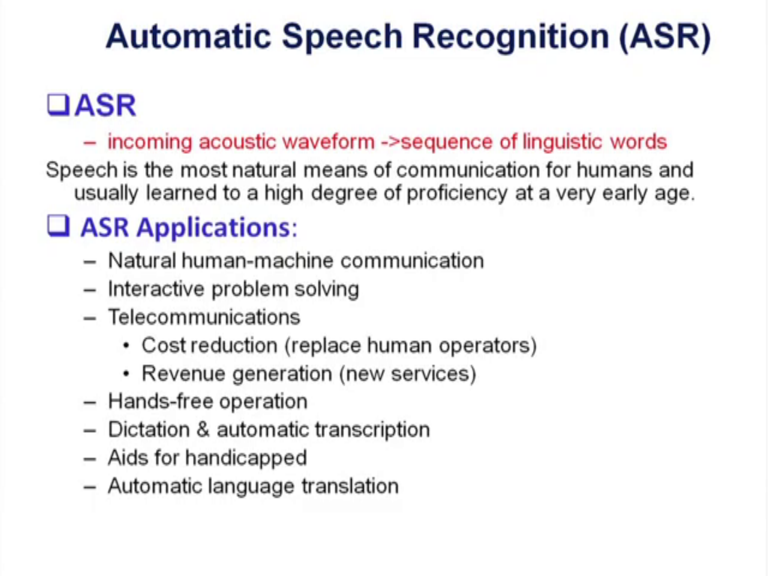
Digital Speech Processing
Prof. S. K. Das Mandal
Centre for Educational Technology
Indian Institute of Technology, Kharagpur

Lecture – 39
Automatic Speech Recognition

So as we said that you have discussing about some speech application. So, my idea is not to details go about that which synthesis which ASR because that if I go for the switch (Refer Time: 00:30) in details then it will take another 4 to 6 lectures. So, I am not going detailed. So, and same things for ASR also a lot of a tricks speech recognition.

(Refer Slide Time: 00:41)



Automatic Speech Recognition (ASR)

❑ ASR
- – incoming acoustic waveform ->sequence of linguistic words

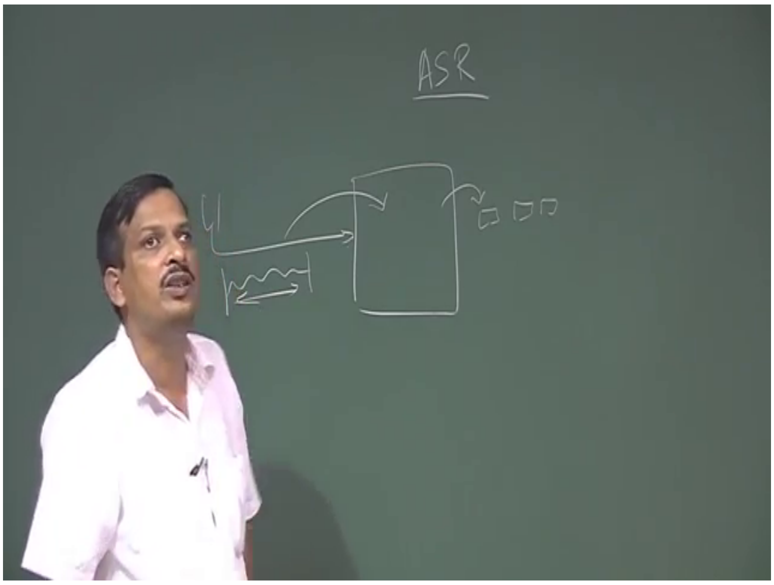
Speech is the most natural means of communication for humans and usually learned to a high degree of proficiency at a very early age.

❑ ASR Applications:
- – Natural human-machine communication
- – Interactive problem solving
- – Telecommunications
  - • Cost reduction (replace human operators)
  - • Revenue generation (new services)
- – Hands-free operation
- – Dictation & automatic transcription
- – Aids for handicapped
- – Automatic language translation

What I will know I will basically describe what is ASR, what are the technology involve and challenges in ASR also I will described. So, that you can take some project and you can as a student you can do some project on that challenges or as a teachers you can gives that project to other students also to pursue in that area.

So, my idea is not detailed discuss about the ASR, but red goes issue which are involved in ASR and rules of project in current scenario also. So, if I said that ASR like the TTA if I say ASR automatic speech recognition.

(Refer Slide Time: 01:24)



So, automatic speech recognition, so what is at the automatic speech recognition means that if I have a machine and I have a microphone. If I talk to the machine should understand not only this is understanding I am not saying understand. Machine should decode what linguistic message is there in the switch signal; that means, suppose I utter a sentence machine should recognize each and every word of the sentence. I am not saying understanding for powerfully.

