Digital Voice and Picture Communication

Prof. S. Sengupta

Department of Electronics and Communication Engineering

Indian Institute of Technology, Kharagpur

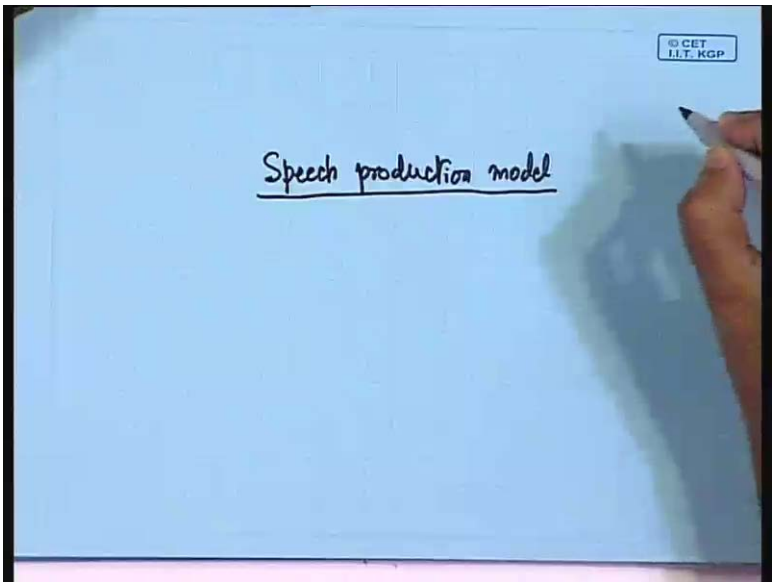
Lecture - 02

Speech Production Model

(Refer Slide Time: 00:00:54 min)

So today we are going to talk about speech production model.

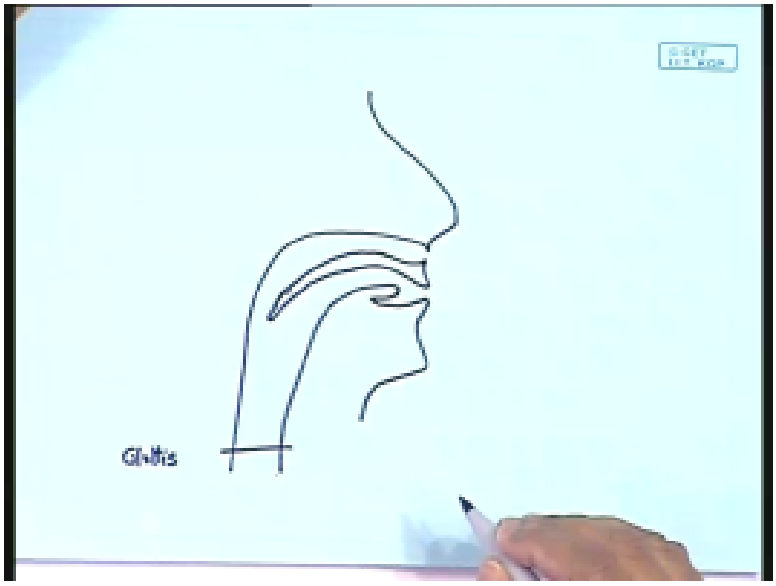
(Refer Slide Time: 00:01:07 min)



It is necessary for us to have a clear understanding about how the human speech is produced and if we are able to understand that properly then it will obviously facilitate us in developing a kind of model for the speech production which in turn they can use when we have to produce speech signals in a synthetic way; in the sense that by having a speech production model we are going to extract the parameters of the speech production model from a speech signal and then we are going to use those parameters in order to have a proper synthesis of speech at the receiver or the decoder end. So basically that is the motivation with which we are learning about the speech production mechanism.

So at first I will begin with a physiological model of the speech production and then we will see that how we actually model that into the proper acoustic behavior about that and what are the properties like the speech and then the excitations that we are going to describe then we will see that how it can be modeled like a source filter model combination for the speech production. So we first begin with the physiological aspect of speech production.

(Refer Slide Time: 00:03:52 min)

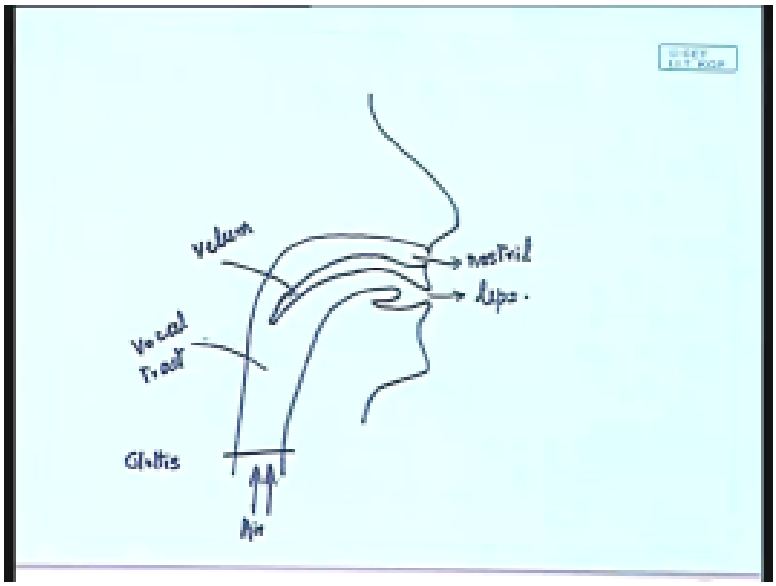


So this is the kind of physiology that we are talking about. Now here we have the glottis (Refer Slide Time: 4:01). Glottis is basically the opening of the human vocal tract. So this whole thing what we have drawn like this, you can see that there is a cavity. In fact there are two openings through this cavity: one is the nostril and the other is lips and this this entire thing forms the vocal tract and here we have what is called as the velum which is basically an articulatory organ which we can move up or down in order to constrict the flow of the air; means either we can produce the sound through our lips, we can radiate it outside through the lips or otherwise we can make an opening through the nostril passage so that we can produce some nasal sounds. Because there are different kind of sounds which we need to produce depending upon what phoneme we are uttering. So essentially what happens is that the whole thing is controlled by the flow of the air. So this is the passage through which the air flows (Refer Slide Time: 5:34) and what I have shown is glottis is effectively the opening that we are having at the vocal chords.

In fact that opening is periodically controlled. So the opening, it opens and closes periodically not exactly periodically rather in a quasi-periodic manner means over a very short time interval we can consider that there is a kind of periodicity that is present in the pulses that is produced through the glottis. In fact what is happening is that there is opening and closing so that

effectively it is a pulse like excitation that is ultimately given to our vocal tract. So this is the vocal tract and this entire thing forms the vocal tract.

(Refer Slide Time: 6:36)

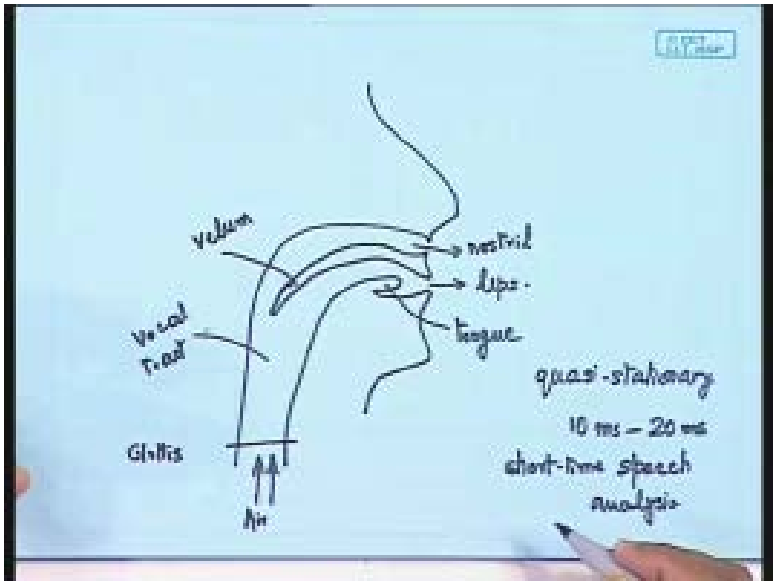


Now, vocal tract is a very interesting kind of an organ in the sense that it is a filter like thing. Means whatever pulses are fed through this glottis the vocal tract effectively does a kind of spectral shaping to those pulses which are produced and that kind of a spectral shaping what we do that varies from time to time means in a short time interval we may be finding that the spectral response of the vocal tract is something and then after sometime the spectral behavior would again change because we are going to utter something else. So there is only a short time analysis that we normally do with the speech signals.

