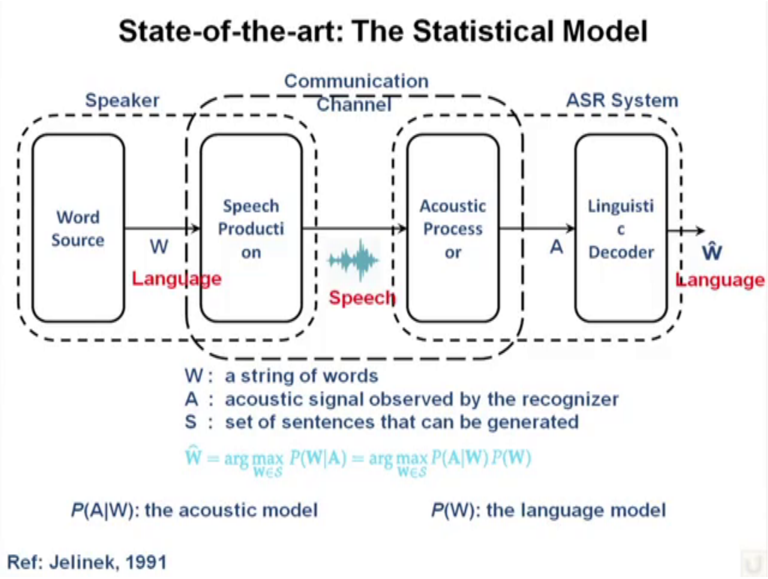
Digital Speech Processing

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

Lecture – 40
Statistical Modeling of Automatic Speech Recognition

(Refer Slide Time: 00:23)



State-of-the-art: The Statistical Model

Speaker

Communication Channel

ASR System

Word Source

W

Language

Speech Production

Speech

Acoustic Processor

A

Linguistic Decoder

$\hat{W}$

Language

W : a string of words
A : acoustic signal observed by the recognizer
S : set of sentences that can be generated

$$\hat{W} = \arg\max_{W \in S} P(W|A) = \arg\max_{W \in S} P(A|W)\,P(W)$$

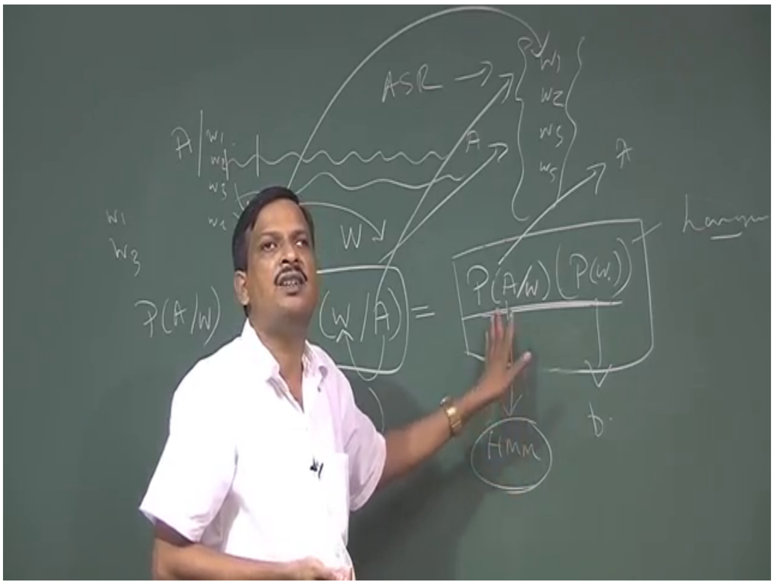
$P(A|W)$: the acoustic model

$P(W)$: the language model

Ref: Jelinek, 1991

So there is the state of what statistical model if you see. This is the block diagram of a state of art statistical model. So, in here what you want in speech recognize the problem in ASR is that.

(Refer Slide Time: 00:34)



What about the spoken I can say the set of set of sentences that can be generated that or I can say that let us I have spoken a sentence, this is the acoustic signal which is measure is A. So, what I think there is a acoustic signal which has frame and parameter are expected and that measurement is called acoustical measurement.

So, I say I have recorded the acoustical signal extracted the parameter as I discussed frame wise, either 100 frame per second or 50 frame per second as you wish. Now that is called acoustical measurement, then what is ASR that from the acoustical measurement, I have to predict which word I have spoken understand or not. So, there is a given acoustical model I have to find out which word I have spoken. So, I can say the problem is probability of w or a given acoustical measurement. So, I want maximize the probability of recognize the word correct word for a given acoustical measurement.

