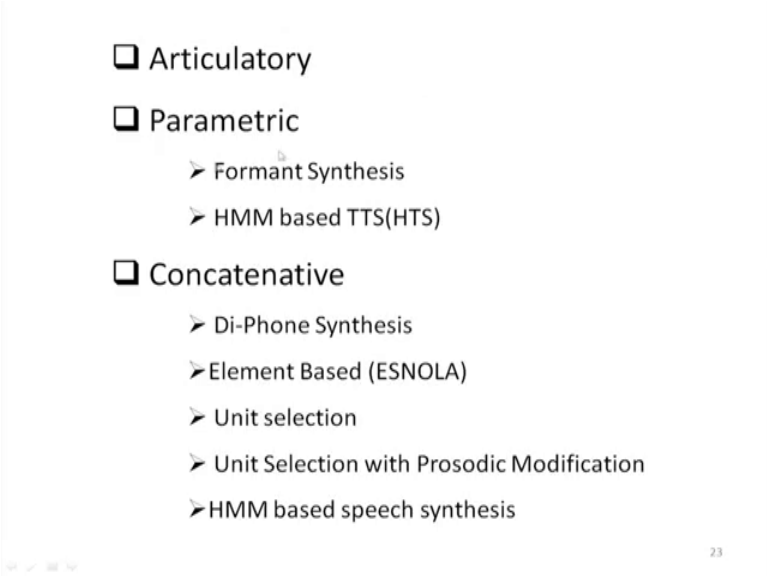
Digital Speech Processing
Prof. S. K. Das Mandal
Centre for Educational Technology
Indian Institute of Technology, Kharagpur

Lecture - 38
Text To Speech Synthesis (Contd.)

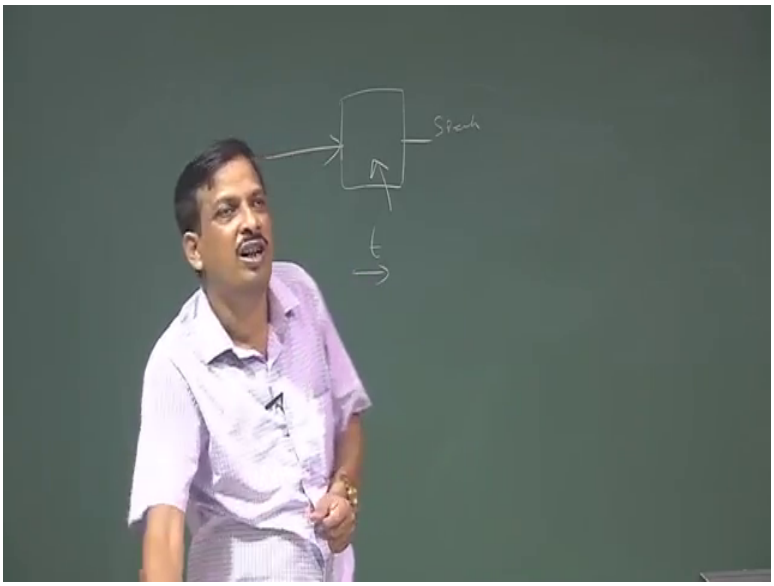
So last class we were discussing about the different synthesis technique. So, if you see that this slides the different kind of synthesis techniques are there.

(Refer Slide Time: 00:28)



Articulatory synthesis parametric synthesis and concatenative synthesis based on the technique 3 types articulatory synthesis parametric synthesis and concatenative synthesis. So, if you see the articulatory synthesis, if you remember the tube modelling class. So, what we try to develop is the mathematical model of the human articulator.

(Refer Slide Time: 00:59)



So, if I able to mathematically model the Human articulator then, if I excite that articulator by a excitation which is call vocal codes and then I can generate the speech signal, but what is the complexity in articulated modelling the complexity is that if I develop that model, some model parameters or you can say that model is changing with respect to time it is not continues if I say Kolkata model is not same for all voice.

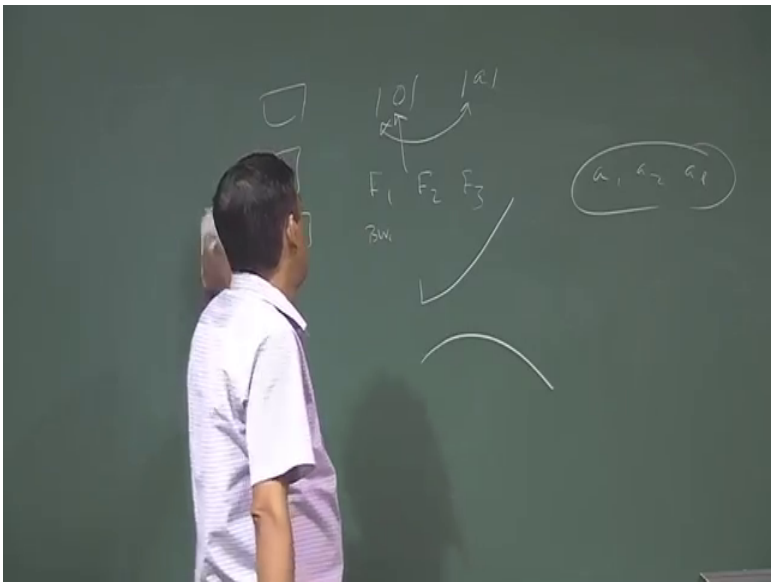
So, once the so against the time the model are changing; that means, (Refer Time: 01:30) area will be change air flow dynamics will be change; so if is all change the model composition is changing. So, this means that I required lot of mathematically you can that super model that which can take this kind of variation. So, that I can generate the articulatory speech, but it is very good things, that if I am able to generate this I able the mathematically model this thing with respect to time and then I can excited that model by a either impels or a noise to generate the speech. So, that is called articulatory synthesis.

