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

Lecture - 24
Speech Recognition Part – 1

So now, let us look at another important IOT application and that is related to speech to text ok. Why is this important right? So, that is the question you may have top of your mind. And why is IOT pitching into this aspect of speech to text. See the thing is if you want to empower; if you want to empower people of different you know let us say abilities different abilities persons with different abilities; then the only way to do this is to use a technology ok.

So, you on one said you can do all kinds of you can create rules, you can do hudtal, you can say equality and all of that is one part of the story. The other part of the story is the creative technology which is so easy to use such that it really empowers people. And they now are you know sort of well equipped and well informed. Whatever be the ability of the person so this is the most important thing.

Take a case where you have a audio system; you have an audio system in a metro train. But the visual system which also indicates the station which is coming has blanked out for some reason it just completely blanked out. So, no but there are announcements that such and such a station is coming. Now how will a person with a different ability we actually get to know what that station is right. So, it is not going to work for him or her.

Therefore, IOT should provide solutions of that nature you should be able to run very simple, but I mean very simple from an application perhaps. But a very sophisticated algorithm behind which has the ability to take that speech information. And then decode into textual form and then print it on his or her mobile phone ok. So, if you are able to do that then you have done something you are empowered people with persons with different ability to you know to actually complete their whatever required you are able to provide support for such people.

So, IOT plays an important role also in the case of empowerment. So, this is the key takeaway from why you need to do speech to text. It is not that speech to text is

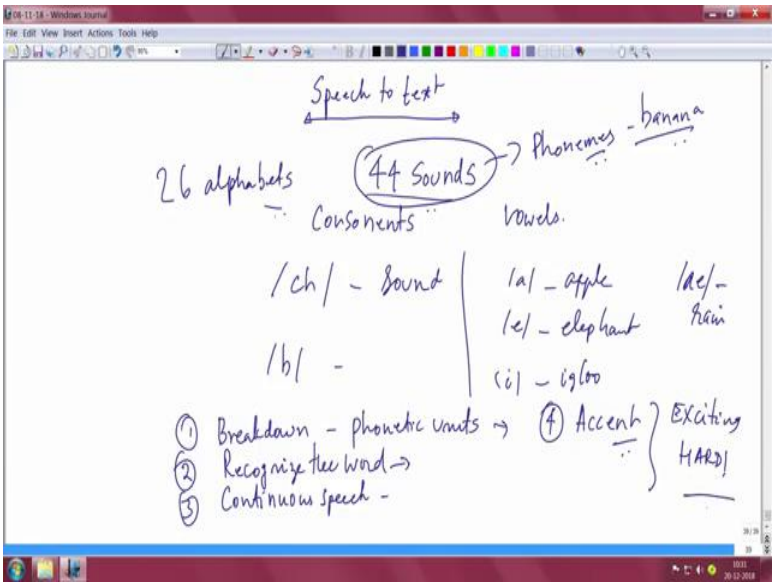
something new ok, so, here is another important thing people have been attempting to decode speech to text way back from the 1940 and 50 bell labs created something called Audrey.

And then from Audrey came something from IBM and then US government put a lot of money in trying to decode a speech. And so much of effort has happens 40, 30, 40 years 50 years people have been attempting to decode speech for a very very long time. It has as I said far reaching applications particularly in this current word and therefore, it is not a new area.

Why it becomes so popular today? Is only because of the fact that the accuracy of speech to text conversion thanks to modern ways of speech recognition has improved drastically ok. The traditional way of decoding speech is in one way and this modern way of doing it has actually improved the speech recognition it is a bit way.

So, let us get into some detail of why it did not work previously or why it had limited application previously and why it has become very popular now. Let us take this thing about what we talked ok, supposing you say banana right; when you say banana the whole English language right when you write on whatever you talk essentially comprises of how many alphabets you have 26 right.

(Refer Slide Time: 04:54)



You have 26 alphabets al ph a right, al ph a b e ts there are 26 alphabets out of this you have vowels and you have consonants right. So, that says clear you know differentiation of the vowels and then the consonants. But if you look at sound; if you look at sound the complete English sound maps in to 44 sounds; 44 sounds maps in to 44 sounds.

