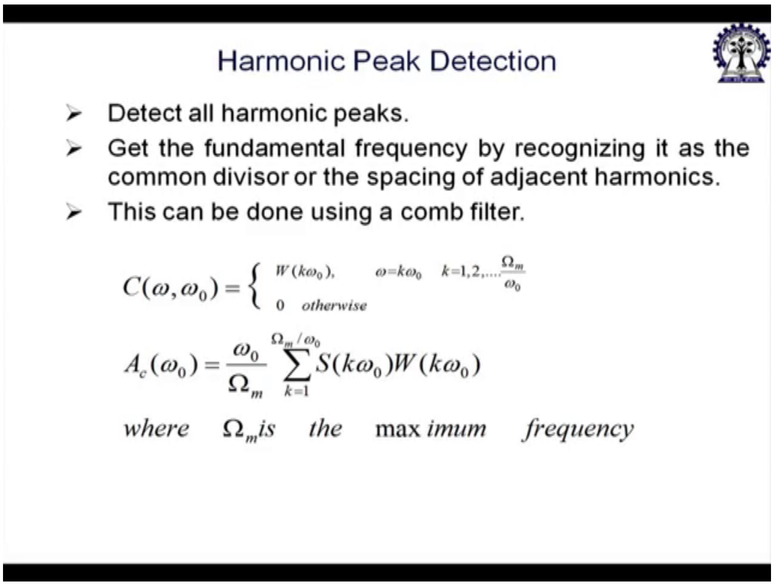


Frequency Domain Pitch Detection Algorithms (PDAs)

- Two kinds of frequency PDAs are:
  1. Harmonic Peak Detection
  2. Spectrum Similarity

Now, then the 2 kinds of PDA are all the mostly used harmonic peak detection and spectrum similarity.

(Refer Slide Time: 11:25)



Harmonic Peak Detection

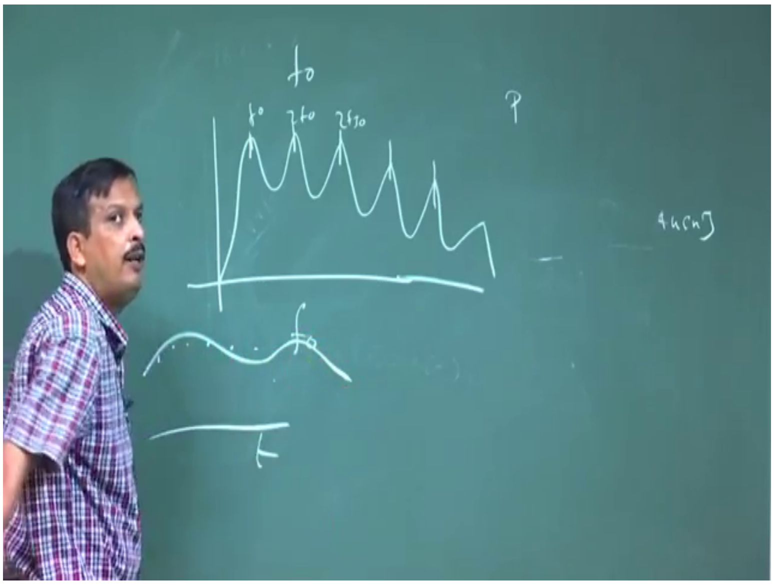
- Detect all harmonic peaks.
- Get the fundamental frequency by recognizing it as the common divisor or the spacing of adjacent harmonics.
- This can be done using a comb filter.

$$C(\omega,\omega_0) = \begin{cases} W(k\omega_0), & \omega = k\omega_0 \quad k = 1,2,\ldots\frac{\Omega_m}{\omega_0} \\ 0 & otherwise \end{cases}$$

$$A_c(\omega_0) = \frac{\omega_0}{\Omega_m}\sum_{k=1}^{\Omega_m/\omega_0} S(k\omega_0)W(k\omega_0)$$

$$where \quad \Omega_m is \quad the \quad \max imum \quad frequency$$

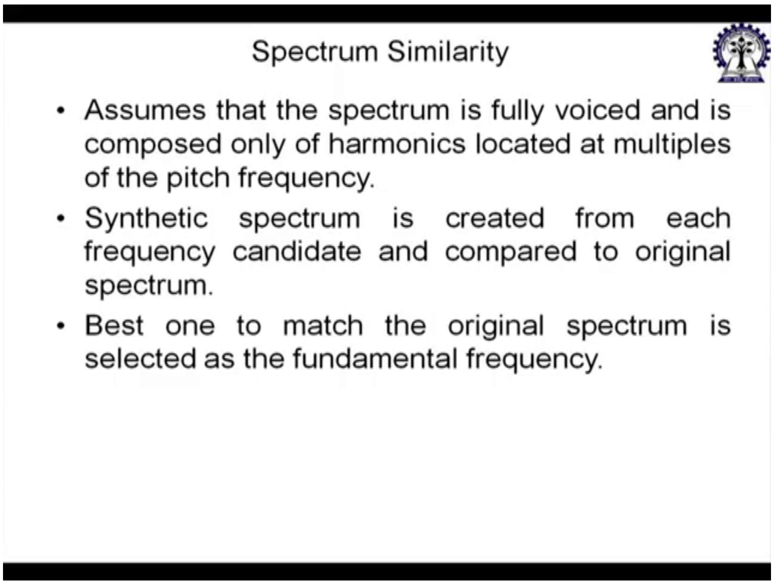
(Refer Slide Time: 11:28)



So, detect all the harmonics peaks. So, suppose I have a f 0, this is my spectrum let us this kind of spectrum is there. So, if you see if this is f 0 this is twice f 0 all harmonics are repeated thrice f 0. So, from there I can find deep this is to f 0. So, I can find out the repetition distance and that give me the f 0 value. So, how do you do that? It is called comb filtering, if you remember the comb, has a peak and gap.

