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

Lecture - 08
Handson On Acoustic Phonetics

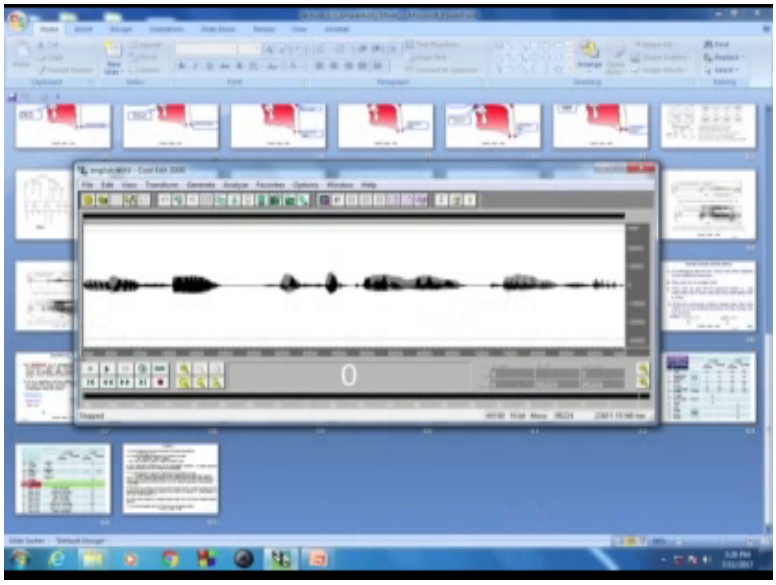
So, last class I have taken about that manner and place of articulation; I have described what is manner and what is place of articulation. Based on the manner and based on the place of articulation; the IPA symbols we have discussed and we assign an IPA symbols. So, and we said that all the vowels are classified based on the tongue position and tongue height; so, f 1, f 3 and f 3 their relationship I have discussed.

For today instead of taking the lectures; today I give you some demonstration to see the voice sound and identify the manner and place of articulation we cannot identify by viewing that sample, but yes we can identify the manner of articulation and from there we can see what kind of consonant it is, what kind of vowels that is, how the formants are moving or kind of things; so, those will explain.

So, if you see there is a number of open source software's available cool edit, kart then I have explained wave circle. So, I will explain in one slides that lot of open source that speech processing software's are available. Now, one of them are the cool edit; if you see the cool edit software that is also downloadable from the cool edit pro is downloadable from the net. And then if you see that when you open a voice; I have recorded that this let us I give you an example.

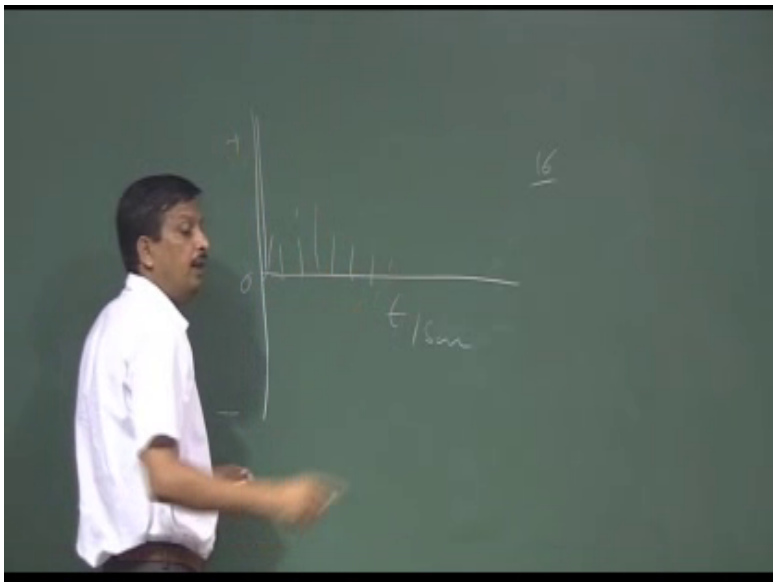
Let us go through that; this is a example of a voice which is recorded for a English speech. Let us listen to the voice; so, this is the voice which is recorded by English speaker English speech; some sentence are there.

