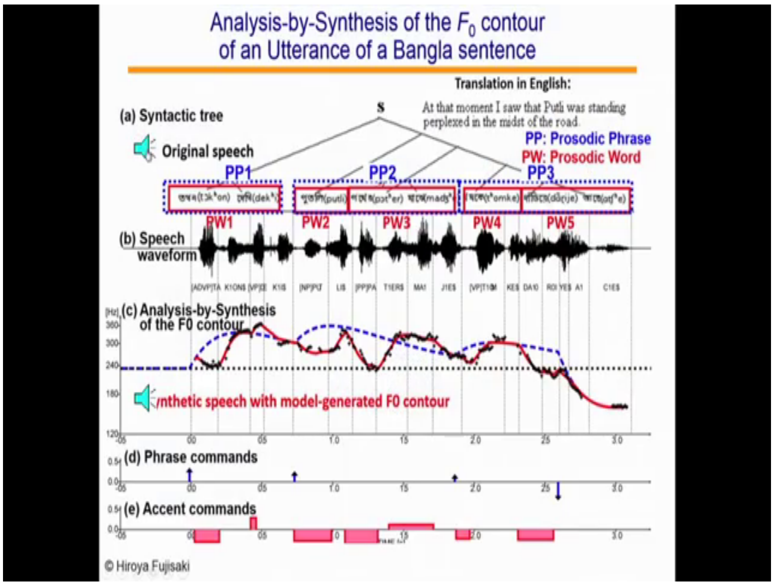
Digital Speech Processing
Prof. S. K. Das Mandal
Centre for Educational Technology
Indian Institute of Technology, Kharagpur

Lecture - 44
Fundamental Frequency Contour Modeling (Contd.)

So, as we discuss about the prosodic modeling, the Fujisaki modeling. We have tried it for Bangla language long back 2010, we have published the paper on Bengali f 0 contour modeling based on Fujisaki model in (Refer Time: 00:32) 2010 and based on that I just present to a one slides where I show you. This is the original speech if you listen that if you listening Bengali if you do not know Bengali then just try to listen the naturalness. [FL].

(Refer Slide Time: 00:48)



Analysis-by-Synthesis of the F_0 contour of an Utterance of a Bangla sentence

Translation in English:

At that moment I saw that Puti was standing perplexed in the midst of the road.

PP: Prosodic Phrase

PW: Prosodic Word

© Hiroya Fujisaki

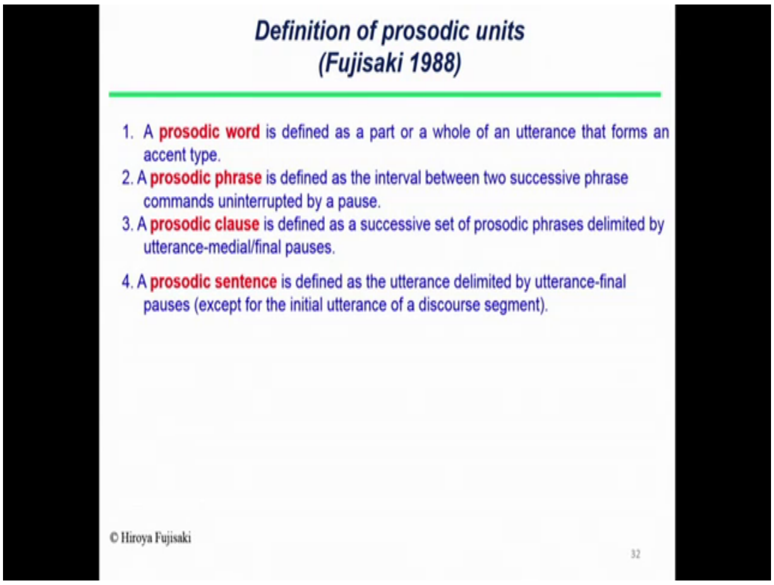
So, there is a Bengali words, Bengali sentence is which contain that [FL]. So, that sentence the original sentence. So, I playing the original sentence. Then what do we have done we have try to find out whether that our generation contour that analysis by synthesis model is correct or not. What do we have done we have extracted the f 0 contour which is represented by the black spot here there is a black mark is there.

So, those represented the original f 0 contour. And if you see there is a blue line once I put the phrase command then the blue line is come, which is the phrase command. And if you see the black line straight line is the base line, and after putting the accent command, you see the red line in generated. So, this red line is actually synthesis f 0 contour. If I synthesis f 0 contour is embedded on the original speech, if you see how the synthesis speech is like this [FL].

So, there are no difference between the synthesis contour and the original contour. So, I can say that quality wise that it is possible to analyze and then synthesis using the f 0 modeling. So, this we have tried for Bengali in many reason and then what we have done that we have developed Bengali the Bengali the thus system based on that STS model and then try to do the f 0 modification using prosodic model. So, I will describe that things in this lectures because that will open up some research in this creative on your own language.

Now, if we see depending on that prosodic model, Fujisaki has the define certain prosodic you can say the prosodic unit definition.

