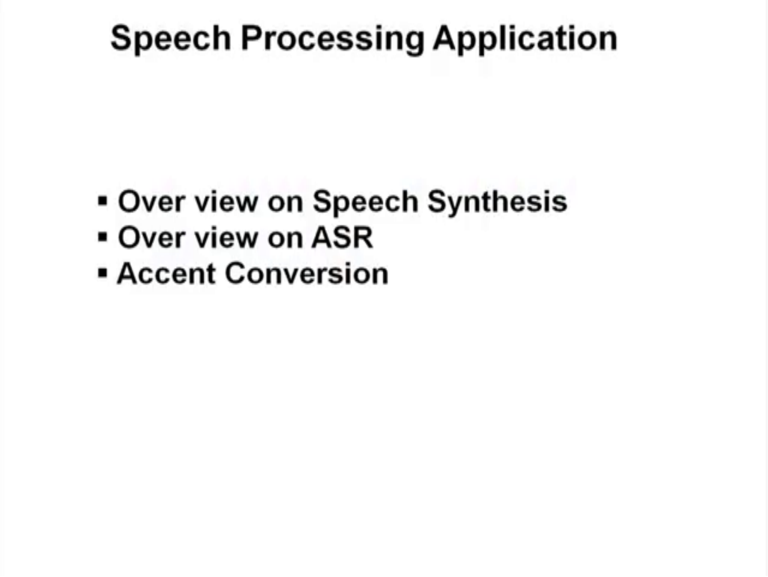
Digital Speech Processing
Prof. S. K. Das Mandal
Centre for Educational Technology
Indian Institute of Technology, Kharagpur

Lecture - 37
Text To Speech Synthesis

So, let start that next week, the new week which is we talk about the speech processing application.

(Refer Slide Time: 00:24)



Speech Processing Application

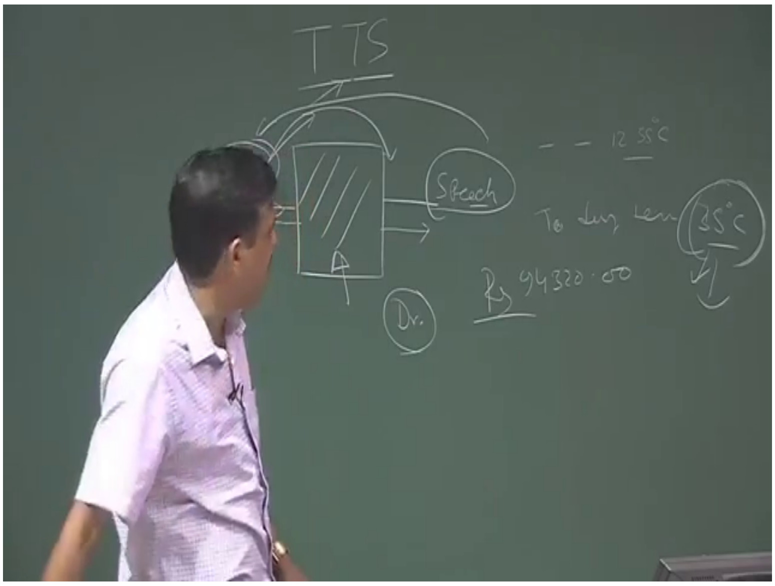
- Over view on Speech Synthesis
- Over view on ASR
- Accent Conversion

Mainly I have chosen three application, but there are lot of applications. So, if time permit in last week I may cover this speaker and that identification that this speaker recognition and identification that part also I covered. So, little bit of GMM I will cover in that topic, and vector quotation I will cover. But today let us talk about that this kind of application like speech synthesis and ASR, I am not detail cover that ASR because it itself can take several classes. So, first I cover that speech synthesis application that how do we develop the speech synthesis application what are the provision closen everything and then ASR we give out that (Refer Time: 01:06) train in ASR, and then I talk about the action conversion which is my dream in research this area. So, lot of research work is going on in the speech action conversion and speech conversion kind of things.

So, we will discuss about all those things in this total week.

So, let us start talk about the speech synthesis popularly known as you know that TTS - Text to Speech Synthesis System, TTS.

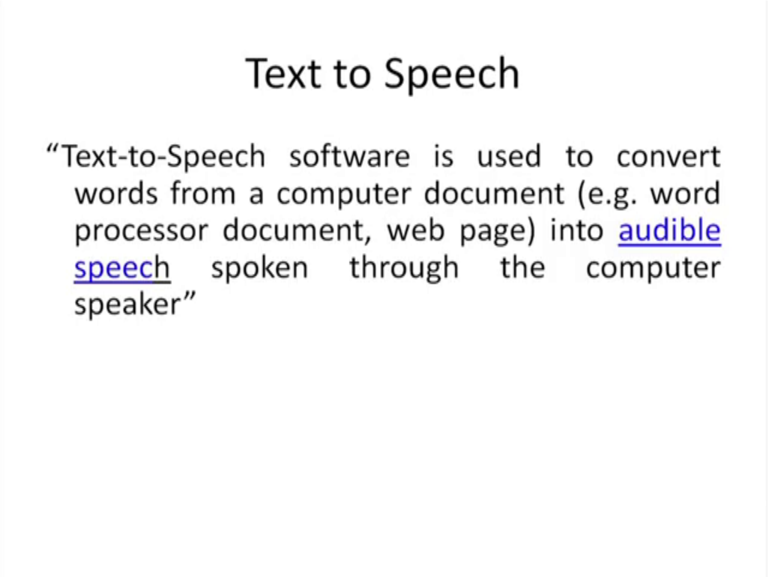
(Refer Slide Time: 01:28)



Now, if I see the name text to speech. So, I can say that I required the machine let us there is a machine or there is a systems, where that system will take input as a text and then output is produce as a speech. Now if I remember, if you remember that speech production mechanism how the speech is produce that message planning, that coding and then execution by vocal tracks and produce the speech. So, speech production this part which convert the text to speech means input text to speech, not only with this production model, but also other text processing model is also important because when we say something it is not just segmental information, it contain the supra segmental information also.

