

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

Lecture – 42
Prosody Modeling

So, let us start that speech prosody. We have said in the 8th week, I will discuss about the speech prosody. So, more or less; all discussion will be related to the Fujisaki model and also some work has been done for Bangla in my lab that is I also discuss. So, think about that; what is speech prosody; first people will ask that what is prosody; what do we mean by prosody. Now if you remember in speech synthesis, what I said that what about the speech synthesis technique, I have used whether it is a I-Phone base or unit selection base the prosody is an important parameters because prosody information is only if you want that speech should be natural communication then without prosody the synthesis becomes unnatural.

So, I want to implement speech prosody means in synthesized speech, I can use speech prosody on other hand prosody can be used in speech recognition also because speech signal not only carry the segmental information, but also there is a prosodic information which is necessary for which can be which may be the meaning full for improve the speech recognition output and prosody may be used in speaker and voice identification all the cases. So, we have discuss about the speech prosodic first then you can think what kind of information processing you should be use for what kind of application.

(Refer Slide Time: 01:54)

**What is Prosody ? --- The Author's Definition
(Fujisaki 1995)**

Prosody is defined as the systematic organization of individual linguistic units into an utterance, or a coherent group of utterances, in the process of speech production.

Its realization involves both segmental and suprasegmental features of speech, and is influenced, not only by linguistic information, but also by para-linguistic and non-linguistic information.

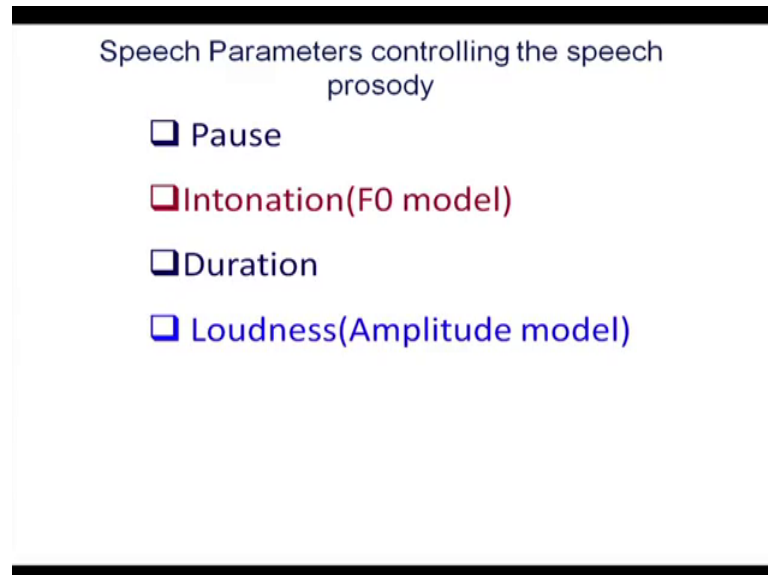
© Hiroya Fujisaki 2

Now, what is speech prosody? The prosody is defined as per the Fujisaki is the systematic organization of individual linguistic units into an utterance or a coherent group of utterance in the process of speech production. So, the meaning is that if I want to say read a paragraph or if I want to say a sentence it is not arbitrary rhythm I have followed, I am followed some rhythm when I pronounce the sentence which is dictate by the synthetic or structures of that sentence. You may say that then people who do not know the takes processing or who do not know the words no alphabet, then also they speak who is the who is the speech prosody yes because this is an acquire phenomena by a human being, but if you want to analyze it you find there is a unique relationship between the structure of the sentence and the prosody.

So, there is a relationship between the speech prosody. So, prosody if I say it is a conveying some information which is may not be linguistic it also conveying the non linguistic and paralinguistic information also. So, suppose if I am; if I say the emotional recognition people lot of people working today in emotion recognition speech emotion. So, emotion information is mainly on this speech prosody because prosodic parameter is change that is why different times; I produce different kind of emotional speech because if you remember the content what phonemes; I am pronounced is the segmental features now all the phonemes come together and with the melody then we can complete speech, but how this melody, how this speech will varies across the sentence, it depends on the

speech prosody. So, prosody is defined a systematic organization of individual linguistic unit in to an utterance or coherent group of utterance in the process of speech production.