(Refer Slide Time: 02:48)



Definition of prosodic units
(Fujisaki 1988)

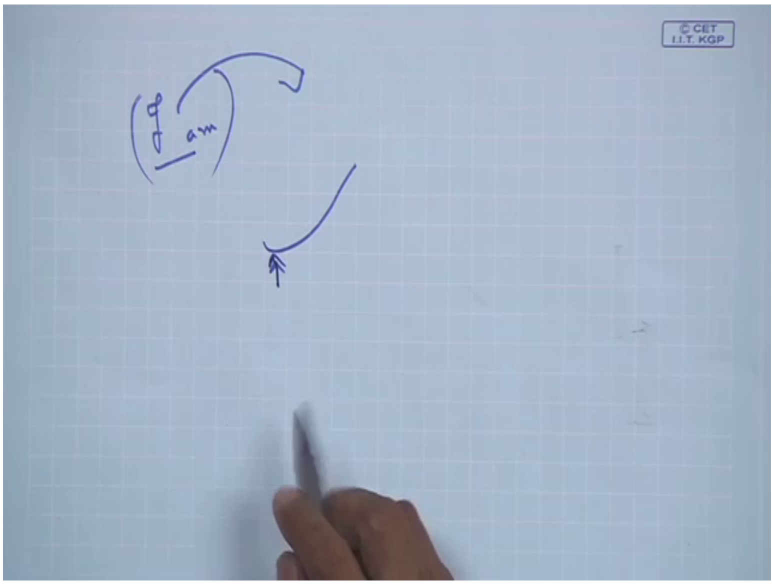
1. A **prosodic word** is defined as a part or a whole of an utterance that forms an accent type.
2. A **prosodic phrase** is defined as the interval between two successive phrase commands uninterrupted by a pause.
3. A **prosodic clause** is defined as a successive set of prosodic phrases delimited by utterance-medial/final pauses.
4. A **prosodic sentence** is defined as the utterance delimited by utterance-final pauses (except for the initial utterance of a discourse segment).

© Hiroya Fujisaki

32

This is basically taken from the Fujisaki definition. He said the written words and the spoken words are different. Because as I said already that in spontaneous speech or in a spoken language there is no specific word boundary, means if there is a sentence let us there sentence I am or I will go to Calcutta tomorrow. So, you say that every I then there is a gap then there is a am then there is a gap. So, in written language this gap identified that what boundary.

(Refer Slide Time: 03:29)



Ok, but in case of spoken language if I say I am then there is a no gap between I and am. There is co-articulation effect of I and am. So, I cannot say this boundary is exist in spoken language, but yes, that is a certain word boundary exist in the written spoken language which is identifiable by perception. If you see if I perceive the origin of speech of Bengali sound this one [FL] if you see that boundary are clearly indicated, but it is not as per the lexical word boundary. So, then Fujisaki defined that prosodic unit, first one is called prosodic word. So, instead of written words we say prosodic word, is define as a part of whole of an utterance that form an accent type.

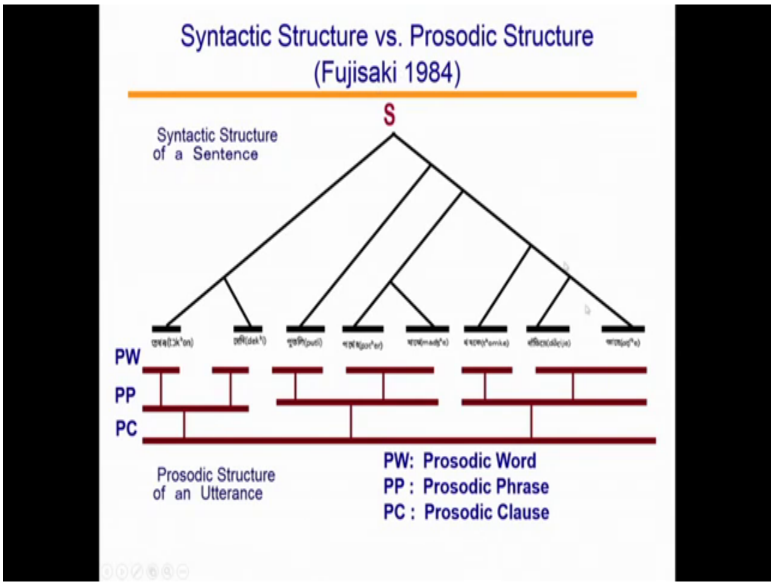
If you see here, one accent type if you see [FL] is from accent type. So, the up to this second end of this [FL] is one word which is call p w one prosodic word one. Second accent type third accent type [FL] is p w 2 [FL] one accent type then [FL] one accent type [FL] is another accent type. So, he define the prosodic word is defined as a part of whole utterance that form an accent type is call prosodic word.

In Bengali we found that the every accent or every prosodic word begin with a negative accent. If you say we said that the Bengali is a bound stress language; that means, at the beginning of every prosodic word f 0 contour is rising if you see the f 0 contour in black line beginning of every prosodic word f 0 contour is rising. So, it is not that prosodic word is consist of a single written word it may be several written word may be form a prosodic word. So, prosodic word is defined in spoken language not as per the written language. So, 2 x one accent type is defined as a prosodic word. Then he define the prosodic phrase is defined as an interval between the 2 successive prosodic phrase command.

If you see this one this blue line up to here up to here up to this line is a single phrase one. So, this is the phrase one. So, this is call phrase prosodic phrase, then this is the second part blue line phrase 2 prosodic phrase 2 and last one is prosodic phrase 3. So, every phrase command it is the beginning of the phrase. So, at this phrase command is end of the first phrase and beginning of the second phrase. So, if you see actual boundary may be leading; that means, see that here is an muscle control things.

