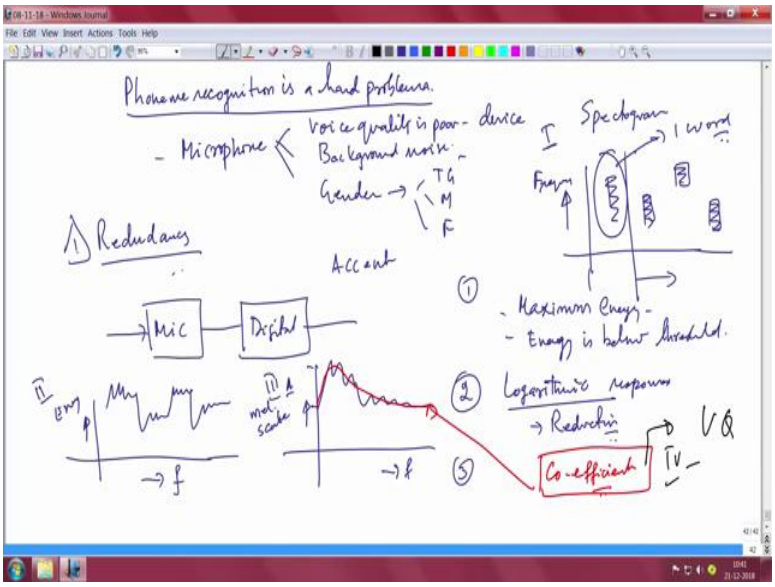
Advanced IOT Applications
Dr. T V Prabhakar
Department of Electronic Systems Engineering
Indian Institute of Science, Bangalore

Lecture – 25
Speech Recognition Part – 2

Alright, so what we will do now is we will try to go a little more practical in trying to try to sort of build a acoustic model. And we will try to do Speech Recognition; which means if you want to do that you have to build the acoustic model you have to build the language model and so on. There are little bit I would say at the moment a little bit abstract let us come to real brass tacks; what is the real issue in all this speech recognition problem.

(Refer Slide Time: 01:02)



I put down one line here which says phoneme, phoneme sorry phoneme recognition is a hard problem. The question is why is it hard ok? If you ask yourself that question you will see that starting from the microphone ok. You start with the microphone, you may have different qualities of microphone each one giving you different sensitivity levels. That means, the voice that is recorded or the voice that you want to or the speech that you want to recognize itself is because of poor quality electronics.

So, voice quality is poor because of the electronic or because of the device itself device related. The second thing is there can be lot of background noise I mentioned to you

already this problem ok. There will be already a lot of background noise. It depends on from where you are trying to do this speech recognition; are you in a mall? Or are you in a railway? Are you in a train which is moving? Or is it a bus? Or is it an airplane?

So, all kinds of problems are there for you to recognize that speech so there is a second thing is background noise. Third thing is gender ok; you can have transgender voice, you can have male voice, you can have male voice or you can have female voice ok. And then you can have accents a c c e n accent c e n t accent right. Accent can be another thing banana there is a banana then somebody else is banana right.

So, if you take these the FFT of that which is essentially the spectrogram as we call. The spectrogram if you take the spectrogram which for the word banana you will see several different types of spectrograms for the same word; which means some people some people are you know sort of emphasizing on some words and some people are not emphasizing on the other words.

For example, if it is a banana you are not emphasizing here on the starting ba right. Other and some other person is a banana so there is a lot more stress there. So, the spectrogram itself will look very different so the energy content in this spectrogram is going to be; is going to be very is wide is going to be very wide. Now what is that spectrogram actually contain? It just contains the word banana.

If you want to do speech recognition irrespective of any accent, irrespective of any gender, or background noise, or any kind of microphone system that you are actually recording the voice. You should be able to decode the word banana correctly. I am just using the word banana because it is that is what we started with ok. So, it is nope not no bias for bananas.

So of course, I like bananas though that is another story. So, here is the hard problem; so you have to keep that in mind that this is what you want to do; which means for sure for sure there is going to be a lot of redundancy that this word the spectrogram actually contains ok. So, I will use the word redundancy, I am just going to mark it redundancy.