So, if I say that this acoustical say this ASR is designed for this set of word w 1 w 2 w 3 w 4 w 5 like that, then I can say for a given acoustical word or acoustical measurement what is the probability of any word of this given list. Now this theory I can apply these can be equal to probability of A given a w into probability of w divided by probability of A joint probability.

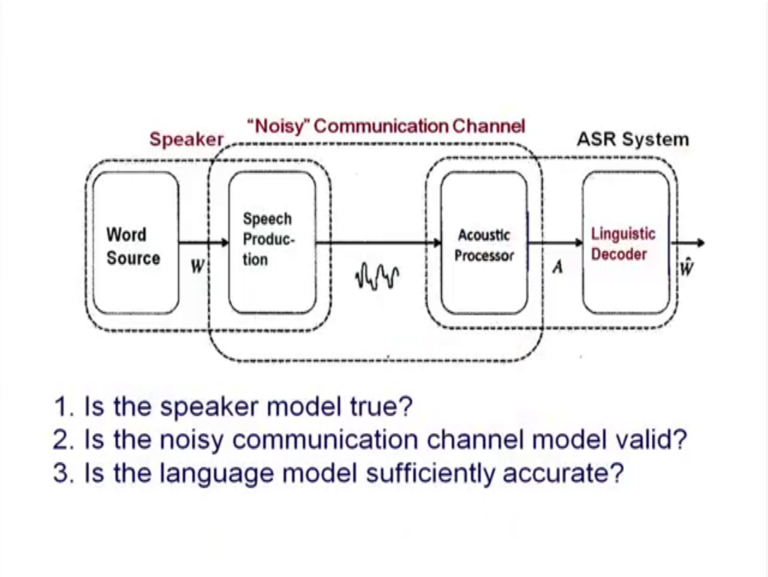
So, if I want to maximize this probability this can be treated same problem I can state that I have to maximize this one. Now probability of A probability of acoustical measurement all the time I have taken the acoustical measurement. So, I can ignore this is not equate, because all I have all equal that I have acoustical measurement is there. I am not saying I have not measure the acoustical acoustical measurement will be not there, but I have do given that I would recognize the word.

So, probability A of A I can omit this all I let us say 1. Now the maximizen problem word down to the maximize the probability of A for a given w and probability of w. So, what are the meaning? I have said that for a given acoustical measurement find out the maximum probability which is the a is belong to which word or which word, it which word give me the maxima maximize this problem it a max of this one. So, this acoustical measurement are maximally mapped to which word, or maximum mapped to which word. That converted to for a given set of word now the word is given for a given set of word I have to maximize this acoustical observation. So, probability of this acoustical measurement is maximum for which word, and multiplied by probability of w.

So, this is acoustical measurement. So, this is call acoustical model. And this is called language model. So, this is the statistical model in ASR. So, developed an acoustical model. So, what I have to do? That let us I have a given that w 1 w 2 is there. So, I said whatever I have measured in acoustical measurement A, what about the parameter? I have extracted find out the maximum probability of A for a given w 1 w 2 w 3 or w n which is for a probability of A for a given w. So, org max of probability of A for a given w and probability of w. So, I have to maximize this one and maximize this one, or I can say the product has to be maximized. This is called language model this is called acoustical model.

So, acoustical model is done based on the HMM and language model done based on the trigram bigram trigram or unigram model. So, this is language model bigram trigram and this is HMM model for acoustical say. I have not details discussed how this HMM model is done. So, this is nothing but a maximizen of the probability. Now I come to the issues what are the issue is there.

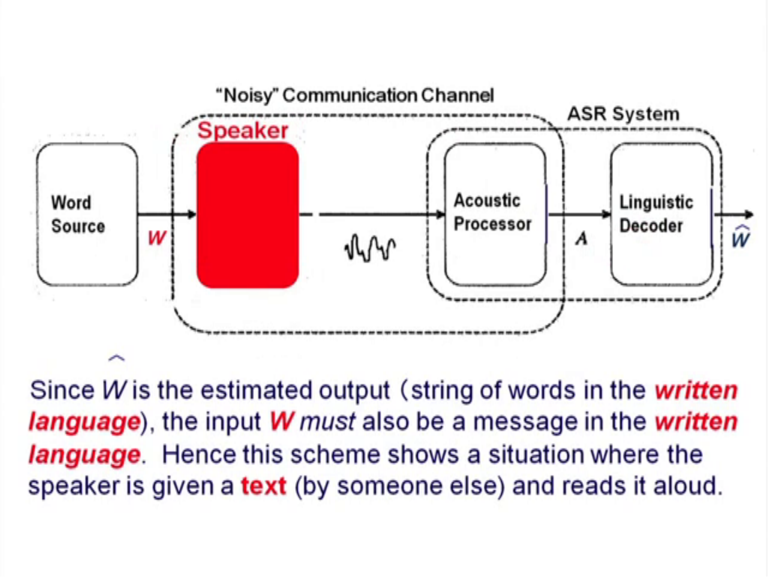
(Refer Slide Time: 06:41)



1. Is the speaker model true?
2. Is the noisy communication channel model valid?
3. Is the language model sufficiently accurate?