So, the effect will the command is executed then effect will come certain after certain time interval. That is why if the actual f 0 rising is here command may be executed in here it takes some time to reach this effect. So, that is why if you see it maybe some leading time. So, this is call prosodic phrase then he define prosodic clause and prosodic utterances. Now if I that is per this definition, if I just draw the prosodic utterance structures.

(Refer Slide Time: 07:36)



Syntactic Structure vs. Prosodic Structure
(Fujisaki 1984)

S
Syntactic Structure of a Sentence
PW
PP
PC
Prosodic Structure of an Utterance

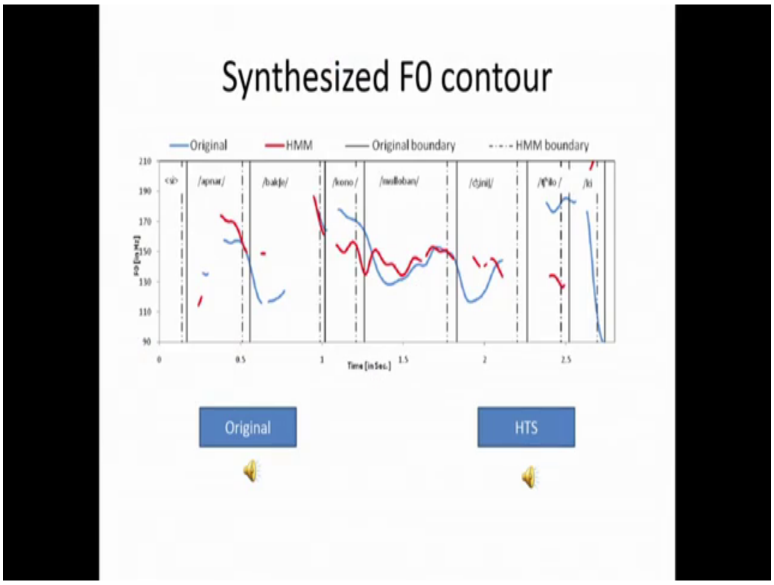
PW: Prosodic Word
PP : Prosodic Phrase
PC : Prosodic Clause

Suppose this is in prosodic utterance [FL] these an prosodic utterance. Then I can say if I draw the syntactic graph syntactic tree of this sentence which is nothing but a depending on the word the how far this is a distance. Then you see the prosodic word p w is defined like this [FL] is a prosodic word. So, there is a some mistake. So, p w [FL] is a prosodic word p w will be connected.

This will be connected. So, this is single prosodic word [FL]. So, there is a mistake this gap will be not there then [FL] second prosodic word [FL] third prosodic word [FL] 4th prosodic word [FL] fifth one if you see fifth one. So, both are the prosodic words. So, there is a mistake in here this will be one prosodic word then prosodic phrase. It may coincide with prosodic word, but there is a prosodic phrase. So, this prosodic phrase is defined like again by that phrase command and then prosodic clause. So, this is call prosodic structured.

Which may differ it may not coincide with one to one corresponding to the syntactic structures. So, based on this idea when you develop the HTS engine, I am not discuss in this by the way already we said again.

(Refer Slide Time: 09:05)

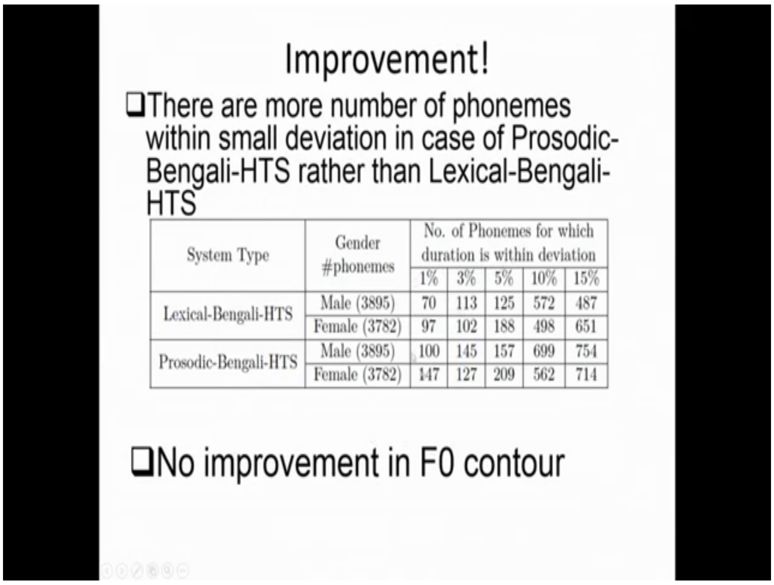

Synthesized F0 contour

So, when we developing our bengaline HTS the thus what we have done, first we find out that develop the Bengali HTS. Then if you see this blue line is indicate that original f 0. Red line is indicate the HMM synthesized means without doing any f 0 modification. I have trend the HTS system and I synthesized the sentence, then red line is that synthesized sentence f 0 contour. And if you see that solid line this is the original boundary dotted line is the HMM pit, what boundary?

Those are the what boundary. So, there is a I am not showing that phoneme alignment every phoneme alignment I can show. So, that phoneme alignment, well I am not showing here if you listen the original speech [FL] if you see the just HTS after training I generate the sentence [FL]. Then we thought why not we train this HTS same system instead of leveling at syntactic origin over boundary I should train it with the prosodic word prosodic structures. So, spoken corpora is level instead of syntactic structure.

