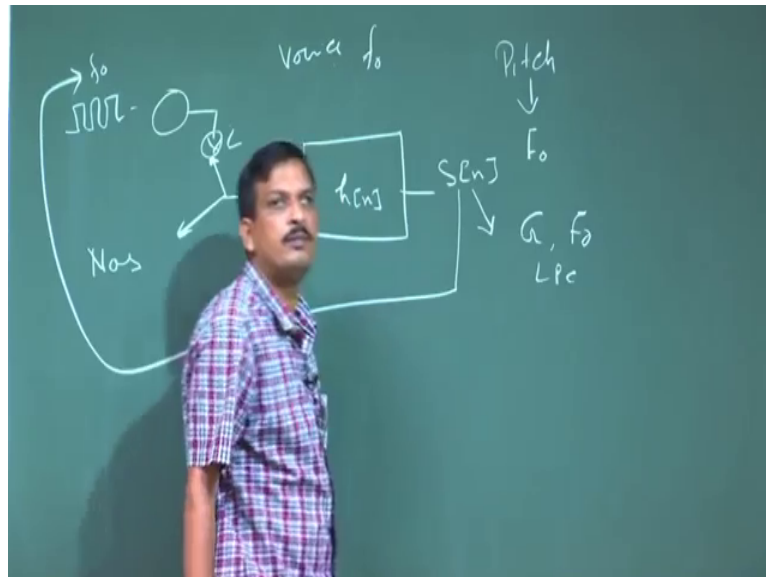


Digital Speech Processing
Prof. S. K. Das Mandal
Centre for Educational Technology
Indian Institute of Technology, Kharagpur

Lecture - 35
Fundamental Frequency Detection of Speech Signal

So, we have discussed about the frequency domain parameter extraction and LPC parameter extraction already covered. So, there is a one important parameter which has to be extracted from the fundamental or the speech signal is the fundamental frequency or F_0 , sometime it is referred to pitch Extraction.

(Refer Slide Time: 00:35)



Pitch, pitch extraction, but pitch is not a physical parameter. It is a perceptual parameter. That is why you call F_0 extraction, fundamental frequency extraction. Now if you think that if you remember that my tube modelling lectures and LPC lectures.

So, what we said that while human being produces the speech. So, there is a vocal tract, that vocal tract function which is h_n . And it is modulated at the input is call got all excitation. So, got all excitation, either vocal code is closed or vocal code is open. If vocal code is closed then there is will be a vocalic sound or bias sound. So, if it is a bias sound that can be modeled as a impulse strain modulated by a got all transfer function, and then multiplied by the gain.


So, when we produce the speech, either speech has a voice segment or unvoiced segment. If it is voice speech then it will be connected to here. If it is a unvoiced speech then there will be noise sound and source and it will be connected to here. Then I get the speech signal s_n . So now, what is f_0 ? So, f_0 is part of the vocal code oscillation. So, if it is vocal code oscillation. So, f_0 we will be here the impulse by which we are simulating the vocal code vibration has a period. So, that period is responsible for fundamental frequency of the speech signal.

Now, think about the reverse thing. I want from the S_n to extract f_0 . I have the worst signal which is recorded, I want to know if it is os, what should the input if I it is generated using a impulse response. So, impulse strain what is the fundamental frequency of the impulse strain. Or if you remember the LPC synthesis, from the speech signal we extracted gain fundamental frequency and LPC parameters. So, fundamental frequency is required to generate the impulse by which frequency impulse should be going through the vocal tract modification got all filter and produce that got all source.

So, that fundamental frequency is part of the (Refer Time: 03:33) excitation. It is not part of the vocal tract transfer function, understand? Some of the speech or even this from the speech signal how do I find out the fundamental frequency. That is called f_0 extraction. f_0 extraction has many application.

(Refer Slide Time: 04:00)

Fundamental Frequency Utilization

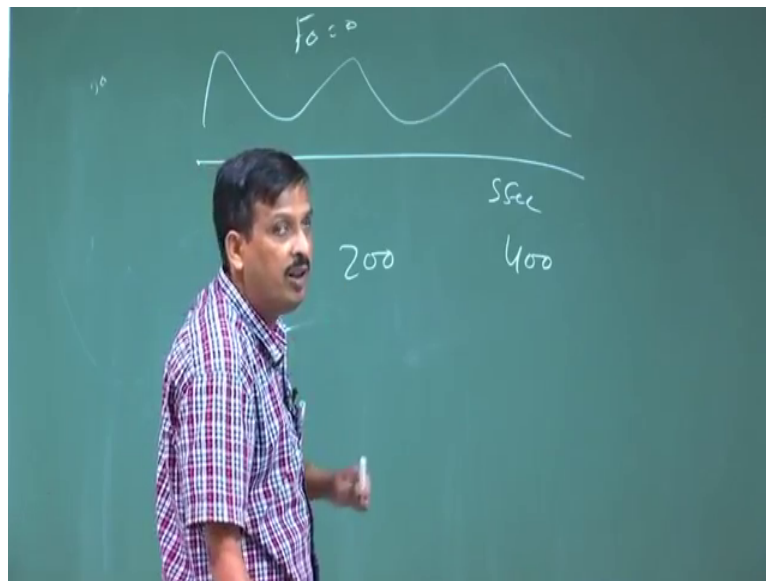


- ❖ Speech synthesis – Prosody Modeling.
- ❖ Speech Recognition
- ❖ Speech coding
- ❖ Voice conversion
- ❖ Spoken language learning

Mainly, when I discussed about the next week speech application, there is a called speech synthesis. There the f_0 is a main important part to produce the melody in the speech ok.