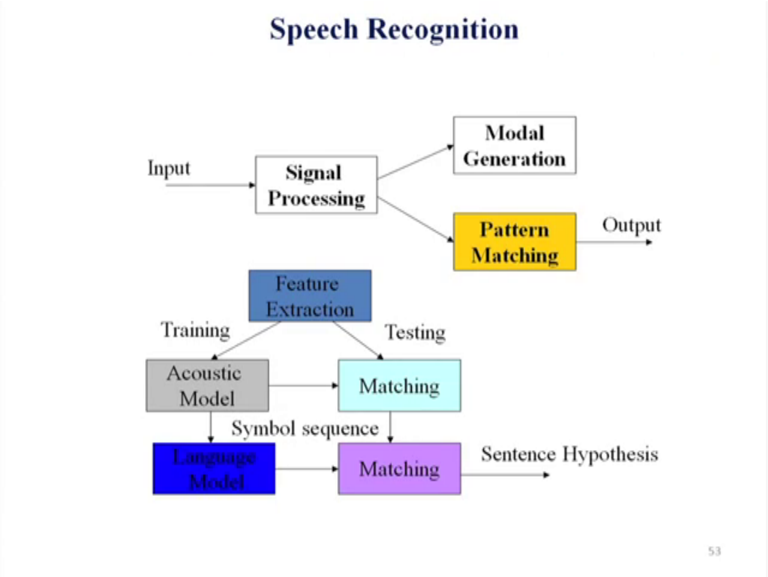
Please understanding in different issues. So, that issue I will come. So, ASR means I provide a input which is nothing but a acoustic signal, and from the acoustic signal it should convert the linguistic information, only the linguistic information like the word which is contained in this utterance, or sentence which I have spoken. So, either I based a single word may be a sentence, or several words together w 1 w 2 w 3 w 4 may form a sentence any from sentence. I can say I can say that open this folders or I can say opened. So, either it may single words or it may be a (Refer Time: 02:49) of words. So, that should understand by the machine that is ASR.

Now, application of ASR there is several application. Challenge that most common application if you see in your mobile phone, there is an ASR if you say home then it will dial home. If you say friend if you score in number friend number then the mobile dial friend. So, there may be single words common recognition. Or if you if you seen in the speak in the Google search engine there is available. In the Google search engine if you say hurt this or that restaurant name then the Google will search the restaurant name. So, you the restaurant name and show you the restaurant name or restaurant place and other things also.

So that means, it provides a hand free operation with the machine. Or it is a man machine communication man communicate with the machine. So, that may be a command and that may be diologue that may be a even may I prefer that something like that the name drawback in Indian language, contained in internets if you see the typing of Indian language is not So easy because it is component script. So, if I develop ASR I just speech on the quantum the machine, machine will type what I am speaking. Then it is first operation. So, man machine communication even somebody does not know how to write, but the noise speaking we got the spoken language is the natural language for the human communication.

So, even I do not know the how to write the language I can speak. So, if I speak front of the computer and computer understand what I what, then nothing like that application. So, all kinds of application different kinds of ASR is required, but basic technology how it develop the ASR that difference the some acoustic signal I to find out the word what I am saying. So, any how I have to do that things. So now, if you see the basic idea that what is speech recognition, that there is a speech signal I have to find out the next one w 1.

(Refer Slide Time: 04:55)



Speech Recognition

Input

Signal Processing

Modal Generation

Pattern Matching

Output

Feature Extraction

Training

Testing

Acoustic Model

Matching

Symbol sequence

Language Model

Matching

Sentence Hypothesis

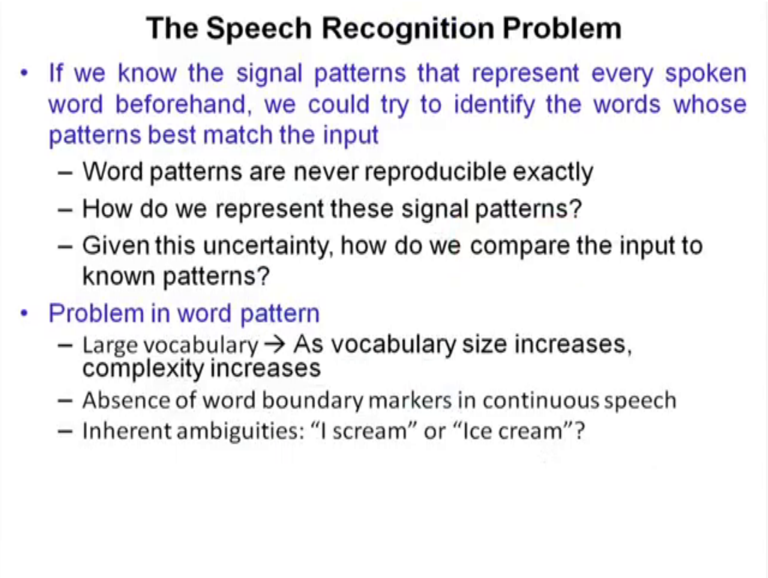
53

So, it is nothing but a that I should know which word I have spoken. So, machine should long that. So, machine should long the words with the acoustic signal. So, from the acoustic signal of the training data, I run the machine and which is called model generation I will learn the w 1 model is generated or machine is learn, then if I said that word based on the learning of the machine, machine should math the pattern and say you have said these word.