So, lot of research is going on in articulatory synthesis because it not only provide me the speech synthesis, also I can simulate the human vocal track, also if I able to construct mathematically. Or I can say instead of this if I able to make a some instrument which can simulate this thing is fantastic. So, that people are doing this articulatory modelling lot of reach is going on in articulatory modelling, but this is very tough because with

respect to time. Suppose I calculate the cross sectional area for o one frame next frame the composition will be change. So, mathematical model is change.

So, those kind of compute a heavily computational complex problem is there, but research is going on that is called articulatory synthesis, then there is a parametric synthesis if you remember we said in that that a speech event or I can say whether it is o or.

(Refer Slide Time: 03:18)



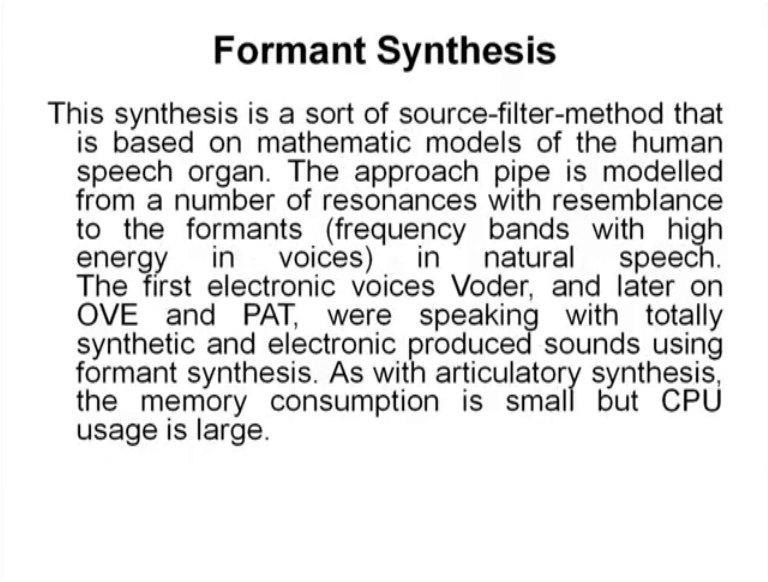
Whether it can be determining by the format frequency, if you remember the format frequency f 1 f 2 f 3 and format band width. If I know the format frequency and format band width I can generate o. So, that is called parametric synthesis. So, if I know the format parameter and format brand width for that particular speech event, I can generate the speech event using articulatory synthesis that you know the lpc model. If I know the lpc feature vector a one a 2 lp then I know you designing a filter, I can generate that event, but the problem is that if you see that once we speak the movement of the speech format is, it is not static that this frame it is x this frame it is y it is continuously moving.

So, speech format are moving continuously. So, co articulatory effect when it is moving from one phoneme to another phoneme how the format is moving is very important. So, that format dynamics I have to know. So, that format dynamics if I able to apply, then I can generate the speech using this kind of parameter that is why it is called parametric synthesis.

So, format synthesis is called parametric synthesis, which is develop based on set up parallel filter because nothing. If I know the format frequency, I can design that filter and based on the gain of that format I can pass imposes to particular and can generate that speech signal. So, that is called format synthesis then there is a hmm synthesis using vic order we develop that hmm based speech synthesis which I will details, I will talk at the end hmm or which is called hts recently mostly used tts system is hts hidden mark of tool kit base speech synthesis model. This is mostly use tts engine currently then there is a concatenative synthesis which is you can say that before, hmm base tts is come this concatenative synthesis is was that highestly use synthesis technique that initially it was develop diphone base synthesis. Then we have develop one synthesis technique call element base synthesis technique and during my recent during 2 thousand I think 3 to 2008, I was in (Refer Time: 05:58).

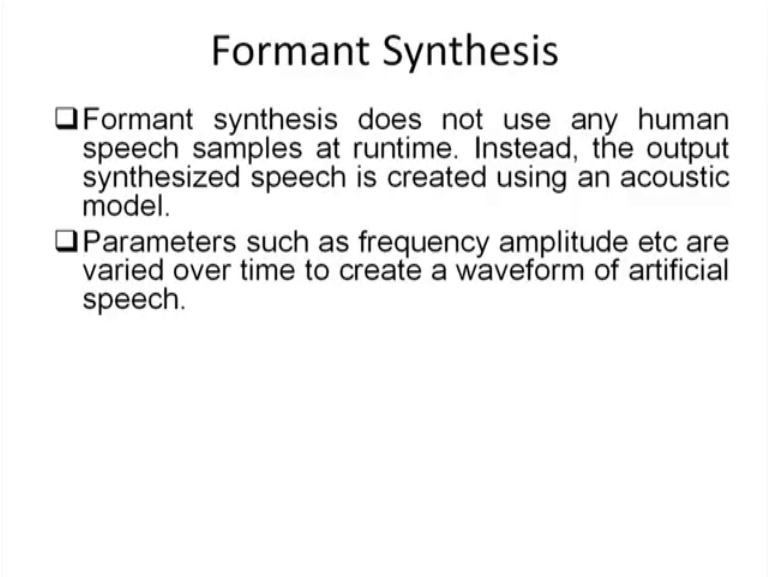
So, that time we develop that esnola this is call if you know Dutta ak Dutta myself we are developing that esnola technique speech synthesis which is nothing but a element we use the speech element as a unit and that time that, I phone base synthesis was there then there is a call unit section the unit section with prosody modification. And last one is hmm base synthesis also it is 2 types the one is called hts one is parametric another is non parametric only the choose choice of the segments are detected by hmm model. So, that is hmm base speech synthesis which is also use in concatenative synthesis.

(Refer Slide Time: 06:49)



So, now, I not dis discuss about this thing already discuss format synthesis does not use any human.