So, when I speak if you see, when I speak it is not a robotic voice. If I speak a sentence the fundamental frequency is changing, it is not fixed all throughout the sentence fundamental frequency is same. So, suppose I have a speech sentence.

(Refer Slide Time: 04:37)

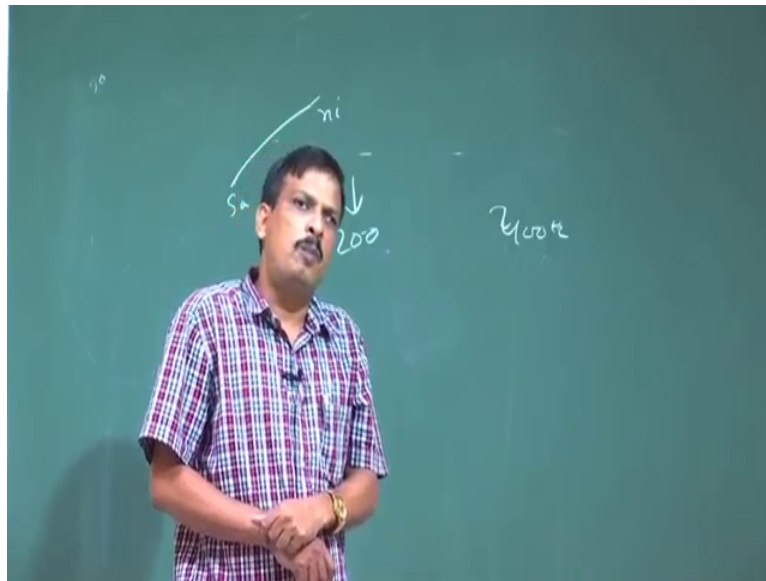


Not throughout the if it is 5 means 5 second signal not throughout the 5 5 second f_0 is constant. f_0 is moving that is called fundamental frequency control. Now if I remove this f_0 movement from the speech the speech become robotic. So, the melody of the speech communication is f_0 is one of the main factor for speech melody. In singing also if you think about the singing what is sa re ga ma pa da ni sa. If you think about sa is the best frequency next sa is the octave past octave. So, if it is my best sa is 200 hertz.

So, sa to sa it is 400 hertz next sa is 400 hertz. And then this 200 hertz frequency is divided in several scale if it is sa re ga ma pa da ni sa there is 7 scale and this can be a uniform or this can be non uniform, if it is uniform then you can say equip temper scale. Now if you see that there is you can say that when I say the raga of a speech song or in a rag different ragas like biravi rag biravi, how people are saying that you are not touching the upper node you are not you are missing the node.

So, those kind of command you have seen from that lot of TV that that that the song competition in the TV channel, and you find that the judges are produce a comment. You are your upper tone upper upper upper tones are not touching or you are not touching the upper nodes what is meaning? So, meaning is that the movement of the f_0 is controlled during the production. So, when I singing a some song I am controlling the movement of the f_0 . So, if it is sa then it is ni.

(Refer Slide Time: 06:59)

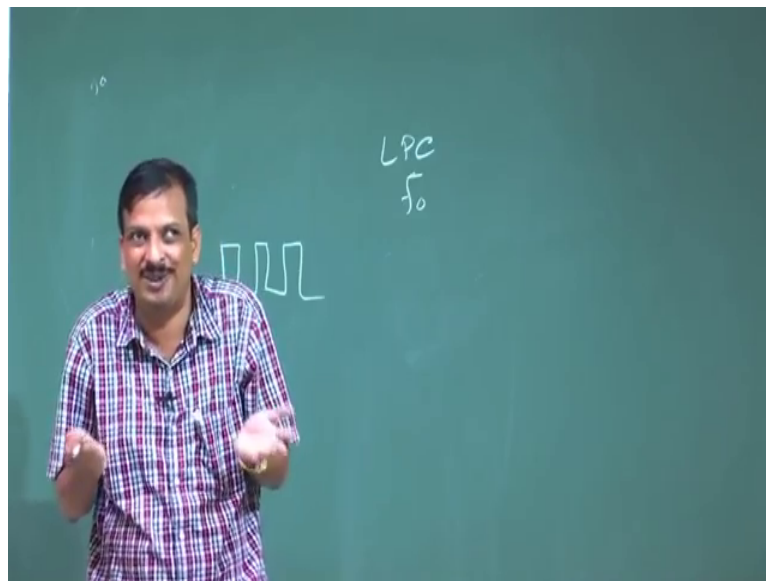


Then sa to ni ni da pa ma. So, movement of the f_0 is actually defined by that sa re ga ma pa da ni sa. So, maybe while they will moving from lower node to upper node, he may not be touched the upper node just falling down because you know sa to sa.