(Refer Slide Time: 02:10)



Now, if you see that this X axis which is written the sample; number of sample and you discuss in signal crossing that once you recorded the digital speech that this X axis is sample number of sample or it can be time or sample then Y axis is the amplitude.

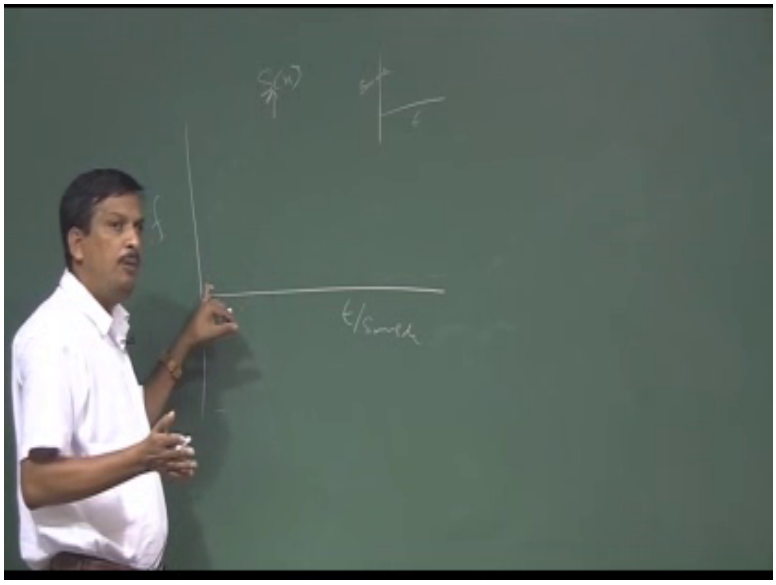
(Refer Slide Time: 02:21)



If you see this amplitude; now if you see this portion, no speech that amplitude of the samples are very low almost close to 0 line; this is 0, 0 amplitude line. So, 0 line if you see this is recorded with 44 kilowatt 16 bit mono. So, each sample is you can say that it is quantized using the 16 bit. So, some value of the 16 bit is there; based on that value, it is Y axis is that value. So this is 0 axis; 0 axis, this is minus, this is plus and samples are all samples are plot and this is look like a sample speech signal.

If you see this portion is a silence portion and here there is a voice portion; there is noise portion, noisy kind of signal; I can magnify it. So, if you see that this is noise kind of things; if you listen it also, some noise signals are there. Now, if I see that this is a time domain representation of the speech signal; now there is a another kind of representation which is called spectral view; look like a you can see; this look like a mostly X ray plate kind of things. So, here what is there it is called 3 dimensional plot of the speech signal; what it is done, if you are not signal crossing background; then broadly you can think about like this that; suppose I want to know that frequency analysis of the signal.

(Refer Slide Time: 04:05)



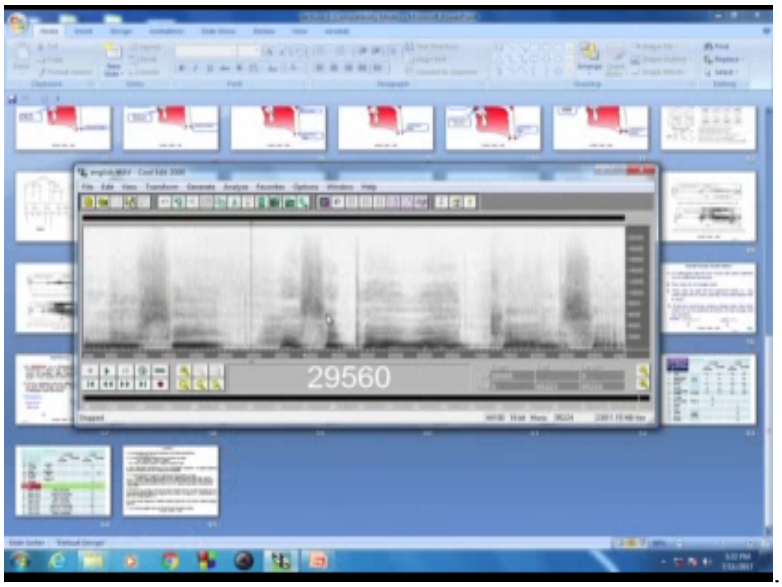
So, I have a sign domain speech signal S N; which I plot time versus sample value or amplitude. Now, I want to plot what are the frequency contained in this signal; so, I say the my X axis is the; let us X axis is the frequency and I want to know the each frequency content amplitude. And then the power of every single frequency; so, this is a 2 dimensional plot; frequency versus power.