So, what boundary are all boundary are marked this is the word is based on the prosodic word based on the prosodic phrase and based on the prosodic clause, then we see that is the word error rate is reduce.

(Refer Slide Time: 10:54)



Improvement!

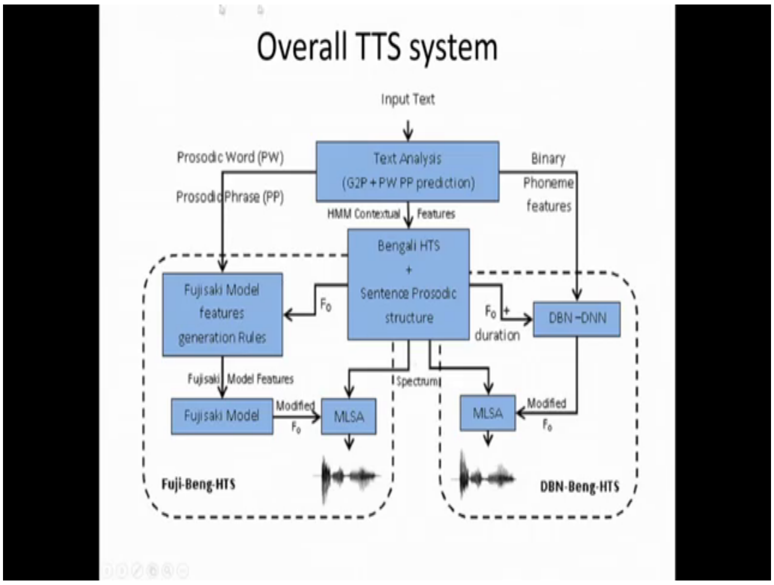
- ❑There are more number of phonemes within small deviation in case of Prosodic-Bengali-HTS rather than Lexical-Bengali-HTS

System Type	Gender #phonemes	No. of Phonemes for which duration is within deviation				
		1%	3%	5%	10%	15%
Lexical-Bengali-HTS	Male (3895)	70	113	125	572	487
	Female (3782)	97	102	188	498	651
Prosodic-Bengali-HTS	Male (3895)	100	145	157	699	754
	Female (3782)	147	127	209	562	714

- ❑No improvement in F0 contour

So, result is there in the paper, if you see the there is a journal paper on HMM base Bengali speech synthesis system based on the paper is there you can search that paper. Or you can say the Bengali HTS based the thus that our paper will come you can read that paper.

(Refer Slide Time: 11:12)



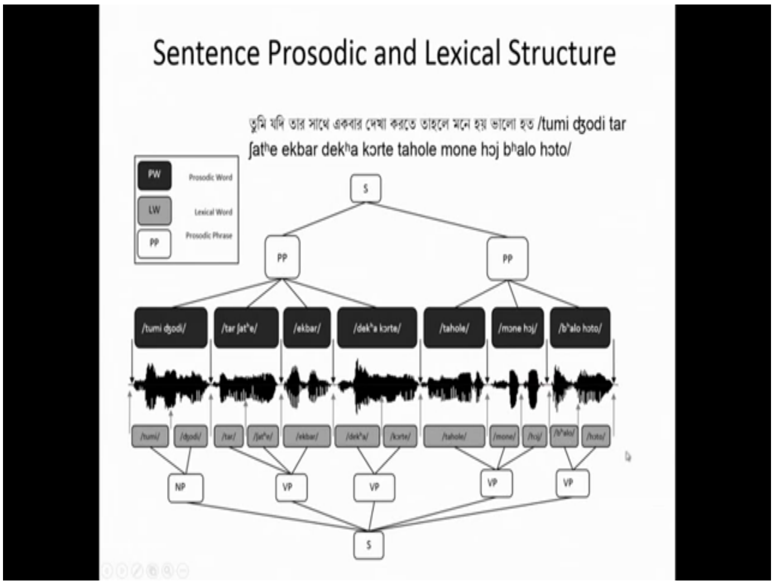
Overall TTS system

But if you see our proposal is like that way that we take the input text which is level input text and that is that the speech which is level based on the prosodic structure instead of syntactic structure. Then we have done this thing that HTS part. Then we said since f 0 if you see the HTS engine is nothing but a equator one is call segmental information another is call supra segmental information which is nothing but a f 0. So, what we said we extract that f 0 contour form HMM TTS and modify as for the language requirement. So, f 0 contour which is generated by train HTS system is nothing but a f 0 contour which is best match based on the HMM modeling.

But then that may not be coincide with the required f 0 contour targeted f 0 contour which is detect which is defined by the structured of the sentence, or syntactic structures of the sentence. So, what we done we developed a f 0 contour modeling based on Fujisaki model, and another one we have try with a DBN network DBN DNN methods. So, that method also we try, so based on that we modify the f 0. So, this f 0 is come and from that input text we analyze the prosodic word and prosodic phrase, and then we modify which generate the f 0 contour and supply to the decoder.

So, f 0 is modified. So, duration information and that segmental information is come from the syntactic you can say HTS engine. So, we are not doing anything on the duration and syntactic. So, duration is based on the prosodic structure training whatever we get using HMM modeling, but f 0 is modified based on the 2 model we have done this work and I can show you the this is the prosodic structure I am not detailed discussion.