So, if it is my first sa is 200 hertz next upper sa must be 400 hertz. So, when I sing if it is upper sa, then I have to reach my I have to change my fundamental frequency from lower sa to upper sa. So, fundamental frequency actually responsible for the melody of song you can say the this ga ma this define in any song, and also in voice communication when is talked about when I speak it is not always fundamental frequency is same. Fundamental frequency is moving to provide a smoothing communication between human beings, unless it will be a robotic voice. Then people will initially people are thinking that fundamental frequency has a no use in speech recognition, or maybe they have used the lcp parameter whatever we extracted that is sufficient.

But really in modern speech recognition systems people are thinking, that why the fundamental frequency can be used effectively to improve the speed recognition rate. The new dimension is coming, which is called spoken language processing. I will discuss I will discuss some issue in speech application, spoken language processing. That whatever we are doing ASR, it is not ASR because spoken language is different from written language suppose discuss and I will in the next week lectures. Then speech coding yes already we have said that if we even if I basic LPC.

(Refer Slide Time: 09:09)



And coding f_0 has to be provided in the receiver end, because I have to know what should be the f_0 of the input impulse, which I will to generate that voice signal.

So, f_0 is very, very important part for speech coding. Then voice conversion very important if you see this is new technology that voice conversion action conversion. Suppose I want we will discuss the details in a speech application classes, that suppose that this is my dream that can I develop a technology, somebody is speaking in American English and I am listening whatever English I am producing. I am not saying speech to speech, I am not saying the English is converted in my mother tongue I am saying I am converting the same English speech which is spoken by an American speaker to my and I am I am listening in my own English whatever English I am producing, because my English is not American English think about the Japanese professor giving a lectures. If I

am listening because my I cannot understand the Japanese English that clearly. So, my understanding level will be going down.

Now, if I able to discover a device that Japanese professor speaking his own accented English, but I am listening in my own accented English, then I think the problem is solved. So, that is a technology. So, voice converse and speech conversion action conversion then a spoken language learning even if singing. Suppose I want that if I want computer based singing learning of singing. So, I can say somebody is singing let us there is a guru sing the song. So, he touches the node 1 2 3 4. So, all node movements are showing now as a disciple I am singing that same song, and computers showed to shown me that you started here, but your tradition from this node to this node taking this much of time and you are not touching this node. So, next time I am producing next way then I you are gradually touching this next node.

So, that can be used as a tools for learning, similarly language learning also. So, suppose I that if you name any language has a tone as a phonemes or I can say the trace as an important parameter for speech clarity. So, those cases f_0 is very important. So, there are numerous application for f_0 . So, extraction and f_0 frequency or fundamental frequency is an important issue. Even if not only speech, but also music signal also. Lot of music research is going on and they are also the extraction of the fundamental frequency is very important, to find out the movement of the node.

So, ragas is defined nothing but a movement of the node. So, you to find out the nodes are move from the signal itself. So, there are a number of things. So, there is a number of problem.

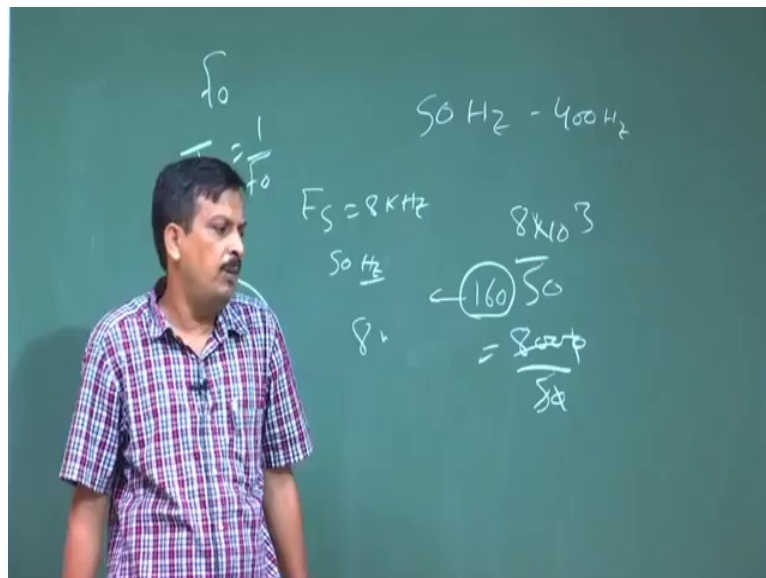
(Refer Slide Time: 12:22)

Fundamental Frequency Characteristics

- F_0 takes the values from 50 Hz (males) to 400 Hz (children)
→ If $F_s = 8000$ Hz these frequencies correspond to the lags $L = 160$ to 20 samples. It can be seen, that with low values F_0 approaches the frame length 20 ms, which corresponds to 160 samples
- The difference in pitch within a speaker can reach to the 2:1 relation.
- Speech is called quasi periodic signal small changes between the period. These small shifts are called "jitter"
- F_0 is influenced by many factors – usually the melody, mood, distress, etc.

And number of characteristics also in the fundamental frequency if it is produced by a human being then you know the f_0 can be values from 50 hertz to 400 hertz this is a variation of f_0 child speech can have a 400 hertz frequency.