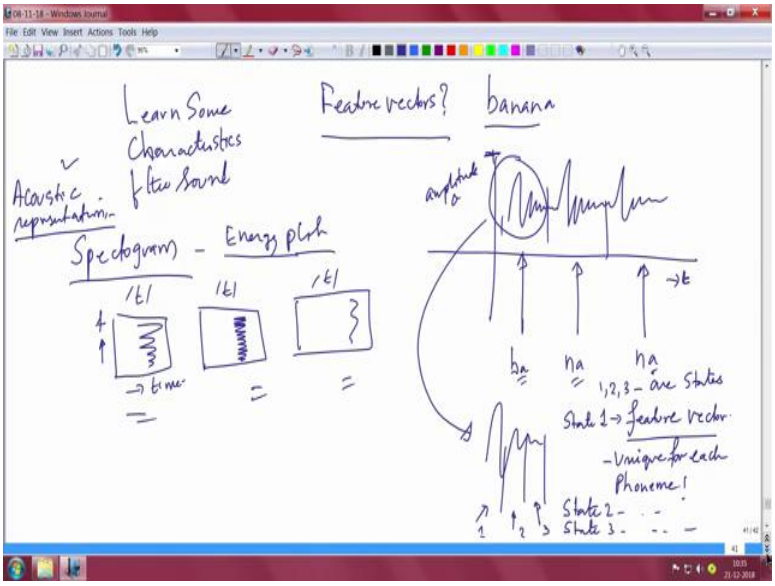
So, what is that redundancy all after all the word is bananas; each one is trying to use different accents, and stretch the word, and emphasize, on some parts of the phoneme, and de emphasize and other all kinds of things happening. So, therefore, there is going to

be a lot of redundancy. The one way that you can do reduction in any you know if you want to reduce this into something that is useful for any processing you want to do.

You need from the mic, from the mic you capture the analog signal and you convert it into some form of digital right digital signal you can you do. So, you have an analog to digital conversion block; which now gives you a digital signal which essentially contains this word I just taken a word as an example this word called banana.

How do you remove this redundancy that is the question. And I also told you that the word itself the spectrogram looks so different. Spectrogram let me show you again what the spectrogram actually has; it has energy components in it and it has the frequency component.

(Refer Slide Time: 06:02)



We already I think I should pick up the previous pictures. You have the frequency component I showed you all these different spectrogram for the same word te right. So, essentially all this energy out there should be you should be able to pull out. I know all the redundant parts you should be able to pull out.

So, what they do is they do an FFT and they try to sort of look at what part of it there is maximum energy wherever there is maximum energy is the it is where the actual phoneme is located and remove those which do not have too much of energy. So, one is

you have some sort of a threshold mechanism and remove energy which is below some sort of threshold below some energy is below threshold you remove it.

This is one way of doing reduction, but that is not going to take you very far. The second thing that you should perhaps do is you will have to after all when you talk when you do speech recognition; it is for the human ear ultimately you should be able to should shoot the human ear. The ear itself has a logarithmic response. So, you have a logarithmic response ok. So, in this logarithmic response of the ear when you if you have to adjust what you do is low frequencies you boost high frequencies you cut off ok.

So, that kind of energy adjustments you will have to do not only pick significant parts in the spectrum in an energy spectrum, but also do an adjustment of the logarithmic adjustments of this whole to suit the human eye you put it on a logarithmic scale. And then you will see that there is a significant reduction. Now, that is so what they do is. So, so this is the second step the third step is; so let me just show you some let me actually show you draw a few pictures so that might help you to understand a little bit.

So, let us first start with this spectrogram I have already shown you this let us say a word or a sentence has this energy here and then there is let us say some other energy here and some part here ok. So, I am just taking a large word. And in that what I will do is I find it difficult to work with this long part. So, let me just take one simple; let us say this is a sentence this is a spectrogram of a sentence I will just take one word in that one word I have taken take one word.