(Refer Slide Time: 04:32)



So, if I say the prosody then it is across the segmental boundary that is why all the prosodic parameters are called supra segmental parameters. So, what are the parameter which control the prosody mainly these 4 parameters, I am not going to the quality the speech quality sometime also referred to a speech prosodic parameter, but I am not going that if I say the speech prosody depends on the pause means sentence where I have stopped.

If you see in between a word in continuous speech there, is a no boundary because there is no silence, we never said one word then we pause then say the next word we never say that way, but pause in appropriate position has an important role to convey the information you know that if I put the pause in arbitrary position then speech become unnatural and also the meaning of the speech may also change. So, pause as a important parameter in speech communication pause means where I have break the utterance.

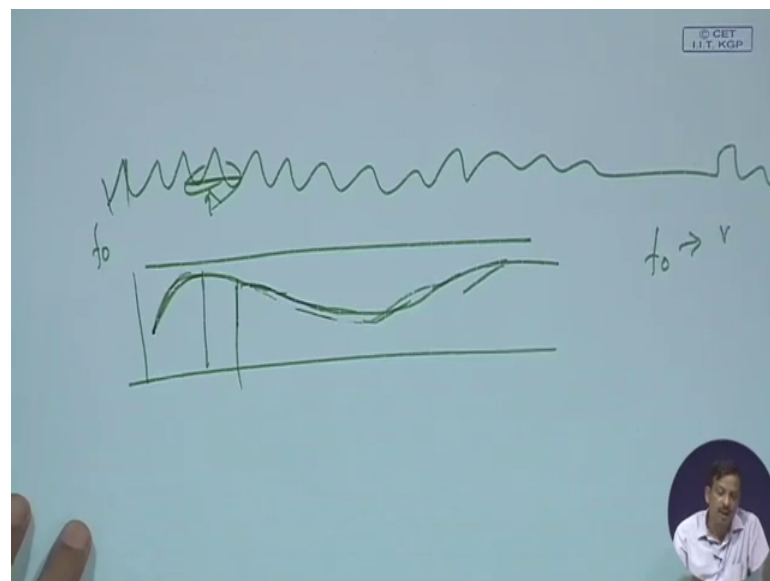
So, definition of yes it is very important is to know that sentence and utterance sent in written language we say this is a sentence in spoken language we say it is utterance; that means, it is not true that all are these one sentence is belongs to one utterance it may not be all true that utterance means within a single without resetting this whole system. Whatever I produce is called single utterance at upper each and every utterance the

system is reset means all prosodic parameters start from new things. So, that is prosodic break I can say prosodic break is the utterance.

So, if there is a long gap generally if it is gap me silence is more than 300 millisecond then you find that prosody prosodic parameter is reset. So, I can say this is a utterance. So, suppose I record a long sentence it may contain 3 or 4 utterance because there may be a utterance and after the utterance there is a more than 400 milliseconds pause then we can say this is single utterance although this is the part of same sentence. So, in spoken language the utterance is the chunk in written language we take a sentence as a chunk. So, this is the sentence level and utterance level you can say. So, within the utterance there inside there is a pause. So, those pause has carry an important information and the placement of the pause in spoken language is not arbitrary there is a depending on the structures of the utterance language structures of the utterance pause is determined by the human being itself.

So, when you see why do we say that we do not put the pause in arbitrary position if we put it arbitrary position then the speech is unnatural and meaning may be change then there is a intonation or F0 modeling if I say an utterance throughout the utterance F0 is not constant fundamental frequency of the utterance is not constant.