So it is nothing but a pattern matching algorithm. I can say language model all kinds of things are in come later on. So, nothing but a acoustic signal machine will how the child is develop that (Refer Time: 05:46) recognition system is developed. So, this speech perception not speech production. So, when the child learn learning some vocabulary you see he learn he listen it and time and time listen it and then it is memorize. So, you learning that take one word by one word and then once he learn then he can whatever I say he can recognize. Similarly machine also machine helps to be learn the would wall. And then what sentence phone whatever you said and then machine helps to recognize from the input signal of (Refer Time: 06:22).

So, that is ASR?

(Refer Slide Time: 06:26)

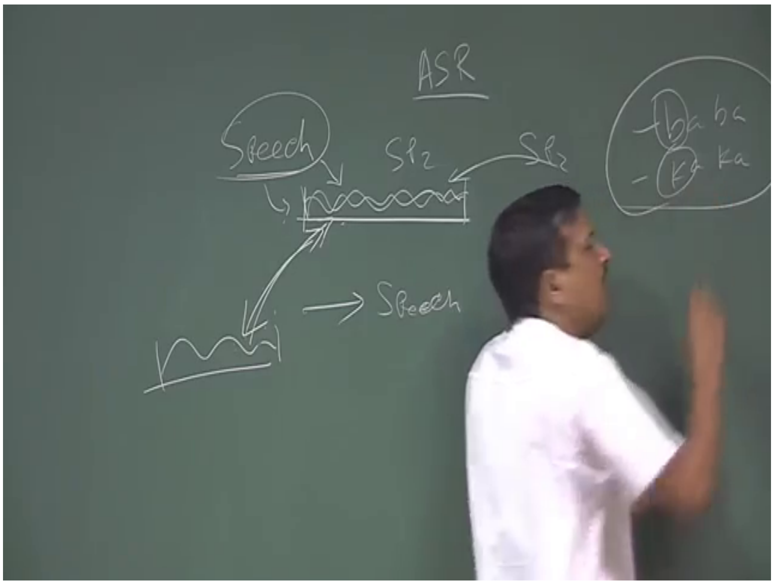


The Speech Recognition Problem

- If we know the signal patterns that represent every spoken word beforehand, we could try to identify the words whose patterns best match the input
  - Word patterns are never reproducible exactly
  - How do we represent these signal patterns?
  - Given this uncertainty, how do we compare the input to known patterns?
- Problem in word pattern
  - Large vocabulary → As vocabulary size increases, complexity increases
  - Absence of word boundary markers in continuous speech
  - Inherent ambiguities: “I scream” or “Ice cream”?

Now, what is problem? If I say let us think about only simple word recognition problem. Suppose I spoke a word let us speech.

(Refer Slide Time: 06:38)



S p e e c h then I sink there will be acoustic wave form of the word speech. Now let us machine learn this or machine store this then I said speech, then there are acoustic output and (Refer Time: 07:03) match to these acoustic what that machine is already is stored acoustic signal, and if the signal is match then I say I said speech. So, speech that word is stored and then there is A. So, how why match the signal that is certain matching that are come later, know how I match the signal. That is a I or pattern matching artificial intelligence pattern matching classification h b m neural network all kinds of things are there pattern matching.

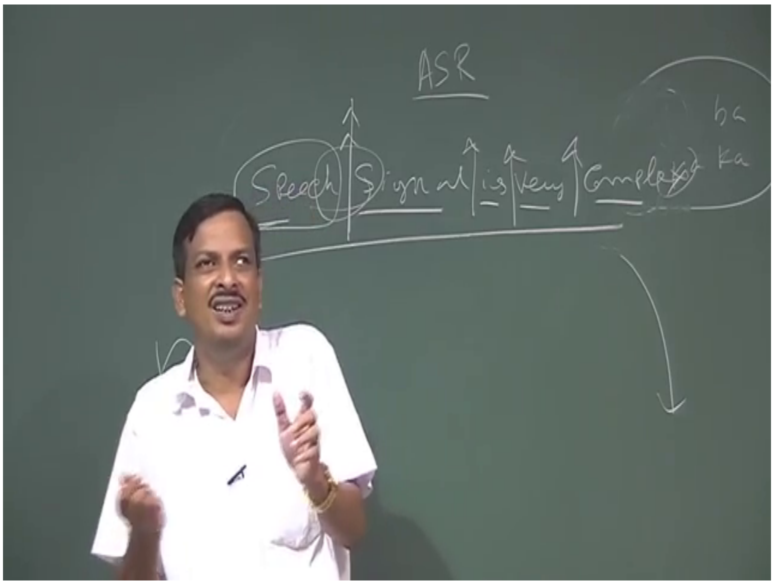
I will not going to the pattern matching directly. Let us say there is a speech signal is there speech spoken how of the speech is there that mean acoustic signal of speech. And I if I take speech it compare with the store machine store acoustic signal, and any how clam algorithm is used and if match this 2 things and give the output speech.

So, purpose is solved, but the what is the problem? Problem is that if you said speech this time if I next time I say speech the 2 waveform are not identical because machine cannot or man cannot produce to identical waveform. Or this time I said speech this next time I said speech. So, there is a length of the word has also a problem. So, that problem to do not kind of variability is the problem. Now I stored only speech first speaker one if a speaker 2s and speech then also have a problem. So, that uncertainty how do you compare the input will loan pattern that has an problem. How to solve this kind of problem? Second thing is that what pattern matching large vocabulary how much how many word I am stored, but any language there is a huge number of words. If I stored many wave words and then there is a probability of becoming same word the you can say the politically similar word will be increase have a sense.