This is also time versus sample amplitude; now the problem is that if I plot this way, then I do not know the time information; I want the time information should be there. So, what I do, I do a spectrographic plot; how do I do it? I discussed in the signal crossing class. So, in X axis let us this is time or sample and Y axis is the frequency.

And then the amplitude or power of the particular frequency is represented by a color; it may be a color or it may be a black and white. So, black means the power is high; white

means there is no power. In this case, black means power is high white means there is no power. So, I am not going details about the how the spectrogram is made that I will discuss in the signal crossing class.

(Refer Slide Time: 05:39)



So, if you see in this picture; some portion are very dark those contain the high frequency power of those portion are very high. So, if I say this Y axis is the frequency axis; if you see there is a frequency axis. So, this is 2 kilohertz, 4 kilohertz, 6 kilohertz, 8 kilo hertz; so, sampling frequency is crossing 4 kilohertz.

So, up to 20 kilo hertz; it is there 22 kilohertz if S by 2. Now, if you see the black portion are contain the high power. So, if I say within 2 kilo hertz; most of the sections are black. So, that contains the high power; now if it is noise kind of sound, if you listen it, it is a noise. Noise kind of sound; the noise kind of sound then you see the power distribution of the all frequency are present. So, there may be some dark portion are there, some light are there, but more or less there is all powers are there. So, I can say that the random noise all frequency has some power; so, this is random noise.

So, if I say who generate the random what kind of manner of articulation; generate those kind of random noise. If I say fricative or if I say aspiration, so if it is fricative that were a fixture noise will be there. So, I can easily say this is nothing, but fricative; now if I see here there is no power at all in the frequency. If you see, there is no power at all; so, if you see the time domain, this is 0 sample value.

So; that means, this portion is silence; so, this portion is silence, this portion is fricative. If I say I can zoom out; this is fricative this portion is voiced. So, it may contain voice consonant, it may contain vocalic sound, vowel it may contain diphthong, it may contain (Refer Time: 07:48) I do not know, but it contains some vocalic sounds; so, vocalic sound look like this.

Again if I say; so in this spectrogram asks you just spot out that fricative sound, easily I can say this was fricative sound, this was in fricative sound. I can say find out the vocalic region, I can say here to here is the one vocalic region. Here to here is the one vocalic region, here to here is another vocalic region.

Now, if you see interestingly here; if you go to the time domain, if you see there is a silence and there is a burst kind of noise is there. If you see there a burst kind of things, if I go to the spectrogram see no voicing, then burst, then aspiration, then voicing. So, I can say here either a plosive consonant region; I do not know which possibility is.

But I can say; it is a consonant which is plosive kind of nature because there is a burst and there is a; if I say this is nonaspirated because there is no aspiration in here. So, I can say it is a unaspirated plosive is there; so, by seeing this spectrogram and the time domain waveform of the signal, I can find out some manner of articulation of the phoneme.

I will show you; suppose if just open non on some Hindi recording like this; let us open this one. If you listen it this is [FL] all kind of things is there [FL] there is a lot of noise is there that is why upside frequency are present. So, if I forgot the noise; if I consider only this portion; is a silence portion. Let us I do the noise removal; there is this tool I provided the facilitator I can remove the noise.

So, there may be a 50 hertz noise is in here; so, what I will do there is a position for noise reduction; analyze let us task form noise reduction; subtraction noise reduction it is procedure; I am not going details, but it is there. So, I can say the select the silence portion I can say this portion is noise portion. So, I can transform noise reduction, noise reduction then I can say get profile because this portion; this is not sufficient to get the profile.