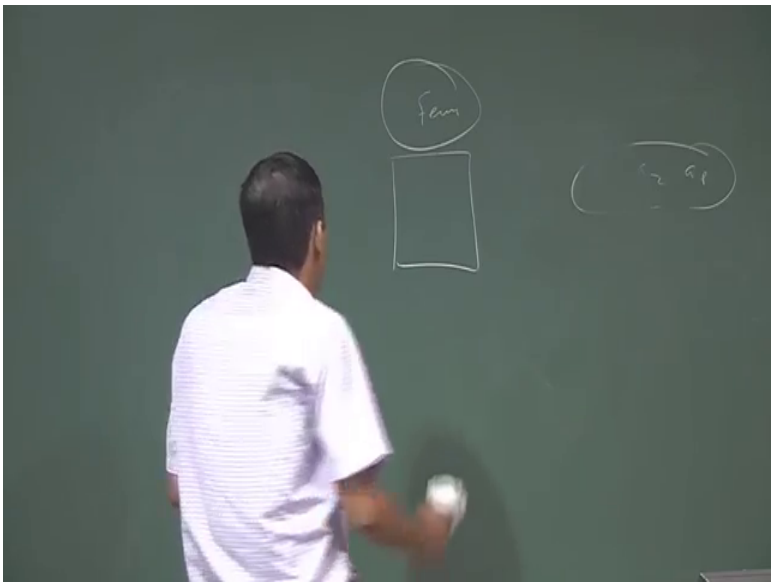
(Refer Slide Time: 06:54)



Speech sample at runtime, instead of we put synthesis speech as a creative using acoustic model parameter such as frequency amplitude etcetera are varied over time to create a waveform up articulate speech then concatenative synthesis, this is a very simple modes of speech synthesis technique.

So, what does it required any concatenative synthesis is required a free required data base.

(Refer Slide Time: 07:23)



If I say the machine let us machine acts as a female voice. So, I required a pre-recorded female data base, if I recode it pre-recorded data base, then for the text which you has the input I can generate that text speech from that pre-recorded data base. Very simple example will come from that that concatenative synthesis you can understand. That, if you go to the train station there is a lot of announcement if you see that I (Refer Time: 07:54) who the train will come at flat form number 2 train come on flatform number 7, if you say that kind of announcement happening. Now if you manually notice if some station had added one flatform extra, after certain time after let us 10 after 10 years there is a ibra system available in that station, but after 10 years there is another flatform is added.

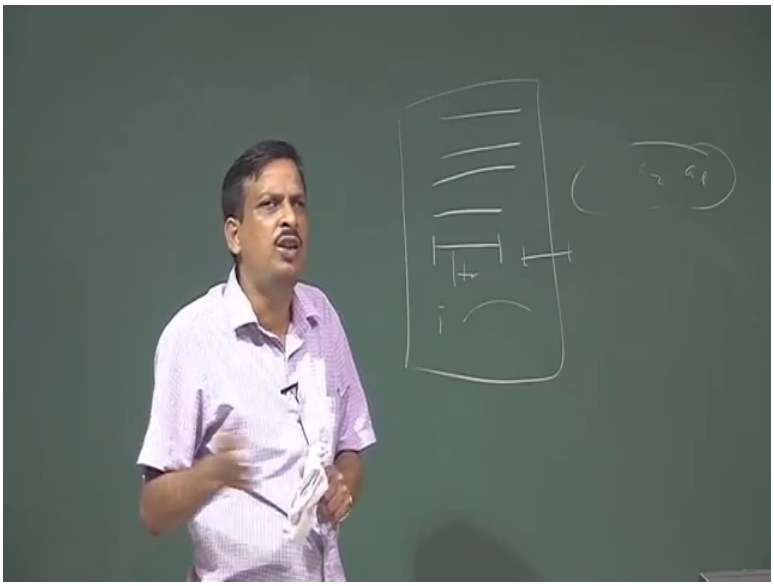
So, let that flat form number 5. So, the recorded voice after 10 years' same voice is not available; so the recorded from different speaker flat form number 5. So, you find that the train is coming on will be one voice flat form number 5, 5 will be otherwise. So, what is there. So, they have a pre recorded pre recorded speech or all information and depending on the train coming and so there are adding some unit.

So, there is a pre-recorded unit maybe train number station number flat form number train number train name all are pre-recorded in the data base. So, depending on that which train is coming they search it added together and play it. So, that is nothing but a concatenative synthesis. So, instead of word; so if say that it is it is very easy because if

the train number of train are very less the number of our you can say the finite number of flat form or also finites.

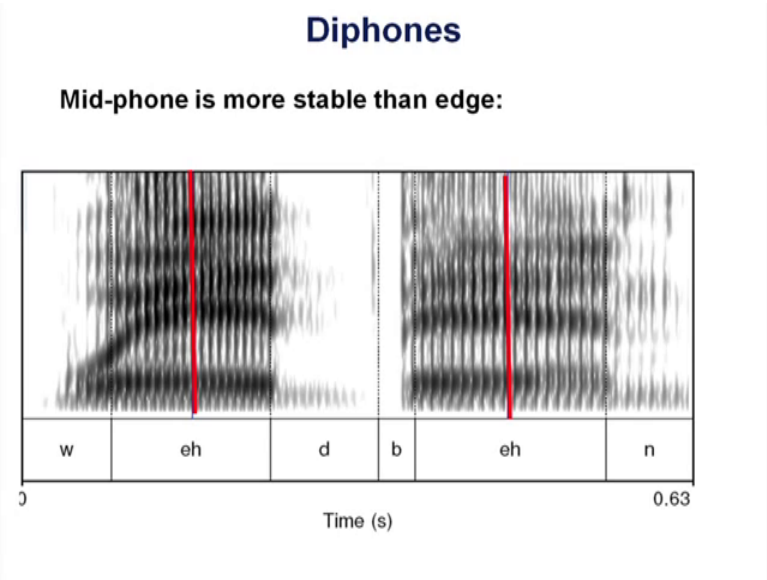
So, I can recorded the finite number of words, but think about the language, I have infinite language words and they can combine in any way.