In fact just to tell you at this point that speech signals are not exactly stationary. So instead of stationary we call it as quasi-stationary process. So quasi-stationary in the sense that the statistical parameters of the speech that remains reasonably constant only within a very short time interval and that time interval is generally of the order of 10 milliseconds to 20 milliseconds it is of that order so that is why we always go in for short time speech analysis. So, short time speech analysis is usually done. Then what happens is that with the air flow then we can use our

velum, we can articulate that in order to produce some constrictions; like, if we produce, if we try to lower down the velum then effectively we are blocking the air flow from the vocal tract region up to the lips and instead we are making and opening more towards the nostril passage so that nasal sounds are emitted. And whenever we are actually uttering something like the voiced sound what is called as the voiced sound there we will be putting the velum we will be articulating it upwards so that the nostril passage is blocked for a bit moment and we can produce the sound out of the lips.

(Refer Slide Time: 9:30)



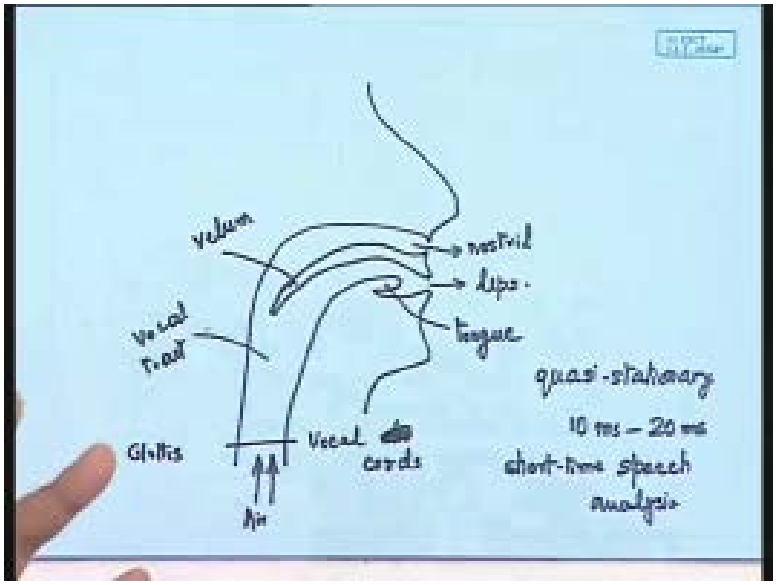
Here we also have the tongue (Refer Slide Time: 9:30). So this is where our tongue is and tongue is also another articulating organ in the sense that, by positioning our tongue we are able to create some constrictions in the lips. Actually without our knowing we do all these things whenever we are producing speech. There are various sounds that we are producing. We are producing the sounds of the vowels, we are producing the sounds of the consonants and there are certain sounds which are kind of unvoiced.

Now you will see that whenever we are uttering any vowel, those are associated with the air flow and the air is thrust through this glottal passage. And in fact here for the production of the vowels

and some of the consonants of course what we do is that the glottal air flow is controlled by the vocal cords. In fact the vocal cords will be located here itself. We write it as c o r d s vocal cords and the vocal cords actually vibrate and that is how it creates the opening and closing of the glottal passage; and what happens during the utterance of vowels is that it is this pulses that produces some kind of resonance in the vocal tract.

In fact using our articulatory organs like the velum, the tongue, we are able to alter the shape of the vocal tract dynamically so that it is not a kind of filter that has got a constant behavior with respect to time; it keeps changing. So it is a source of source of excitation that we can imagine at the glottal passage followed by a time varying filter.

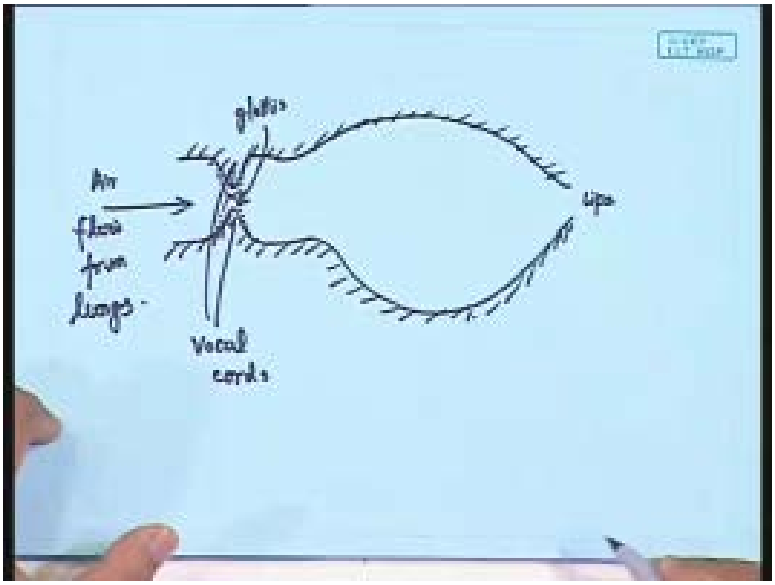
(Refer Slide Time: 11:36)



So in fact we can modulate in a better way in the sense that instead of considering the physiology let us see that if we can use something like a model like this that we now show how the passage will be created. So this is our glottal opening (Refer Slide Time: 12:10) so this is lips. This is not a kind of the diagram that one would like to draw considering the physiological behavior. But you see why I drew this diagram like this will be clear to you. In fact let me mark the different portions of it.

Here this is the glottis that we are talking of (Refer Slide Time: 12:36) and in fact this is the passage, this is the glottal passage but here we just use some dotted lines and the dotted lines indicate that this set is actually the vocal cord and the vocal cord as I told you they vibrate. Now, as a result of the vibration in the vocal cord this is how the glottal passage may get closed and then again the glottal passage can be opened. So it is a quasi-periodic opening and closing of the glottal passage that I was talking of and here we have the air flow from the lungs.

(Refer Slide Time: 13:23)

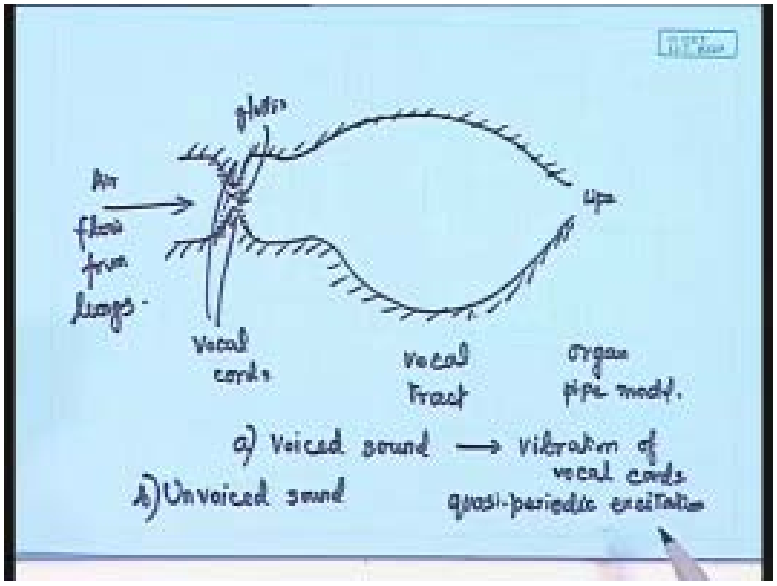


So, now we are going into so here the air that enters through the glottal passage into this vocal tract is effectively modeled as a cavity. This is a vocal tract cavity that we can imagine and if our lips are open then you can imagine it like an open organ model as if to say that you are thrusting some vibration through some open pipe. So, we can in fact consider the model of an organ pipe. So very crudely speaking, we can use the organ pipe model for that and then we will be seeing that how the organ pipe model would shape the pulses or shape the excitations which are produced from here.