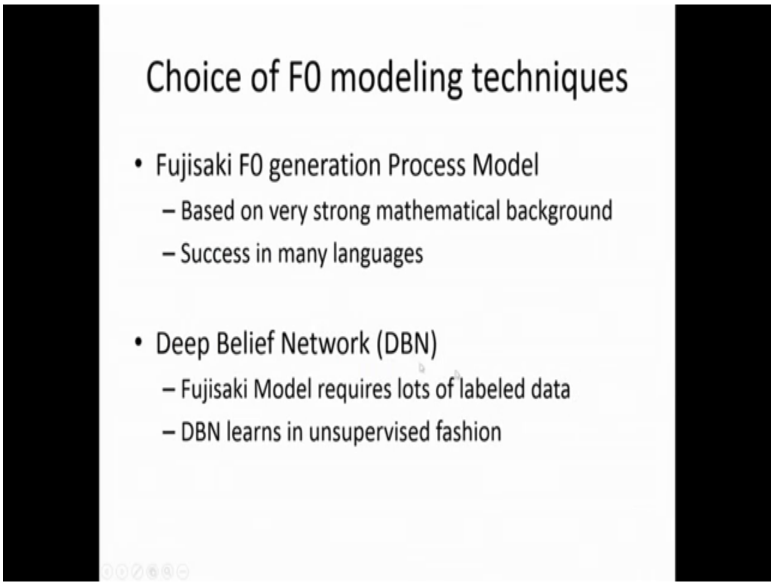
(Refer Slide Time: 13:16)



Sentence Prosodic and Lexical Structure

তুমি যদি তার সাথে একবার দেখা করতে তাহলে মনে হয় ভালো হত /tumi ʤodi tar ʃatʰe ekbar dekʰa kɔrte tahole mone hɔj bʰalo hɔto/

(Refer Slide Time: 13:19)



Choice of F0 modeling techniques

- Fujisaki F0 generation Process Model
  - Based on very strong mathematical background
  - Success in many languages
- Deep Belief Network (DBN)
  - Fujisaki Model requires lots of labeled data
  - DBN learns in unsupervised fashion

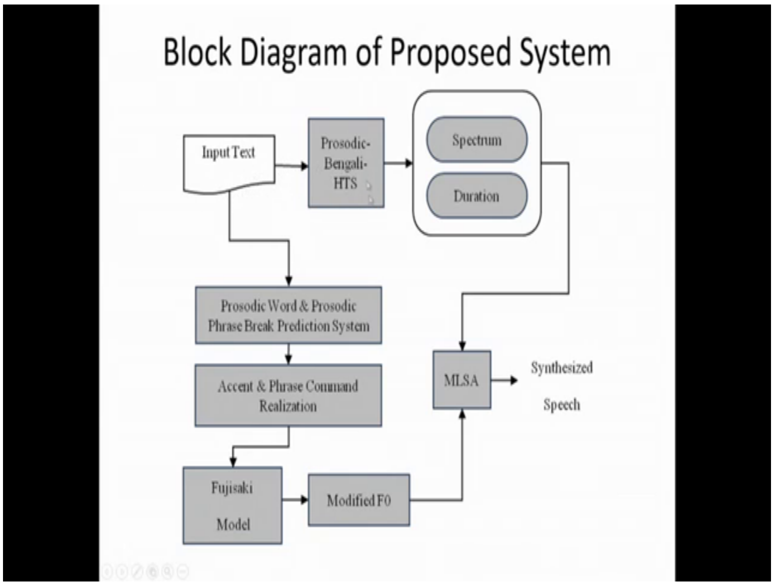
So, I will show you that for there is a 2 kind of modeling we have done one is called Fujisaki f 0 generation process model another is called deep belief network base, because where the problem is Fujisaki is that Fujisaki model is successful model and it is a generation process model. So, it is very good.

But only problem in this model this requires a lot of level data to find out the rule for that a generation process model. If this kind of sentence come this is a structure of the sentence then based on that structure or from the rule what should be the height of the action command phrase command all kind of things I have to define. So, that require some level data for training, but DNN require un DBN is learn in unsupervised fashion it not that much of level data is required. So, it tried with that also, I will show you the result in the end.

(Refer Slide Time: 14:12)

F0 Modeling using
Fujisaki F0 Generation Process
Model (Fujisaki Model)

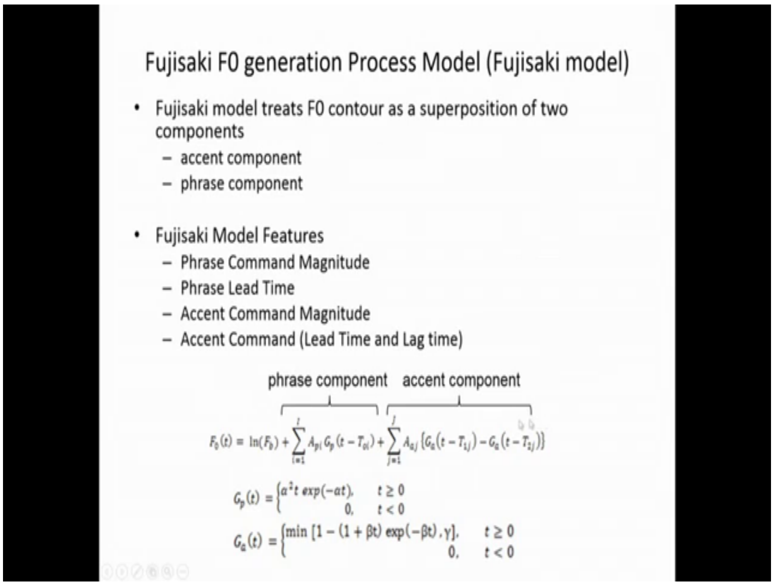
(Refer Slide Time: 14:18)



Block Diagram of Proposed System

Then f 0 model Fujisaki generation process model we have done. So, what I have say I have already explain that spectrum information and duration information with training from the HTS engine. And with stuck that f 0 contour and modify the f 0 contour based on the accent command or phrase command by the Fujisaki model and then we synthesize the speech.