So, if I see the basic block diagram of the speech synthesis, then I can say that text make contain some. So, what contain information in the text, some linguistic information, non linguistic information parallel linguistic information. So, there is a linguistic information, there is parallel linguistic, there is a non linguistic information all those information have processed in our mind and executed by this vocal track system to produce the speech. So, the speech signal if I say it contain all the linguistic information, non-linguistic information and parallel linguistic information all those three are there in this speech signal. Now if I want that a system or a machine or an algorithm should develop such that it will take input text and covered to speech. So, that all kinds of text processing and that signal processing are require in case of development of TTS.

(Refer Slide Time: 03:44)



Text to Speech

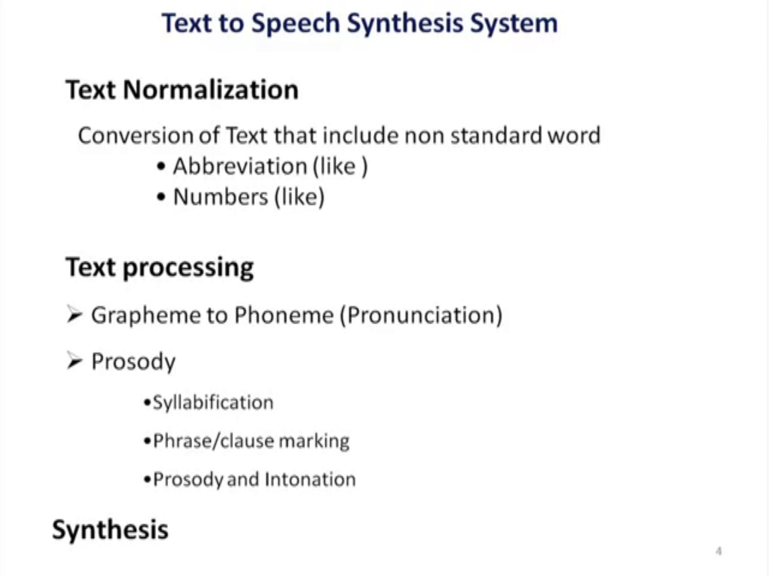
"Text-to-Speech software is used to convert words from a computer document (e.g. word processor document, web page) into audible speech spoken through the computer speaker"

So, if you see in the slides if I show you the text is definition of text to speech; text to speech software is used to convert words or sentence from a computer documents into audible speech, spoken through the computer speakers.

So, TTS is nothing, but a input is text and output is page. Application if I have a good TTS think about that it is provides the hand free operation, it remove the digital divide. Suppose I cannot there is a lot of people are there they cannot read the text, but they can speak, because speech is you can say the spoken language there is a use for the communication. The person who does not know anything about that written text he can also speak if you go to the village area lot of people they cannot read, but they can speak. So, what I want that suppose I have an information stores or text information stores in computer you see the many forms, many graduate, newspapers all are textual information if I want to disseminate that information to those kind of people how cannot read then the TTS is must; that means, machine can read the text, and machine can talk because they understand the spoken version of the language. So, there are comfortable with the spoken version.

So, that is called text to speech synthesis and their application.

(Refer Slide Time: 05:18)



Text to Speech Synthesis System

Text Normalization

Conversion of Text that include non standard word

- Abbreviation (like)
- Numbers (like)

Text processing

➢ Grapheme to Phoneme (Pronunciation)

➢ Prosody

- Syllabification
- Phrase/clause marking
- Prosody and Intonation

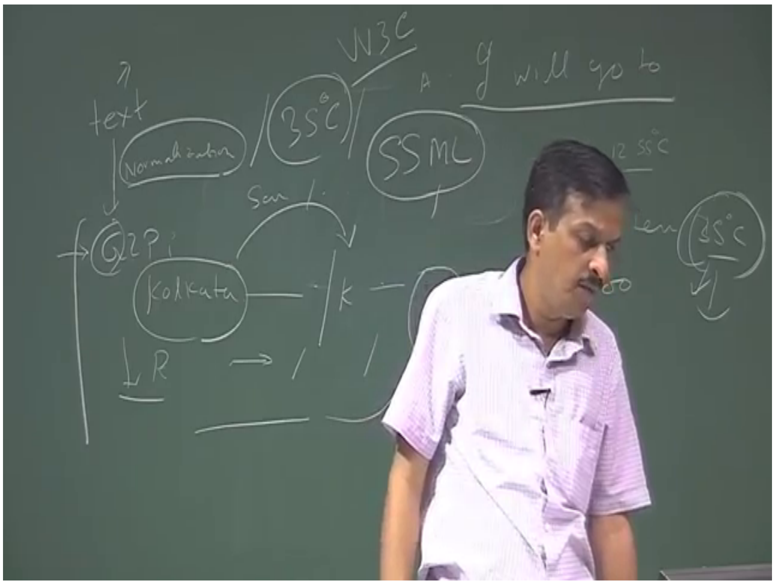
Synthesis

4

Now for typically text to speech synthesis system there, the first block is called text normalization. If you see if I say open any page whether it is English, whether it is Bengali, whether it is Hindi, whether it is Marathi if you read any text any text messages there are lot of normalization is require. Suppose there is a some text contain text. So, text is running and there is a number 2.35degree centigrade, today's temperature is 22.5 or two days temperature are is written 35 degree 35 degree centigrade when you speak you never say 35 degree. So, should 35 degree centigrade. So, if you see this information if I say translate in text then it is a 35 degree centigrade. So, those words, instead of their numerical 35, I want to convert 35 degree centigrade similarly suppose I have a phone number 94320. So, see that here we never said 9, so 1, 2, 3.