(Refer Slide Time: 08:47)



So, how do we recorded all speech it is very trap to recode all those think. And also if I recode it then matching it is very tuff because the some and part of sentence will be same some other way next part will be sentence speaking some other way. So, all kinds of problem will be generated. So, what will they said instead of recoding word sentence all kind of things initially they thought if I recode only the diphone. So, for a particular language if I have a you know that phoneme, how many phoneme are there in English how many phonemes. So, for Bengali suppose there is a 33 consonant and there is a 7 vowel and I can generate how many diphones will be there, I can generate how many possible diphone will be there.

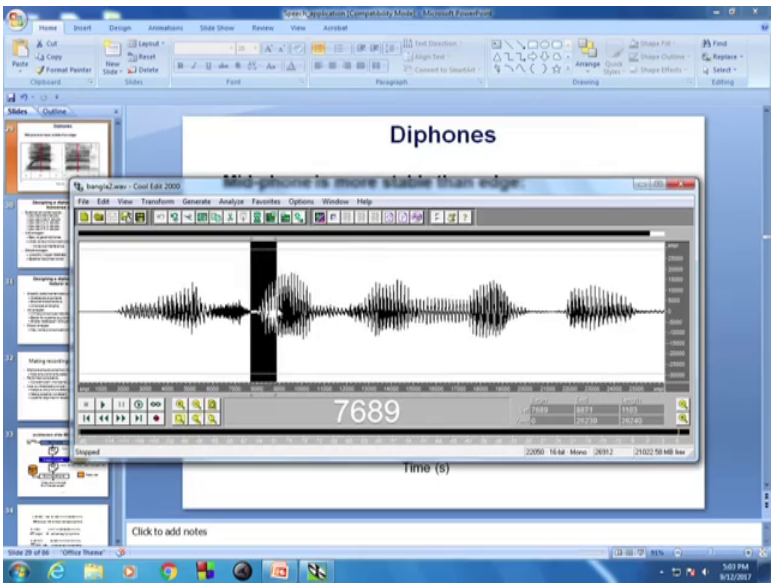
(Refer Slide Time: 10:36)



So, what is diphone is basically if you see diphone is define like mid phone is suitable, then the edge said the mid of the phone is suitable for the edge. So, that is why it is called diphone both the phone. So, how it is recorded?

Now, if you remember or I can show you in here, suppose I open and waveform then here let us this is the waveform.

(Refer Slide Time: 11:08)

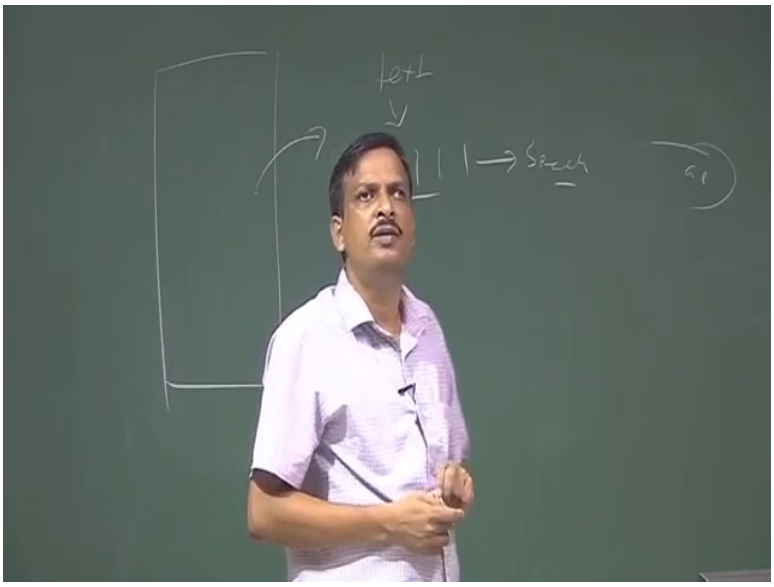


If you see that if I cut the waveform in here there may be some erroneous result will come. If you remember I said the steady state vowels are more lightly same, if I zoom

this potion if you see the vowel perioder look like same. So, if I cut it here in one of the (Refer Time: 11:09) potion of every vowel one vowels, then from here to middle of this to middle of this. Then I can say it is a diphone and when I added them different diphone the error will be minimum. So, because the steady state is nothing but a steady potion of the vocalic reason because it there is no dynamic are there I cannot say which one is sow, because it maybe the pure consonant, but consonant vowel transaction is important. So, that part also contain the consonant information.

So, I cannot selected that transitory part I can select the steady state part. So, steady state part to steady state part can cut I can make it a dictionary. So, all possible diphone for a particular language, I cuts steady state middle part of the steady state next middle part of the steady state and add ppra dictionary. Signal dictionary which contain all the diphone.

(Refer Slide Time: 12:31)



Now if you see any text will come any sentence or any text will come that is consist of some combinational those diphone. So, I can pick up those diphone and added together to combine them and I produce the speech. That is called diphone synthesis technique diphone synthesis technique.

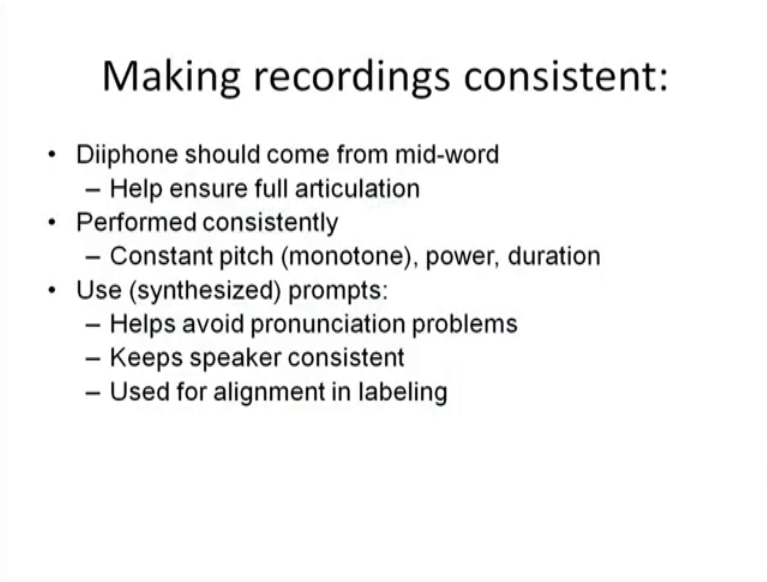
So, why I cut in the middle because that part is the steady state vowel, I cannot cut here because there is a transitory portion which is part of the sow also. So, that is why I take the diphone middle to middle that is why it is called diphone; so this diphone data base. So, training choose unit. So, what is required I have to recorded all diphone. So, I can