Suppose I said as I give you in that problem the problem in I think in week one lecture. The best on the manner of articulation and identify the word, but if the similar kind of what are there like baba and kaka is there, baba and kaka this kind of word. Then I said the different between the 2 word is only this is ba and this is ka, but if you say ba is also a stop consonant ka is also a stop consonant. So, identify between these stop consonant by only deeper by only place of articulation. Because this is also voice occasion period is voice, but this is stop consonant this place is bilabial this place is less mailer. So, how the identify the place. So, identifying the different between the 2 signal full very less. So, if I if my vocabulary side is increases the similarity what will be increases and accommodation will be fall down.

So, complexity of pattern matching will be increases, then absence of what boundary yes this is very important. If I develop the word model, but if is say I will go to calcatta, but I have the speech like speech signal is very complex.

(Refer Slide Time: 10:40)



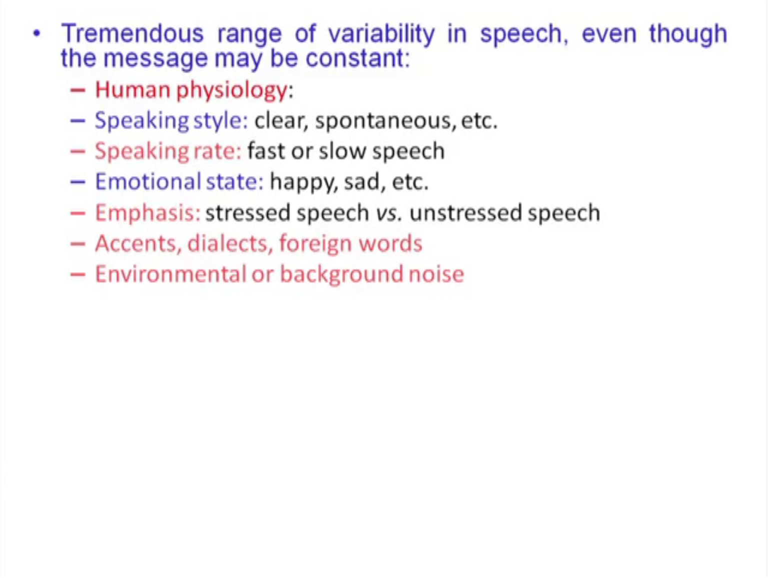
If I said speak signal, signal is very complex let us now if you say, beginning the word this is a word this is a word this is a word this is a word. In a written the word boundary are definite word boundary are there, but when a speech when I speak this one if I open that acoustic signal of this we will boundary (Refer Time: 11:10) this because this is a (Refer Time: 11:11) even these signal is very complex.

So, there is a no definite word boundary in the spoken language. So, I cannot say speech in a word how do I identify speech is a word from a spoken word spoken sentence, but if it is isolated what is I can no there only word is spoken, but if the quantity was speech what boundary are absent, but in case of isolated word also I said the single I cannot reproduce exactly same word twice. Then if I say there is a 2 word like that ice cream and ice cream. Both pronunciation are same ice cream and ice cream which were I want to said I do not know the acoustically they are similar.

So, those are the problem in speech recognition. There are other problem also. Human physiology, physiological change. Suppose today I say some word then I got coal and cup. And I say the same sentence, but the acoustic signal will be differ even there is a physiological change in happens in here. So, there will be different similarly the word spoken by a male voice and female voice in the problem because men for may be a a higher side male female men for may be a lower side female for may be higher side. So, all kind of variation will be there. Then speaking style every, every person at different speaking style the same word speak I say for speech somebody who has speech somebody who has speech signals are different.

Speaking rate I may said first somebody said very slow. Then emotional said if I happy my signal will be different if I sad signal will be different emphasis. English is a bound stress language I will come that later one bound stress less, but bengali is an english is a contrast (Refer Time: 13:16) language bengali is a bound stress language. So, if I am saying what about the English I am pronouncing it is not exactly bound stressing list. Similarly emphasis and there is a accents dialects foreign word variation, if see the dialects and variation very complex issue. Same word if there is a dialects and variation the pronunciation is different accent different.