Now there are two kinds of sounds that are produced. One is what is called as the voiced sound and the voiced sound are actually produced through the vibration of the vocal cord. So, voiced

sounds are produced through vibration of vocal cords. And in fact this is This can be modeled like the quasi-periodic excitations that goes on at the glottal passage so this is modeled as quasi-periodic excitation. Its input rather is modeled as a quasi-periodic excitation but then it will be shaped within our vocal tract accordingly. that aspect we will see shortly. And the other kind of sound that we produce is what is called as unvoiced sound.

(Refer Slide Time: 15:33)



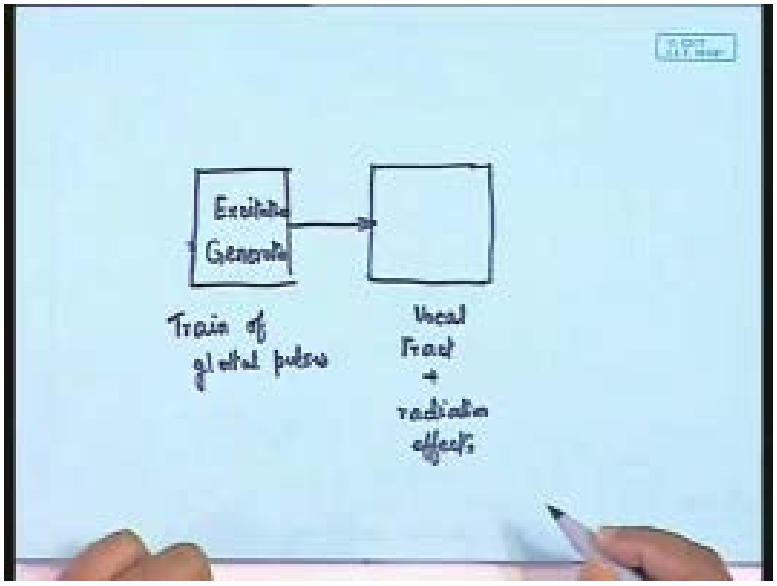
Now, unvoiced sounds are actually not produced as a result of the vibration of vocal cords. We would imagine as if that for movements when the vocal cord makes an opening into the glottal passage there you have a constant air flow that can take place. So we assume that there is a steady air flow but then we use our articulatory organs as I was mentioning over here (Refer Slide Time: 16:08) that we use our articulatory organ in order to produce constrictions in the passage and the constrictions actually produce some turbulence in the air flow so that gives rise to a kind of noise like behavior is there and we produced such unvoiced sounds for some of the consonants; like say for example when we say (yes) what are we doing we are actually using our.....I mean just think over, we are using our tongue in order to constrict the steady flow of air which we are introducing and that steady flow of air is constricted by controlling our tongue so we produce the sound (yes) and whenever we are saying (yes) we are not using our vocal cord

for that but whenever we are producing vowels {a i o} like that just think that the sound is actually coming through our glottal passage. So we produce both this things: voiced sounds which are associated with vowels and some of the consonants and then also unvoiced sound.

But as I was mentioning that unvoiced sounds since they are not produced by the vibration of vocal cords. We can we should not model it like the quasi-periodic excitations. But instead we should put a kind of a noise generator model. So the source of excitation is something which interestingly it changes. At times we have the quasi-periodic excitations as the source of excitation in the vocal tract and at times we will be having a noise like model for the production of sound. So that is what I was also telling you that the entire speech processing model is so dynamic; it changes with time and it is only a quasi-stationary behavior that helps us in producing a model because there is a quasi-stationary behavior that is why we can have short time speech analysis and at least try to extract the speech parameters.

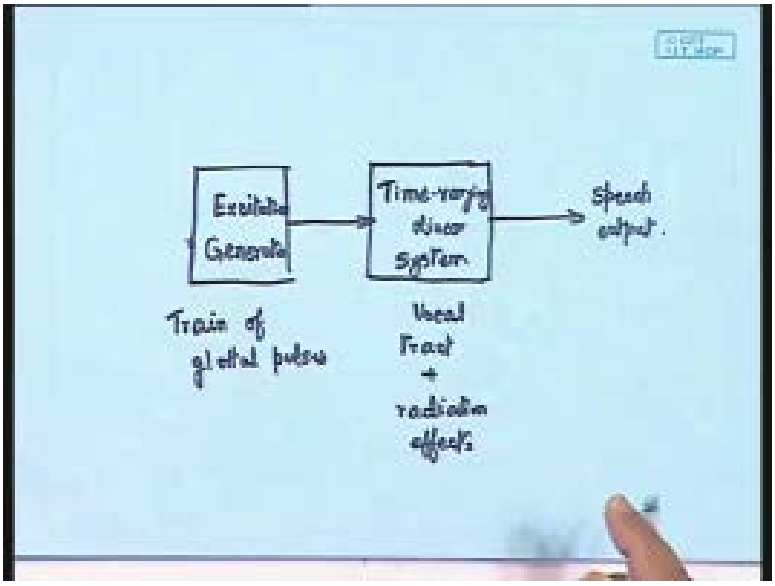
now what are the So a very crude model that you can think for the speech production could be described by something like this that you can have that there is an excitation generator so there is a signal source which we call as excitation generator and this excitation generator as I was telling you that it can be of two types: one is that it produces a train of glottal pulses. Why train, is because, as I was telling you, the glottal passage opens and closes quasi-periodically so you will be producing a kind of train of glottal pulses and at times your excitation generator is just a noise generator. then this excitation generator output or rather the signal that you are producing, the pulses that you are producing that goes to the vocal tract so here you have the vocal tract and also you consider some kind of a some kind of radiation effect because after all what is it that we are doing during our speech production is we are ultimately producing some vibration in the air medium like the air flow which is ultimately coming out of our what you say which is which is coming out of our lips ultimately it is radiated and it goes up to the listener's ears and at listener's ears it produces vibration and those vibrations actually produces signals to the brain and that is how we are able to make out that what sound the person, the speaker is making and then from that sound our brain is ultimately able to infer; we do a kind of linguistic perception and then we are able to make out what was the word or what was the sentence that the person was speaking.

(Refer Slide Time: 21:11)



Now, vocal tract, along with that we also have to consider the effects of radiation. So we call that this is vocal tract plus radiation effects. And as I was telling you that this acts like a filter but this is a kind of a linear system and we can model it as a time varying linear system. So this is essentially time varying why because its parameters keep varying with time as I was telling you as I was emphasizing again and again so this is a time varying linear system and this ultimately produces the speech out it. So this is the simple model.