So, if I say this is f 0 beginning twice f 0 3 f 0 4 f 0 then I can find out the average difference between this gap, rap average difference that give me the average value of the f 0, that is the idea.

(Refer Slide Time: 12:20)

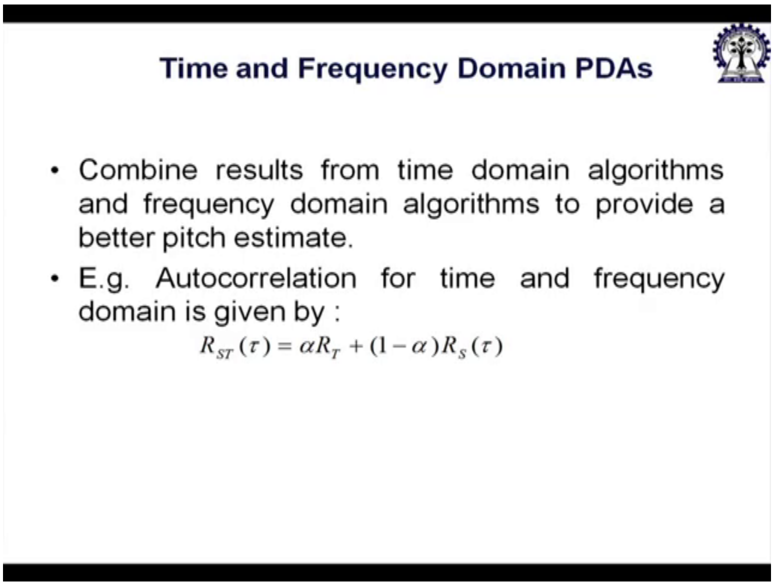


Spectrum Similarity

- Assumes that the spectrum is fully voiced and is composed only of harmonics located at multiples of the pitch frequency.
- Synthetic spectrum is created from each frequency candidate and compared to original spectrum.
- Best one to match the original spectrum is selected as the fundamental frequency.

Spectral similarity I have a spectrum I can simulate this is my f 0 this is twice f 0 this is thrice f 0 and try to similar, find out the correlation whether it is matches with the original spectra or not once it is matches you said this is my f 0. Spectral simulate device f 0 detection. There is a time frequency domain original detection also a lot of algorithm on time frequency based f 0 detection.

(Refer Slide Time: 12:44)

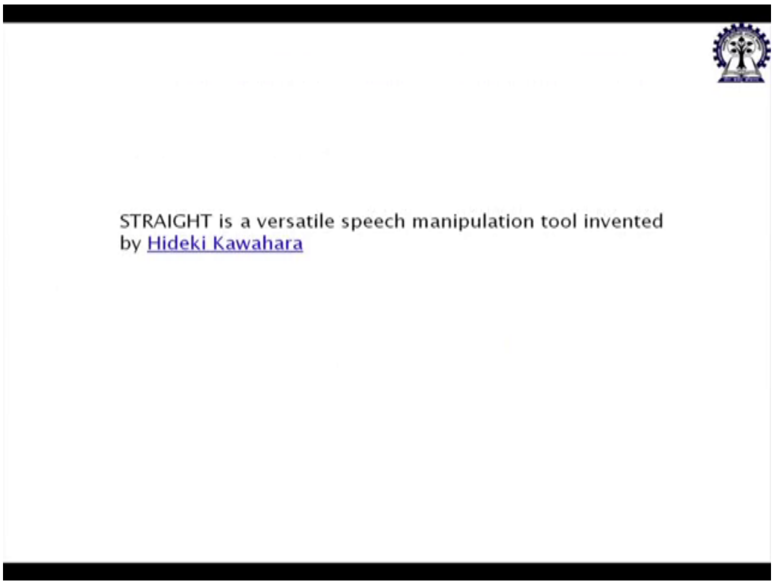


Time and Frequency Domain PDAs

- Combine results from time domain algorithms and frequency domain algorithms to provide a better pitch estimate.
- E.g. Autocorrelation for time and frequency domain is given by :

$$R_{ST}(\tau) = \alpha R_T + (1-\alpha) R_S(\tau)$$

(Refer Slide Time: 12:48)



STRAIGHT is a versatile speech manipulation tool invented by Hideki Kawahara

But if you see most of the time today recently, the state agents software which is available in the free and open source by which you can extract the f 0 reliably. There is a other software also if you can use the plot also. Plot also you can export by f 0 they plot use the auto correlation based technique to extend f 0.

So, using plot also you can extract the f 0 value. So, this is the f 0 extraction you can say. So now, if I give you a speech signal you can extract the f 0 and you can plot the f 0 against the time. So, for every 10 millisecond again if I got a 0 value I can plot it to find out the f 0 contour. During the prosodic modeling I will discuss lot about f 0 (Refer Time: 13:37) and if you see how this f 0 contour is useful to model the prosody even if f 0 can contour can be useful in ASR also. I will show you that things in lecture number a or week 8, week 8 prosodic modeling will go week 8 and the speech application will be the next week this week 7.

Thank you.