(Refer Slide Time: 07:56)

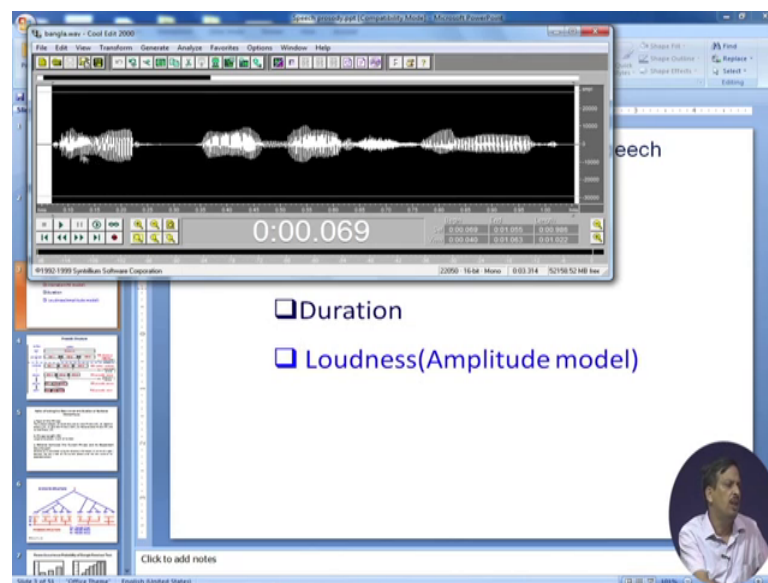


So, if I say suppose I have a speech signal let us I have a speech signal of one second, the if I extracted the F0 for every 10 milliseconds; if I have a 1 second utterance then I

found there will be a movement of the fundamental frequency and the fundamental frequency is not continuous because in utterance there may be a non voice region. So, if the speech signal is non voice there will be no F_0 because F_0 is the part of the vocal cord vibration if the vocal cord vibration does not exist. That means, there is a no F_0 , but if you see that F_0 control throughout the sentence you find there is a movement of F_0 continuous movement of the F_0 in a particular pattern. So, that pattern is called intonation of the utterance or F_0 control of that utterance.

So, F_0 is not flat for all whole the signal; it is moving along the signal and if I continuous if I model it continuous modally using a sine card if I make it continuous then I found the F_0 is moving in a particular ribbon or particular pattern that is called intonation then if you see here if I show you in the speech, let us suppose this is a sentence.

(Refer Slide Time: 09:26)



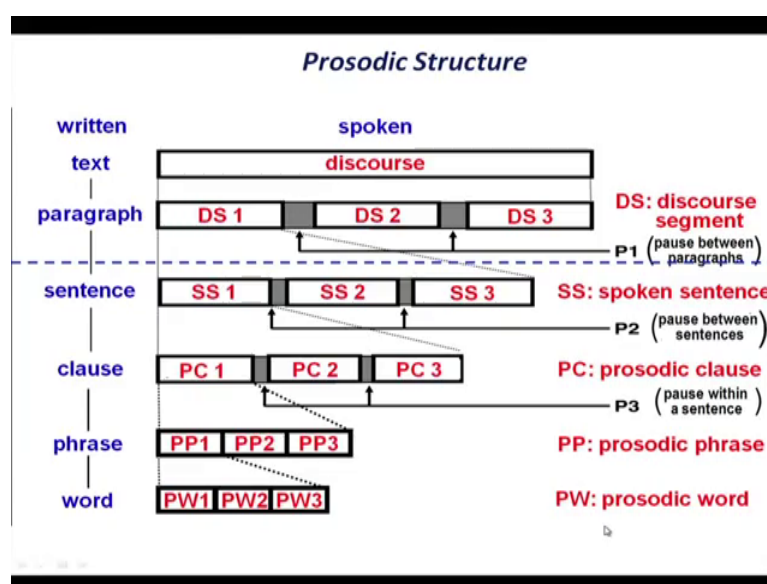
If I found the duration of each phoneme, each syllable instead of phoneme; I can define duration of the syllable all duration of the syllables are not equal throughout the whole sentence.