So, 9 9 lakh 4320 never said we said 9430 like that if it is phone number, but if it is rupees the we said rupees then I say 9 lakh 430. So, lot of text conversion or text normalization is required. Another wise aberration if you say if I say if I write doctor then it has to be pronounced as a doctor. So, there is a lot of short form of also when we pronounced we complete form we pronounced. So, there is a lot of text normalization is required unless it will not on the TTS will produce, but TTS does not know how to produce this. So, this as to converted in normal text format. So, that is called text normalization. So, if I write the TTS. So, there is a input block is text.

(Refer Slide Time: 07:42)



Text may contain aberration number or two suppose 2 by 3 (Refer Time: 07:50) then it will pronounce as 2 by 3.

So, all kind of number aberration will be there in the text and that text has it will be normalized first; normalization. So, that is a TTS this is called text normalization, text as to be normalized first. Once it is normalized then this goes to text processing unit how it is process? Because you know text contain some words. So, suppose there is words and written form of the word and pronunciation form of the word there is a change. Suppose if you write phycology the written form and pronunciation form is different. So, I can say there is a requirement text processing which can convert this is called grapheme to phoneme or it will sometime it will grapheme G to P what is grapheme? You know that if I say alphabet a b c d all are grapheme. So, the word written in form of grapheme that has to be convert to from of pronunciation string IPS string. I said in that first week that I can convert my name in IPS string that depended on the pronunciation.

So, if I write let us write Kolkata, it is a grapheme information is has to be convert to its pronunciation information which is Kolkata this as to be convert. So, this is call grapheme to phoneme conversion. If I write phycology pc like that way it as to be pronounce by written as phycology in pronunciation format which are the phoneme name you know the number of if I pronounce psychology. So, saw is the first phoneme name. So, I write saw if it is palatal saw I write that symbol if it is dental saw I write that symbol, so that way all IPS string. So, pronunciation string as to be represented by a IPS form. So, I get I pronunciation form of the written word. So, that is call grapheme to phoneme conversion details I will come then there is a call prosody if you see if I say text if I say I tomorrow I will go to Calcutta, then is a prosody is involve. So, prosody which contain the melody of the spoken form, you can say the melody of the spoken form I never say I will go like not never said I said I will go to Calcutta.

So, there is a lot of variation in supra segmental parameter. So, those supra segmental parameter is not arbitrary, it depends on the synthesis structure of the input text. So, even supra segment parameter like pause duration f 0 and intensity, all are varied depending on the synthesis structure of the language. So, last week I will deal about the prosody modelling.

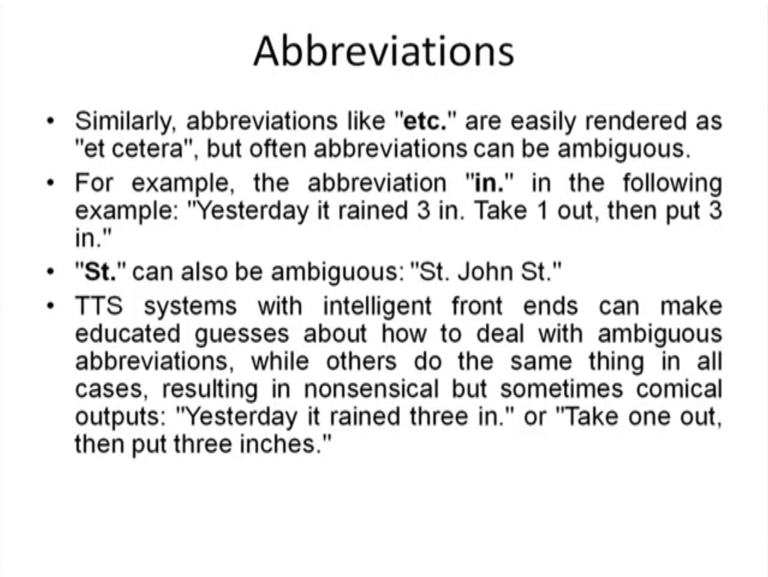
So, for prosody modelling some I have know the synthesis structure of the text, even its can goes to pragmatic structure also. So, I have to process the text using NLP natural language processing you know that there is a field call NLP language processing field. So, I have to use some kind of language processing so that I can model the prosody which is exit in the spoken form. There is a beautiful example suppose if I write I will go to Calcutta even if written language you see there is a gap in between the word, but if you spoken form there is a no gap there is no word boundary. So, this is continues spoken form, but we will lesion some boundary so that based on suprs segmental parameter.

So, those are called prosodic word boundary, I will discusses about in prosodic modelling that part. So, those called prosodic word boundary. So, some kind of text processing I require to model the speech prosody. So, that modelling require text prosodic. So, there is a lot of text processing block two part, mainly one part is converted g two p grapheme to phoneme conversion, second part is that some kind of synthesis sematic and pragmatic information even I have to extract from the text by which I can model the speech prosody. Once I goes those information when supra segmental information and G to P (Refer Time: 13:17) segmental information. So, once I know the segmental and supra segmental information, I can use some signal processing algorithm which is nothing, but a synthesis to produce the speech. So, now, I can use the synthesis to produce the acoustic wave form ok.

So, let discuses about that aberration I have already discuses that you can read many things that text normalization that conversion and aberration. And there is a if you know some if I if you remember there is a W3C standard W3C worldwide wave consortia; as a standard under the voice browser activity group that is called SSML. SSML speech synthesis mark-up language it is nothing, but a SML structure which is use to do this kind of job text (Refer Time: 14:23) grapheme to phoneme conversion and prosody that marking. So, when you develop website if you all the text all the written whatever the written communication is there, if it is tagged using SSML then it is very easy to synthesis that text when it pass to the synthesiser.