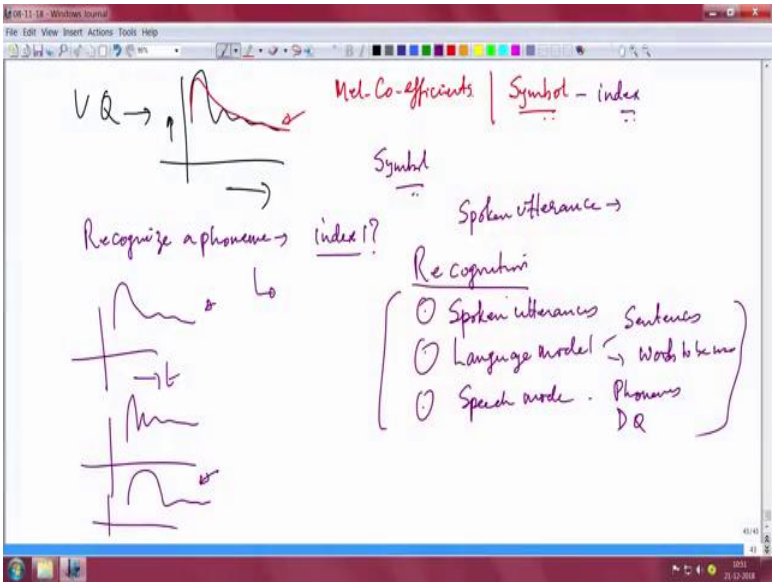
Now, you take this one word and see what all you can do. Look at the energy plot of this one word you will find that the so this is I will put it in roman I so, that you will get to understand what I am doing. So, one I am going to two what I will do is I will plot frequency on the x axis and I will plot energy on the y axis ok. Then I will get some energy then some reduce energy then some energy here and then some energy here ok.

This is what I got with respect to this energy component. What I will do is in the third step I will do further reduction in this redundant data by fitting I as I mentioned to you put it on to a log scale which is suitable for the human ear. This log scale is also called the MEL scale MEL scale I have already mentioned to you that feature extraction is essentially getting to extract the coefficients will come to that point very soon.

So, from here what you do? You fit a system like this you get you get to a again frequency and some emphasis of a low frequency is emphasized as I mentioned because this is what human ear does and certain lower frequencies you are reducing in so this is amplitude right, so this is the amplitude. So, you are emphasizing on lower frequencies and you are reducing on higher frequencies. Now it is easy that I have to do a very simple trick to get this waveform into a nice I will show you with a red color; what I will do is? I will do this I will fit a curve ok.

If I fit the curve any curve fitting kind of activity leads to coefficients right. You should be able to end up with certain coefficients. Moment you get these coefficients you can see now your original voice is gone your original phoneme is gone it has got converted into some form of coefficients alright. I would not stop there I will not stop there what I will do is if I know the coefficients if I know the coefficients I will do another step. So, this is 1 2 3 and I will do another step that is the fourth step and that fourth step is called vector quantization it is called VQ right.

(Refer Slide Time: 11:57)



What does that VQ do? Pretty straightforward VQ simply says if you get let me put back the picture so you got sorry let me put back I am not so good in drawing, but I hope I will be able to. So, if I am able to fit a curve like this which essentially relates to MEL coefficients now I will use start using because this is a MEL range right this is a log scale MEL coefficients ok.

I get some coefficients, I will take these coefficients and map it to one symbol that is all some number symbol means one number that is all nothing else. So, what has happened to your phoneme now your phone him I simply become a symbol. What have you done you have actually reduced tremendous amount of data into just one symbol hmm. Now the thing is it is now you can now you see the game what you do you give it to some symbol or I will simply call it some sort of an index.

Ok it is some sort of a index now it is quite straightforward. If you want to do recognize something if you want to recognize a phoneme you want to now let us start you want to do a recognize a phoneme. So, your task is recognize let me write it a little better recognize a phoneme ok. For you as an electronic engineer that phoneme recognition means; you are to get to the right symbol you have to get to the right index you have to get to the right index. What is the index? Is the question.

Now, so you can do a simple search and then you list what are all the closest matching symbols to the existing spoken utterance. You would have the spoken utterance spoken utterance ut te ra u tt er u tt er utterance which is an analog signal. You do FFT, then you do log scaling you get two coefficients, and you do vector quantization, and you end up with a with you end up with a number ok.

Now you say hey I got this number this guy is talking something ba na na how does this ba match closest to which vector symbol does it match to because you do not know right you all you are seeing is this word ba and some index you got you were to start matching this index appears closest to this index. So, what you are trying to do is you will take this waveform like this which is what you saw right. And you will start matching to those which are like this and perhaps to something which goes like this.