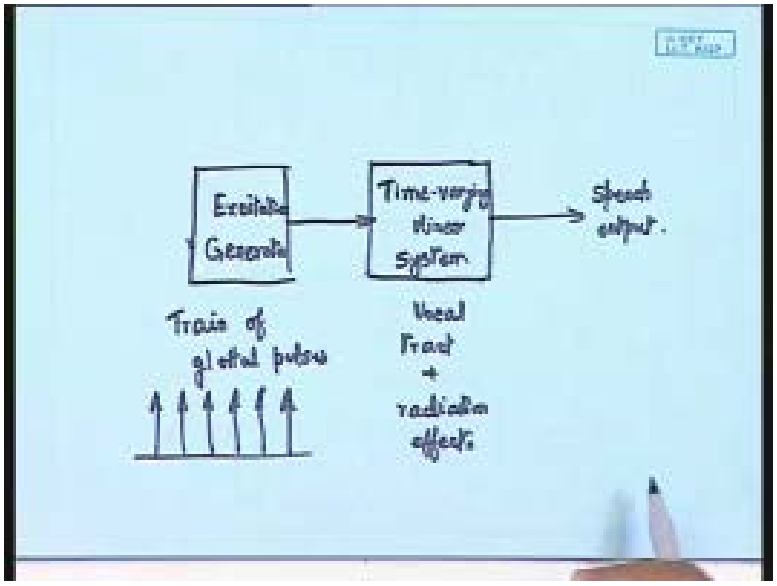
(Refer Slide Time: 21:59)



Now how to now having known the excitation generation part; if it is a train of glottal pulses we may represent it as something like this say these are the pulses (Refer Slide Time: 22:15) which are produced periodically or better to call as I have been always telling you that it is quasi-periodicity because it is not that you can assume that it is perfectly periodic, it is just a quasi-periodicity that can be assumed but this train is the cause of the excitation in the time varying system.

Now, whenever you have such kinds of pulses you can imagine that that really produces a very flat spectrum. But how the actual speech outputs spectrum..... the actual speech output spectrum would not look like a uniform spectrum; it will have some kind of shapes in the sense that we will be having, that there are certain frequency components which are highly dominating and there are certain frequency components which are not so dominating. Let us first talk about the voiced sound.

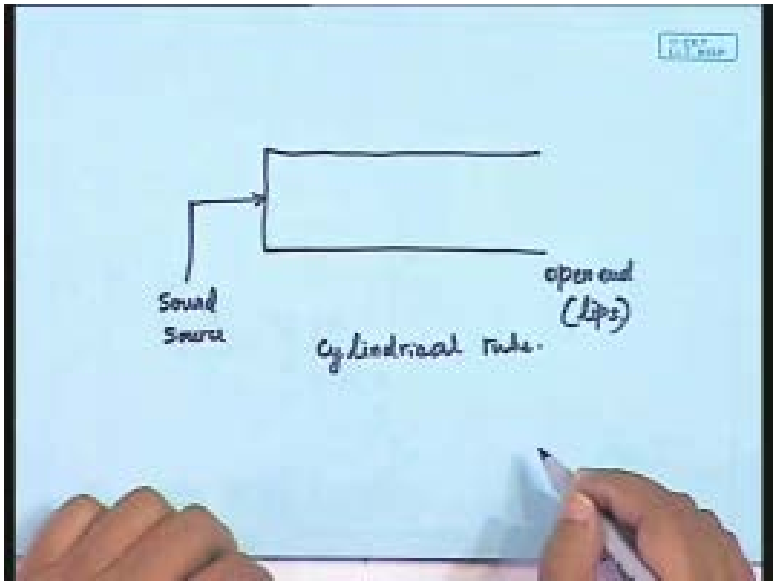
(Refer Slide Time: 23:23)



So, voiced sound; talking about the voiced sound is easy in the sense that there the source of excitation is always a quasi-periodic pulse..... I mean a train of quasi-periodic pulses like this and then that effectively excites the organ pipe model that I was telling you.

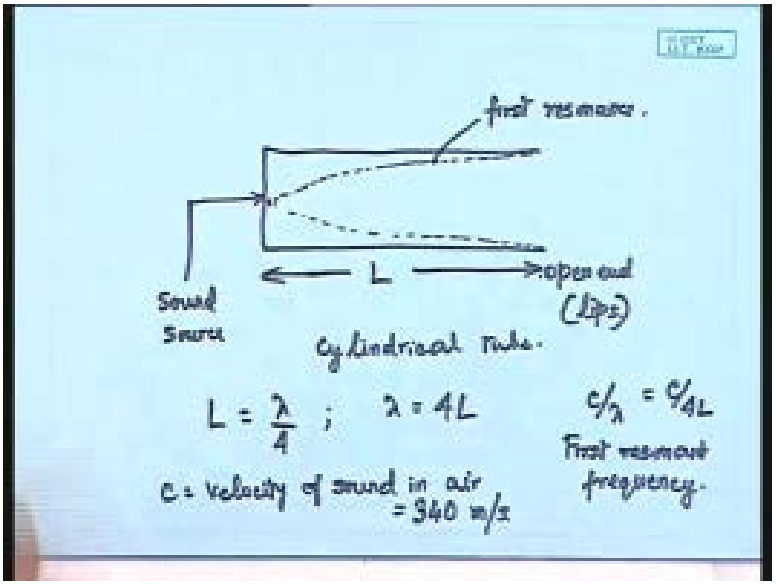
So let us now talk about the organ pipe model just to model our vocal tract system and then we will see that how we can really consider the variation in the spectrum shape that is going to take place in the vocal tract. So we now model the vocal tract in a very simple way. So we consider a cylindrical tube; an organ pipe like cylindrical tube. So this is a cylindrical tube and here there is a sound source. So we have the sound source or rather the excitation produced on this wall and then actually we can say that this end is open so you can assume that we are having an open end that is at the lip end we have the opening

(Refer Slide Time: 24:59)



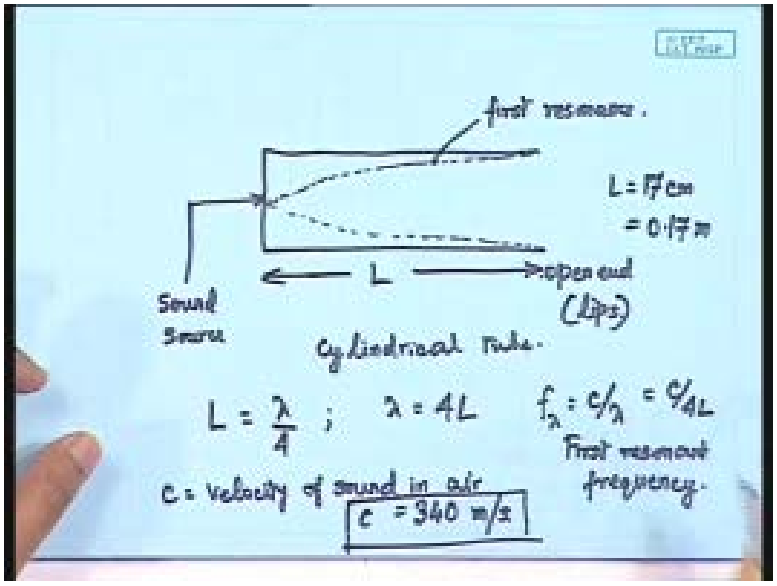
This being a cylindrical tube you can imagine it like an organ pipe and in fact its first resonance would occur like this. So you can see that here a node will be created and here an anti-node will be created. So as a result of that you can see that we will be having the first resonance which would look like this. So this is the first resonance position (Refer Slide Time: 25:39) and corresponding to this we can have..... actually if we assume that the organ pipe is of length L then you can imagine that corresponding to the first resonance we will be having L equal to lambda by 4. So, rather to say we should write it in the other way that is to say the wavelength of the first resonance would correspond to lambda equal to 4L and if we express this in terms of the resonant frequency it will be C by lambda or it will be C by 4L so this will be the first resonant frequency.

(Refer Slide Time: 26:35)



Now some typical values we can put forward. So here C is the velocity of sound in air. So C is the velocity of sound and let us take a value equal to 340 meters per second for this which is quite a typical kind of value that we can take for velocity of sound in air. So it is 340 meters per second and what is L; L is a effectively the distance between the larynx and the lips; and a typical value of L let us say that we take the value of L to be 17 centimeters. So C is equal to 340 meters per second and L is equal to 17 centimeters or rather 0.17 meters.