So; that means, duration of the segmental property is changing along the utterance I can say that utterance instead of sentence. So, along the utterance duration profile is changing that is called duration of syllable or phoneme whatever you can say phoneme duration in terms of you can morally written syllable also. So, syllable duration a

changing across the time. So, duration is important parameters which convey the speech prosody if I demonstrated you if you repeat this if you copy this duration and paste it, again if you check the duration you find the prosody is changing.

So, duration modeling duration is also an important parameter which control the speech prosody loudness amplitude if you see the whole utterance amplitude is not fixed amplitude if you take the average amplitude of every 10 millisecond if I plot it instead of unvoiced region, all voice region; you find that the amplitude is moving sometime it will fall or it will fall and again rise if there is a emphasis in the last 2 words. So, there is a amplitude movement also, but people say that amplitude is not that much of important for prosody because even if amplitude is equal if you able to model that duration F 0 and pause, then I can convey the prosodic information of the speech. So, that prosodic parameter for the speech are pause intonation or F 0 duration and loudness amplitude.

(Refer Slide Time: 11:44)



Now, we discuss about one after another in modeling. So, here in we have done some work in pause modeling, if you see that suppose there is a para; think about a paragraph, readout this course that I have reading a story written in or I can; I am reading in a page; something is reading in one language, it may be English it may be Bengali, it may be Hindi, I am reading that text, if I read that text, you find this kind of textures whole text has an larger discourse every paragraph has an discourse segment. So, whole text has a discourse depending on that discourse, my voice quality will be changing. Suppose in

some paragraph; somebody is died in a story; next paragraph one, I am describing this that by the intonation will be changed or by prosody will be changed on the next paragraph because of that pragmatic information.

So, discourse is an upper level spoken discourse that depends on the whole story that depends on the whole sentence whole paragraph the whole text kind of things then there is a discourse in the segment each paragraph has its own discourse and after every paragraph pause is mandatory every paragraph, we have a pause.

Now within a paragraph; a paragraph consists of certain sentences; few sentences, let us say, there is a 10 sentences; after every sentence, you know, there is a pause in speech and each sentence is consist of certain clause after every clause pause is mandatory if this pause is more than 300 milliseconds, then I treated that clause is an utterance or if it is less than 300 milliseconds, then I can say those are not single utterance, but there is a pause in between the clause. So, after every clause pause is mandatory a clause may consist of several phrases. So, I am working down in the written language processing and I try to correlate with the spoken language. So, every clause has a phrase. Now every phrase can be modeled I will show you in the one prosody model.

So, suppose there is a phrase; phrase 1, phrase 2, phrase 3; it might be often noticed that even though in written language there is a phrase one and phrase two, but in spoken language it might be only consist of 2 phrase 1 and 2 merged together consist make a 1 phrase and phrase 3 another one. So, we can say those are called prosodic phrase that I will come later on and those is called prosodic clause and those is called prosodic sentence. So, prosodic clause may have about utterance. So, utterance I can say this is a utterance if the this is see more than 300 milliseconds if it is not then clause one and clause 2 make a contender prosodic utterance then phrase suppose prosodic phrase every phrase may consist of several words.

So, there will be written word W 1 W 2 W 3, but you may found that suppose there is a prosodic the prosodic phrase that this is the linguistic phrase PP 1, PP 2, PP 3 that is phrase based on the linguistic analysis of the written text, but if you analyze in spoken text you will find there may be a PP 1 and PP 2 merged together to form a single phrase spoken phrase which is called prosodic phrase then this is PP 1 and this is PP 2.

Now, every phrase may consist of a several word; W 1, W 2, W 3, W 4; now both at the region word in spoken form you may found that W 1 and W 2 pronounced together I will come later on what do you mean by pronounce together pronounce together and form a prosodic word. So, a prosodic utterance we say utterance I start from utterance in text it is sentence in spoken it is utterance. So, utterance consists of a prosodic clause prosodic clause consists of a prosodic phrase prosodic phrase consists of a prosodic word similarly text in text clause sentence clause phrase word in text processing. Later on, I will explain it in on diagram now the problem is that the pause the occurrence of pause after every clause is mandatory by the if you found after every phrase pause may not be there or pause may be there.