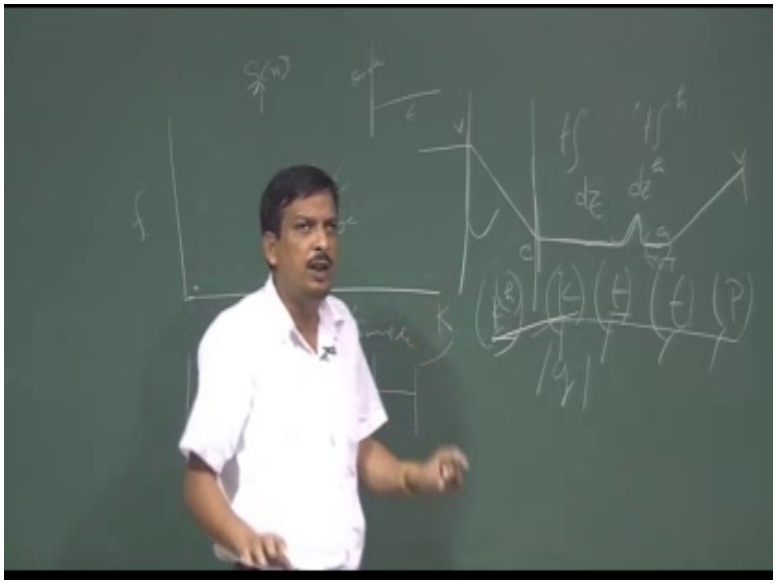
So, I can edit the size; I can reduce and then also it is not sufficient. So, what I will do; I will select much more whatever is possible, then find type try to find out noise reduction this; then also I have to reduce it to relocate, then adjust this one not possible. So, this one is possible get profile I get the noise profile then I close it; then I control a; select all, then task one noise reduction.

So, what it will do; whole signal, this signal is very long that is why it will take time. Whole signal noise will be subtracted as per the sample noise I have given to the system. So, there are lot more details are there, signal processing details are there I am not going that that much of detail; since the signals is noisy that is why I just cleaning the signal fast.

Then I show you how the [FL] is like that; what is the occlusion? What is burst? If you see now; if you see this portion, noise is reduced; almost noise is reduced. So, if I see almost noise is gone; now if I see listen [FL], now it is clear [FL]. Now if you, see [FL] so he pronounced [FL]; so, [FL] followed by a vowel, again [FL] followed by a vowel, again [FL] followed by vowel; so, I can say if I zoom this portion only so, I can say if you see just little bit of un zoom.

If you see this portion; to this portion, this portion is transitory part. So, this is steady state vowel part; if I show you. So, this black color is gradually increasing and here if you see them almost steady. Then, again if you see this is looking just follow the mouse again starts getting. So, I can say this portion is nothing, but a transitory portion when the vocalic; while we want to pronounce after the vowel; I want to pronounce again [FL]. So, this is the transitory portion transitory between vowels and consonants.

(Refer Slide Time: 12:47)



So, if I say I am in here vowel; I want to produce a consonant, then there will be a vowel to consonant transition; that means, that articulator is producing vowel [FL] then it want to produce vowel [FL]. So, what is the effect? The vocal cord has to be open and the articulator has to be closed. So, vocal cord opening is not started, but articular is moving to close to produce the [FL] so once this articulator moving to [FL] the air flow is reduced; if you see the formant structure is changing from steady state vowel to consonant.

Then if you see after here; no noise, no signal if you see there is no signal; almost silence. So, this part I can say occlusion part; then if you see there is a burst here there is a burst, there is a burst. So, burst is happening then there is a VOT; if you see voice sounds set time, voice sound set time is very less; if you see here, if I zoom it and then I show you; if you see the after burst, vowels are not yet started. So, there is a delay between the vocal cord vibration and articulatory opening.