(Refer Slide Time: 14:36)



Fujisaki F0 generation Process Model (Fujisaki model)

- Fujisaki model treats F0 contour as a superposition of two components
  - accent component
  - phrase component
- Fujisaki Model Features
  - Phrase Command Magnitude
  - Phrase Lead Time
  - Accent Command Magnitude
  - Accent Command (Lead Time and Lag time)

$$F_0(t) = \ln(F_b) + \overbrace{\sum_{i=1}^{I} A_{pi} G_p(t - T_{0i})}^{\text{phrase component}} + \overbrace{\sum_{j=1}^{J} A_{aj} \{G_a(t - T_{1j}) - G_a(t - T_{2j})\}}^{\text{accent component}}$$

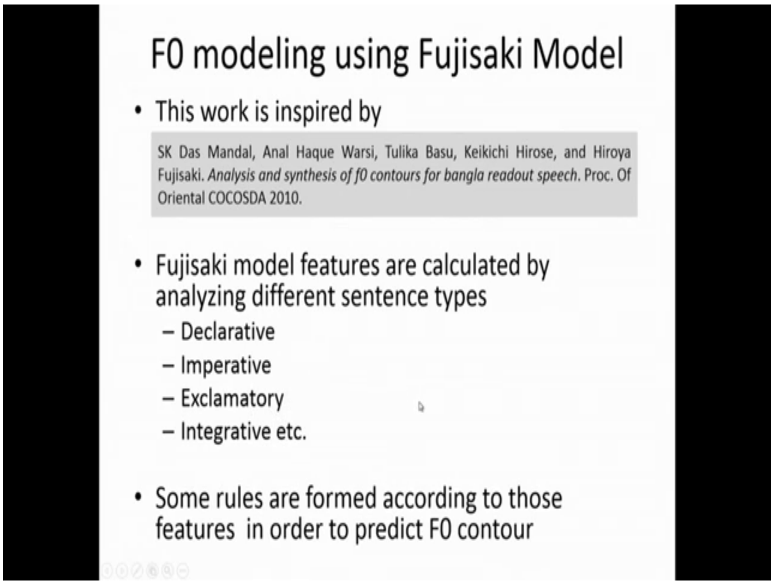
$$G_p(t) = \begin{cases} \alpha^2 t\, exp(-\alpha t), & t \geq 0 \\ 0, & t < 0 \end{cases}$$

$$G_a(t) = \begin{cases} \min\,[1 - (1 + \beta t)\exp(-\beta t), \gamma], & t \geq 0 \\ 0, & t < 0 \end{cases}$$

So, in that case Fujisaki generation process model, so accent component and phrase component rule has to be generated. So, what is there you know that if this is the Fujisaki equation, you see the phrase component this is the accent component. So, phrase component magnitude and phrase lead time. If the original phrase boundary is here while it is started. So, that lead time is required ok.

So, that lead time and then phrase accent command magnitude, and accent command lead and lag time what is the accent command. So, if you see in here you see there is a phrase command there is a timing occurrence of the phrase command timing, and if you see the accent command there is a beginning and end command beginning and command end. So, we generate some rule for that on some paper is published based on that and then we generate those rule using that Fujisaki model and using those rule We synthesize that generate the f 0 contour.

(Refer Slide Time: 15:31)



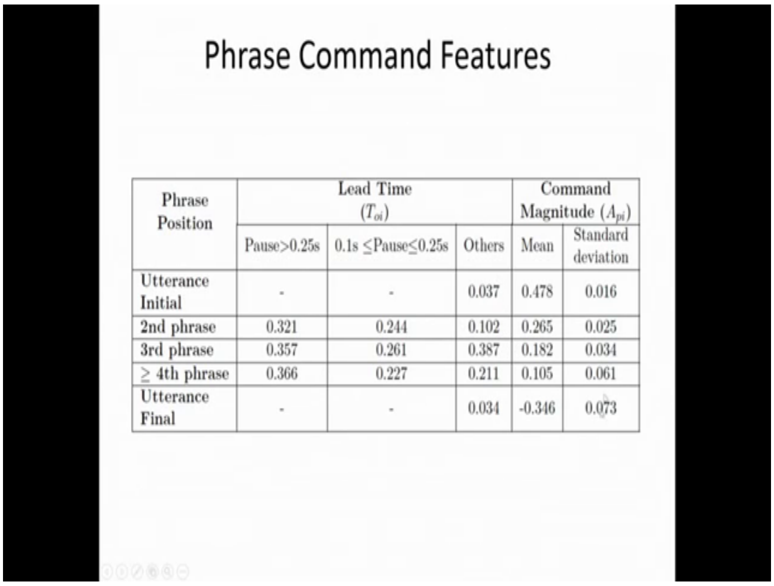
F0 modeling using Fujisaki Model

- This work is inspired by

SK Das Mandal, Anal Haque Warsi, Tulika Basu, Keikichi Hirose, and Hiroya Fujisaki. *Analysis and synthesis of f0 contours for bangla readout speech*. Proc. Of Oriental COCOSDA 2010.