Now, issue first issue is the speaker model is to forget about these slides.

(Refer Slide Time: 06:45)



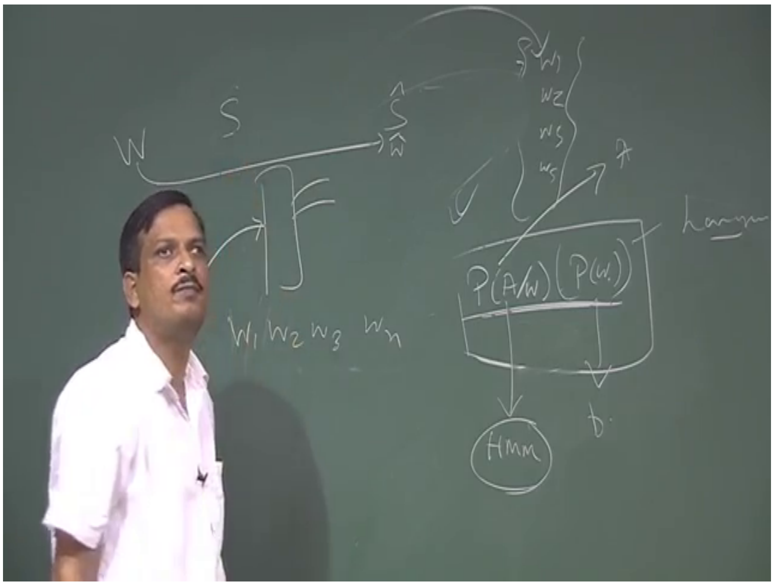
Since $\hat{W}$ is the estimated output (string of words in the ***written language***), the input ***W*** *must* also be a message in the ***written language***. Hence this scheme shows a situation where the speaker is given a **text** (by someone else) and reads it aloud.

Now, this is (Refer Time: 06:46). So, what I said that acoustic signal is related to set of w. So, what I said if I to do that way that one source w is there speaker has spoken that word production of that word from the production signal, I measure the acoustic parameter from the acoustic parameter I want to detect this w.

So, I am estimating this w. So, this is ASR system now I have to say I am I am assuming the acoustical signal is related to written text w; that means, I am assuming w cap is estimated output of the word written language. The input word must also be a messaged in written language. So, speak suppose I am speaking some sentence; that means, if I speak spontaneously. So, how the when speaker said some sentence how they generate that message. There is a message planning idea message planning production. Now my ASR model told me, that whatever the production signal is there that is nothing but a w 1 w 2 w 3 w 4 or whatever the some set of words. Which is which can be explicitly expressed in written form.

But there is a catch spoken word and written word are not identical. When we will listen the words we can identify spoken so.

(Refer Slide Time: 08:47)



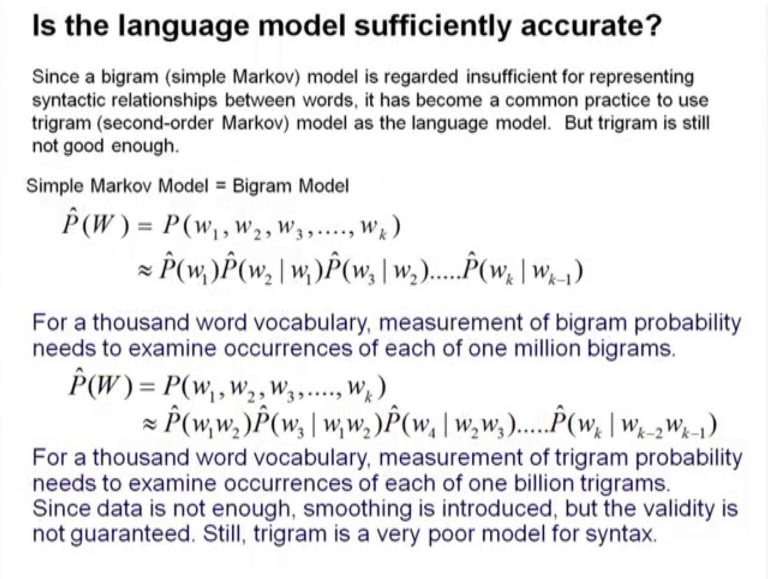
So, I am saying that speaker when it produce the speech I have prompted the speaker that you have to set this w 1 w 2 w 3 w 4 w n. So that means, something which I am reading text is given speaker is read out the text, for that signal this as r model is valid. Now if I say it is a spontaneous speech I am speaking, when I speak I do not care about what is written in there. Some words omitted many thing you record it record any just continuously read even it continuously read one paragraph, not that first time reading. If you the practice 10 time then read it then you find many cases many syllabol many words are dropped.