Here is the big problem; if every sound had an alphabet problem solved right problem is complicated because every sound is not mapped to an alphabet and therefore, this mismatch creates one definite hurdle for us. Take banana as I said with banana; banana is a consonant basically its a word. And it consists of if you take the first sound ba ba like that right say ba is a ba is actually a sound which comes.

Quite like that if you say fish that comes right that fish that is one type of sound right. So, like that you can have sounds coming from independent you know consonants or independent alphabets or a combination of alphabets also right. You take child right this is ch so ch essentially is a two alphabet sounds so this is one type of sound.

When you say banana b ba ba you say ba; so that is single alphabet sound. You can have de like dinosaur you say de dinosaur right or you can say ge like guitar is all single sound, then you say hand ha ha comes from here right hand you say so he he like that you say. So, these are all single consonant kind of kind of sounds.

So, all these sounds if you sort of sum up you will get about 44 look at also. So, these are all consonants you can also look at sounds which are coming from. So, these are all consonants conso co n s o n a n t s consonants right. You can also have sounds coming from vowels right, you can also have vowel sound vo w e lss vowels, you can have apple is essentially a vowel sound right you can say a apple.

You can also be talking of elephant a right you say elephant this is apple like apple and this is like elephant. You can also have so, aei you can also have for i can be like igloo e e e igloo right you say i igloo so it comes like that. So, you can go on for each one of these vowels you can also have combinations again there like rain ray ae ae you say rae.

So, a and e you can also have a and e combination coming from rain. So, you see all these sounds if you just do a summation you start counting them you will find that the it will be about 44 different sounds. And I already gave you the motivation these 44 sounds are not mapping into 44 alphabets they are mapping into 26 alphabets that is the

problem. So, you should have a way of sort of ensuring that there is a proper mapping between these things.

Therefore, this issue of speech to text recognition is hard because these 44 sounds also called phonemes pho n pho ne mes. These phonemes have to you know be recognized correctly. And after recognize the phoneme correctly banana this ba na na there is ba na na ne is there na again banana.

You must have a mechanism of splitting this into a particular sound map the sound of banana in to correspond recognize this sound first, recognize this sound and map it to the. So, you know the phoneme and if you know the phoneme you will be able to put down those two letters ba ba banana bana you will be able to put.

Again na na ba ba na na you should be able to split the phoneme. Here this phoneme is the sound from the sound you get back the letters put them together and make it into a word. So, then you will be able to write banana right. So, all this is something that by training the computer you should be able to put down this word banana right.

So, essentially speech to text is about able to get to you know recognizing the phoneme correctly and then creating a word correctly this is I would say a solved already there is no problem at all this you can do very simple tools are there and words you can easily get the trouble is not there. The trouble is in continuous speech and often when I am talking let us say you are interested in you know converting online whatever I am talking into sentences; which means there is a it is a continuous speech.

And then there are very difficult to say when is the first word stopping and when is the next word coming. So, that gap between word is also gone is diminished, which is a clear indicator, recognition of continuous speech has been a hard problem. And that is where maximum amount of algorithmic work has happened in the last few decade or so and so from the 40, 50, 60 are whatever has happened; they are all about trying to say I could get a 100 words correctly then someone would say I used a nice model and I was able to get 500 words correctly I was able to get 1000 words correctly.

That is not the game anymore it is not about how many words you got correctly it is about able to recognize the complete continuous speech sentences that are spoken at quite a high baud rate may be somebody like me. And you should be able to put down in

real time it should be able to put down that text directly right. So, that is why the hard problem is. So, the trouble with rather the motivation of this whole problem of speech to text is; break down of the word of the sound into individual phonetic units it has been hard ok.

Banana if you are able to so we are coming again back to the beginnings banana you should be able to get the sounds correctly the ba na na that sound you should get correctly. And once you get to that so that have already been hard there is that problem 1st problem. 2nd problem is related to putting it together because you can have even if you recognize it correctly that it not does not mean that you will be able to form the word right so that the problem is still there.

And therefore, putting down that word because na can occur in many many instances right. So, nest for a ne ne you say there so ne will come there, natter if you say natter na comes there na natter right. So, that so, is it banatter or is it banana? So, you be able to put down that word properly. So, therefore, phonetic breaking down into phonetic units is one part of the problem. Creating the putting it together into forming a word so recognize the word is the second part of the problem.