- Fujisaki model features are calculated by analyzing different sentence types
  - Declarative
  - Imperative
  - Exclamatory
  - Integrative etc.
- Some rules are formed according to those features in order to predict F0 contour

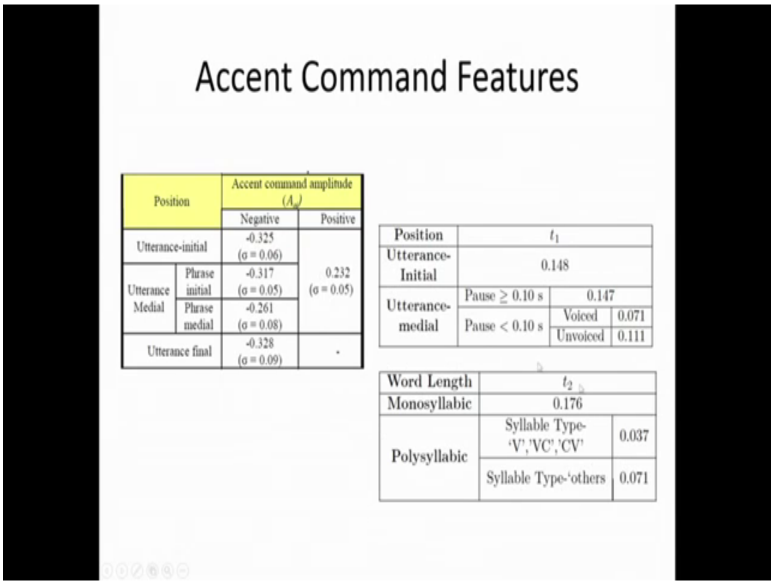
(Refer Slide Time: 15:33)



Phrase Command Features

Phrase Position	Lead Time (T_{0i})			Command Magnitude (A_{pi})	
	Pause>0.25s	0.1s ≤Pause≤0.25s	Others	Mean	Standard deviation
Utterance Initial	-	-	0.037	0.478	0.016
2nd phrase	0.321	0.244	0.102	0.265	0.025
3rd phrase	0.357	0.261	0.387	0.182	0.034
≥ 4th phrase	0.366	0.227	0.211	0.105	0.061
Utterance Final	-	-	0.034	-0.346	0.073

(Refer Slide Time: 15:35)



Accent Command Features

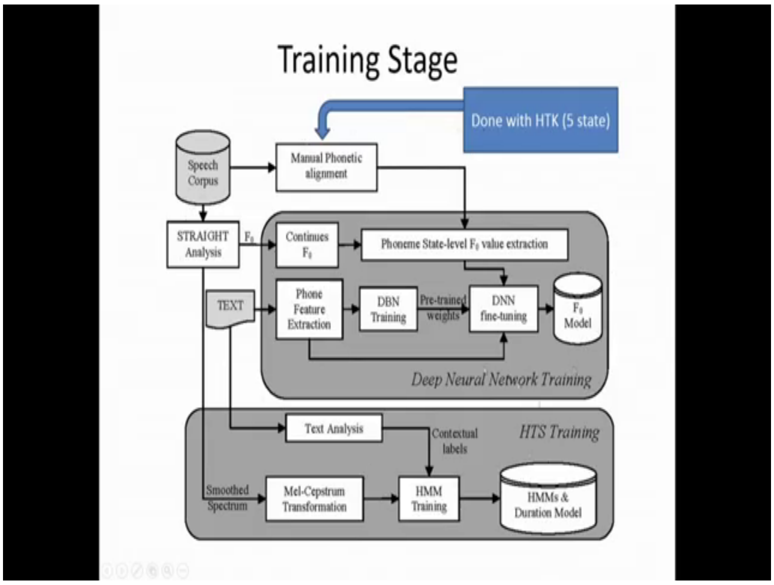
Position		Accent command amplitude (A_a)	
		Negative	Positive
Utterance-initial		-0.325 (σ = 0.06)	0.232 (σ = 0.05)
Utterance Medial	Phrase initial	-0.317 (σ = 0.05)	
	Phrase medial	-0.261 (σ = 0.08)	
Utterance final		-0.328 (σ = 0.09)	-

Position	t_1		
Utterance-Initial	0.148		
Utterance-medial	Pause ≥ 0.10 s	0.147	
	Pause < 0.10 s	Voiced	0.071
		Unvoiced	0.111

Word Length	t_2	
Monosyllabic	0.176	
Polysyllabic	Syllable Type-'V','VC','CV'	0.037
	Syllable Type-'others	0.071

So, then we synthesize the speech I will show you the result in then later on.