Or I can say when human being communicate we never say complete message word by word, or even if in continuous speech all words are not explicitly mentioned. If you find in continuous speech always end of the sentence the amplitude is go going down even, I have seen many cases in end syllabol is not there acoustic signal is not there. So, this ASR model does not model this speaker mine words or you can say what speaker want to say that is prompted by the input text only.

But if it is not prompted all kind of variation of spoken language will be there, and my model is not true for that. So, this is the challenge, I will details. I will discuss next challenge is the understanding the input message because what speaker does he understand the input message and then spoken. So, what about the speech production output is there it is not w, it is a spoken or it is a spoken. And whatever I estimated I want to estimate these w only, but acoustic signal is only content is spoken information not written information.

How they are different I will come. So, similarly there is a problem in language model also.

(Refer Slide Time: 11:29)



Is the language model sufficiently accurate?

Since a bigram (simple Markov) model is regarded insufficient for representing syntactic relationships between words, it has become a common practice to use trigram (second-order Markov) model as the language model. But trigram is still not good enough.

Simple Markov Model = Bigram Model

$$\hat{P}(W) = P(w_1, w_2, w_3, \ldots, w_k)$$
$$\approx \hat{P}(w_1)\hat{P}(w_2 \mid w_1)\hat{P}(w_3 \mid w_2) \ldots . \hat{P}(w_k \mid w_{k-1})$$

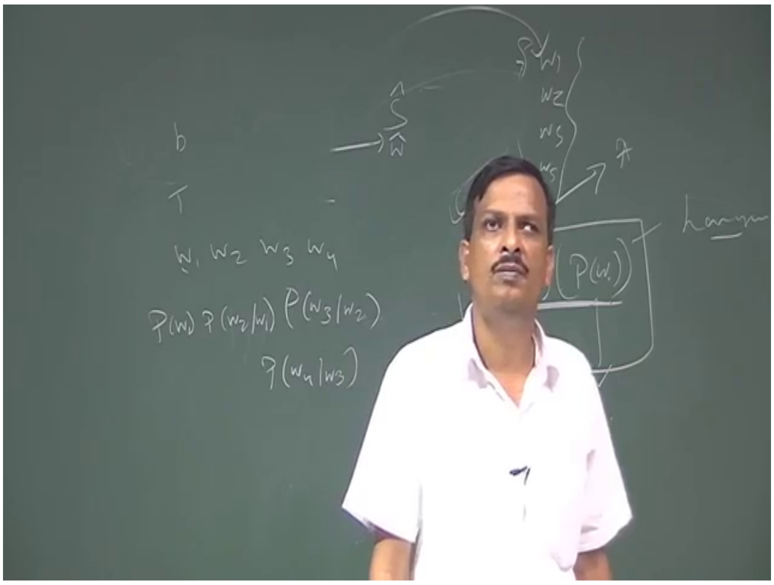
For a thousand word vocabulary, measurement of bigram probability needs to examine occurrences of each of one million bigrams.

$$\hat{P}(W) = P(w_1, w_2, w_3, \ldots, w_k)$$
$$\approx \hat{P}(w_1 w_2)\hat{P}(w_3 \mid w_1 w_2)\hat{P}(w_4 \mid w_2 w_3) \ldots . \hat{P}(w_k \mid w_{k-2} w_{k-1})$$

For a thousand word vocabulary, measurement of trigram probability needs to examine occurrences of each of one billion trigrams.
Since data is not enough, smoothing is introduced, but the validity is not guaranteed. Still, trigram is a very poor model for syntax.

What I said that language model I have e w is nothing but a language model which is which can be developed by bigram or trigram probability. What is bigram probability? Bigram model that for a P w P given w 2 s o if I have a sentence w 1 w 2 w 3 w 4 I can say the bigram model is p w 1 p w 2 for a given w 1.