(Refer Slide Time: 12:30)



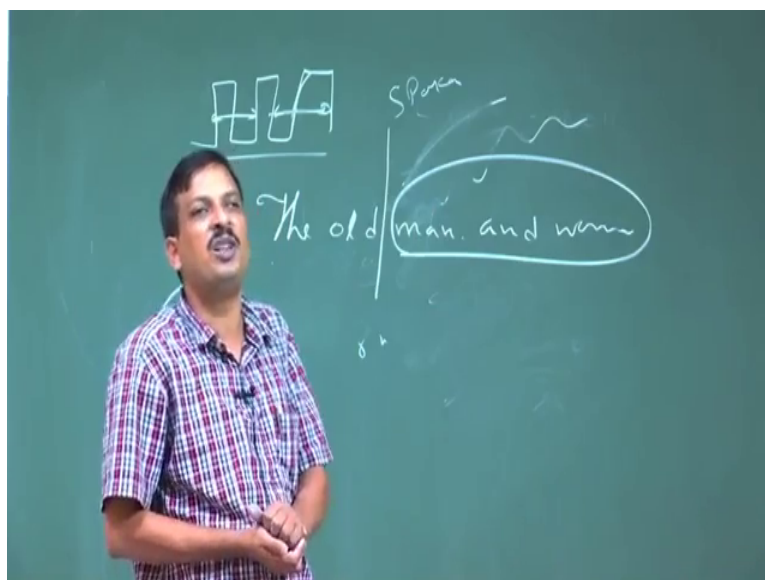
Now, if you see F_s if sampling frequency is my 8 kilo hertz let us 8 kilohertz. And if I say that length if f_0 is 50 hertz, then how many samples will be there. How many samples will be there? My f_0 is 50 hertz and F_s is 8 kilo hertz. So, I can say 8 k 8 into 10 to the power 3 divided by 50. So, it is nothing but 8,000 divided by 50, 160 sample.

160 sample in 1 f_0 period. So, if I say 160 sample how many milliseconds? 8 this sample is 10 milliseconds.

So, duration if it is 50 hertz then I have a 20 millisecond duration fundamental frequency. So, if f_0 then what is the division t_0 is nothing but a $1/f_0$ 20 millisecond length signal is one single f_0 . Similarly if you see if I have a best frequency is 50 hertz my f_0 can varies up to 200 hertz within a single sentence can varies because octave second one is 2 is to 1. Next speech is called quasi periodic, if you see the speech signal it is not exactly periodic. So, speech signal is not exactly periodic.

So, when I generated the speech signal using unit impulse response the f_0 .

(Refer Slide Time: 14:43)



Between the 2 conjugative period is not same, they are some differences is there. And that small differences is called jitter. So, suppose I generate an f_0 I have extracted an f_0 I cannot say all the f_0 will value will be same for some segments. The f_0 is continuously moving oh there may be a difference forward backward that kind of (Refer Time: 15:16) So, it is not a smooth f_0 , it can be look like this kind f_0 control. So, next frequency next period it is less next period also jitter is randomly varying that is why it is called quasi periodic signal not exactly periodic signal. Then f_0 is influenced by many factors. I cannot say that my f_0 will be same for all time, this time I say one sentence, next time If I say same sentence f_0 may not be the same. f_0 may be different, because it

depends on my emotion who is doing on emotion analysis or find out the speaker characteristics f_0 has a part, emotion.

Most important is information is communicated by f_0 only. Even if you see if I say spoken language spoken words and written words are different. If I say the old man and woman the old very interesting example, the old man and woman the old man and woman if I say the old man and woman, then the man and woman both are old.

If I say the old man and woman man is old woman is not old. So, in spoken language if you see that if you if you if you recorded it and if you see the f_0 movement you find the f_0 define the boundary whether man and woman is aroused together whether the old man an women; that means, man and woman both are old. So, in that case maybe the word boundary will be here which is called spoken word boundary. Details I will discuss in that prosody when I do I prosody the analysis and prosody use of prosody information in speech processing in the 8th week lecture ok.

So, f_0 has an important part, but extraction f_0 has a challenge also.

(Refer Slide Time: 17:35)

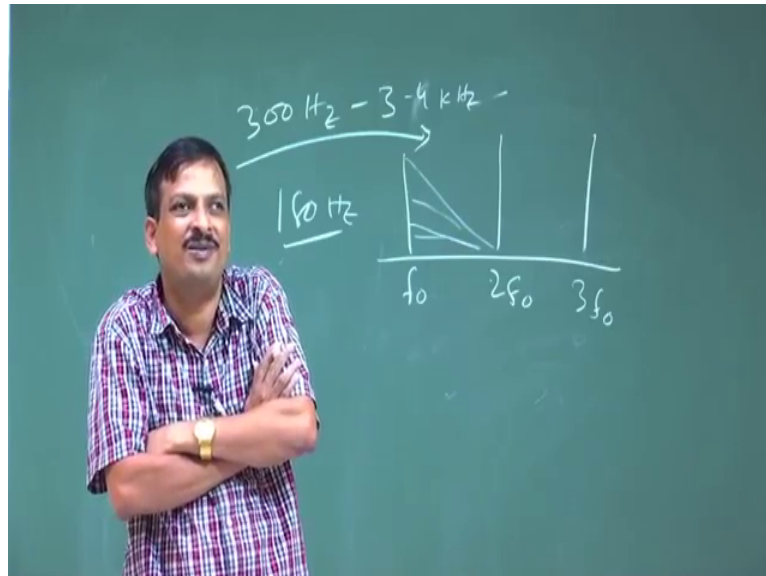
Issues in Fundamental Frequency Detection

- ❖ Purely voiced or unvoiced excitation does not exist either. Usually, excitation is compound (noise at higher frequencies).
- ❖ Speech is called quasi periodic signal
- ❖ Difficult estimation of pitch with low energy.
- ❖ High F_0 can be affected by the low formant F_1 (females, children).
- ❖ During transmission over land line (300–3400 Hz) the basic harmonic of pitch is not presented but its folds (higher harmonics).