(Refer Slide Time: 15:44)

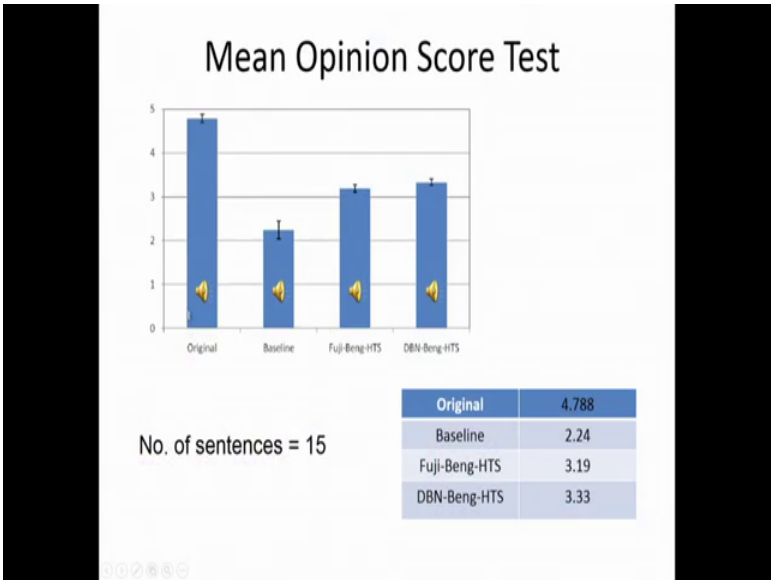


Training Stage

Done with HTK (5 state)

And DBN instead of Fujisaki generation process model. We try to find out that f 0 contour modeling using deep neural network training. So, I am not going details on DBN, because DBN is the topic deep neural network is a topic written this. So, I am not going details of the deep neural network training and all kinds of things you can go through the paper because that is not also that much require for this course. So, I am not going that. So, if you see that we have done and if you see the result this is the result.

(Refer Slide Time: 16:19)



Mean Opinion Score Test

No. of sentences = 15

Original	4.788
Baseline	2.24
Fuji-Beng-HTS	3.19
DBN-Beng-HTS	3.33

This is the original sentence [FL] this is the best line means where only HTS is train based on the prosodic structures not syntactic structure prosodic structure [FL] ok.

Now, I modify the f 0 based on the sentence structure using Fujisaki model. [FL] now based on the DBN. [FL] both of the cases it is shown that Fujisaki model and DBN is almost same result is found, but Fujisaki model is generalized model. So, even if the there is a success rate is high for even if the sentence type is completely unknown for this training DBN. So, in that case also because the rule is exist. So, it is better, but both models are comparable and this is already published paper. So, you can try on your own language using Fujisaki model.

So, what I said that 2 (Refer Time: 17:29) here we have done. One is called train the HTS engine with prosodic structures, and modify the f 0 contour based on the Fujisaki model or DNN model. So, if you remember I have explain one another application that accent conversion. So, my proposal is that can I not modify the f 0 contour and segmental information based on that target language accent, using some DNN or deep neural network or some rule base method. So, we have not yet done that we have started that work. So, it is possible that accent conversion can be done using because it is ultra after all HTS synthesis nothing but equator. So, using the same principle we can do the accent conversion, so doing that research on this area.

So, this is the end of that prosodic modeling using Fujisaki model. If you have a specific query or if you interested to do pursue the research working your own language then you contact me I will I will give you whatever help is possible for me I can give you the help. And I think most of the Indian language that there is a lot of TTS is available in Indian language, but yes there is a problem in prosody modeling. Because prosody modeling not only require that generation of the f 0 model, but also language analysis is very important parameter. Because if you see in this our this research also include that (Refer Time: 19:08) speech tagging and all kinds of things are there. Because of you until unless you find out because the in TTS input is only the text.

So, I do not know the f 0 contour. So, from the text I have to generate the f 0 contour. Yes, I can text some trading from some sentences, but if you see that syntactic structures of the depending on the sentence specific prosodic can be given. So, that kind of things is required because I have to I have to I have to you can say I can parameterize that synthesize or you can say target f 0 contour best. On some correlation between the linguistic parameters and acoustics parameter has to be trained. So, that requires that extraction of that linguistic parameter is very important. So, I have to extract those linguistic parameter, which parameter has an correlation with the f 0 contour I have to find out that is while established you can see that this syntactic structures of the sentence is important. And if you see that pause modeling although in HTS engine cases pause modeling is not separate it is within the training there, but yes we can definitely train the system better way if we know when the pause will occur how the pause is related with syntactic structures.

So, that information can be in cooperate in HTS training and that can generate better result. So, this is the last lecture in prosody modeling. So, this will be the my last lecture in speech this course, but after reviving all the lectures whatever I have given in this as things up to 7th week you can say those are related to or up to 6th week is related to signal processing. So, only one things I have missed which is called GMM Gaussian mixture modeling I have not covered, but yes I will cover one lecture I may uploaded on GMM the basic GMM Gaussian mixture modeling vector (Refer Time: 21:15) I have trust upon, but I am not detail. As I beginning I said these course is not contain that e I part or sub computing part. I am purely concentrated on signal processing part and see only the not signal processing last 2 week I said something about that modern application in speech research, which can you taken care and you can pursue that speech research or you can So that you can technology is developed. So, that is why I touches this prosody modeling as important issue today.

So, you know that that there is a international conference on only prosody speech prosody. So, speech prosody there is prosody modeling action conversion and emotion is nothing but the part of a prosody. So, emotion recognition emotion synthesis all are nothing but the prosody modeling. So, if you know the prosody modeling then work on emotion area will be very easy. So, that is why I touch the prosody modeling part. Yes, I can include 1 or 2 lectures on GMM. So, that it can give you some over view what is Gaussian mixture model why we have use that and that hmm, I have said that hmm is nothing but ai algorithm you can find lot of lectures on HMM modeling. If you want then I can if you particular query if you have a particular query you can directly contact me so that I can explain that things whatever I know.

Thank you very much.