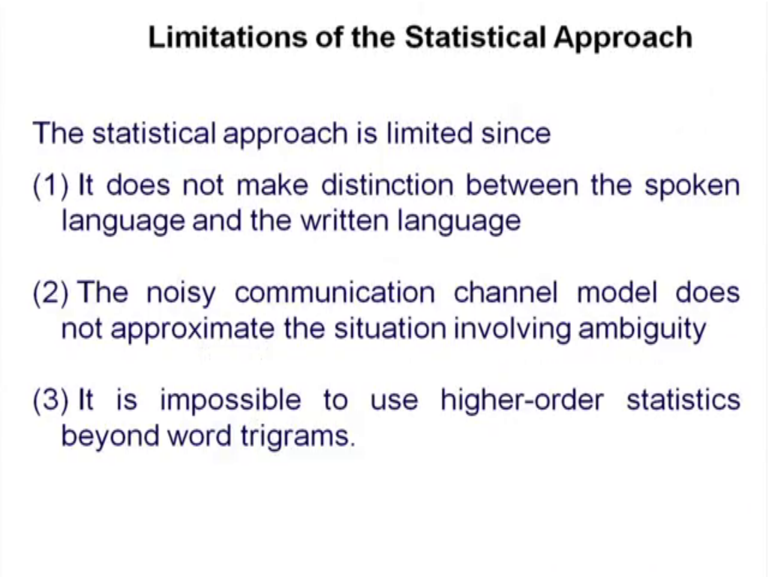
(Refer Slide Time: 11:37)



Probability of w 3 for a given w 2, then probability of w 4 for a given w 3 like that way, there is bigram model. Similarly trigram model, probability of w 1 w 2 then w 3 for a given w 1 and w 2, but up to trigram model I can go if I go beyond above trigram model the complexity will be arised.

The required training data size will be very huge. So, even if develop the trigram model also the equated data size required very huge. So, for a less resource language whose computerized resource are very less, those languages development of trigram model and acoustic models is very difficult. So, this statistical problem it is true the statistical means if I have a enough data then my statistical inference may be very good. If I have a limited data then I have a statistical inference is very limited.

(Refer Slide Time: 13:18)



Limitations of the Statistical Approach

The statistical approach is limited since

(1) It does not make distinction between the spoken language and the written language

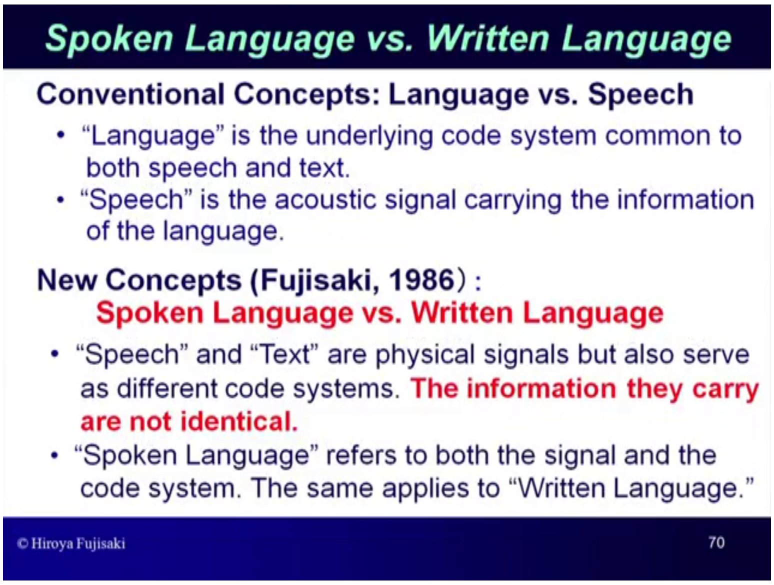
(2) The noisy communication channel model does not approximate the situation involving ambiguity

(3) It is impossible to use higher-order statistics beyond word trigrams.

So, statistical model the limitations are it does not made distinction between the spoken language and written language, what I have described. That we have defined the ASR for a given w, but it not distinction between the written language and spoken language.

If the noisy communication general model does not approximate the situation involved in ambiguity, that is I will discuss written ambiguity means. There is a some ambiguity in written language by ice cream I do not know whether it is ice cream or ice cream. Then it I a impossible to use higher order statistical model or language model, beyond trigram I cannot use, but that trigram model not give me the enough language model information. So, what is I do? So, they leads till there is a open challenge, which calls spoken language recognition.

(Refer Slide Time: 14:14)



Spoken Language vs. Written Language

Conventional Concepts: Language vs. Speech

- "Language" is the underlying code system common to both speech and text.
- "Speech" is the acoustic signal carrying the information of the language.