So, what is done; this portion there is no signal then the signal is started, then what is happening; again it want to produce [FL]. So, again articulator has to be go to position of [FL]; So, there is a transitory movement to restart a statistic. So, I can say there is a VOT; then again there will be consonant to vowel transition. So, now if I see; I can recognize; if I give you a signal you can say; this portion is nothing, but a transitory portion. This portion is nothing, but a occlusion, this is the burst then, this is a VOT and then this is again consonant to vowel position; again steady state vowel position. If you look the waveform of the steady state vowel; almost all period are same.

So, I can say; a two peak each period; so I can say this to this is period; this is a period or I can say here to here or may be a period, here maybe a period. So, I can say almost all period are same. Now, if I see the transitory portion; if you see the structure is not yet completed. So, I can say here to here is nothing, but a consonant to vowel transition that part vowels in steady state part; while vowel steady state.

Then again it will start from steady state to consonantation. So, by seeing a wave form or by seeing a spectrogram; if you see the formant structure I am not showing you the formant because this time I am not introducing the formant here to see the formant in spectrogram later on I can show you.

So, but if you see this black portion is started breaking in this portion. So, I can say this portion is nothing, but a consonant to vowel transition. So, after seeing a spectrogram; I can know the structure of a stop contour or proceed [FL] there is a occlusion burst VOT, then again vowel again occlusion burst VOT. So, [FL]; now if I say [FL] with [FL]; if you see [FL] if it is [FL], now if you see zoom it this portion if it is [FL] if you see again consonant to alteration occlusion period; burst, then if you see aspirated VOT. So, long VOT; you see compared to [FL] it is long VOT which is totally aspirated; noise like of sound aspiration is there; noise kind of aspiration is there.

So, if you see there is a long aspiration; then I say this is [FL]. So, this is unvoiced aspiration that is why; I say it is [FL] to differ h. So, this aspiration; so, difference between the [FL] and [FL] if I say what is the difference between the [FL] and [FL] is unvoiced VOT is unvoiced, but in case of [FL] VOT is; sorry VOT is unaspirated and in case of [FL] VOT is aspirated by seeing I can say.

So, I can say I do not know if I not listen the sound I cannot say while it is a [FL] or [FL], but I can seeing this spectrogram; I can say at least I can say this is your some aspiration plosive consonant; seeing the spectrogram I can say this is a unaspirated plosive consonant. I do not know the place of articulation because place of articulation, if I want to know then I have to know how the tongue is moving. Because the difference between the [FL] and [FL]; what is the difference? [FL] is bilabial dental plosive [FL] is dental plosive and [FL] is velar plosive.

So, if I say [FL] and [FL] only difference is that tongue movement from [FL] to [FL] and [FL] to is different from [FL] and [FL]. So, it is one with the velar opening velar closure

anther one is the tip of the tongue touches the teeth to produce the [FL]. So, cavity structure will be different in that time. So, seeing this spectrogram I can at least say what kind of manner of that consonant. So, this is [FL] again I can go little detail then if you can see there is [FL]; let us see the [FL]. So, [FL] is nothing in hindi [FL] Bengali [FL] or [FL] is [FL] is whether it is English, Hindi, Bengali; [FL] is voiced; voiced unaspirated plosive.

So, if you see the voiced; that means, there will be voice in during the occlusion. So, if you see the occlusion period; there is a voicing. So, I can say the there since voicing is there then it is voiced. If you see there is a burst is also there; there is a burst then VOT is less if you see the [FL]; let us I show you the [FL] again if it is [FL] you see the aspiration will be there again VOT occlusion period is voiced burst is there and then the see the aspiration is voiced. So, this pronunciation as voiced aspiration [FL] then if I say I do not know; this is maybe [FL] I think it is [FL] if it is [FL] then if you see occlusion friction then VOT; very little VOT.

Then if you see occlusion, friction, occlusion, friction; now if I show you [FL] let us come to the [FL] only difference is that it will [FL] is if you see affricate. So, if you see [FL] the aspiration this is the aspiration part; this is the friction part and then followed by an aspiration is also there [FL]. Now if you see the [FL] what is the difference? Only the occlusion part should be instead of unvoiced, it is voiced and then there is aspiration understand. So, [FL] then [FL] so [FL] place of articulation I cannot say, but seeing a spectrogram from the manner of articulation, I can say what kind of consonant it is.