say I can say put I what I do, I recode them, I cannot recode a single diphone by pronunciation. So, I can say I can I can generate some nonsense words which contain all the diphone. And I pronounce that word and I can cut those that diphone from word signal now or I can recorded the natural speech.

And I collect or I can see have a large text purpose is there. I can analysis that text purpose to find out how many sentence is require to cover all the diphone which I required I can search in the purpose, I can find out the sentences and take that sentence recorded that sentences and cut the diphone from that recode, but the problem is that if it is natural sentences then every sentence has it own prosody depend on it is you can say the the structure of that sentence.

So, if I cut a diphone from the beginning of the sentence and the same diphone, if I cut from end of the sentence 2 diphone will be different because of prosody is changing. So, if I do not want that prosody is not required, when I just simply synthetize the segmental information, then I can say I discard the prosody I recorded a recorded the word with a carrier sentence which is neutral carrier sentence and then I can cut the signal from that nonsense word. So, that is called diphone sentence.

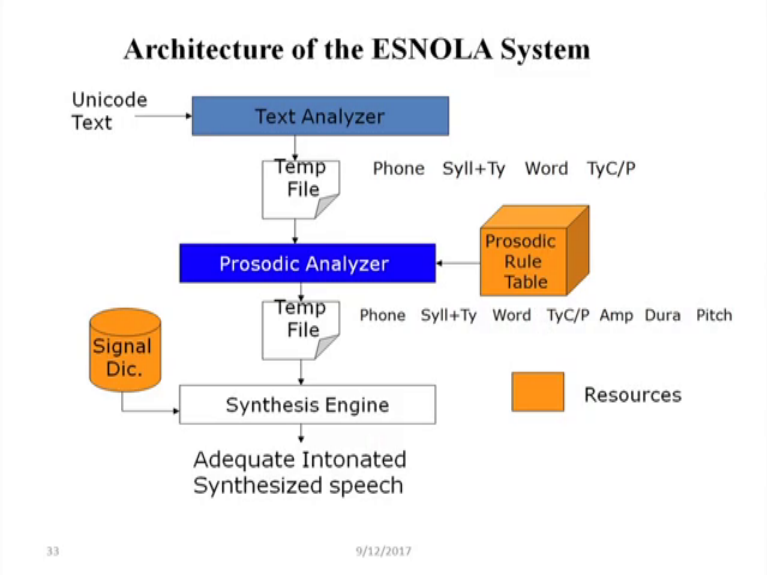
(Refer Slide Time: 15:26)



Now if you see all kinds of diphone at the given the advantage then the this advantage making recoding consistent diphone should come from mid words not take the beginning words because beginning words voice is started.

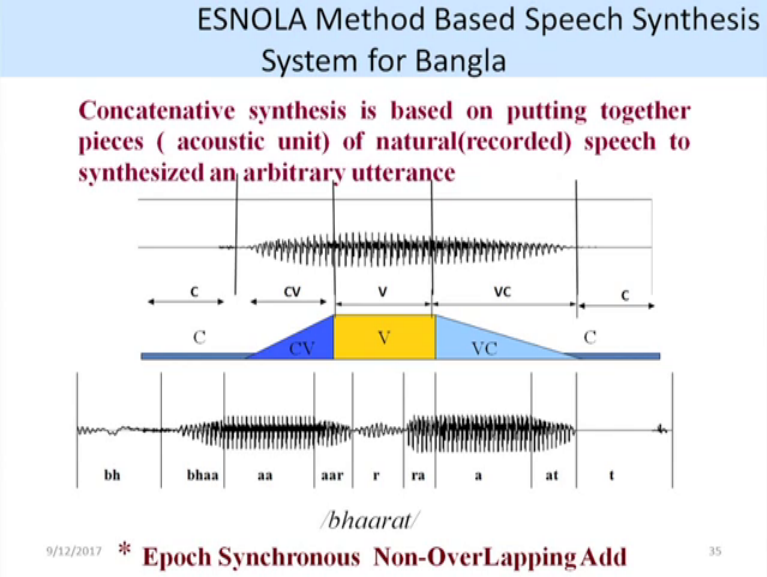
So, beginning word I do not want I want mid of a word that diphone should be taken.

(Refer Slide Time: 15:40)



Then that same time we are also developing one technique which is call esnola (Refer Time: 15:46) non-overlapping add method base tts which text analysis all those are same only what you want we instead of diphone we call it is a part name. So, I am not explaining the prosody part I am only explaining the signal part which is call part name dictionary, what is part name content if you see I can go to the here.