(Refer Slide Time: 27:36)

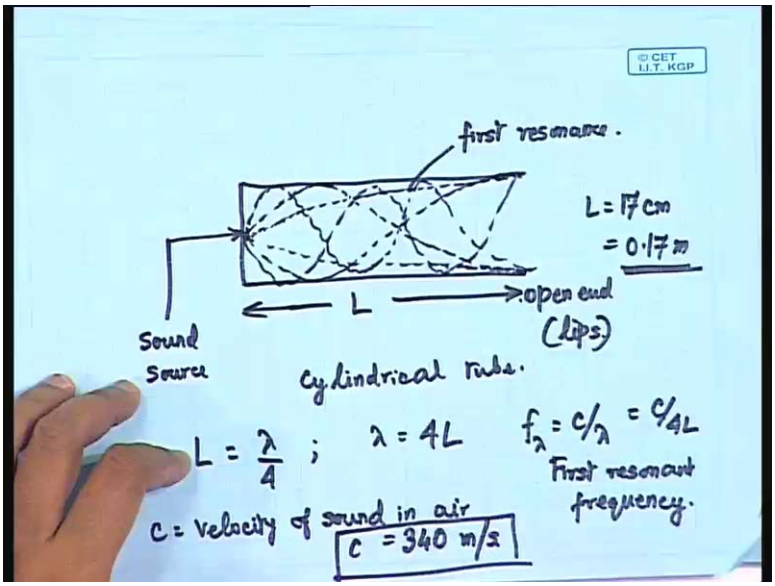


So if you substitute the value over here So if you call this to be the first resonant frequency f lambda and you substitute this value of 340 over here; so if we a write f lambda as 340 meters per second and this we write as 4 into 0.17 expressing this also in meters then we will be getting the frequency in Hertz and in fact this works out to how much? this is actually of course this This result you will be getting in Hertz and in fact this corresponds to 500 Hz just see it is 340 divided by 0.68 which means it is 500 Hz. So 500 Hz becomes the first resonant frequency which we can say provided L is 0.17 meters. Well, that is a big question that how can we take L to be 0.17 meters always.

In fact as I was telling you that the vocal tract being highly articulatory kind of an organ we will be having some changes in its shape always. So this 17 centimeters first of all this can vary from one person to the other that is to say the nominal distance that you can take between the larynx and the lips that itself can vary from person to person and not only that even we can see that even for a person it changes with time. So, given this value we have 500 Hz but for some people it may be 400 Hz, for some people it may be 600 Hz and sometimes that 400 or 500 may come down all the way up to 300, 350 sometimes it may go about to 600 or so.

So, actually speaking, we can have that kind of resonant frequency, there has to be some range, we should not just say that the resonant frequency is 500 Hz rather than that we should specify a wide range and let us say between 200 Hz to 2000 Hz for that range considering the variation from person to person and also considering the variations that take place with respect to time. So we have to consider a range like that.

(Refer Slide Time: 00:30:35 min)



Now this being the first resonance that is produced there will also be other harmonics which will be associated with it. So you can see that the next harmonic that you can generate is like this (Refer Slide Time: 30:54) so this will in fact generate all the odd harmonics because this happens to be an open end. In such a situation we would have here..... that the next harmonic would have so this is where it is having the open end. So if you take this curve; if you take this distribution then your resonant frequency that you are getting corresponding to this that becomes three times of this. So the next resonant frequency will be three times f lambda. So, if it has been computed as 500 Hz then we will be having this as 1500 Hz or 1.5 kHz.

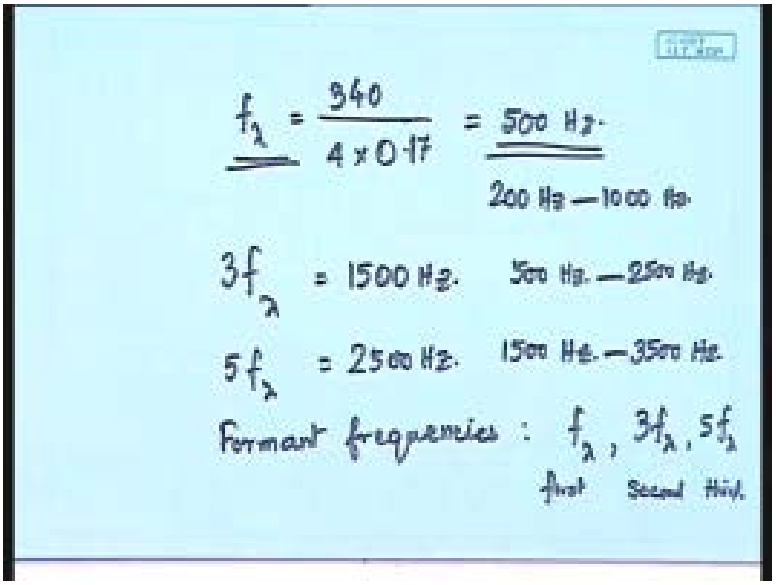
Again for this also, again for this resonant frequency also we should specify a range. Let us say that here this may vary from 500 Hz all the way up to 2500 Hz. The next resonant will occur at

the fifth harmonic. so there you can imagine that we will be having a distribution like this (Refer Slide Time: 32:22) and corresponding to that we will be having the resonant frequency occurring at 5 of f lambda which according to this calculation assuming these values it is equal to 2500 Hz but actually from practical point of view we should have again a range of values and that we can say as 1500 Hz up to 3500 Hz. These are the kinds of variations that are possible.

Now these frequencies that we are getting what we are calling as resonant frequency this is actually the characteristic of the vocal tract and we call these resonant frequencies as Formant frequencies so we call this as the Formant frequencies and the first formant frequency is what we have corresponding to f lambda so f lambda is the first formant frequency so this is first formant frequency. The second formant frequency is 3 of f lambda which is the second formant frequency; the third formant frequency is five times f lambda and so on.

Actually more than that more than three formant frequencies we generally do not consider because you can see that if you consider the fourth formant frequency according to this calculation itself it goes to 3500 Hz and considering its wide range of variation it will be well above 4 kHz or 5 kHz so generally the fourth formant frequency is never audible.

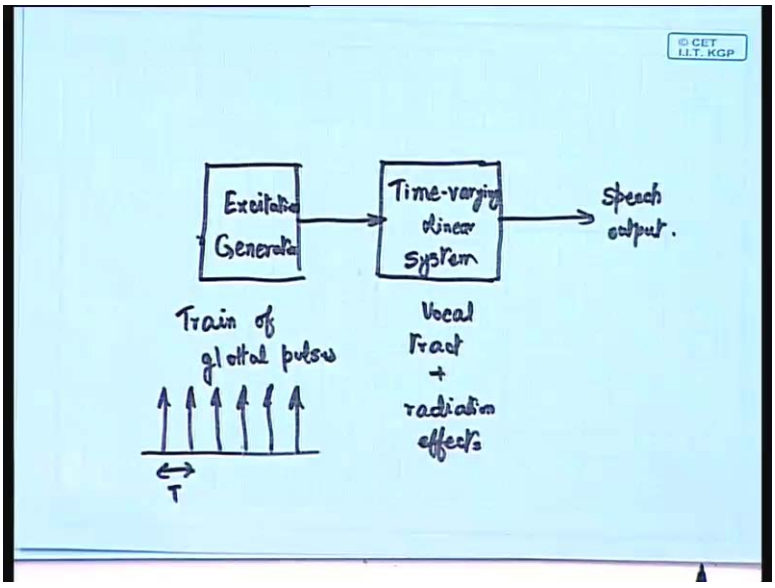
(Refer Slide Time: 34:19)



In fact it varies from person to person. For some people it is only the first and the second formant frequencies which are very dominant. For some people it may be that the first formant frequency which is dominant and the other formant frequencies are relatively lesser in amplitude. That usually happens with the persons having low pitched voice and again with persons having high pitched voice the situation may be somewhat different in the sense that for high pitched voice we may not be able to hear the first formant properly and in that case the second and the third formant may be more dominant. So there is definitely a kind of variation that takes place from person to person depending upon whether it is a low pitched voice or the high pitched voice. Now, yes, talking about this aspect, before proceeding further we should have some idea about the pitch that we are mentioning.