So, SSML mainly design to tag you input text such a way that it can take by a synthesis engine and process that text because the text is structure and it is process, it can produce the better synthesis speech. So, SSML there is a lot of tag sets are there. So, if you where there is a study is available you can go through the W3C SSML, you get a lot of document, and that document contain what are the tag set are there. So, in text normalization they normal they use say as suppose if I write 35 degree centigrade, if I treat this is a word then I can say for this word this grapheme, I can say as 35 degree centigrade; then there is call I will come to that PLS that that things also there abbreviation also.

(Refer Slide Time: 15:42)

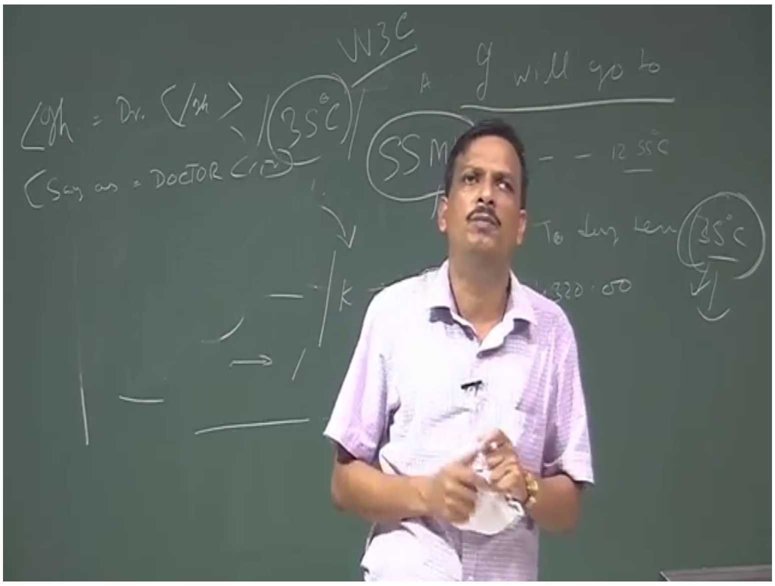


Abbreviations

- Similarly, abbreviations like "**etc.**" are easily rendered as "et cetera", but often abbreviations can be ambiguous.
- For example, the abbreviation "**in.**" in the following example: "Yesterday it rained 3 in. Take 1 out, then put 3 in."
- "**St.**" can also be ambiguous: "St. John St."
- TTS systems with intelligent front ends can make educated guesses about how to deal with ambiguous abbreviations, while others do the same thing in all cases, resulting in nonsensical but sometimes comical outputs: "Yesterday it rained three in." or "Take one out, then put three inches."

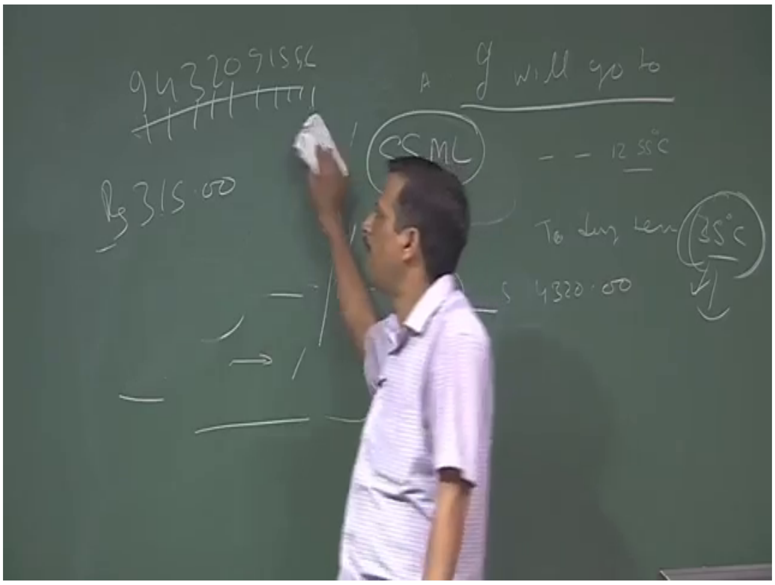
If it is doctors then I can say it is grapheme information is doctors. So, I can say that text the grapheme information is let I write gh is grapheme information.

(Refer Slide Time: 15:54)



So, I can say grapheme information is doctor Dr dot and end of grapheme, then I can say it can be say as doctor. So, I can define that things. So, if I define that things SSML take care about that or I can developed a dictionary base approach, where all the aberration there full form will be there. So, once I get aberration I run to the dictionary and get the full form and that text normalization for number it is difficult because you do not know where what kind of number system is required because some time if I write 94 32091556 is a phone number, then I can say this have to read this as read as 9432091556 we never said like that our that systems.