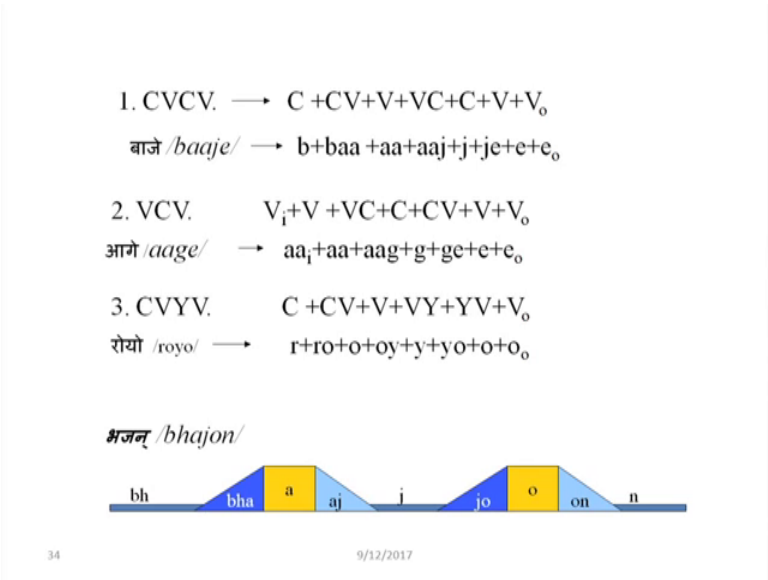
(Refer Slide Time: 16:17)



If this is the this is the recoding of a consonant vowel and consonant signal like I recorded and if it is ka you know this is accusation period, this is burst this is bot and this is transitory part. So, I can say after burst to the steady state vowel is important because that is a dynamics part of the signal. So, I can say it is consonant to vowel transitory parts cv segment. Then I can say put to bust beginning a accusation period to the end of the bust, I can say it is a consonant then that steady state vowel then again from steady state vowel to consonant transitory part.

So, my dictionary content only c cv v may content or be may not required also I can generate this v potion depending on the duration, I will come later on that then vc and then c. So, for all the consonant and vowel I generate those c cv v vc those kind of signal then suppose I want to produce. So, it is nothing but a consonant which is occlusion past burst then to transitory part then steady state vowel part then to transitory part then is a single consonant part then to transitory part steady state vowel, then to transitory part then I can generate any word will come I can divide those word like that way and I can generate.

(Refer Slide Time: 18:05)

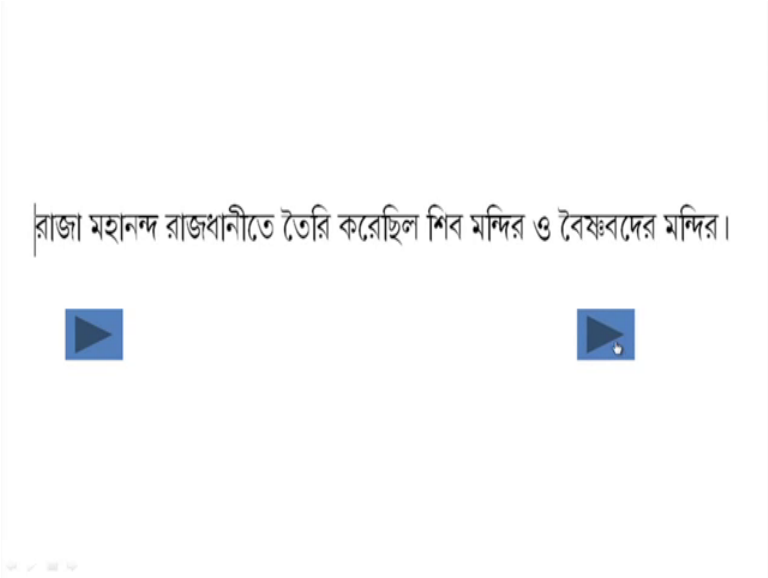


So, here is a Hindi bhaj Hindi is there. So, to a transitory part to then consonant then o then o n it will pronounced. So, the synthesis root is simple cvc if you see the I can generate c cv v vc c v and fed out of v o means fed out. So, if you see the there then vcv

if the segment is vcv then I can generate like this way if it is cv yv in can generate like this this is. So, those rules are the synthesis part.

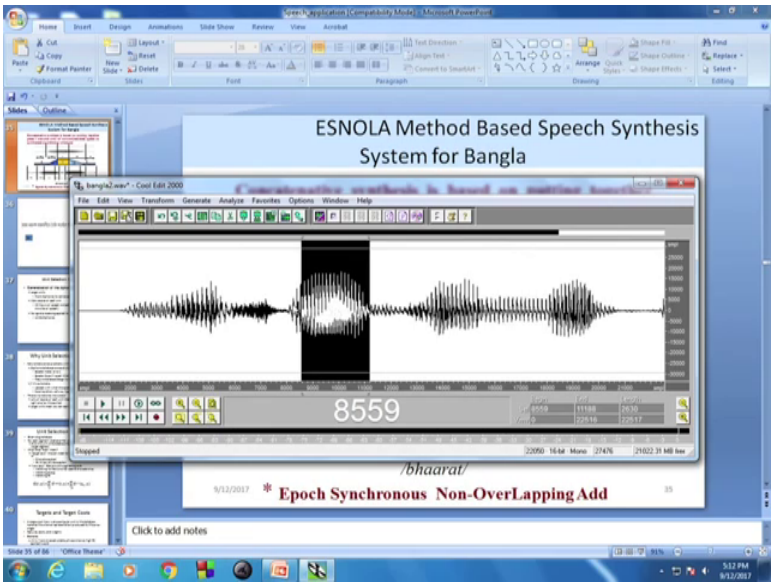
So, dot is called this is called esnola base synthesis technique, I can show the quality for Bangla.

(Refer Slide Time: 18:56)



So, prosody is not there if you see it is flat, I can say it is only all part name I have collected and all part names are concatenative together to produce the whole sentences. Now if you show I told you that steady state vowel no. So, if you see the steady state vowel if I show you in signal.

(Refer Slide Time: 19:26)



If you see the steady state vowel, see the vowel periods are almost identically. If I say if this is this is the beginning of the vowel period. Then the up to here is a one vowel period this is called epoch the beginning of the gotal impulse, if you analysis it will come in here. So, here to here is a vowel period. So, if I suppose this steady state vowel I do not know how much long is required. So, depending on the duration requirement, I can repeat this vowel portion to generate the I will, I can show you here also suppose if I cut this vowel I do not know whether you are able to listen, it is audible I can select the device I think device has to be loaded.