Now, pitch is basically, in terms of that train of glottal pulses you can say that it is the repetition frequency of this glottal pulses.

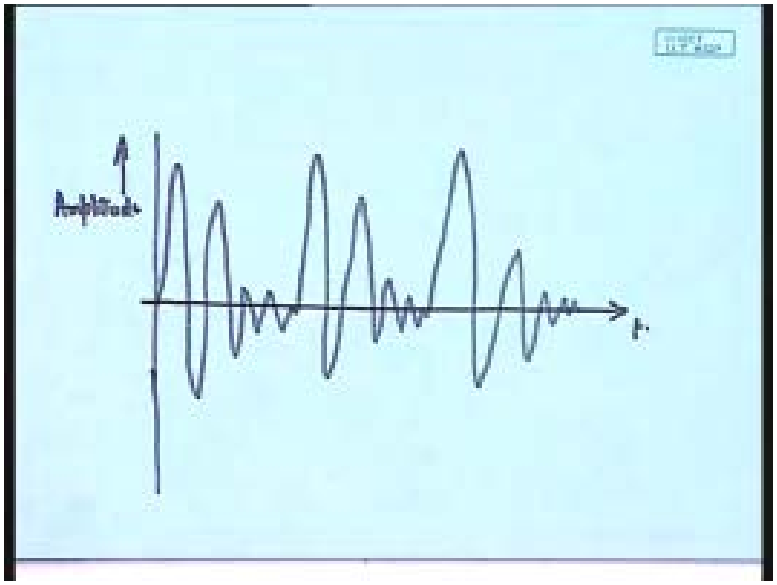
(Refer Slide Time: 00:35:46 min)



So if you have a time period corresponding to this as T if the train of pulses is repeating after a time T then one upon T what you are getting is a kind of frequency that you can say as the pitch frequency. But how the pitch frequency becomes effective for our speech production? Here we

are having a source of excitation but the excitation is actually taking place in the vocal tract (Refer Slide Time: 36:23) which in a very crude sense. Of course it is not a very good model but in a very crude way we can say that..... I mean, assuming it to be an open-ended model of an organ pipe where at this end you are producing the sound source (Refer Slide Time: 36:45) and here you are having the lips and considering like that. So we are essentially having the presence of different formant frequencies.

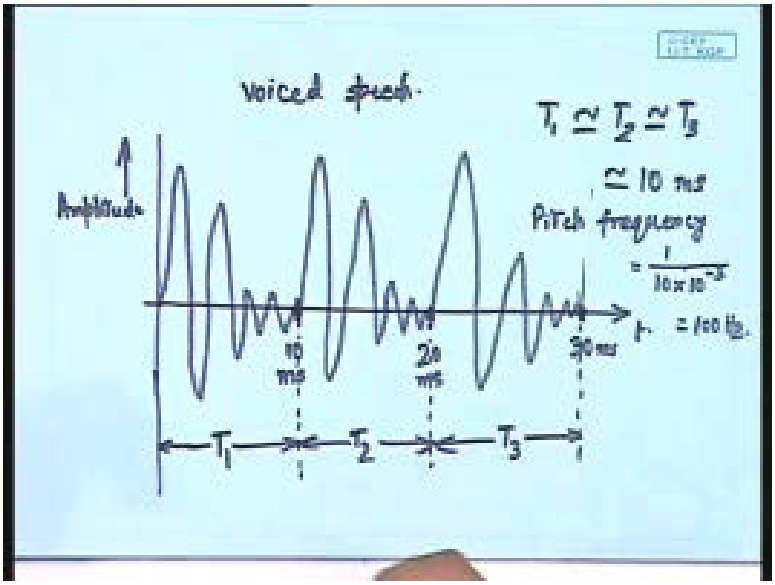
(Refer Slide Time: 37:48)



Now the actual speech signal that is produced would look as something like this. You can say that if we plot a typical speech signal versus time it would be somewhat like this. So this axis is time and then this axis is the amplitude so it will be like this (Refer Slide Time: 37:40). Can you see now a kind of periodicity? Well, if you take the waveform from this instant to this instant and call this time as T; again from this instant..... say we call this time as T 1 and then let us say that from this to here we call this time as T 2, from here to here because here there will be the beginning of the next time in variation like here we are seeing that here it varies somewhat like this so again this starts from here so this we should take as the period T 3 (Refer Slide Time: 38:38). So we normally observe that these periods T 1 T 2 T 3 they remain reasonably constant over a very short interval. Some typical values could be like this that say in the time scale if we

say that this is the 10 millisecond point and supposing this is 20 millisecond approximately, say this is 30 milliseconds so effectively we have shown the speech signal variation; some voiced speech signal of course so this is a voiced speech a typical voiced speech that we are showing for 30 milliseconds and within 30 milliseconds we are assuming a kind of quasi-periodicity because then reasonably T 1 is equal to T 2 approximately and that is also approximately equal to T 3 so that some kind of a quasi-periodicity can be assumed and you see that in this particular example the example waveform that I have drawn, there this time period is approximately equal to 10 milliseconds.

(Refer Slide Time: 40:27)



So now you see that the time with which this is repeating is this 10 milliseconds so if we have to consider the pitch frequency the pitch frequency would be the reciprocal of this 10 milliseconds. So reciprocal of the 10 milliseconds so you can say that it is 1 upon 10 into 10 to the power of minus 3 which means to say that we have a 100 Hz pitch frequency. Typically the pitch frequencies are of that order; it may be 80 Hz, it may be 100Hz, 120 Hz, 150 Hz somewhere in this range we will be having the pitch frequency. So, within the quasi-periodic time interval like in this case we have taken 30 milliseconds we are observing that there are three complete quasi-periodic cycles of speech sound that we have over here.

Now, here this being the pitch frequency being 100 Hz we can say that 100 Hz as if the source of excitation. So in our excitation generator this T if we take this to be approximately 10 milliseconds (Refer Slide Time: 41:20) then it produces so it is corresponding to a 100 Hz repetition rate. But now our question is that if 100 Hz excitation is applied then why are we having this sort of an effect. Well, the answer lies that again this sort of effects (Refer Slide Time: 41;44) are appearing because of the resonances that is produced in our vocal tract and the resonances are actually triggered by the excitation that is produced and there are presence of not one but up to three different resonant frequencies are there and that also may be changing with time so we cannot exactly predict that what kind of waveform will be generated but assuming that it is a it is something like an unrestricted air flow in the vocal track.