New Concepts (Fujisaki, 1986) :
Spoken Language vs. Written Language

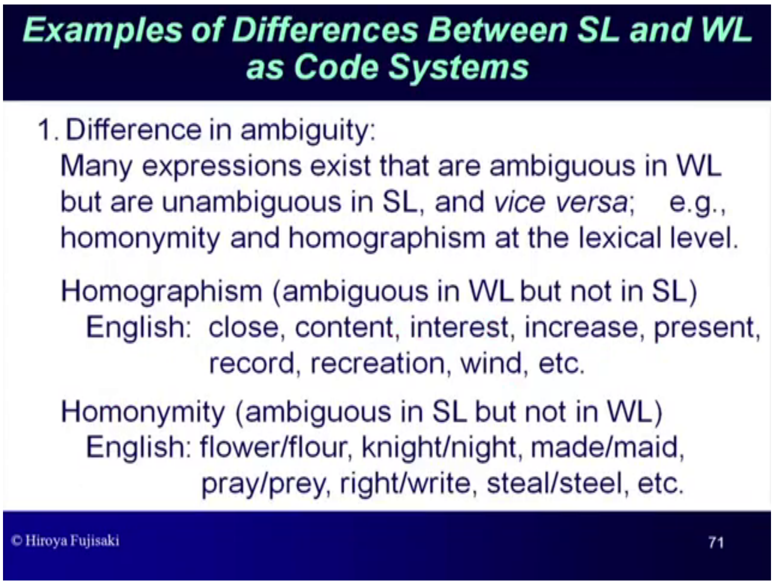
- "Speech" and "Text" are physical signals but also serve as different code systems. **The information they carry are not identical.**
- "Spoken Language" refers to both the signal and the code system. The same applies to "Written Language."

© Hiroya Fujisaki 70

So, I have to develop some system, which is spoken language recognition. Spoken language is deform from written language. Speech and text are physical signal. So, fujisaki, if you see the slide is taken from the fujisaki (Refer Time: 14:33) fujisaki.

He said that spoken language versus written language spoken new concept in spoken language, speech and text that physical signal, but also saw that the different code system. The information they carry or not identical. Spoken language refer to both signal and both system same applies to written language.

(Refer Slide Time: 14:56)



Examples of Differences Between SL and WL as Code Systems

1. Difference in ambiguity:
   Many expressions exist that are ambiguous in WL but are unambiguous in SL, and *vice versa*; e.g., homonymity and homographism at the lexical level.

   Homographism (ambiguous in WL but not in SL)
   English: close, content, interest, increase, present, record, recreation, wind, etc.

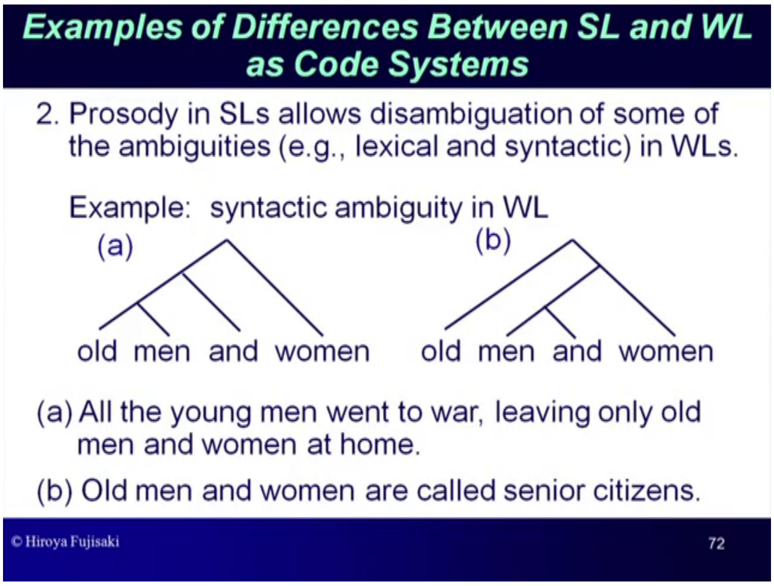
   Homonymity (ambiguous in SL but not in WL)
   English: flower/flour, knight/night, made/maid, pray/prey, right/write, steal/steel, etc.

© Hiroya Fujisaki 71

Now, if you discuss difference in ambiguity. So, there may be something is difference spoken language in ambiguous, but written language is not, but something in written language is ambiguous, but spoken language it is not. That is homographism and homonymity. Ambiguous in spoken language, but not in written language flower, but pronunciation are same.

But there spelling are different written language they are not same, similarly pronounce written language they are same, but pronunciation, but written ambiguous in written language, but not in spoken language, close, interest ok.

(Refer Slide Time: 15:40)



Examples of Differences Between SL and WL as Code Systems

2. Prosody in SLs allows disambiguation of some of the ambiguities (e.g., lexical and syntactic) in WLs.

Example: syntactic ambiguity in WL

(a) All the young men went to war, leaving only old men and women at home.

(b) Old men and women are called senior citizens.