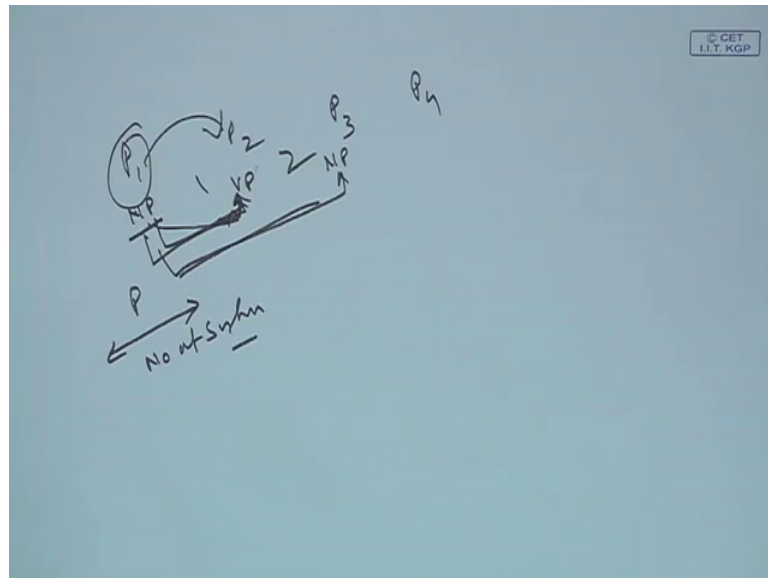
So, I have to make a model which can detect after which phrase pause is necessary and how much duration. So, I have to detect the occurrence of the pause in after phrase and the duration of that pause if it is there is a pause then what will be the duration no what do you mean by pause you can see what do you mean by pause let us see the sentence I am then. So, this is a sentence if you analyze a sentence you find after some utterance there is a pause there is a silence region of silence. So, that can define a pause there may not be a region of silence, but what happens if you see there is a la there is a large movement of all prosodic parameters.

So, maybe there is a deep F 0 counter is going down and then again going up. So, there will there will one kind of resetting of prosodic parameter duration F 0. So, that can indicate that there is a break in speech chunk. So, that is called prosodic phrase. So, the occurrence probability of the physical pause there may be a physical pause there may not be a physical pause, but there may be a break of co-articulation or there maybe break of prosodic parameters. So, if there is a prosodic parameter or break of co-articulation may also be there that can also define as a pause with geo duration, but there is a co-articulation break that identify when play the speech that phenomena identifying that this is the chunk of the speech.

So, you after we have studied that kind of things we found that factor effecting the occurrence and duration of a sentence medial pause we call sentence medial pause because at the end of the sentence pause is mandatory at the end of the clause pause is mandatory, but within a clause after which phrase I should put the pause with how much duration is called sentence medial pause modeling. So, take this parameter type of

phrase, length and distance between the current phrase and dependant counterpart. So, what we have done for Bengali, it may not be true for other language also I do not know we have not tried it.