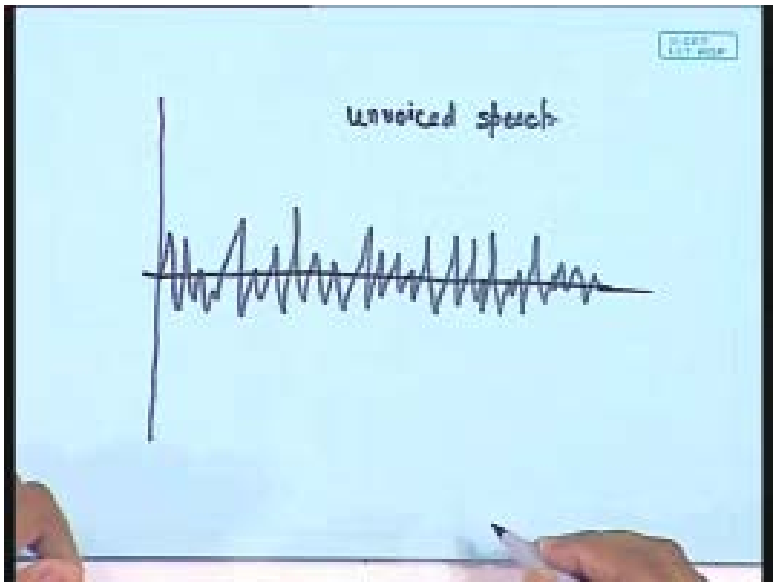
Mind you that the model that we have considered (Refer Slide Time: 42:27) is valid only when there is unrestricted air flow. But not always that is possible because while producing different speech sounds we invariably produce some constrictions and the constrictions actually make this model somewhat different.

Now these things which are appearing like..... you may be observing that within this you have got five different peaks. So five different peaks if they are occurring almost at regular intervals then what does it mean. That if you are considering from this distance to this distance in the time scale you can say that as if that this is one fifth of the time period T 1. So one fifth of the time period T 1 means corresponding to something like 2 milliseconds and 2 milliseconds means that it is corresponding to 500 Hz and why 500 Hz so again the answer lies over here (Refer Slide Time: 43:34) that 500 Hz is the fundamental formant frequency or the first formant that we have considered which again as I told you that it varies; it may vary over a very large range. but this is to say

So, effectively the time waveform that you are getting is a resultant of the excitations for the different formant frequencies may be that up to three formant frequencies are audible so depending upon that the waveform that we have is a resultant of that and we observe something like this. So this is what we have for the voiced speech whereas for the unvoiced speech the waveform would look something like a..... noise like observations will be there. We will

having no regular patterns so it may be the signal could be something like a random pulse like, I mean a random voice like thing; so this is for the unvoiced speech.

(Refer Slide Time: 44:50)

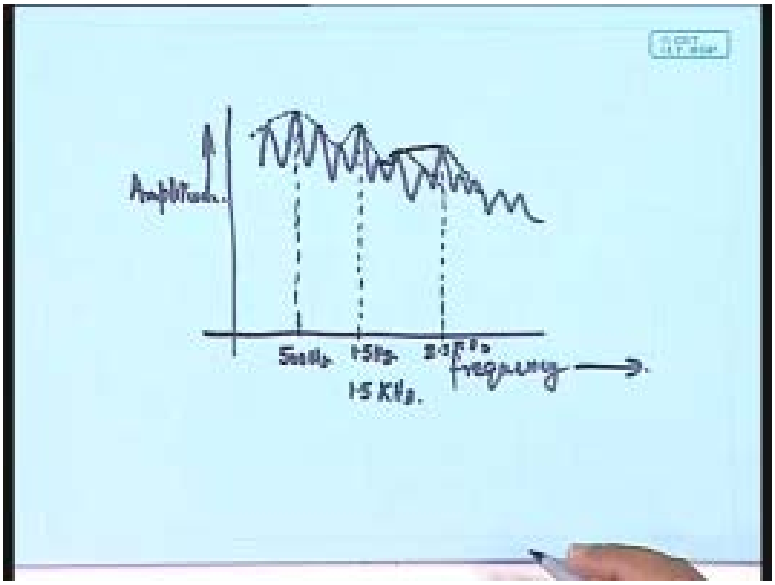


I mentioned little while back that unvoiced speech is the one that is not produced from the vocal cords; that is not produced through the glottis; it is just as result of constriction that we create in our vocal tract. This is actually the kind of model that we will be following for the voiced and the unvoiced speech. And if we look at the frequency spectrum of let's say the voiced and the unvoiced speech it would look something like this that we would be having say amplitude verses frequency if we plot say this axis we have the frequency and here we plot the amplitude (Refer Slide Time: 45:55). now you will be seeing that if the pitch is let us say 100 Hz then you will be observing that corresponding to the multiples of 100 Hz..... and how the multiples are generated again is because it is the reach in harmonics so we will be having something like this.

So you can see that what happens is that here we have an envelope; here we have a different envelope. and here we have a So you can see that corresponding to this there is a peak in this envelope, there is again another peak over here (Refer Slide Time: 46:57) so you may observe that this peak is occurring at 500 Hz. of course I did not draw according to the scale so please

excuse me for any confusion that it might create because of not drawing it up to the scale. But I suppose that you can get the idea behind it that supposing this is 1.5 kHz and supposing the next formant frequency as I may draw it like this and here the next peak is occurring at 2.5 kHz something like this. So there you can see that this 500 Hz would correspond to the fast formant frequency 1.5 kHz, this should be 1.5 kHz should correspond to the second formant frequencies, 2.5 kHz should be the third formant frequency and so on.

(Refer Slide Time: 47:53)

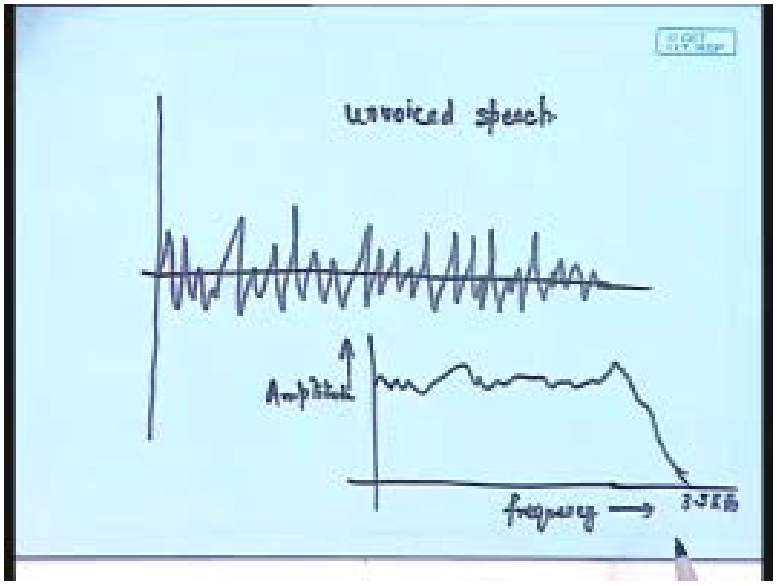


Now what is happening is that in between these formant frequencies you also have the variations that are arising out of the pitch frequencies. So the amplitude spectrum would be something like this. Whereas if you are considering the excitation model directly your signal I mean your excitation signal is having a very flat spectrum. but after the pulse shaping to the vocal tract because as I have been telling you again and again that vocal tract acts as a filter and very precisely it is a time varying filter that it is doing. So this would give rise to a spectrum like this and in fact this spectrum also what we are observing is definitely a short time observation because the spectrum again will change from one quasi-stationary period to other.

If we compute the spectrum for the first 30 milliseconds for the next 30 milliseconds our spectrum would be entirely different. It may have the formant frequencies somewhat shifted, it may have even the pitch shifted, even pitch is also changing with respect to time and that is why we produce some intonations in our speech otherwise had it been a flat pitch, if we talk at the constant pitch always you will not be finding any intonation with that speech signal at all and it will not be very pleasing to hear. You would not like that kind of speech. But the speech where there is a variation that takes place in the pitch you definitely find those kind of speech signals to be more easy to follow, more easy to enjoy and such things.

For the case of unvoiced speech when you have a waveform like this (Refer Slide Time: 00:50:11) then you can well imagine that if you are taking the frequency spectrum of that you do not expect much of a variation. You would definitely not expect the formant frequencies because as I was telling you that in the first place that unvoiced speech signals are not produced through the through the excitation of the glottal pulses, it is just a result of the constrictions which produce noise like disturbances. So, because it is a noise like thing you will be observing more of a flat spectrum. There will be some variations but a flat spectrum is something like this (Refer Slide Time: 50:59) that if this is at say 3.5 kHz in between it will be having some kind of a flat spectrum if this is the amplitude; no such definite peaks would be observed for unvoiced speech. But for voiced speech definitely it is the formant frequencies which will dominate.