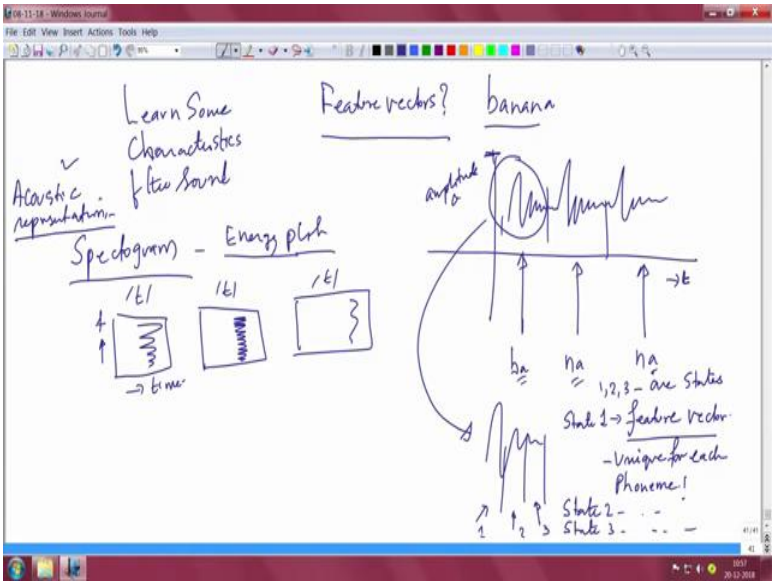
And the 3rd thing perhaps is this problem of high baud rate by people like us where we are talking about continuous speech recognition right so we should be able to do it in high speed. 4th problem perhaps is if you take South Indian accent we say banana in one particular way you go to some other regions say banana somebody says banana and another person says banana something like that.

So, the fact that accents are different creates a problem. So, accent a c c e e e nt accent is a problem. So, all these things of these four problems break down into phonetic units, recognizing from phonetic units to a word, continues speech, that is all the word should come in real time and taking care of accents has created a this whole area of speech protects as exciting as well as hard ok. Now, this is the motivation for IOT and this is a reason why it is important to learn this area fine.

So, now, the trouble is we already mentioned that this conversion from sound to a particular phoneme is not easy. But it is always good to get some feel right you want to know this phoneme representation; how does it look if you take a sound clip and take it inside mat lab and to do some speech processing on it use the speech processing toolbar

toolbox. And try and see how this phoneme a phoneme looks there is a nice representation that you will get a feel for and that indeed is called the spectrogram ok.

(Refer Slide Time: 16:16)



The spectrogram is nothing, but an acoustic representation it is nothing, but an acoustic representation of the phoneme. I have taken a single phoneme you see te te I have taken this te as a single phoneme. And you will see that this is an energy plot essentially we have plotted energy. What do I mean by that? I have plotted on the x axis time and on the y axis I have plotted the basically all of this is so I will say this is the frequency and this is the time ok.

This is nothing, but the time actually this is not required you can remove this. And you will see some intensities right some is going you can see the pattern this is as pattern a this is pattern b and this is pattern c. Clearly the same phoneme is represented in three different ways; the energy associated with each one of them clearly is different. Actually you can do this in you can take you can record speech and you can take and then take it inside any standard tool toolbox processing you will see this energy plot.

Clearly phoneme recognition is hard because this you can see that it looks different in the three of them correspond to the same phoneme. Therefore, it is important for you to realize that there are there is a process by which one should be able to get to phoneme recognition by just looking at the spectrogram it is impossible for you to make out anything just gives you a feel of the energy that is about it.

So, what is the engineering way to capture and understand the sound to phoneme that the next problem. So, obviously, you have to motivate yourself in that direction. Let us go back to our example of banana ok, how is the acoustic representation of this banana I put it down here; its a very poor representation because I am not good at drawing. So, you can see I have split this so again this is time and this is frequency ok.

I have not plotted the this is not the energy plot this is not a spectrogram, but this just the time series of the word banana ok; where you are essentially talking not just the different frequencies with which it is associated. It is also talking about how much so there is indeed a frequency component of this signal. And also there is a component of. So, I would say it should be to be more precise it is more time it is more about. So, this is about how much of strength is there.