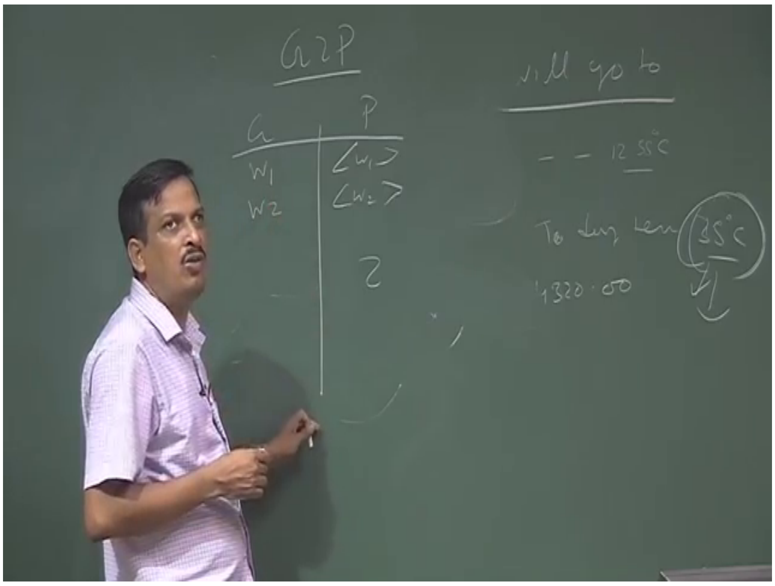
(Refer Slide Time: 16:39)



So, but if I write 3315 rupees 300 an then we never said 315, we say 315 rupees. So, depending on the context you have to say what kind of normal text normalization you should apply that is called text normalization.

Now, for G to P if I detail discuses G to P grapheme to phoneme conversion. So, text to phoneme or I can say grapheme to phoneme conversion there is a lot of papers you find that for every language the there is a several kinds of G to P engines are available grapheme to phoneme.

(Refer Slide Time: 17:30)



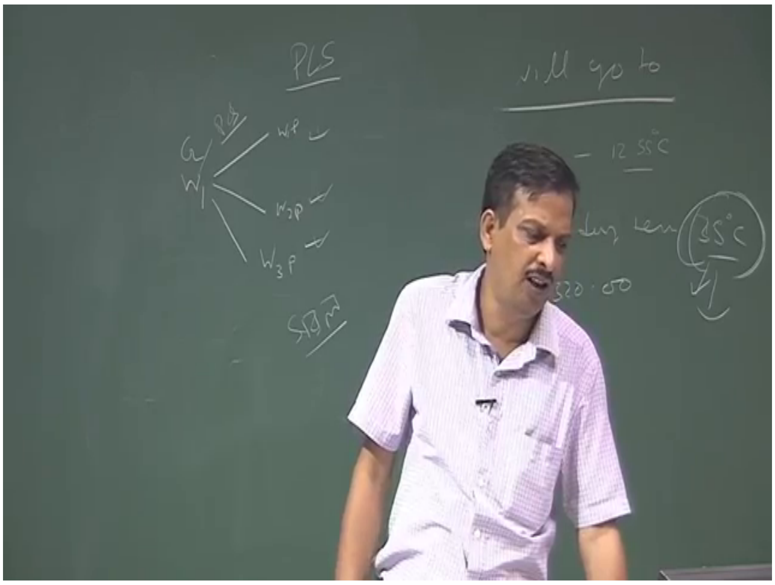
Conversion where dose system will take input as a text which is in grapheme information and converted its pronunciation information (Refer Time: 17:52) buffer Bangla, if you search in the night grapheme to phoneme conversion for Bangla you find my paper which is publish I think 2008 or 9 that grapheme to phoneme conversion for Bangla and that is available you can see that we have develop the TTS engine that grapheme to phoneme conversion. But sometimes it is very difficult to develop the role base of there is a you can say there is a lot of approaches are there either I can develop dictionary based approach, where I can say I have a two table here grapheme information here is pronunciation information. So, there is a W1 word and there is a pronunciation of W1.

So, there will be a W2 word, there will be a pronunciation of W2 words. So, this is one kind approach this is called dictionary base approach where I required the pronunciation dictionary of all the words because I do not know which text will come in my input. So, all the words pronunciation dictionary is required, but you know the words list of words are infinite if you say that there is a some kind of combine word. So, you can intelligently develop a root word dictionary, and you can form a combine dictionary based on the on the fly requirement, but combining also there is a problem.

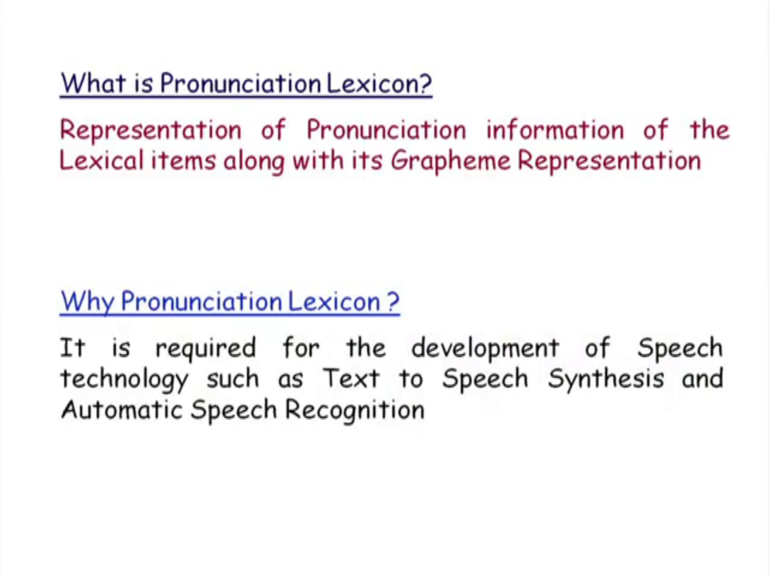
Means, in Bengali sometime if the two words are combine, the middle vowels maybe pronounced like that we say we say Rajputra, that Rajputra if it is Rajputra then it is middle vowel is deleted because raj is a word and Putra is a word, they are combine that is why Rajputra is this followed every word pronunciation if I combine I get Rajputra. But some time vowel is not deleted like it is not one (Refer Time: 19:47) it is not one (Refer Time: 19:48). So, that the we have pronounce (Refer Time: 19:51).