If you see the retroflex; I do not know whether that retroflex is [FL] this may be [FL] then that that may be [FL] I think so, if it is I do not know it may be [FL]. So, [FL] see the again there is a burst unvoiced occlusion. So, I seeing these pictures waveform I cannot say whether it is a [FL] whether it is [FL] whether it is a [FL], but I can say yes it is a unvoiced, unaspirated plosive.

So, at least manner I can say. So, if you look like the record your different kind of consonant with vowel and look like what kind of it is look. So, you can see the look, but if you closely watch the spectrogram; then you find in case of retroflex, the formant movement that only a formant structure. Here I cannot show you, but in plot it is possible to draw the formant also. If you see the formant movement will be different, so a

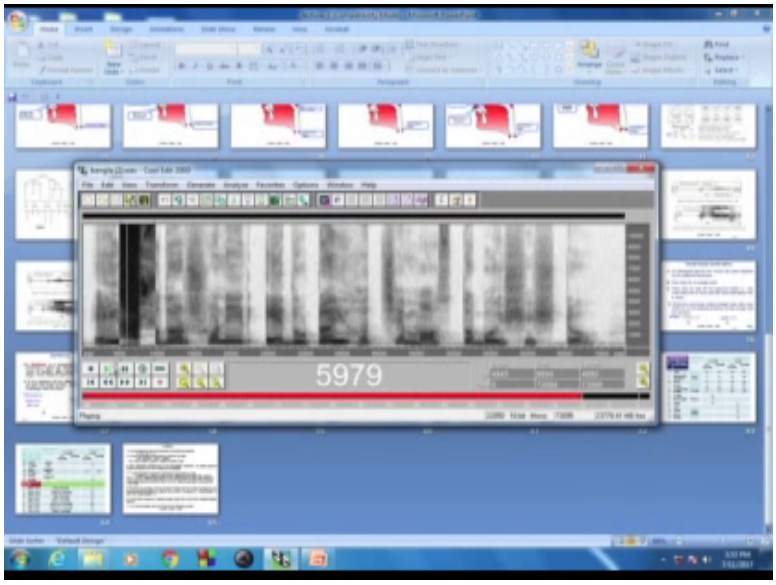
consonant whether it is [FL] whether it is [FL] whether it is [FL] whether it is [FL] is only differentiate while the transitory movement from consonant to vowel or vowel to consonant.

So, if I say the difference between the [FL] and that [FL] only lies between the vowel to consonant transition and consonant to vowel transitory part where since the movement of the tongue is different that is why transition movement of attention will be different; frequency movement in the transitory movement will be different. So, only relyhave to rely on this thing to find out whether it is [FL] or whether it is a [FL] or whether it is a [FL]. So, if I have a lot of confusion maybe happened in signal crossing between this group and [FL] is different by the [FL] is by only aspiration. So, if I able to find out the aspiration; then I can say yes I can differentiate between the [FL] and [FL] because else where there is no signal occlusion is silence.

So, seeing the manner you can learn what kind of signal processing I should use for what kind of application; then you have to know what kind of signal I am getting and what are the phoneme say it how it is behaved. So, I cannot say the result consonant recognition result of English and consonant recognition result of Bengali may be different. Because Bengali has more number of stop consonant compared to the English stop and plosive. Even Hindi has more number of stop consonant compared to that English. So, identifying the voice consonant is easy because voice there is a sound signal are there sound frequencies components are there. But if you see the difference between the [FL] is only the transitory part; this portion is silence there is a burst, there is a VOT and difference between the [FL] and [FL] is I rely only on aspiration.

So, if I want to find out the high accuracy we were not get it within the [FL] and ka. So, that now there is a another aspect is that co articulatory effect. I am not showing you this is the nonsense pronounce in the [FL] the structure is very means very well behaved, but if it is continuous pronunciation; if you see I can show you in Bengali example if I say I show you Bengali example.