And you say hey there is a good probability that this guy matches to the original this is the one that seemed to match to the original. So, you will assign some probability and say this is the probability association of that particular vector quantized number that I got so this is first part of the problem. So, the closest to the input is what you want to arrive at and assign it certain probabilities ok; this is part 1.

So, if you look at the overall idea this digital waveform you want to do you know you want to remove redundancies and all that. Clearly means that FFT does something some small reduction it does, then you do MFCC MEL frequency MEL frequency cepstral

coefficients you attain MFCC coefficients you attain features essentially features of the phoneme or the word.

See you have to be very careful where you can have features for the word the word can have can be split into phonemes. And you can do at phoneme level or you can do at the word level you need to do both in order to arrive at the word and you need to do a crossword so that you can form sentences right. So, this is when you will ultimately say that this is the spoken speech. So, you can think of a hierarchical situations where there is a sentence, sentence comprised of many words, each word contains a set of phonemes.

And each phoneme can have states ba ba is there is an emphasis somewhere there is a reduction somewhere. And then you have na is another phoneme it can have a stress somewhere na can have accent problem can have variability in a gender problem, can have na can have a problem with respect to background noise. So, all these issues will have to be considered. And then finally, you will come up with you know trying to match the right number which is the vector quantized number closest to the one that was uttered.

So, all this essentially is what I wanted to say about the phoneme part of the recognition. Now, if you want to get into another state another important step which essentially is about the point related to recognition. I will now come to recognition; you just did some matching and you got hold of some part of the some part of the phoneme by checking out your indexes. But all that has to fall into place because you are doing things in real time right.

You have to get to sentences, you have to look at words, you have to do this whole process. So, when you talk about recognition you need to look at spoken utterance, utterances. Then you have to look at the language model you have to look at the language model, essentially you have to talk about you will be talking here about sentences and the words that you can use and all that words to be used.

And you will be talking about what is known as the speech model right you need also the speech model. Essentially you are talking about phonemes and you will be talking about the DQ quantization and all of them vector quantization essentially leading to these symbols ok. So, one issue you may be wondering is why are we doing all this ok.

So, my take on this is; if you want to improve anything with respect to speech recognition the problem is what is it you want to use IOT for? You do you want IOT to optimize a solution for my voice or you want IOT to do speech recognition for speaker independent voices this problem will come. So, the way to improve robustness is another problem right.

Robustness when you say it should be able to decode under a lot of background noise as well. So, all these issues mean what is it you want to ultimately focus on to improve accuracy. Accuracy improvement is the key there is a metric for it which is called word error rates and all that, but keep the following in your mind. You may want to say I want to do speech recognition across all types of people I do not care.

Gender should not be an issue it should be banana whether it is Prabhakar or anybody else it should be able to decode well right; which means you are saying I am not I do not want optimization to any specific user or any specific speaker. Typically the speaker related improvement in accuracy is done when you want to do let us say a dictation right. You want a computer to you know type out something which I am talking then you may want to say I do not want I am not interested in any student or any of my projects associates to talk into it and do a dictation only Prabhakar's voice should be decoded correctly.

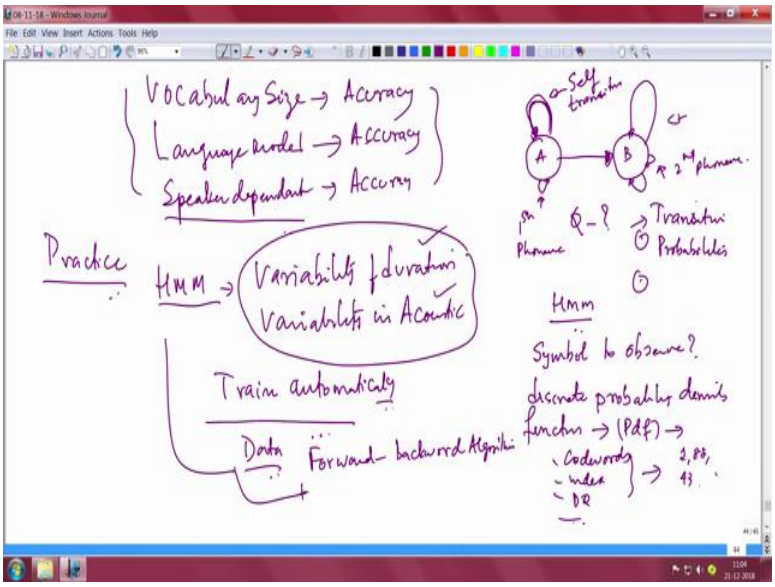
If you want to pick take that decision then you do it with you can improve accuracy by training the system into recognizing only my type of you know voice ok; that is one way. The other way is that is not what I want to do; I want to do telephone kind of speech where any two individuals are having a conversation. So, it is a telephone speech so you do not know whose are talking to whom and all that.