(Refer Slide Time: 13:41)



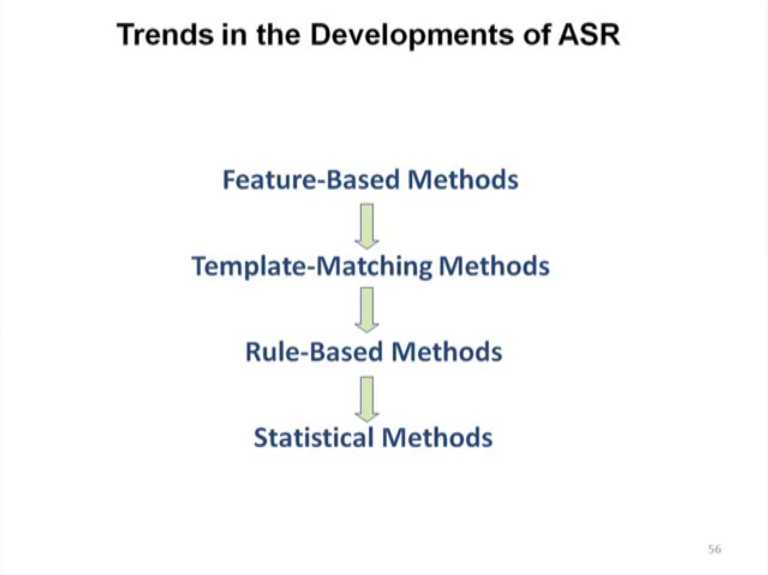
- Tremendous range of variability in speech, even though the message may be constant:
  - Human physiology:
  - Speaking style: clear, spontaneous, etc.
  - Speaking rate: fast or slow speech
  - Emotional state: happy, sad, etc.
  - Emphasis: stressed speech *vs.* unstressed speech
  - Accents, dialects, foreign words
  - Environmental or background noise

Foreign word, if I say suppose there is a Bengali parts and his pronounce zoo, how the pronounce the pronounce zoo?

If it is spoken by an you can say the (Refer Time: 13:53) look at it performance re pronounce zoo, but a bengali person who first language bengali he said zoo zoo. Although it is a foreign word, but when I pronounce I pronounce in my style not that what about said by that foreign language style. So, that all variation will be there same thing there is happen that I developed a english in the english recognize here based on that let I take that one kind of dialect dialectal english like that. Let us I tend that using tmit it data base you know the data base tmit it data base I trend a ASR, then I if I trend it by tmit it data base and then I the try to recognize the bengali person whose first language bengali as speech as spoken in English.

Then I there is a lot of miss conception is, there means recognition is happening what are what error rate is increased. So, all kinds of accent dialects foreign word variation is environmental and background noise I recorded speech in a studio environment and on. I testing is I testing with my mobile phone I recorded the speech in student environment and machine is trend that and after when I provide in the machine the test with that is recorded on a factory environment. So, there are lot of background noise how the how the eliminated the background noise? So, all kinds of problems are there in ASR that is why there is a you see there are lot of long history of ASR development. So, history if you say then this is the history.

(Refer Slide Time: 15:26)



Trends in the Developments of ASR

Feature-Based Methods

Template-Matching Methods

Rule-Based Methods

Statistical Methods

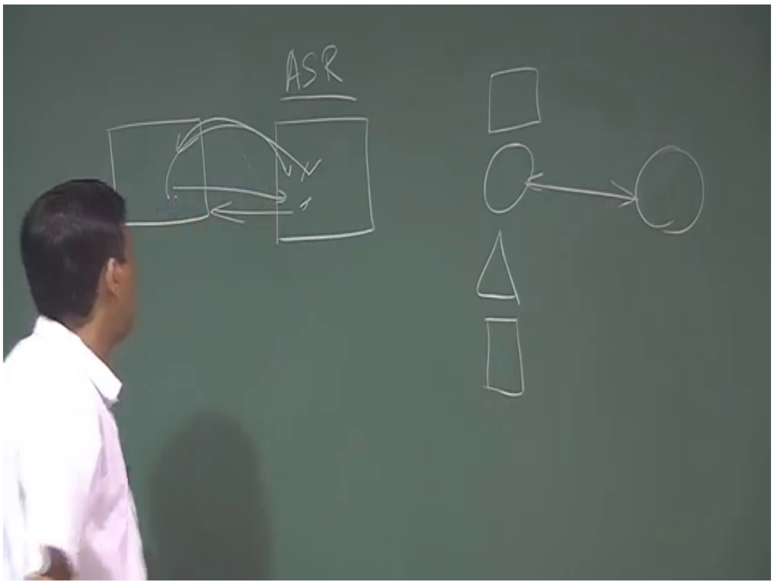
56

So, trend started with feature based methods then template matching method rule based method today statistical methods is most usely methods, and right now there is a research trend in brief neural network based a subsequent.

People are trying to develop the subsequent using the deep neural network method. Every method has an pol and pol statistical method rule based method template matching method.

So, I just in just go one or 2 method, so that if understand the differ issue of the ASR. Let us template this method this let us I describe the template base method. So, what is there which is nothing but a template means. So, template means that suppose I have let us let us say the I have a rectangle is it template.