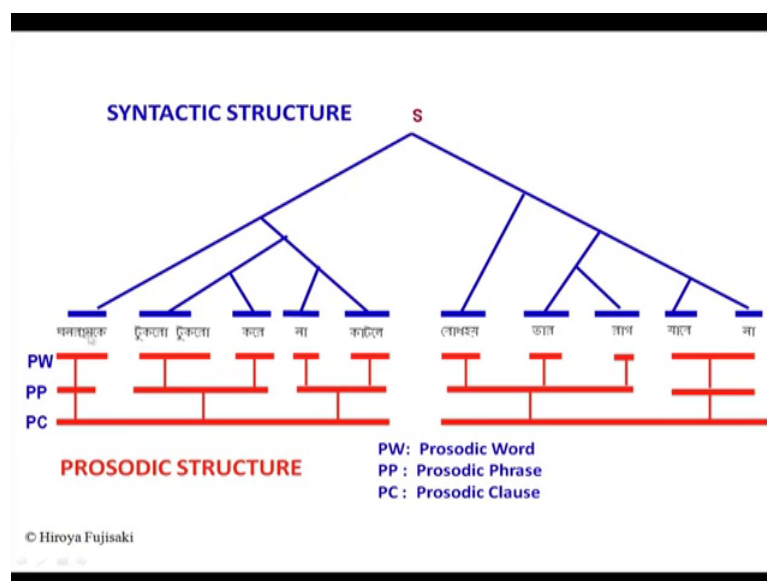
(Refer Slide Time: 19:50)



So, suppose I have a sentence; so, I have a phrase P 1, P 2, P 3, P 4, I have a 4 phase phrase then it may be mp it may be V P; N P. So, I do not know. So, I have detected that there we find out that type of phrase an important parameter then length of the P 1 in term of number of syllable number of syllable we measure in term of number of syllable and also whether this P 1 dependant on P 2 or not means if you see I will show you we have draw a what kind of binary curve where we can say whether this phase is dependent on this or not if it is far away it is far away dep. Suppose this phrase is dependent on this phrase and this phrase is depends on this phrase then the distance in case of this the distance will be one and this case of this distance will be 2. So, if the distance increases there is a probability of pause increases.

So, you have take 3 these parametric kinds of parameters then we analyze like this I can give example like this. So, this example is by a Bengali example you can see.

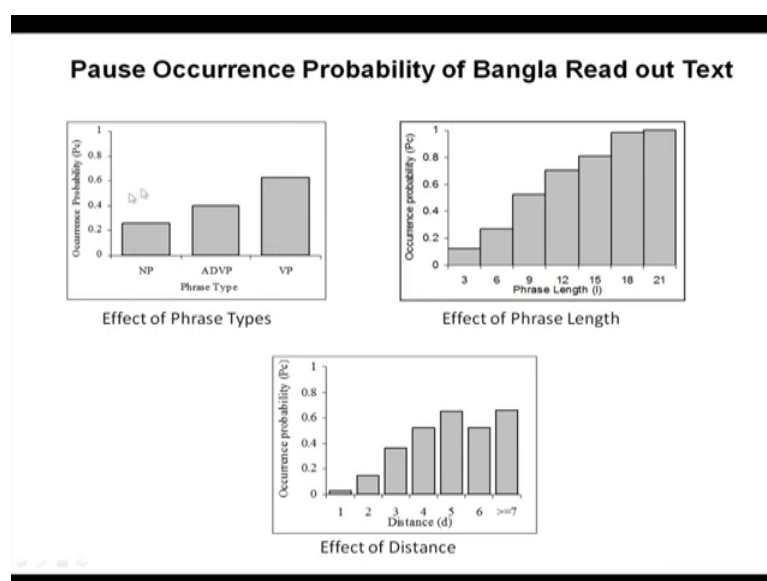
(Refer Slide Time: 21:10)



[FL]. So, it is a Bengali example. Now if you see this is W 1, W 2, W 3, W 4, W 5, W 6, W 7, W 8, W 9, W 10, now if you see we have said this words is related to directly related to [FL]. So, this sorry this word related to [FL] and this was related to [FL] this was related to [FL]. So, now, if you see this word is related to this word. So, distance is high. So, there is a high chance of probability of pause in here this boundary this word related to this word is very closely relation. So, I cannot say that this probability of pause in here will be very less then this word is related to this word. So, probability pause here is high then probability of pause here is less like that way we have define we have done that things you can read this paper the paper is available Bangla duration modeling you can read this paper.