So, no there is vowel o is added, those kind of complexity are there. So, if I have, but simplified that develop a pronunciation dictionary and or you can say it is call a look up table in computer programing. So, I can say I have a pronunciation dictionary and I can compel once I get the grapheme word in can pick up the pronunciation word from the table. But there is a if you know the W3C they have a standard call PLS pronunciation lexical lexicon specification PLS pronunciation lexical specification pronunciation lexical specification.

(Refer Slide Time: 20:32)



(Refer Slide Time: 20:38)



What is Pronunciation Lexicon?

Representation of Pronunciation information of the Lexical items along with its Grapheme Representation

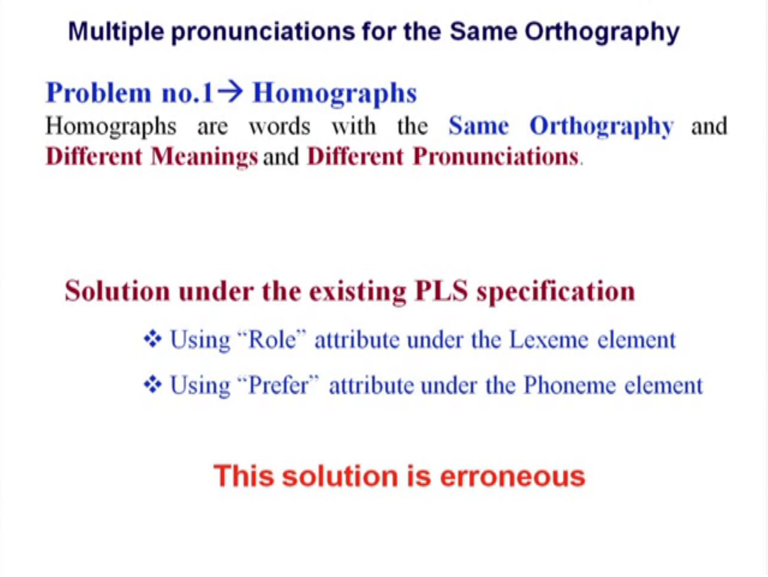
Why Pronunciation Lexicon ?

It is required for the development of Speech technology such as Text to Speech Synthesis and Automatic Speech Recognition

Detail I have already study the details of PLS W3C and we as said that some problem for step present PLS. So, this is nothing, but you can say it is required by both speech recognition speech synthesis both I will come. So, the there is lot of tag set are there in PLS standard, if you see there is a lexeme; that means, lexicon is started then there is a metadata then lexeme, then grapheme, the phoneme, then alias and the example.

So, they said, but the problem is that they said the some words if have a multiple pronunciation use the prefer active we choose the correct pronunciation, but it is not always true. Some time same word may have different pronunciation even more than one pronunciation more than one means more than two words also pronunciation, but both pronunciation are the valid pronunciation in that language it is not dialectic variation. So, we can say I have race that write that paper that is called homograph problem.

(Refer Slide Time: 21:46)



Multiple pronunciations for the Same Orthography

Problem no.1→ Homographs

Homographs are words with the **Same Orthography** and **Different Meanings** and **Different Pronunciations**.

Solution under the existing PLS specification

- ❖ Using "Role" attribute under the Lexeme element
- ❖ Using "Prefer" attribute under the Phoneme element

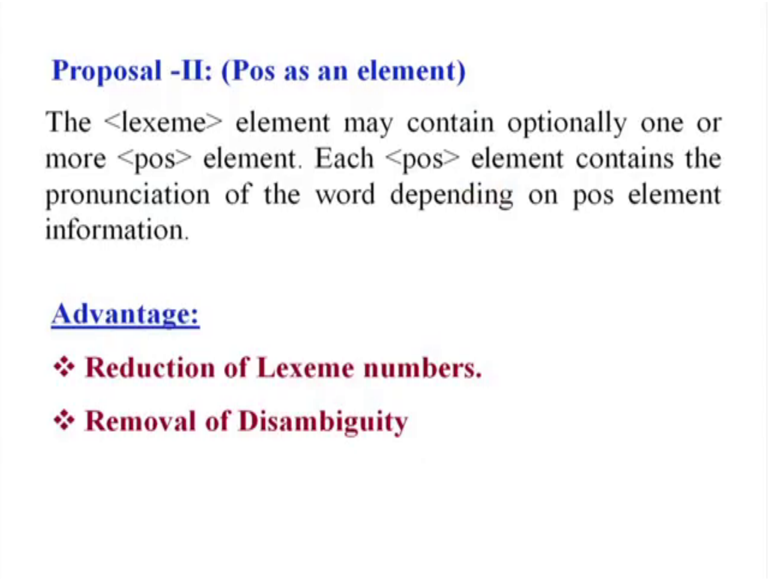
This solution is erroneous

That homograph are the words with same orthography, but different meaning and different pronunciation meaning is not important pronunciation important. So, homograph; homograph means orthography grapheme grapheme all are same. So, if the pronounce if the graphemic representation words let us it is W1, but it map to two pronunciation one is W 1P and may be the W2P, the pronunciation is two may be three pronunciation is the same word can have a three pronunciation depending on the context where I use that.

So, that is called homograph grapheme is same, but pronunciation is different. In case of Bengali that I can give you example in Bengali we are [FL] if you know the Bengilo script then you know the [FL] may have a pronunciation two pronunciation, either [FL] or it may be [FL]. So, if it is verb it is [FL] if it is adjective, if it is adjective then it is simple [FL] means simple then it is [FL]. So, all kinds of homograph variation as to be solve. So, it was found that for so speech information of this word somewhat solve that homograph problem, but not always through. Some time same word in case of honorific in nature maybe pronunciation is some things and if it is non-honorific nature may be pronounce in the different ways. So, those kind of variation may not be; that means, sematic variation, there may be a something information that if it is honorific then the pronunciation is something like Bangla [FL].