(Refer Slide Time: 25:37)



That continuous things; let us say Bangla this one [FL]. So, if you see that the speak is which is continuous. So, it maybe a [FL] so in word if when I write if I write a or [FL]. So, there is two words up there what is you see the continuous speech there is a no word boundary is there only continuous speech.

So, that may be a vowel, that may be a [FL] that maybe a if I say this is what kind of consonant; again it may be a [FL] because occlusion period is voiced there is a burst and there is a transition; I do not know which kind of vowel it is then I have to listen it; I can see it. Then I can say this may be a [FL] because is a friction. So, Bengali has a only pronounce the probab. So, [FL] is there palatal sound is mostly cases we pronounce the palatal [FL] so, it may be palatal [FL]. So, there is a if you see; there is a no word boundary; there is no gap between the two words when we say in a continuous, but when we writing it there is a gap in a words.

If I show you English example; if you see there is a no gap I can give you the number all you can understand. So, if I show you number 7, 5, 4, 6, 1, 9 equal numbers. Now I can ask you based on the manner and seeing the spectrogram; can you identify the number? So, this kind of problem people may ask; you that let us we do the serial crossing later on find out the speech crossing later on. But seeing the number; if I say that find out what should be the number this one and what should be the number this one? So, if it is a 7; it should be start by a friction.

So, that way I have to find out whether there is a lot of difference or not. Now, one things I can again show you; let us see this speech [FL]. Now, when you learn when you heard it; you cannot find out if we do not know the Japanese you do not know what it is spoken or you can recognize the word sentence all kind of things. But if you heard the basic sound; you can understand the basic sound. So, I can say that some language so that is why past I have said that to decipher the language is important to find out the sequence of phonemes to produce a meaningful pronunciation.

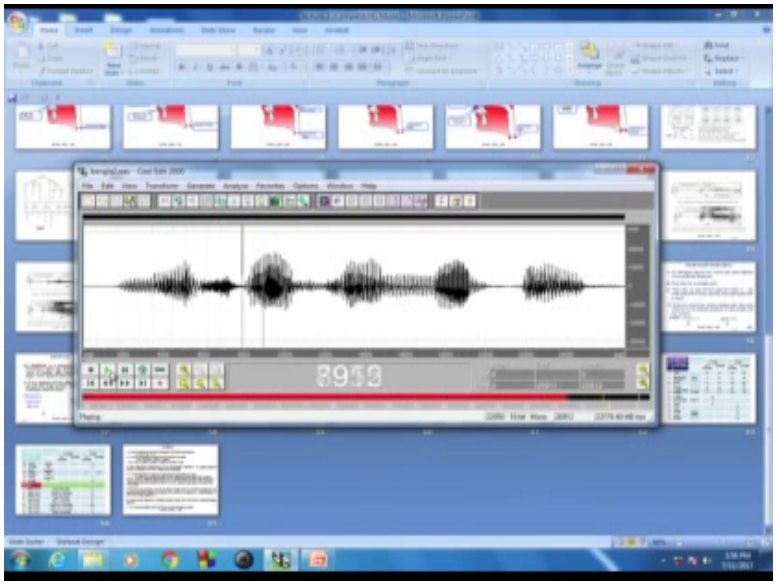
So, all the sounds are there; all basic sounds will be same; you can you can also see it. I will share this all sound file you can see it, all the basic sound there is a may be [FL] there may be a [FL] there may a [FL] all kinds of sounds are there, but if you see once I play it; I cannot recognize it. Because I do not know the language, but sound structure is same that is why we say IPA; International Phonetic Alphabet; which alphabet represent the sound only, it does not depends on the language.

So, [FL] it may in Bengali exist [FL] may exist in Hindi [FL] may exist in Tamil that sound may exist in Malayalam, that may exist in Japanese, that may exist in Chinese, that can exist in German also. So, phonemes are the sound sequence of phoneme produce the message. Now, production of phoneme involve this part; so, there is problem that problem is that some phoneme say it maybe exist in Bangla, but it may not be exist in Hindi or it may not be exist in English.