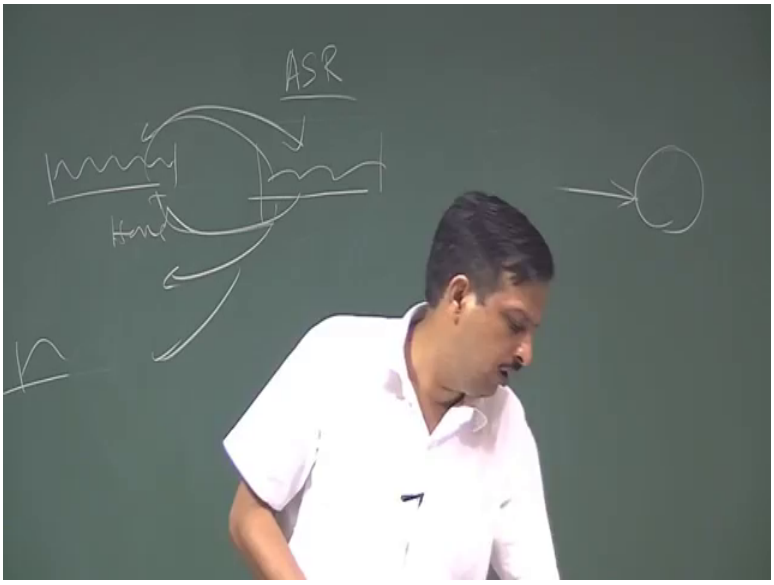
(Refer Slide Time: 16:20)



Now I produce similar rectangle, and then I say this is a rectangle because this match to this. So, this is a template which is store in computer or store in machine, and I have driven at a test template this one. So, they map store template with the given template if their match sound match then declare this given template is nothing but a this because this has matched.

So, I have a let us I will in machine, I will have a rectangular template then circular template triangle template, and let us like this kind of template I thing then if I provide a test template with circular. Then we try to match to with this we will match then say it is a circle because the template is said circle. Similarly in ASR further said plus there is a some words which is template base matching is word using in word base like that whatever you doing in mobile that when you say home, when you record a when you store the home at the number these are pronounced home. So, that home what template are there in the mobile phone.

(Refer Slide Time: 17:32)

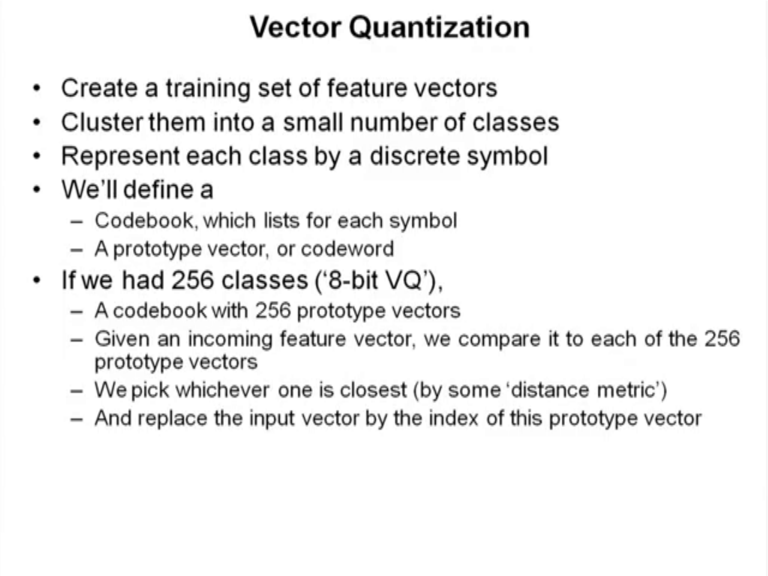


Now one I produce home in microphone and I give this acoustic signal to here machine, there machine try to match further this template match to this template onward.

So, there may be a home, there may be a hane, there may be a likes school all kinds of what is are there. And once I get this template try to match with template of all words and the best matching template will declared you have say this one. That is called pattern reorganization matching between the 2 things is pattern reorganization. So, how the match there is a different kinds of algorithms are there they are easily initially use makes are quantization, dynamic time working hidden mark of model and then is today neural network, then art artificial neuron network, then there is a today deep learning will be week topic to let me using deep learnings for template matching, or any kind of specification suppose vector machine.

So, any kind of specification technique I can find out whether the 2 templates or match or not. There quantization initially this was very much use in speaker identification.

(Refer Slide Time: 18:47)



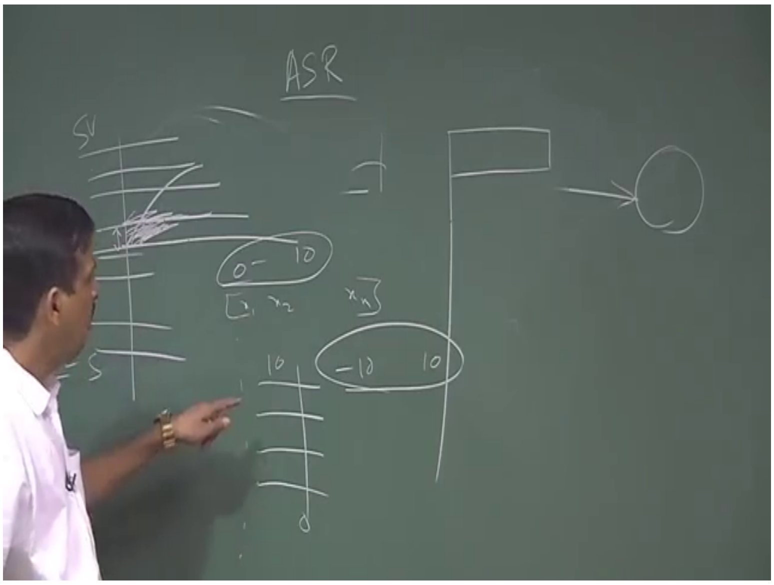
Vector Quantization

- Create a training set of feature vectors
- Cluster them into a small number of classes
- Represent each class by a discrete symbol
- We'll define a
  - Codebook, which lists for each symbol
  - A prototype vector, or codeword
- If we had 256 classes ('8-bit VQ'),
  - A codebook with 256 prototype vectors
  - Given an incoming feature vector, we compare it to each of the 256 prototype vectors
  - We pick whichever one is closest (by some 'distance metric')
  - And replace the input vector by the index of this prototype vector