So, those kind of pronunciation variation will be there homograph problem, then homograph phone problem (Refer Time: 23:47) speech may solve it you can go through this document I have already explain it the for. So, speech then there is a problem is called. That homophone problem homophone problem is just opposite. That different orthography had same pronunciation. So, orthography representation is same.

(Refer Slide Time: 23:53)



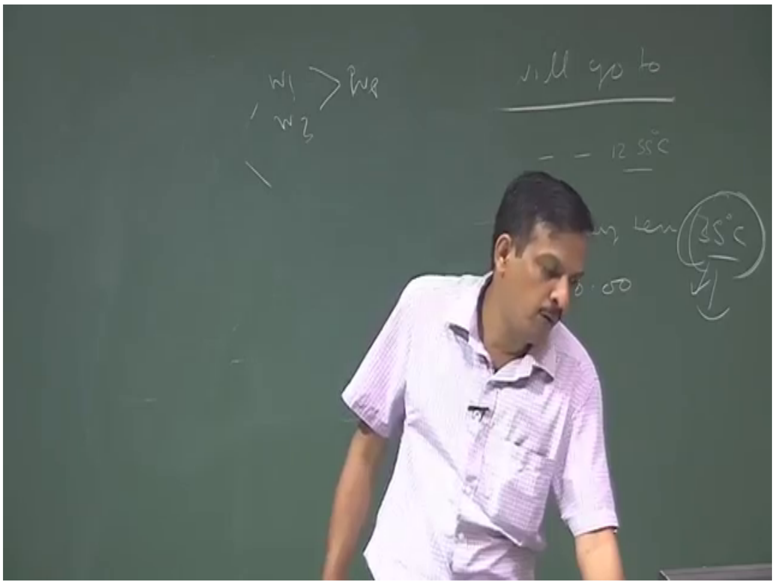
Proposal -II: (Pos as an element)

The <lexeme> element may contain optionally one or more <pos> element. Each <pos> element contains the pronunciation of the word depending on pos element information.

Advantage:

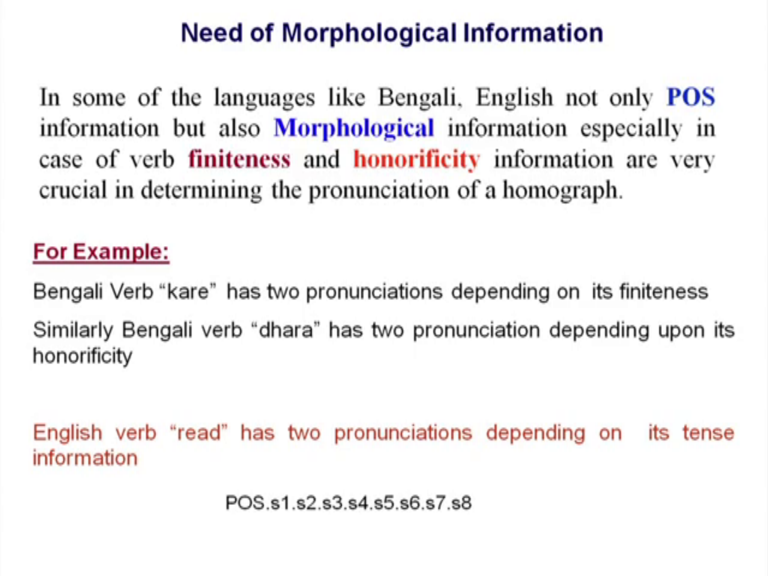
- **Reduction of Lexeme numbers.**
- **Removal of Disambiguity**

(Refer Slide Time: 24:14)



But pronunciation maybe have sorry different orthography maybe W1 and W2, but the pronunciation is same like that right right write in English both are pronounce as write, but depending on the context I can say which write I have pronounced. So, those problem will be there in PLS.

(Refer Slide Time: 24:44)



Need of Morphological Information

In some of the languages like Bengali, English not only **POS** information but also **Morphological** information especially in case of verb **finiteness** and **honorificity** information are very crucial in determining the pronunciation of a homograph.

For Example:

Bengali Verb "kare" has two pronunciations depending on its finiteness

Similarly Bengali verb "dhara" has two pronunciation depending upon its honorificity

English verb "read" has two pronunciations depending on its tense information

POS.s1.s2.s3.s4.s5.s6.s7.s8

So, those problem we have propose some solution and we have made that paper to submitted this paper to PLS then we have taken it. So, you can go through the detail paper is available in the net.

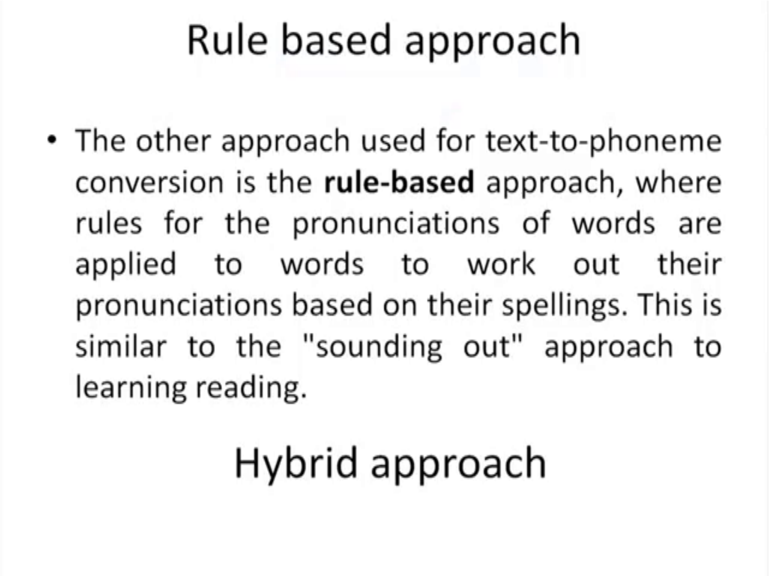
So, there is a lot of details examples are there lot of analysis I have done honorific, then we say the then there is dictionary that you can morphological information is important many country also rise this issue that morphological is important for Corian also morphological is very important like that (Refer Time: 25:14) in case of Bangla said that finiteness and honorificity information may be required. So, you suggest some kind of morphological analysis for solve the pronunciation problem in Bangla.