So, what I use to say that depending on those 3 parameters you have done you have analyzed the what is the occurrence to W t of pause in a read out text of Bangla. So, Bangla you have collect read out sentences because we want the textual information must correlate with the spoken information. So, what we said we take that written text we can say that we have taken some sentences those sentences are read by some speakers and then we analyze the pause.

(Refer Slide Time: 23:05)



(Refer Slide Time: 23:12)

Modeling of Pause Occurrence Probability

Pause Occurrence Probability can be model using a linear Model like:

$$R_x = al + bd + c \quad (X : NP, ADVP, VP)$$

Model Parameters for Pause Occurrence Probability for Bangla

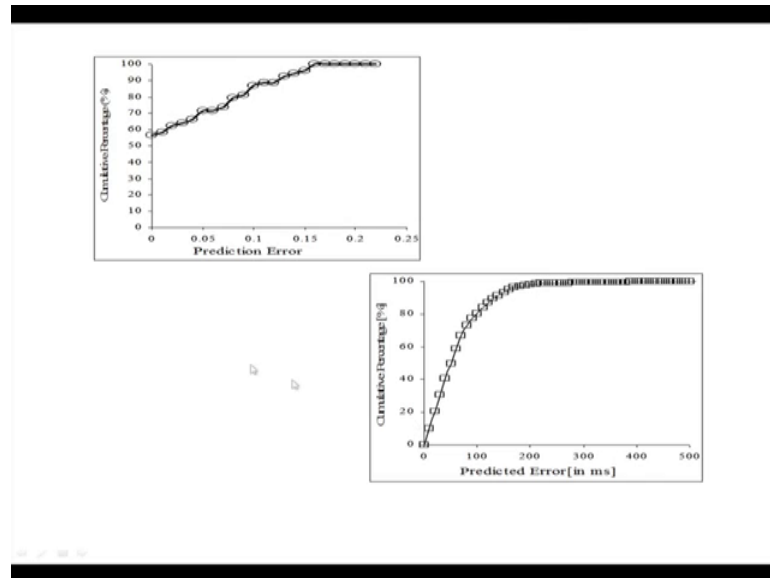
Phrase Type	Coefficients	Value	R square value
NP	a	0.077	0.902
	b	0.122	
	c	-0.436	
ADVP	a	0.072	0.813
	b	0.139	
	c	-0.486	
VP	a	0.078	0.791
	b	0.162	
	c	-0.467	

And then we try to correlate; what is the relations between those pause with this 3 parameters then we realize this 3 parameters then after analyzing we try to develop a linear model using these 3 factors so that you can predict the pause.

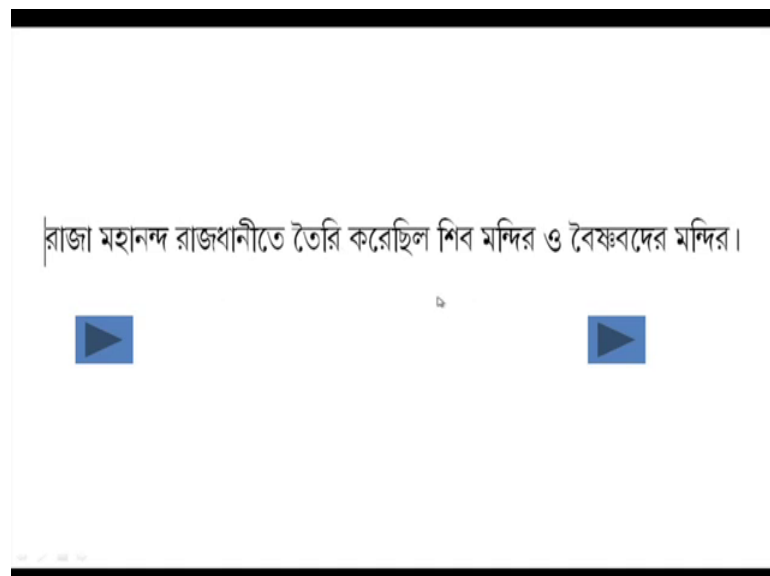
So, one is called occurrence probability prediction whether for a after this word after this phrase there will be a pause or not whether there will be a pause or not whether there will be a pause or not. So, that we probability we have calculated then we try to linear regression analysis we have done linear modeling we have done may be it is; it can be