Some phoneme which is exist in English, but not in Bengali. So, what is happened suppose; I produce the English, the speech which I produce in English language sequence of sound is such a way that it follow the English language grammar, but what happen is that my English cannot be say; it is a British English, how the British is pronounced. Because in case of British English; my mother tongue is first language is Bengali. So, I learned that Bengali phoneme production when I learn English; I try to copy that production with the English language. So, that is called l 1; l 2 acquisition.

So, details I am not discussing; details are there. So, there is another part also segmental, super segmental that part I will come later on. So, what I am saying is a segmental property that in a segment; what is the look like this. So, this is called segmental property. So, I request all of you to find out to records or to analyze one number of speech signal; see it in spectogram or time domain try to find out manner of articulation

of that production. If you see, there is a vowel; vowel transitory effect also there if I say I have a Bengali sound; I think I have; I will show.



Sample English that bangla two [FL]. So, if you see there is a word one word is end and another word is began, but there is no gap it is a co articulate it likes a complete single word that details I will come later on that.

So, there is no gap main things is that there is no word boundary which is define speech signal; there is no silence it cannot say that this word is defined by this word is there is a gap in written things. So, there may be a silence in speech no, all words are related with each others with a co articulate effect then there is a prosody will come then the silence will come based on the prosodic structure. So, I am not coming that part then I can say the phonemes are have a co articulate effect [FL] followed by [FL]. So, there is a [FL] then [FL] to a transition then to then o transition. So, all kind of transitory movement will be there; now if you see there will be exist some combination which cannot produce by me of which cannot produce which cannot producible. So, that combination will not exist in that language.

So, if you see there is no you cannot find a case where there is a two aspirated sound will be club together; that is not pronounceable. So, if it is two aspirated then one will be unaspirated; one will be aspirated. So, depending on the how tongue movement restriction,n some combinations components does not existing for a particular language or; so, that is not valid. So, those kind of things also you learn; so, then there is a called

phonologies that the written words not offer with the pronunciation so, that part I will discuss in the TTS; when I talk about the TTS; phonologic that word to phoneme conversion.

So, now I will request all of you to record the sounds and listen that sounds and find out what kind of consonant it is; then you can say, you can expertise your skill. I said one of the outcome of this course is that for a given signal, you should able to level the signals. So, for that skill development you should practice it; where is that this kind of example I sometimes given that sometimes I given that some I will give you some spectrogram and some words; next I say [FL] or I can say that 1, 1, 2. Let us seven see what is; I have given now if I give the spectrogram of the three words and told you identify which is 1, which is 2, which is 7.

Now, if you see all three are vocalic; this one is nasal murmur to vowel. If you see here starting is a stop consonant of plosive; then there is a 2 vocalic; 2, then if you hear there is a fricative sound in here; then there is also nasal murmur ending and say then there will be a voiced consonant here. So, I can say seeing this is first one is fricative then I can say this 7; first one is plosive two first one is vocalic one that kind of things I can ask you. Another things is that I have just closed it again I open it if you see that I can show you how to see the pitch; fundamental frequency let us see here. So, this is the vocalic region detail signal crossing I will discuss.

If you analyze the frequency; if you see that this is frequency analysis. So, I say lean instead of linear view, I can make it log view. Now, if you see this is the frequency parenthesis and this is the DB in power. Power of the particular frequency, now if you see this let us scan it; with little up of high things.

So, if you see this is the first peak; so, first peak maybe is called fundamental frequency. So, I will discuss the find out the fundamental frequency of a signal those methodology I will discuss why this is the one of the methodology to see the fundamental frequency. Now, I can move this cursor to here and I can see the scan again fundamental frequency will be change. So, the movement of fundamental frequency will be there. So, that I will discuss later on let us not do it.

But you should practice those things; visualize different consonant find out I can say the level of occlusion period, level the burst, level the VOT, level the steady state portion of

the vowels, then level a diphthong, then the level a vowel; vowel combination. So, those things are there in the slides and once you practice it; you can able to relate it. If you say that or if you can give the feedback that you have some difficulties; then I will again take a tutorial to show you those kind of consonant and those kind of vowel combinations.

Thank you.