If I say I cannot get a purely voice signal, there may be a some noise is included quasi periodic in nature in speech difficult to estimate speech with low energy. High f_0 affected by a low form and yes this is very important issues. If you see that if you remember the first formant frequency is in and around 500 hertz. So, it can reduce and it

can be goes to 300 and 60 hertz also, so if it is 300, 60 hertz in the first formant frequency and if f_0 is in around 300 hertz.

(Refer Slide Time: 18:35)



So, f_0 can be codified with the first formant frequency with the issue. Next challenging issue is that due to the band limitation. If you remember in telephone speech we say telephone speech is 300 hertz to 3.5 kilo hertz 4 kilo hertz if it is that. So, below 300 hertz no signal no component. So, suppose I am speaking about the telephone my fundamental frequency is 180 hertz f_0 is not present fundamental frequency is not there, but still receive the fundamental frequency.


Because the property of fundamental frequency says that the spectral information will repeat if you plot it you know this, this is f_0 next will be twice f_0 and all the spectral component will be repeated here next will be thrice f_0 . So, f_0 is not there, but spectral repetition tell me what is the f_0 of that signal. So, sometimes this is also important that we have ever looking for not only f_0 f_0 may not be there, but $2f_0$ is there $3f_0$ in there.

So, the repetition of the spectral information from there I can find out that I perceive they have f_0 frequency. So, if this kind of band limited signal is come then it is very difficult to find out f_0 also. So, there is a lot of technique for f_0 extraction, from the beginning there is some time domain methods.

(Refer Slide Time: 20:05)

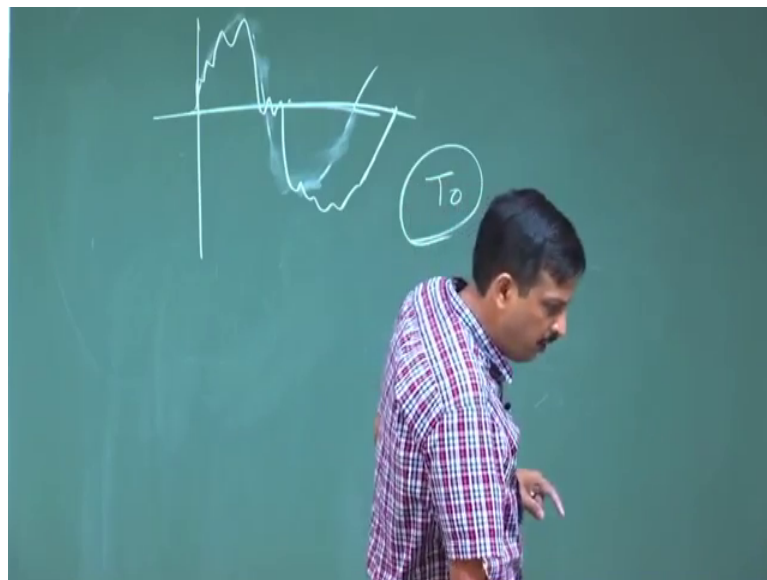
Time Domain Methods

- ❖ Zero-Crossing Detection
- ❖ Autocorrelation Function
- ❖ Average Magnitude Difference Function



Some frequency domain methods. So, let us start with some time domain methods. Now if you remember during time domain speech processing class, I said 0 crossing, 0 crossing.

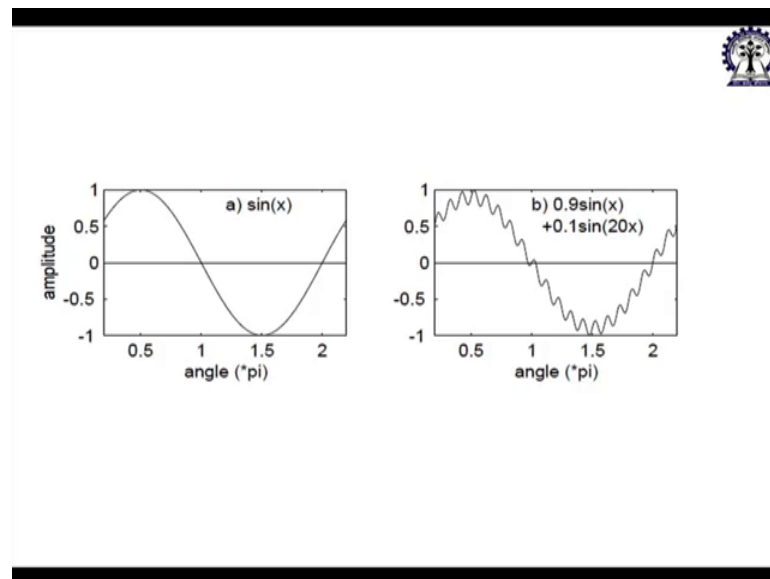
(Refer Slide Time: 20:24)



So, we say that if it is a pure sine wave signal then within a single period signal will cross the 0 line 2 times. So, from that philosophy I can find out f_0 of that signal f_0 by f_s by f_0 , we have calculated that time number of 0 crossing per second.