And last week I will take one class or single identification because I will just describe the problem, and then I will the I will to say the how the vector quantization is used in there. So, basically vector quantization is nothing but a creating a you can say that training base someone if I said speech, speech, speech, speech 10 time. Every time there is a different speech I produce.

They should I store all trends speech in machine, or I can say from the all trend speech manage a representative vector that is speech. So, so (Refer Time: 19:28) quantized what the issue is same as quantization, issue the quantization issue is there if you remember in my vector the Digital quantization class.

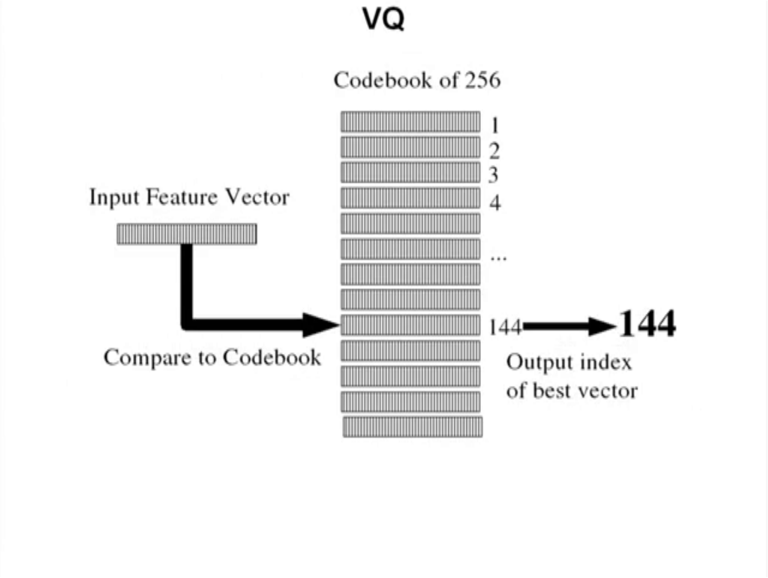
(Refer Slide Time: 19:39)



What I said I have I have say I divided the let us first 5 volt in let us 8 bit then I divided this plus 5 volt to minus 5 volt in 266 level and each level changing by single bit if it is sign then 127 plus and 127 minus. And I said that all this voltage level this delta we may be presented by only single representation that line. So, what about the voltage within this limit may be I say it is fall in here one 6 quizzer, then I say little represented by here this line. So, I quantized that continuous 5 volts you some sort of a level. So, same things is happen. So, suppose I have a feature vector x 1 x 2 up to x n it is a feature vector I have collected all kinds of feature vector, then I said x 1 has a variation from let us it is a x 1 varied from 0 to 1.

I said generate high vector within the variation of 0 to 10. So, I can say 0 to 10 is divided in 5 level. Or I can say 0 to 10 let us variation of x 1 is from minus 10 to plus 10. So, 20 is the variation why say there is a 20 variation if there, let us equally divided that variation with a 5 then 5 division will be state. So, that is vector quantization if you read this slide you can understand that codebook and vector quantization details I will discuss later on in a speaker identification problem.

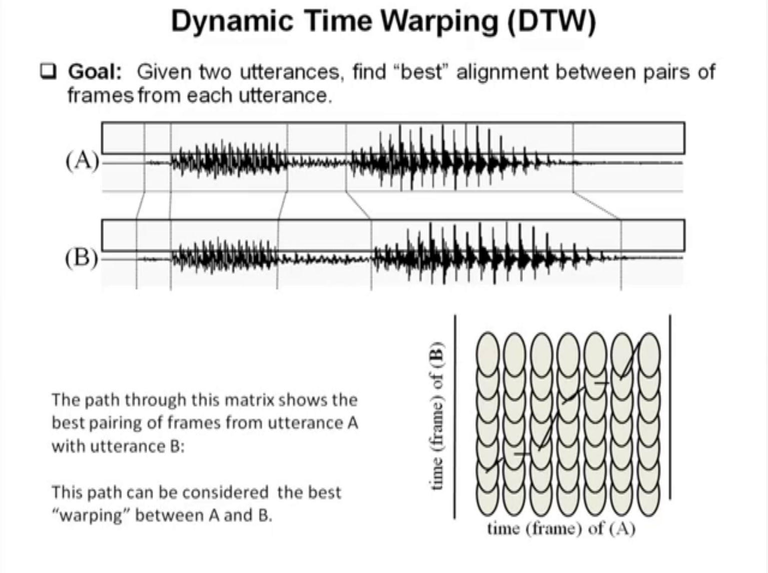
(Refer Slide Time: 21:14)



VQ

So, it is nothing but distance measuring.