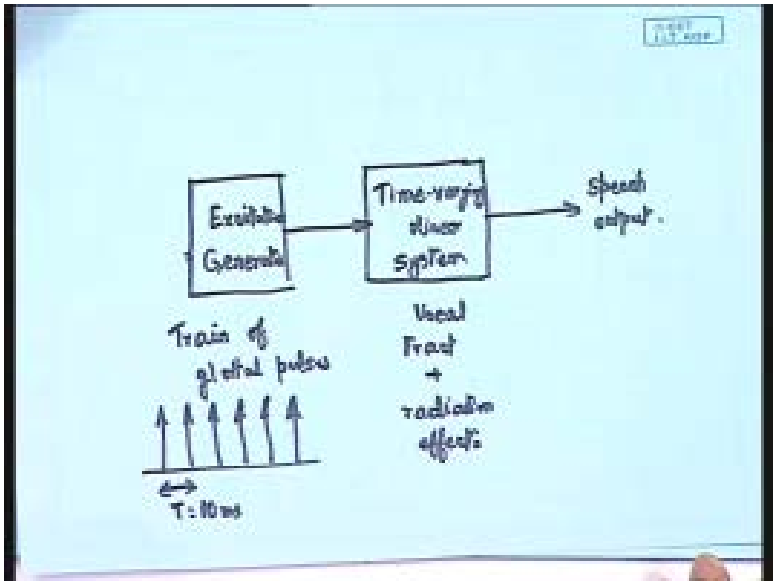
(Refer Slide Time: 51:21)



So, what is actually the story of the day? I mean, why at all we are doing this sort of analysis and why do we study these formant frequencies well?

The reason is that if now given the kind of model that we have here: excitation generator, then the time varying linear system and that produces speech output then in a similar way if we can artificially produce this excitation and have a filter to model this vocal tract then we should be able to produce speech in a synthetic way. But definitely synthetic speech will never be anything pleasing to hear. Perhaps some of you must have got experience in listening to some synthetic speech outputs where there is a great deal of artificiality which exists. In fact they are the they are very simplistic models are assumed because of which you will find a lack of naturalness in those type of speech signals.

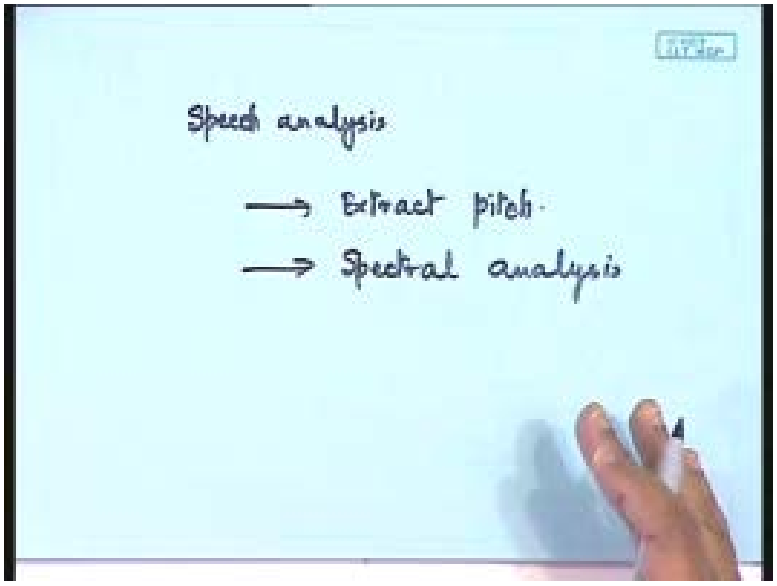
(Refer Slide Time: 52:51)



So if you want to make it more like natural what we need to do is that we should not just artificially create this excitation because artificially how do we create excitation? A train of glottal pulses of constant frequency; well, that is not the case because as I was telling you that the pitch varies that is why we should not be considering the train of glottal pulses just like that. In fact it should be derived out of our speech signal. So we should do a kind of speech analysis.

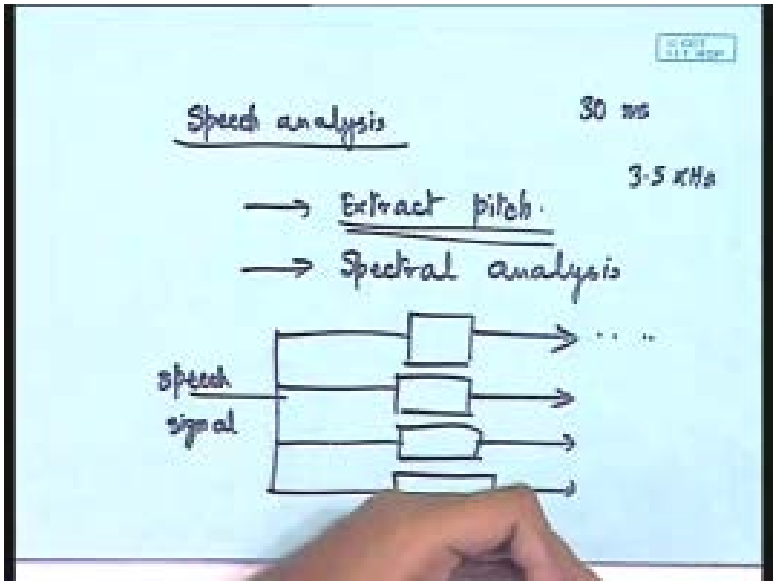
We definitely have to do a speech analysis and the purposes of the speech analysis should be that we should first of all extract the pitch. If we are able to extract the pitch then we will be able to determine that when the pitch varies that also will be known to us because extraction of pitch would be from our speech signal itself and then also we should do a kind of spectral analysis.

(Refer Slide Time: 54:16)



Now, in spectral analysis what we do? We analyze the speech and we divide it into a set of filters. So if this is your incoming speech signal (Refer Slide Time: 54:32) then you will be feeding this signal to a bank of filters and each filter..... it is a kind of a band pass filter that you can assume and this individual band pass filter output..... I mean, we can compute the energy for each of these band pass filter outputs for a given quasi-periodic for a short time interval. If we take some time interval like 25 milliseconds or 30 milliseconds let us say, then for 30 milliseconds we should be able to extract the energy that corresponds to the different bands.

(Refer Slide Time: 57:15)



So, at least individual band's energy if we have, individual band's sound energy if we have and if we also extract the pitch information reliably then we can have an encoding after our speech analysis results. How? First of all we will be encoding the pitch information that what is the pitch frequency we have observed and we will also encode the energy of the different filter bands. We need not have to distribute that into a large number of frequency band because after all the frequency is only up to 3.5 kHz 3.4 3.5 kHz so that will suffice. So generally may be that 10 or maximum up to 20 filter bands would suffice and for individual filter bands we have its energy.

So what we will be doing is that using this speech and using these different spectral components that we get out of the filter outputs we should be able to create the speech. And in this speech analysis process what happens is that we will be able to extract the we will able to.... since we are extracting the pitch information from our speech itself we will be able to model this speech appropriately and it will produce more naturalness.

And not only that, there is also one more point which you should keep in mind that by having a speech analysis and only giving this speech information and the individual band energies..... and also of course..... you have to give another flag information

whether it is a voiced speech or unvoiced speech; if it is a unvoiced speech then one has to switch on a noise generator in order to do the speech synthesis. So this sort of information representation also will be more compact.

So in the next class we will be seeing a kind of vocoder design that follows this kind of speech analysis followed by a speech synthesis using this model. So thank you for now.