So, we can find out the number of 0 crossing per second, and then I find out that t_0 value, for which the signal cross the 0 line 2 times. That is called 0 crossing base f_0 detection, but what is the problem? In this case it is now suppose there is a high frequency speech is not a monotone signal. So, it is there is a high frequency component, now you see the number of 0 crossing signal this is the final fundamental frequency, but number of 0 crossing it change within one period, I can show you one picture is there.

(Refer Slide Time: 21:33)



So, in that case I fail to detect the f_0 . So, 0 crossing rate base f_0 detection is not Very good solution or good methodology for finding out the f_0 .

(Refer Slide Time: 21:47)




Zero-Crossing Detection based F0 detection

- Based on a direct application of the definition of periodicity
- Counting the number of time that the signal crosses a reference level
- If the spectral power of the waveform is concentrated around F_0 , then it will cross the zero line twice per cycle
- Mostly Inexpensive in computation
- Weakness against noise
- Presents weakness when used to analyze signals with energy in high frequencies

But if you see the detecting of number of 0 crossing is the easiest algorithm. So, time complexity of computation for 0 is very less, but it is not at all noise (Refer Time: 22:08)
So, I can say method of 0 crossing detection where the f_0 detection methods is very not that useful, or you can say that is not reliable.

(Refer Slide Time: 22:25)



Autocorrelation Technique

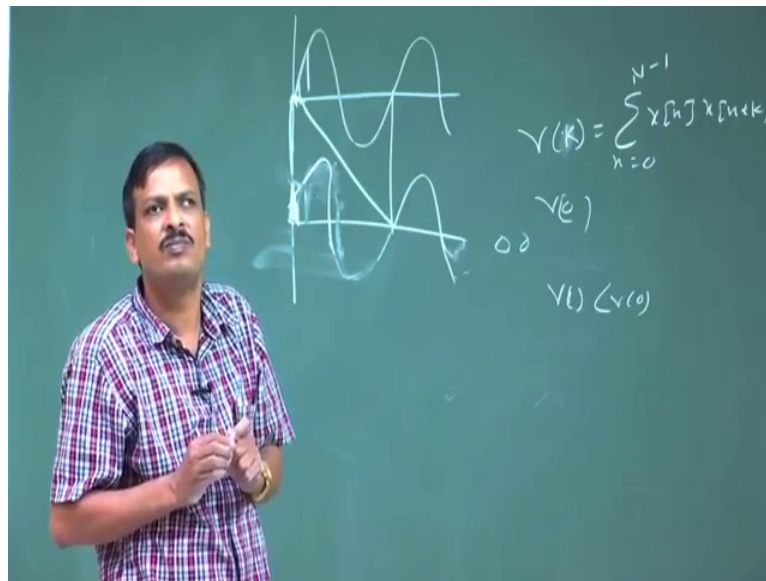
- Autocorrelation is a cross-correlation of a signal with itself.

$$\phi(\tau) = \frac{1}{N} \sum_{n=0}^{N-1} x(n)x(n + \tau)$$

- ❑ The maximum of similarity occurs for time shifting of zero.
- ❑ An other maximum should occur in theory when the time-shifting of the signal corresponds to the fundamental period.

Next technique is called autocorrelation technique, which is mostly used technique for f_0 detection. I am not reading the slides, you can read the slides. Now what is if you remember the, what is autocorrelation?

(Refer Slide Time: 22:41)

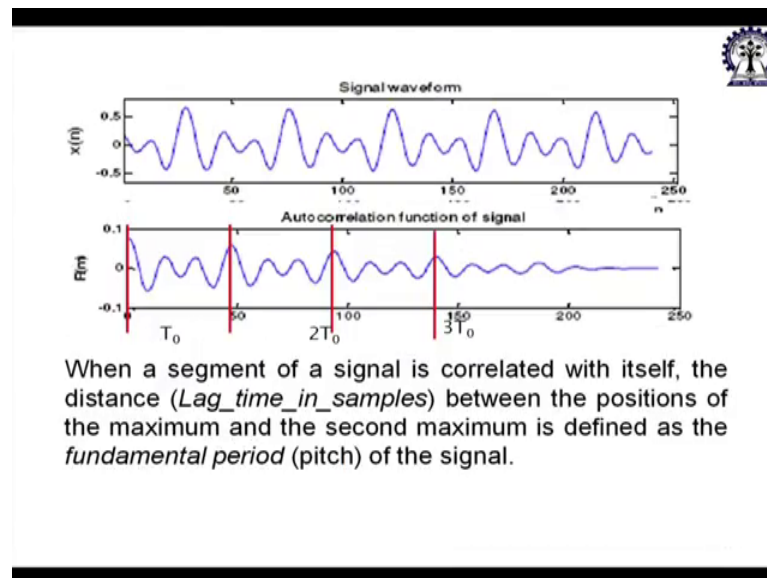


So, suppose this is my signal. Think about the pure sine wave. So, autocorrelation means correlation the signal with itself. So, let us I again draw the same signal. Now autocorrelation let us autocorrelation told me the equation is that r tau is equal to summation of n equal to let us I writing in digital domain, or you can say n equal to 0 to n minus 1 x of n x of n plus k let us write instead of tau I write k (Refer Time: 23:33).