So, I will say it is not only frequency it is also amplitude right this is also a frequency I could not see there is a repetition. So, there is also a frequency component associated with it. So, we can split this word banana into three parts; you have ba, na and na. This na and that na should look identical, but it does not look identical perhaps in practice also it does not look identical. For the simple reason when you say banana the second na has a much more bigger stress compared to the third now right so it should look different.

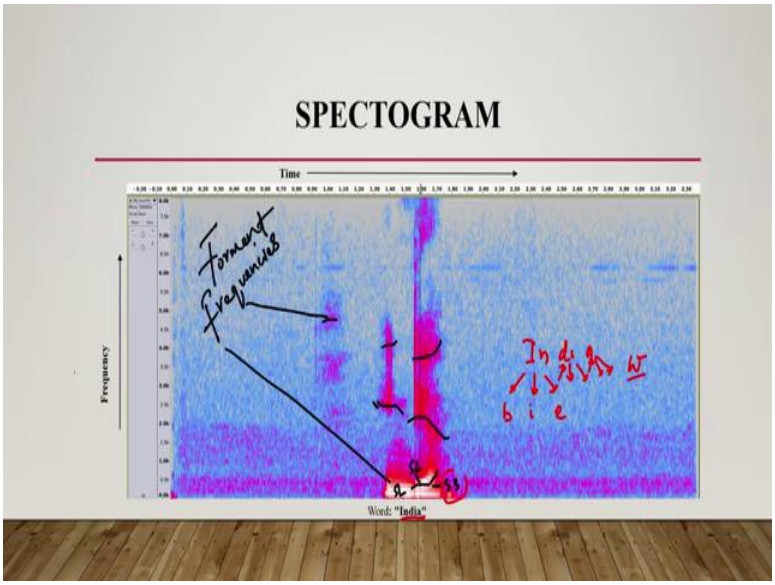
Therefore, this indeed could be a very close representation of the word banana. Problem is this ba right this sound that you hear you should be able to map it into the correct phoneme. How will you do that you take this first part this first part ba and again split it into multiple states I put it down here; state 1, state 2, and state 3. It turns out that each state has a feature vector which unique ok.

So, you have a unique feature you will have a feature vector which is unique for a given phoneme. And therefore, you can with certain probability say that this was indeed ba. So, that is a good nice thing about the fact that you are able to map a given phoneme sound that you hear into a phoneme with a certain probability and say ya this must be ba. Similarly you do that for na then you do also for the other na then you will get a set of feature vectors. You will see that there are three states for each phoneme that is you have three.

Then you have three for the second na then you have three other another three for the other na. So, you have 9 states essentially and all these 9 states put together give you

features with which they give you the feature vectors by which are quite unique and with certain probability you will say the word uttered indeed is banana.

(Refer Slide Time: 21:36)



So, our project staff Payal Rahul and Praveen have created a spectrogram which we just discussed of the word India you can see this word is India here. And this how it looks the x axis here is time, the y axis is frequency. And this color here actually is also indicative of the energy associated overall energy. We can see India right when you say India you have something starting with some energy peaking up all of this is all the peaked energy.

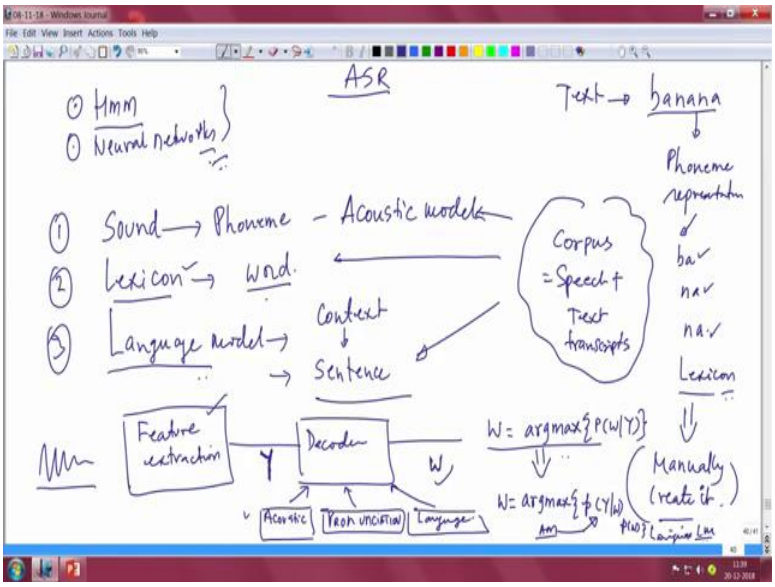
And then as you are as the word dies down India that lower part appears to be here. You can imagine that this essentially is captured is all the power that is there. And this spectrogram is very special to the kind of accent that people have used it is definitely based on the accent. However, this word India can definitely be split in to I would say nice sound to phoneme translation if you divide it into three phonemes.