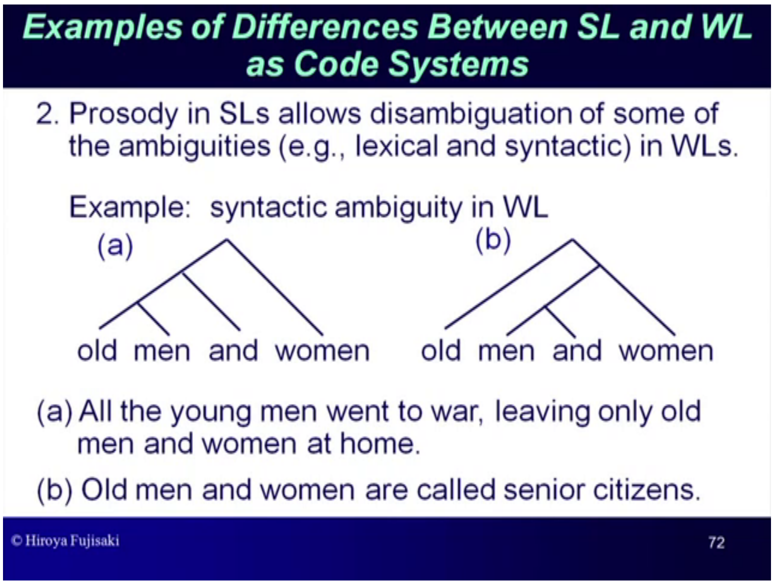
© Hiroya Fujisaki

72

Now, if I come the prosody, this is the important issue spoken language. Not only contain the words linguistic information. So, as said linguistic information or not words is not only the linguistic information there is a linguistic information called prosodying.

So, prosodying is message how the message is plan rhythm is important. So, I said the spoken written language what boundary are not exist in spoken form, but there is a boundary in spoken language which is spoken language boundary, not written language boundary. If I say in english yes there is a 2 example, old man and woman if I say the old man and woman. Then man and women both are old if I said the old man and woman the man is old, but women is not.

(Refer Slide Time: 16:37)



Examples of Differences Between SL and WL as Code Systems

2. Prosody in SLs allows disambiguation of some of the ambiguities (e.g., lexical and syntactic) in WLs.

Example: syntactic ambiguity in WL

(a)
old men and women
(b)
old men and women

(a) All the young men went to war, leaving only old men and women at home.

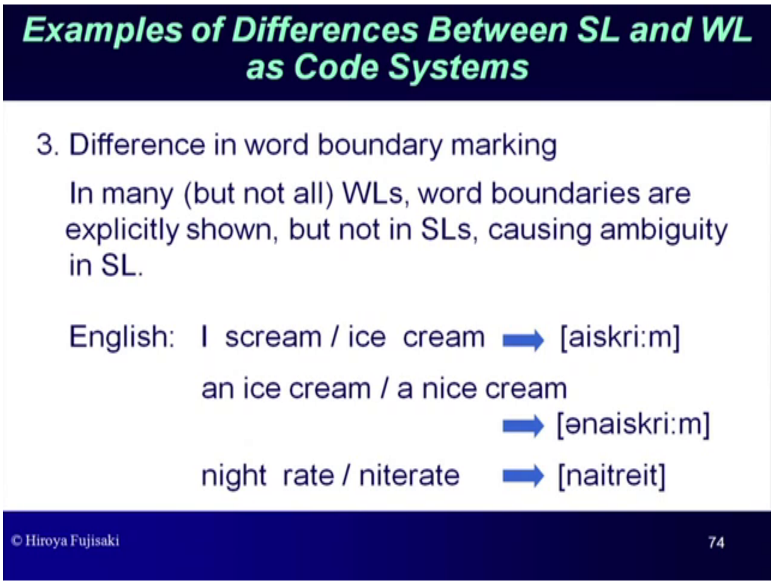
(b) Old men and women are called senior citizens.

© Hiroya Fujisaki

72

Similarly, in Bengali I will give you interesting example [FL], if I say [FL], I can say like that way or I can say [FL], if I said that the meaning is different and if you see the what boundary positions. The position are boundary at different [FL] is same [FL]. So, prosody is ambiguity is there now, but the prosody information is very important now for spoken words identification is very important.

(Refer Slide Time: 17:19)



Examples of Differences Between SL and WL as Code Systems

3. Difference in word boundary marking

In many (but not all) WLs, word boundaries are explicitly shown, but not in SLs, causing ambiguity in SL.