Now for r 0; that means, if this is my signal this is my signal first coefficient is nothing but the product of the all samples are like this is this sample this sample this sample multiplied by this sample. So, I get the maximum value. Next time I shifted this signal by one sample in here. So, instead of this signal my signal will you start from here. So, I take this sample with 0. And n then 0 will be padded with the length of the signal. So, what you get I will get an autocorrelation coefficient r 1 which is less than r 0. At 2 less than r 0, if you say the signal will become in the in page where it is reaches here.

So, if this much of amount of shift is there then the 0th sample will be multiplied with 0th sample. Again I get maximum r value. Yes, I get the f 0 by 2 also maximum r value because it is pure sine wave. If it is not pure sine wave, then I will get maxima when the sample is matched exactly with it is period. So, if I see that plot, you see this plot.

(Refer Slide Time: 25:12)



So, if you see this is my signal. And this is my value r value $r_1, r_2, r_3, r_4, r_5, r_6$. Now if you see at around 50 sample, let us this 45 sample this again come maxima. If it is 45 sample next to also come at 45 plus 45 maxima and particle plus mix. So, this is twice f_0 this is thrice f_0 . So, I can say from here to here this 45 sample is the t_0 .

(Refer Slide Time: 25:52)

$$T_0 = \frac{1}{f_0}$$

$$f_0 = \frac{16000}{80} = 200 \text{ Hz}$$

$$V(k) = \sum_{n=0}^{N-1} x[n] x[n+k]$$

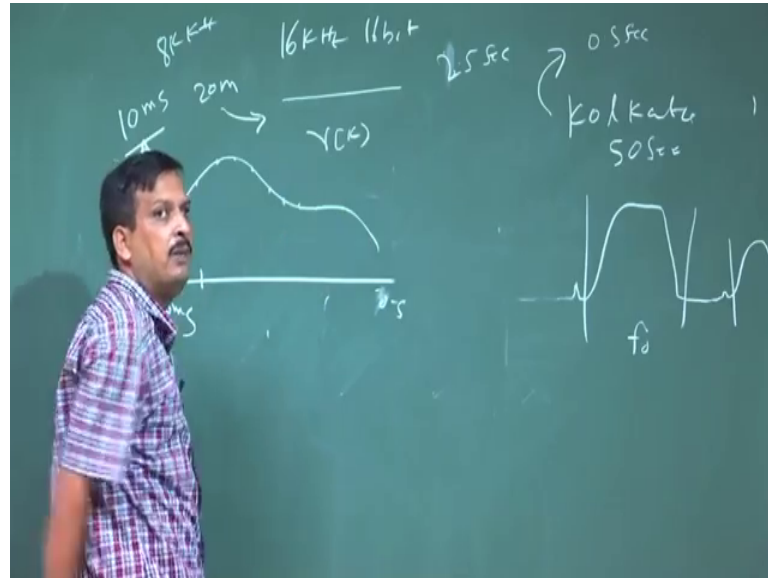
$$V(0) > V(1) > V(2) > \dots$$

So, what is f_0 ? Is nothing but $1/t_0$. Suppose this signal has a sampling frequency is 16 kilo hertz, then I can say f_0 is nothing but a 16 divided by t_0 let us 45 sample. So, let

us 50 sample 300 hertz. So, around 300 hertz, I can calculate the f_0 . Now how you process it? This procedure is So, how he process it?

If you remember, that suppose I have a signal, I have recorded my name at 16 kilo hertz 16 bit.

(Refer Slide Time: 26:52)



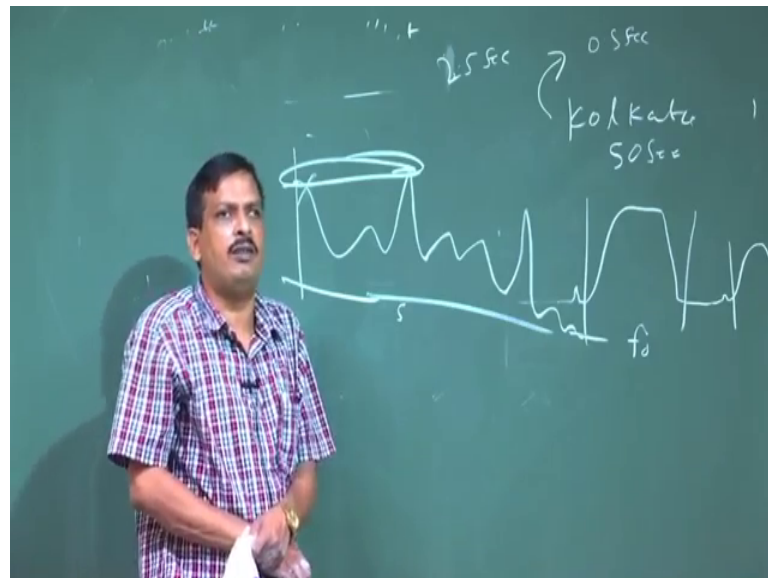
So, my name may be 3 point 5 second signal my name is long. So, it may be less not 3.5 second let us 2 0.5 second signal. So, how many samples are there? 16 k plus 16 k plus 8 k. So, around 40k samples of there 40,000 samples are there. So, I said when I say my name the fundamental frequency is not same throughout the signal, it is moving. And also the fundamental frequency only present if the signal is voice.