So, G to P is not that simple many language it may be simple that they in case of Indian language yes it is somewhat simple, but Bangla is not that simple, but in case of Hindi we found that it is little bit of simple compare to other language because since we have the syllabic language then the pronunciation whatever we write we almost pronunciation the pronounce the same, but it is not always true there may be some variation. In Bengali there is lot of variation in written script in case of English there is also part of variation.

So, but if you see the unfortunately English pronunciation dictionary is available in the net because that is available, but unfortunately in the Indian language pronunciation dictionary are very rare cannot find any pronunciation other language. So, there is very top challenge to develop that TTS in Indian language, but yes we have done something for Bangla some other group also there who are doing for that Hindi, Tamil, Telugu all language TTS are write now there, but biggest problem in G to P grapheme to phoneme conversion which is very important block for the TTS engine. So, W3C as a PLS speciation even we can started a community building on that how do we built community the pronunciation dictionary, there is a website we have develop a web base engine where you can develop the pronunciation dictionary is PLS standard W3C PLS standard. So, that is website is available in IIT Kharagpur I think it is (Refer Time: 27:04) main like that I conduct domain name.

So, at the end I will give you that wave address where you can use that website to develop you own pronunciation lexicon in the W3C format.

(Refer Slide Time: 27:18)



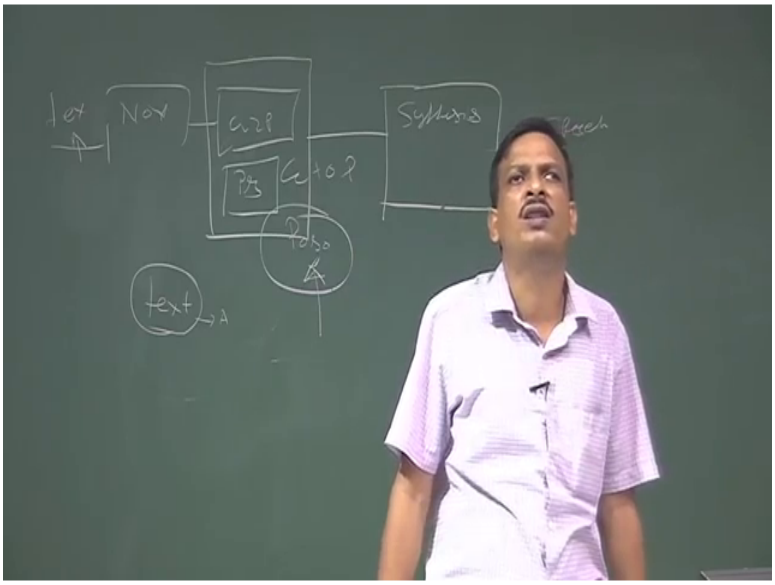
Rule based approach

- The other approach used for text-to-phoneme conversion is the **rule-based** approach, where rules for the pronunciations of words are applied to words to work out their pronunciations based on their spellings. This is similar to the "sounding out" approach to learning reading.

Hybrid approach

Now there is a other kind of G to P approach also they have the rule based approach, that depending on the grapheme position we can develop some rule where rule base approach will pronounce the grapheme to pronunciation conversion will be done must most mostly use approach are hybrid approach, that some simple rule will develop and then we use the dictionary for all exceptional cases. So, that is called hybrid approach; then once I done that that text normalization and G to P.

(Refer Slide Time: 27:52)



Let I am not discussing about that text processing required for prosody modelling because for prosody modelling I will discuss in the next week, the last week I will be detail discuss about the prosody modelling of different languages. English there is a prosody modelling. So, there is lot of prosody model available (Refer Time: 28:14) command responsible model to be model all kind of prosody are there.

So, prosody modelling is one of the important issue will discuss later. So, if text processing for prosody modelling we excluded the G to P is sufficient. So, I can say the text, text input text will come and then it will converted normalized and announce it is normalization has done in war normalization done then go to G to P, once is go to G to P and prosody modelling that text processing for prosody is combined together call language processing and then it goes to speech synthesis.

So, I will synthesis there are lot of formats are there will discuss details on synthesis because it is speech processing class. So, we discuss about the synthesis part os output is called the speech. Now the one problem is there that text I think many of you know the Unicode I think so, because if I write text it is in English. So, it is already there in (Refer Time: 29:24) code system, but if I write in Bengali, Hindi, Tamil, Malayalam all other languages those codes are come in Unicode.

So, text input text is in the form of Unicode. So, you have to develop you have to take care about the Unicode processing. So, one Unicode same thing, but every grapheme has a unique Unicode. So, I do not have any problem in processing. So, Unicode processing is there then best on that you can develop the dictionary normalized and G to P all kinds of thing based on the Unicode of that text. Then synthesis there is lot of algorithm for there in this is articulatory synthesis, parametric synthesis and concatenative synthesis.

So, in the next class I will detail discuss about the each of the model articulatory synthesis, parametric synthesis and concatenative synthesis which is the main part of that TTS engine. So, there is a lot of others approach also there. So, then we discuss we show some something is available I can show you, I can lesion you I can show you the voice all kind of things.

Thank you.