English: I scream / ice cream ➡ [aiskri:m]

an ice cream / a nice cream ➡ [ənaiskri:m]

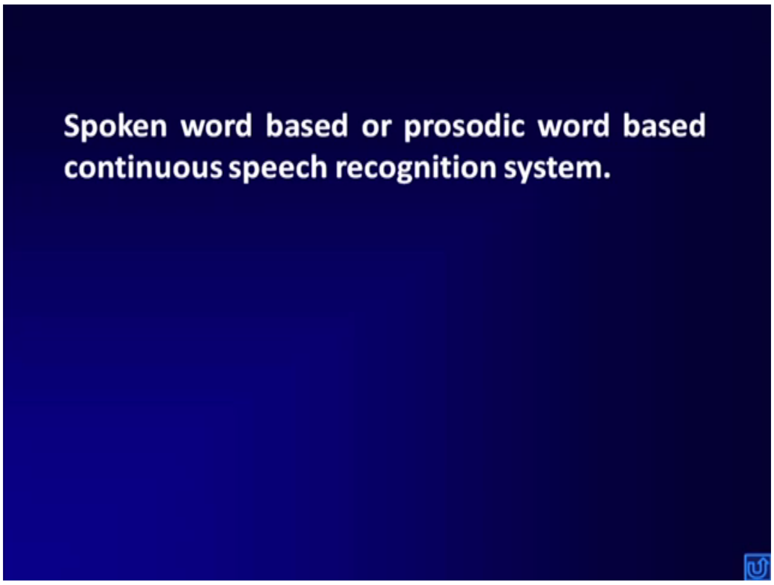
night rate / niterate ➡ [naitreit]

© Hiroya Fujisaki

74

What boundary marking? What languages spoken word written language or ice cream ice cream I say ice cream ice cream. Spoken language ambiguity. Then I am not going details of the written language and spoken language.

(Refer Slide Time: 17:37)



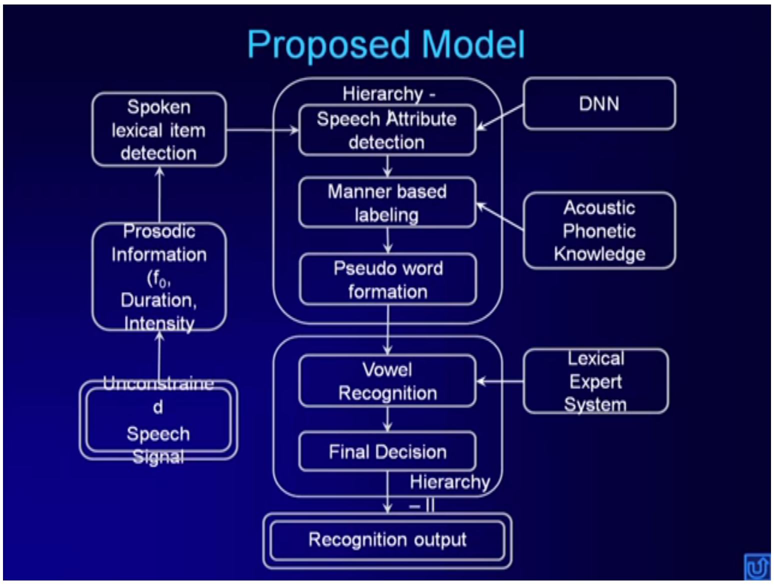
Spoken word based or prosodic word based continuous speech recognition system.

So, then we have in our team are doing some research on spoken word or prosody word based continuous speech recognition system. We said the written language words are different from spoken language word.

So, instead of trading the same system with written language can I trend the system with prosodic word or prosody that spoken word language spoken word. So, you developed a system which trend on spoken word for bengolis now. Identification of the spoken word is important. So, spoken word is identifies for a spoken language database. So, I can say I record this I collect this continuous speech. And identify the prosodic information based spoken word position, and I said those words I can create a dictionary and I told the recognizer those of the spoken word, not w. So, I can say dictionary. So, I can say that my a statistical model only estimate s cap from s, not w to w cap because w is converted to s.

So, I have saying my thing is that I have a s I estimate s cap and then try to convert to w using language model ok.

(Refer Slide Time: 19:06)



Proposed Model

Spoken lexical item detection

Prosodic Information (f_0, Duration, Intensity

Unconstrained Speech Signal

Hierarchy - Speech Attribute detection

Manner based labeling

Pseudo word formation

DNN

Acoustic Phonetic Knowledge

Vowel Recognition

Final Decision

Hierarchy – II

Lexical Expert System

Recognition output

So, if you see that there is a lot of research in this area, this is the proposed model. That prosodic word are boundary are identified there is a lot of this 2 1. Thesis is there my own research is also in this direction then manner based leveling pseudo word formation and then if you want the details. Then you can contact me I can share you the all details information of this I am not detail discussed this kind of things this is available. So, this is complete of speech recognition speech, speech technology speech recognition things. So, next class let us discuss about something on speech base learning or my dream research action conversion. So, next class may be 10 to 20 minutes we try to discuss in the next class ok.

Thank you.