So, suppose I say Kolkata, let us say Kolkata. If you remember that manner based manner of the signal co is an unvoiced signal, then there will be a worst then consonant to (Refer Time: 27:54) transition then steady state (Refer Time: 27:55) o then o 2 law is a voiced consonant then again it will be stop consonant then again bust. Then again vocalic transition, then again it is a stop consonant, then again it will a bust. And there will be vocalic transition and again end. So, every vocalic part up to here this is the vocalic part I can get f_0 value. From here to here I can get f_0 value and from here to know vocalic part I get the f_0 value. Now how do they know the vocalic? This is very important talk to the important part. So, instead of doing finding out that in signal what I will do let us Kolkata is recorded in 0.5 second.

So; that means, there is a 8 k samples are there. Now I frame the signal 100 frame per second. So, in Kolkata I will get 50 frames signal. So, how each of the frame, which is duration of 10 millisecond maybe window by 20 millisecond will give me a average fundamental frequency value? So, if I extracted that to take that 20 millisecond window and shifted that window by 10 millisecond then I get 100 frame per second, so for every 100 millisecond. So, if I draw the Kolkata in f_0 frame. So, if it is 3.5 second signal. So, there will be a 3.5 second. So, this is 0.5 second. So, there are a 100 frame. So, 0.5 second will be divided by 150 frame. So, every frame is 10 milliseconds. For every 10 millisecond I can calculate a average f_0 value, but next 10 milliseconds.

So, I can get the f_0 1 2 which each points are 10 millisecond apart. So, 20 milliseconds signal I will calculate r_k value and find out the maximum first maximum then calculate the lag 1 by lag is the if I can get the f_0 value. So, this is the autocorrelation base f_0 extraction very easy, but it is time consuming. Now if you remain if you see this signal in autocorrelation f_0 extraction if you see this signal, signal is decaying. R_k value or you can say the r_k value Plot is decaying.


(Refer Slide Time: 30:37)



So, it is nothing but like this way this way, this way, this way, this way, then this is going down and almost 0. So, I cannot get the first peak and second peak are not equal, it is down. So, detection of peak is very difficult sometimes may be erroneous also. So, how do I improve this result? I want this decaying should not be there.

So, if I want the decaying should not be there, then that is called modified or equal to the normalized autocorrelation, normalized autocorrelation. The formula is this one instead of calculate.

(Refer Slide Time: 31:37)



Normalized cross correlation function (NCCF) method

The normalized cross correlation function (NCCF) is very similar to the autocorrelation function, but is better follows the rapid changes in pitch and the amplitude of speech signal.

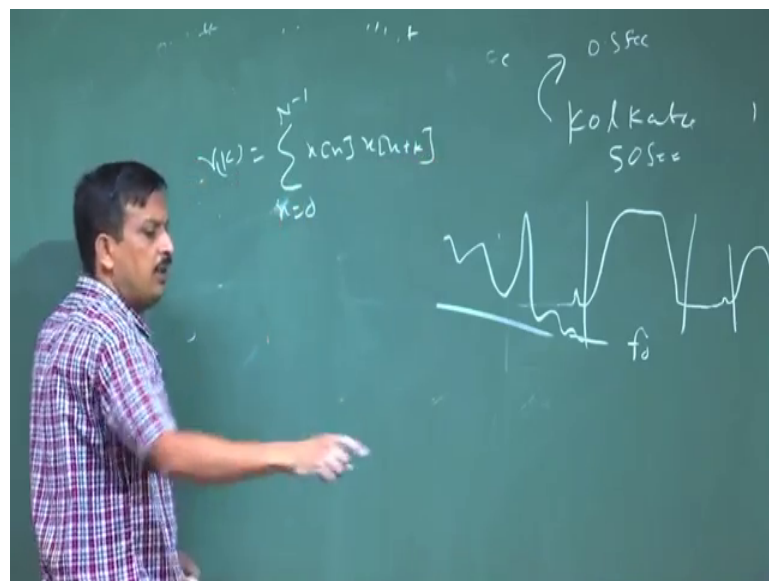
The NCCF based PDA overcomes most of the shortcomings of the autocorrelation based algorithms at a slight increase in computational complexity.

The NCCF function for speech segment $x(n)$, $0 \leq n \leq N-1$ is defined

$$NCCF(m) = \frac{\sum_{n=0}^{N-m-1} x(n) \cdot x(n+m)}{\sqrt{\sum_{n=0}^{N-m-1} x^2(n) \cdot \sum_{n=0}^{N-m-1} x^2(n+m)}}, \quad 0 \leq m < M_0$$

Autocorrelation calculating the autocorrelation using that r_k is equal to n_k equal to n equal to 0 to n minus 1 $x_n \times x_{n+k}$.

(Refer Slide Time: 31:42)



I am calculating different formula, this using this formula.

So, this is called normalized cross correlation function instead of autocorrelation. If I calculate r_k value this NCCF normalized cross correlation function, then if I plot it if you see first one is the signal second one is the normal autocorrelation third one is in ccm if you see the amplitude is not decaying. So now, I can detect the peak easily, and calculate the f_0 easily. So, this is called normalized cross correlation based f_0 vector restriction. So, next class other methods also I will discuss ok.

Thank you.