So, if you able to listen, it I think you are not able to listen, it I will see that what kind of device it is there. So, there is a sentence now if you see steady state vowel I want to increase the vowel (Refer Time: 20:56) in steady state. So, I if I said this is this is the steady state beginning and upto here all are look like same. So, let us I cut it here I will show you. So, this is the one period of the vowel up to here is one period. Then I can copy it, then I paste it here again.

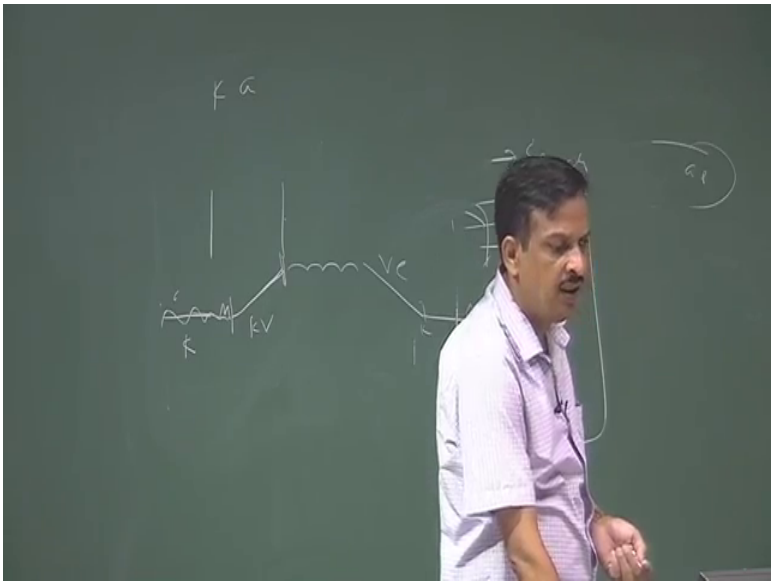
Then I can again paste it here. I can again paste it here I can again paste it here, if you see the vowel length is increases, but quality is same. So, this length of vowel is increases I just only repeating the single period. So, steady state vowel I can change, the steady state vowel length by repeating the same period. So, if you see the example, I can show you this is the less steady state vowels are very small if you see it steady state

vowel duration are very less. If I play this only the steady state vowel portion are lengthen.

So, this technique is used in Bengali tts to produce that any. So, this tts engine can itself take any word you can give it can pronounce to the same clarity, but the main problem is that prosody is not there, prosody is not there. So, prosody has to be incorporated; so that I will discuss later on. So, this is Bangla tts we call esnola epoch synchronous overlap add method and this is the block diagram of the whole tts. If this is the block diagram whole tts engine, and this is the rule we have writ10 down then we generate this dictionary is created. So, we recorded how the dictionary signal dictionary is created we recorded all the consonant with vowel in a non-sense words. So, how the dictionary is created I can told you that problem is that if it is and there it is.

So, I recorded 4 5 times and I take the most appropriate one middle one and then I cut the occlusion.

(Refer Slide Time: 23:33)

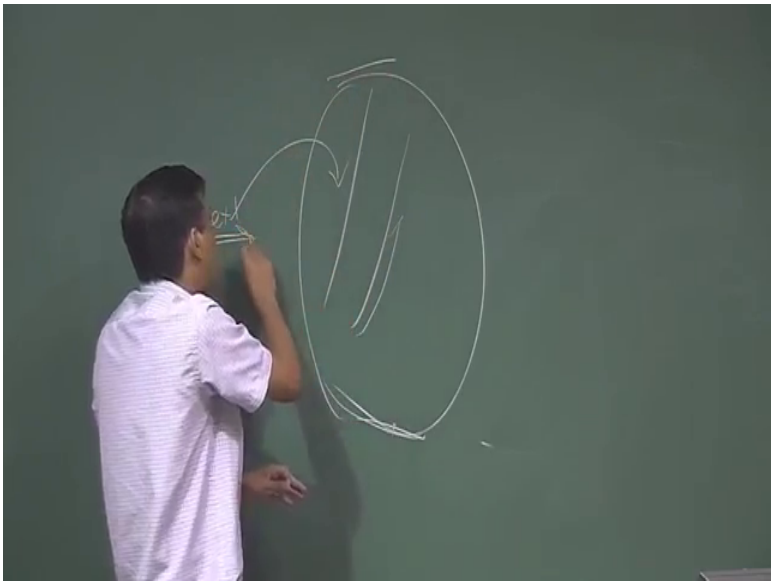


Occlusion period plus burst is my then from burst end of the burst to steady state vowel is my to vowel transition. Then I do not require the steady state vowel, because I can repeat the last period many times how many I required and then there is a vowel to consonant transition. So, it is there will be a vowel and there is a consonant transition. So, vc portion will be there and then again there will be a con the consonant which is occlusion plus burst if it is (Refer Time: 24:04) consonant, if it is vocalic consonant same

up to beginning of the vot and from the beginning of the vot to steady state is called consonant to vowel transition. So, all consonant with vowel combination recorded and cut to create the dictionary and once dictionary is created any text will come first to the rule and it can generate that.

So, this is used in esnola synthesis then there is a unit selection method will come. Because if I see the all diphone base tts and all that part name is tts has some problem in natural prosody because prosody is not there, I have to model that prosody, but modelling of the prosody is not that is simple. So, they said instead of selecting the diphone can I select a larger unit. Because today in memory does not have any problem the computer memory become very easy computational power also is there. So, I can say let us I recorded large amount of speech data, from a single voice because if it is mix then male female will be mixed.