improved by non-linear modeling also we have not done it, but you can do it non-linear modeling also then duration probability, we have done that. Find out the pause duration again linear modeling we have done and we have calculated something and we have calculated the cumulative percentage error for prediction error in the for prediction pause prediction pause for duration more prediction error we have calculated and we are reported in a papers.

(Refer Slide Time: 24:00)



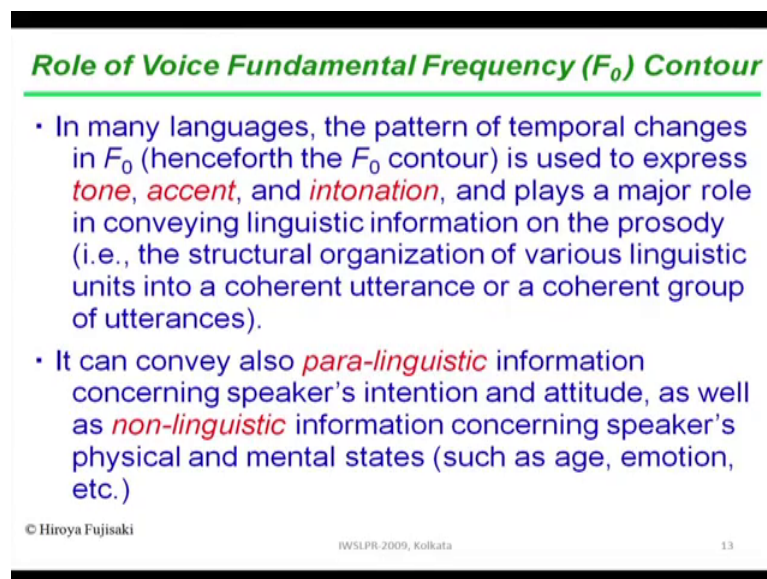
(Refer Slide Time: 24:14)



Now, if you see, there is an example in here also if you listen this example that this thing then you find that there are 2 kinds of sentence I have synthesized one is depending on the pause modeling one is just simple concatenation synthesis then you find that if I use the pause model it improves the clarity. So, I can say my pause model is now somewhat improved the synthesized page, but if you think in today in all kinds of TTS; TTS which is used in HTS or HMMs; TTS system this kind of modeling is not required because we have some the system lot of original data so; that means, within the corpus there is an occurrence of pause which is already taken care by the HMM model.

I will show you HMM base synthesis, this pause and duration are within the and it change the system that is model and also F_0 somewhat it is model, but F_0 also not model correctly I will show you that what is the problem they are having and how can it be overcome.

(Refer Slide Time: 25:38)



Role of Voice Fundamental Frequency (F_0) Contour

- In many languages, the pattern of temporal changes in F_0 (henceforth the F_0 contour) is used to express *tone*, *accent*, and *intonation*, and plays a major role in conveying linguistic information on the prosody (i.e., the structural organization of various linguistic units into a coherent utterance or a coherent group of utterances).
- It can convey also *para-linguistic* information concerning speaker's intention and attitude, as well as *non-linguistic* information concerning speaker's physical and mental states (such as age, emotion, etc.)

© Hiroya Fujisaki
IWSLPB-2009, Kolkata
13

So, this is pause now come to the F_0 which is very important parameter because if you see the F_0 even if pause and pause and duration if you model F_0 itself, then it can increase their clarity of the speech model. So, F_0 modeling is an important part of this prosody modeling.

So, I am discussing about the available this F_0 model and then DTLs; I will discuss about the Fujisaki F_0 modeling. So, next lecture I start that F_0 modeling.

Thank you.