So, it is independent of the speaker. So, you have speaker specificity or specific decoding or you have speaker agnostic or something that is independent of any speaker related thing. So, any speech recognition systems will have to cater to these simple requirements of I would say so basic requirements, but will have to be done. So, so it should not be dependent it should not be independent and all that. Another way to improve accuracy is to keep the dictionary small.

Put a very small dictionary and work with only a small set of words in the dictionary ok; which can be very targeted for a given application. So, you use only that many words

and be done. With those words the set of sentences that can be found are also limited. So, the language model also improves the accuracy. So, you can have dictionary; that means, you can be talking about vocabulary size so let me put down that.

(Refer Slide Time: 23:29)



Vocabulary size improves accuracy. Then the other part is language model improves and because you have reduced the vocabulary size; the language model is also pulled to that size of words set of words that you can use will that also improve accuracy? Yes, definitely Will speaker dependency improves accuracy? Yes, differently speaker dependency also because your targeted application is for speaker dependency.

So, if you have speaker dependency dependent if you do you train the complete system that also will improve accuracy all these things have to be borne in mind. But in any general solution you will not be talking about doing this because you are looking at telephone speech. So, even if you want to do a speaker dependent or anything, you need a model which often people use in speech recognition and this is called Hidden Markov model. Let me give you an intuition for the hmm go back to the word banana ok.

The word banana you will be able to you know use as I said you have variability of duration ok. Then there is variability in the acoustic part variability in acoustics. And you need a system which will be able to train by itself train by train automatically. Any speech recognition if you want to put it in practice, you will have to use a good mathematical model.

What do you mean by this? I will show you how it works variability of duration take one state, take another state ok. Let us say you say banana, baaanana you say banana ok. This long one I have shown because this is not spending too much time in ba. Whereas, if you say baaanana you are this you know in that state time you spend in that state is much larger ok. Now, na come to na banaana banaaana banana small banana even smaller.

So, you see the amount the time you are spending also is a variability, but the word is the same so here is my problem. Remember one thing this arrow can never be back because you won't say naba you will say bana banana. So, this arrow always is forward only. Now you are asking the question what is the question you are asking; if I spend some amount of time in that state what is the probability of transiting to the next state baanana if I say that.

That means, I am spending more time my probability of transiting to b is considerably lower and I am spending a lot more thing in this state a. So, you will be talking about transition probabilities sorry. And you will also be talking about the fact that the self transition this is called self transition right this is called self transition. And this is the other one which is called the transition to the next state this is the this is that say first phoneme this is the second phoneme ok.

Now there is another problem the second problem is when you say baanana ok. One is the transition the time variability, the second thing is: what is the index that you can emit in a given state, what is the index that is another problem. In other words that vector that you got that number that you got rather not the vector the symbol that you got. What is the symbol? Because I initiated you into that symbol in beginning now you are back to the same problem you can have so many symbols coming out in that state a which is that symbol.

So, you are now asking for any implementation; in this model, in this framework, what are the possible symbols that can get emitted when you are in that state? And what is the transition probability to the next phoneme? And what could be those symbols that will get emitted there? This is what hmm is actually trying to do you have something that you observe and you are trying to ask questions related to the amount of time you spend in a state and the symbol that will get emitted in that particular state.

So, let me summarize this hmm and then move on; you are asking the question like what symbol you are likely to observe. That is the first is all say symbol to observe ok. Now for that what you normally do is you have a discrete probability discrete probability density function ok. And which essentially will let you know what are the probable code words. The PDF will tell you the discrete PDF will tell you what are the probable code words or people also call it indexes or some people say DQ whatever.