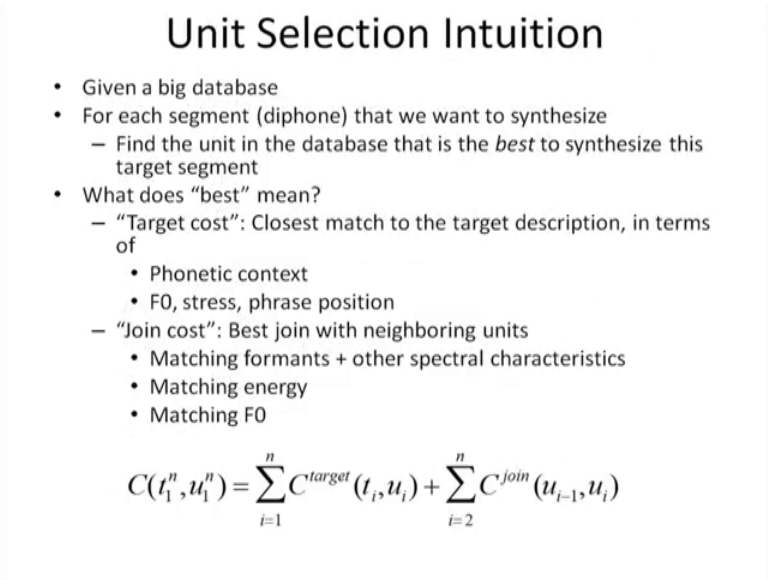
(Refer Slide Time: 25:18)



So, for the single voice I recorded large amount of speech data. Then was the input text is come I search on that large amount of this is the level speech data.

So, this is resource is large amount of level speech data is available once the level speech data is available. What about the input text will come I can search on the level speech data and find out the best possible match part of that corresponding to input text and I can produce that is my synthesis base? So, that is called unit selection bet best speech synthesis technique. So, if it is unit selection.

(Refer Slide Time: 26:16)

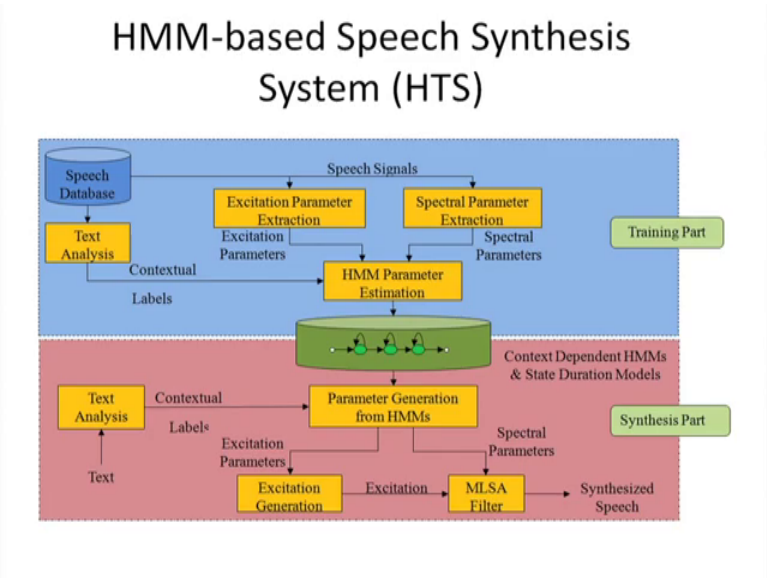


So, if you see only problem, is unit selection is that how do I select the unit. So, given a big data base is there for each segment that we want to synthesis find the unit in the data base that is the best to synthesis this target segment. How the best is measure depending on the target cost. Closest match to the target description in term of phonetic context f 0 stress phrase position and joining cost is that best to join the neighbouring unit. So, when I join it the end of the segment and another end of the segment, how they are joining good and how this whole segment and this segments are good to join together in term of closest match. Target description in term of phonetic context f 0 stress phrase position.

So, in unit selection prosody is in build. I cannot modify that I not use to modify that synthesis base according to the language. So, the prosody which is come in unit selection method which is available in the data base if the text which is given if it is whole text is available in the same context then I will get best prosodic match sentences, but suppose that input text does not have any instant on this data base only some part diphones are there are in here and there, then if I club together then I cannot minimise the target cost and the prosody which will come it is not come depending on the language structures it will come within the available level data prosodic structure. So, prosody is driven by the data, but if the data is big then my synthesis output will be good if the data is less the synthesis output quality will be not that good.

So, target cost and joining cost based on this cost 2 cost function we select the so konig velloni university that is the unit selection tts are there festival engine, you heard about that festival engine. So, that is nothing but a unit selection base tts system I am not details describe the target cost which is available in the slide you can go through it then there is a another tts is come hmm base synthesis.

(Refer Slide Time: 29:03)



This is hmm base synthesis we have develop for Bangla also hmm base synthesis. So, it is 2 type one is called parametric, if I use vic order then it is parametric synthesis because here. What I will do speech data base is there we extract to the excitation parameter. So, if you see the speech signal consist of 2 part one is called segmental supra segmental or I can say it consist of 2 part one is called excitation parameters another is called spectral parameter or you can say, if I have a vocal track vocal track is excited by a vocal code. So, I can say excitation parameter and vocal track parameter.

So, this is called spectral parameter and this is called excitation parameter. So, we have extract we have separate from a large speech data base, with this we have separate the excitation we have the excitation parameter and also we have a spectral parameters based on this 2 parameter. We developed hmm parameter estimation technique now once the text is come input text.

So, text analysis is done contextual labeling is done then depending on that text, we have estimate the parameter which segmental parameter is best match for that input text and

which excitation parameter has the best match with the input text. Once I get the excitation parameter and spectral parameter then we can use the synthesis filter or vic order to generate the speech. So, that is hmm base speech synthesis technique which is used in which is name as hts hidden mark of tool kit base speech synthesis technique.

So, next class I will discuss about that asr some portion of the asr. And then we talk about that action conversion that thinks. And the last week is the prosody.

Thank you.