(Refer Slide Time: 21:23)



Dynamic Time Warping (DTW)

❑ **Goal:** Given two utterances, find "best" alignment between pairs of frames from each utterance.

The path through this matrix shows the best pairing of frames from utterance A with utterance B:

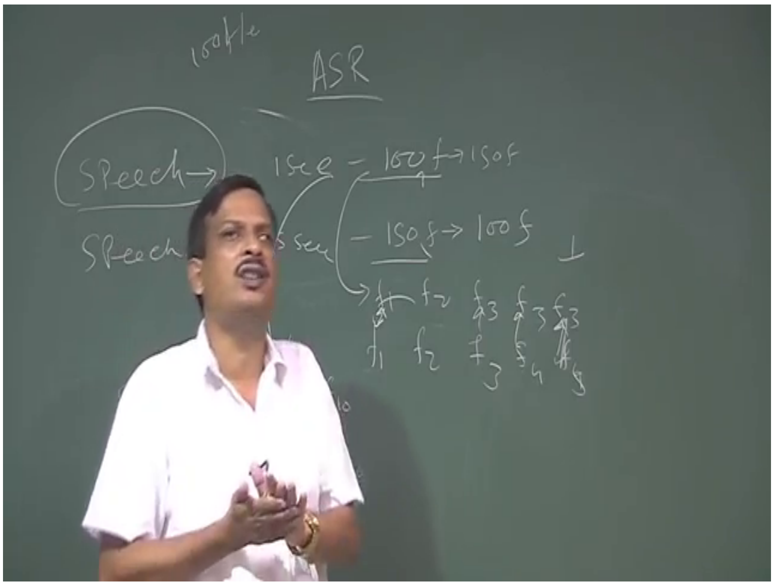
This path can be considered the best "warping" between A and B.

Important issue where the speak recognition is start. Dynamic time warping, which is very important. It is leads to that you can say that dynamic programming from here which is started. So, what is the problem?

Problem if you see there is a 2 utterance A and B, let us first utterance is let us it is a speech of something else something is utterance of one onward then wire then unwire then wire like this. Same word it utters 2 prime the length is differ there is 2 different length. Now if this one is storing my machine, and this one is I want to match with this distinct there cannot be match. Because I cannot match whole over to the whole over what I say and comparing let us 0 to 10 millisecond of here with here 0 to 10 millisecond, then if this portion I compare with this portion this portion is high, so if say no this cannot be the word. So, this is not correct matching have you understand or not.

Suppose I have a look forget about this example, suppose I have a speech.

(Refer Slide Time: 22:43)



This length of this signal is let us 10 millisecond, let us say one second. And next time I pronounce speech I pronounce the length is 1.5 millisecond. So, if I frame it what about the framing I done parameter expression I done, let us p c parameter for 100 frame per second 100 frame per second. So, this give me the 100 frame this give me the 100 and 50 frame. Now I try to match 100 frame with 100 and 50 frame. So, there will be time alignment, if there is a mix match in time alignment. So, I made compare this signal this portion of signal with this portion of signal, but see this signal can compare only this signal. So, what I want the 2 sequence, 2 number although there is a number of framing difference, but I want they the frame should be compare with similar type.

So, I want either this 150 frame should be up to 100 frame, or I can say this 100 frame representation can be up to 150 frame. Any one I have to do so many frame will be same I can say frame one let us frame one is compare with frame one frame 2 compare with frame 2 of this one and then frame 3 of this one is compare with here frame 3. Next time also frame 3 is compare with frame 4 is compare with frame 3, then frame 4 is compare with frame 5 is compare with frame 3. Until unless the distance between the 2 frame is very high. Then I can say where I can say this is the 100 frame 1 to frame 100, but once I compare this is frame one to frame 150 within that time all because I have not jump.

So that is his dynamic time warping. So, time frame of this signal is wrapped at for the whole data. So, from they are the best I want from all if you see the here time sense a and here time sense b I want the best possible match path and maximum, I final the distortion between the 2 things. And if this is within the limit for trend jump will be not happened. If the outside the limit I said the frame jumping happen here within the limit. So, frame jump is not happen frame jump is not happen frame jump is not happen. So, that way I can time work detail you can go through. So, this is dynamic time warping. So, template match with time alloying with testings. So, it is can cater to the radiation of speech rate, but it cannot it I do not know whether it is cater to speaker variation or not.

So, all kinds of other kind of technology I have to use. So, there is a lot of techniques of ASR are developed lot of sophisticated pattern matching techniques are developed is VM support vector machine there is a or neural network, there is a low now today's today in all research or went to they deep neural network. DVN deep neural network these neural network So, all d n n. So, all kinds of things are happening. I have not reading out the slide. Now next class I will discuss this one. That state of our speech recognition system what is they are problem? What are the research issue of they are in state of our speech recognition problem? Issues and there is a less resource what is the requirement of this kind of statistical model. And what is the problem in (Refer Time: 26:57) language what are the alternative solution what are the research going on this lesson I will discuss in the next lecture ok.

Thank you.