What are the possible? What is it going to be 1, is it 88, or is it 43. Or what is this range of code words that are likely to come. And that you essentially use a discrete probability density function and you apply. So, this essentially will tell you: what are the possible code words that will get emitted alright. Now, if you want to build the hmm model and you want it to. So, you have looked at duration and you looked at acoustics suppose you want to train automatically.

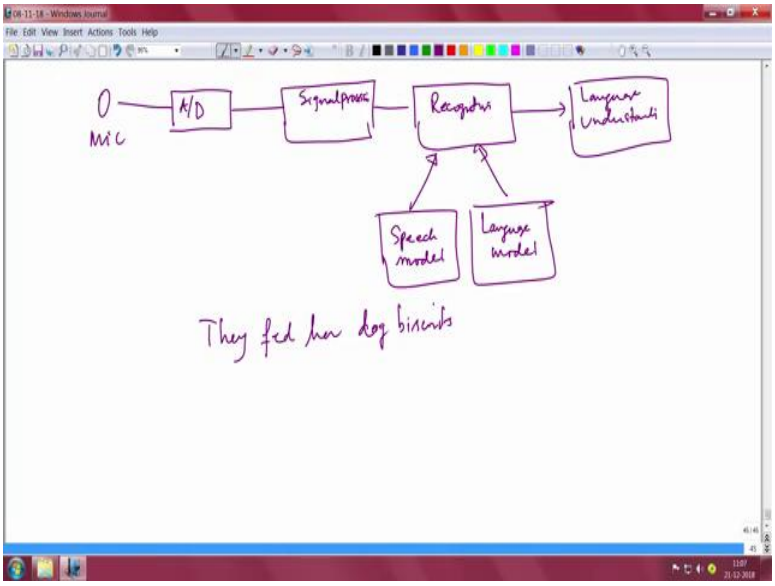
Let me tell you the only way by which you can train automatically is humongous amount of data. You feed a humongous amount of data sentences essentially and keep you know training that system make the system learn automatically. Because it is humanly impossible to catch hold of all this let us say seven billion plus people and ask them to utter the word banana it is not going to work.

You have to have a mathematical way you need a model which can learn automatically this word whatever be the way it is uttered across genders across a noise across you know all the hard issues that I mentioned to you with respect to all of them it should be able to learn that. So, you must get into some form of automatic learning. How is this done? That is the next question what algorithm will you.

So, you have a model no problem for the model. But how will you make it learn the way you learn is you use another algorithm for that and that is called the forward backward algorithm forward backward algorithm. So, you see Hmm in combination with this backward forward algorithm will solve your problem of variability in duration, variability in acoustic, and trying to learn automatically across all genders across everything and it should be able to print directly the speech that has been recognized so that is the key point.

So, in summary we can put down all the blocks that would be of great interest to us which essentially means you have a mic, you have digital conversion analog to digital.

(Refer Slide Time: 34:31)



Then you do this nice thing which I mentioned to you which is signal processing all that you know getting to coefficients and reduction in data and all of that and then you do recognition. And if you want to do recognition you will need assistance from the speech model. And you will also need assistance from the language model correct language model.

And ultimately you will be able to decode the speech here. Language, semantics, language understanding call it this is another story we will come to that if time permits. But semantic is also an important concept here. When you say you know these are really some things which I was looking up on the web to give you a meaning of you know to give you a feel of whole area.

I will give you a simple sentence I will write down the sentence I found this on the web they fed her dog biscuits. You can say; they fed her a dog biscuits; that means, they fed her dog biscuits. The other way to say is they fed her dog biscuits which is right? There nothing wrong in the sentence the lady's dog was given biscuits or the lady was given biscuits we do not know.

There is no way by which you can say anything about this because there is a perfectly good correct sentence. So, understanding the context under which something is spoken is also very important and these are very advanced techniques. So, we will not get into that,

but just to give you a feel that it is high time now for us to roll up our sleeves and try these things in practice ok.

So, let us go into the next step of an implementation. So, that you can try out a few things all by yourself in the creation of acoustic model, in the creation of language model, in the form of trying to recognize voice under extremely harsh conditions perhaps. But we will see what is actually doable given the limited time.

Thank you.