Again in di a when you say India you can divide it into three and this in itself you can represent it as a beginning then there is a beginning stage there is an intermediate stage and then there is an ending stage. So, I would say b i n e all the three can be there. So, again going back to the same example like what we took with respect to banana. Here also you can have India where in starts like this which are with a little bit of energy peaking to some high energy perhaps and then falling down and the next phoneme picking up where this next phoneme perhaps is India e ya indi so this is picking up.

Now again and this again can be put into three parts and then the next phoneme can rise again here. And then it could also have three different states. So, you can have beginning, intermediate, end, beginning, intermediate and end, beginning, intermediate, and end. For this each of these phonemes and then arrive at the word you have to arrive at the word.

So, here is the next part its not going to be so simple as you think that if you get these feature vectors that use those feature vectors and arrive at that word it is not going to be so easy. So, what they normally do is they rely on additional tools to actually say this is the word or these are the set of words. Let us start with simple word. And so the goal really is to arrive at that word using taking assistance of several simple things simple blocks that you can put together.

(Refer Slide Time: 24:56)



I will show you a picture of what I mean; take this sound here ok. You know that you can get from the sound you can get to the phoneme right, and you get to the phoneme because you have done what is known as feature extraction. Feature extraction has given you this phonemes. So, now, you have this phonemes from this phoneme you have to decode this word W. In order to do that you follow a set I see a nice set of steps to arrive at that.

So, what do you do? Start with our example of the word banana right. You have a phoneme representation of banana what is the phoneme presentation simplest is ba na

and na right. You split this phoneme representation and you manually create what is known as the lexicon right which is essentially word from phoneme you arrive at the word. So, this lexicon essentially is from the phoneme you are able to arrive at the word.

So, if for decoding it is now quite straightforward from phoneme representation you look up this dictionary and then you find from the lexicon you actually find that that is actually banana. But you cannot; that means, you should already a prior a priori you should have created this word to phoneme mapping, which is manually done there is no choice you have to keep this ready with you.

And therefore, this keeping ready part of word to phoneme is indeed the lexicon part ok; so this is one thing. Then if you want to actually put back the what is known as the context and then you want to create the sentence. Sentence creation comes from the words right you want to put the words in sequence in a proper sequence to form the sentence. You need assistance of what is known as the language model.

So, the occurrence of these words the putting of these words together to form the sentence has to you must rely on this language model. Now there is one nice thing that this picture is telling you, this feature extraction gives you this essentially it will give you all the feature vectors which is y let us say let us call this y from the feature vectors you are the decoder runs and puts in that word W. So, you may have to apply a simple likelihood function which essentially could be represented as this word is the argmax of P of W given y this is as simple as this.

That means, you are trying to say what is the likelihood of this word given the set of feature vectors right. So, this is a likelihood and this is not easy. So, what you do is you transform it into something which can be easily done in a computer based system and that indeed is could be represented as argmax of probability of Y given W times probability of the word itself right.

If you this is now easy, in order to get to this you need support of the acoustic model right. So, this acoustic model will give you support in order for you to get to this word as easy with these three blocks in place. So, this part P of Y given W comes from the acoustic model. So, I will say AM the acoustic model which is shown here and P of W comes from the language model I will call it LM.

By this method you should be able to get to the word quite easily alright. So, there are other techniques for recognition of continuous speech, you can have techniques such as Hidden Markov Model HMM techniques. Then there are advancements which go beyond HMM where people use neural networks also for the purposes of continuous speech decoding.

I made one assumption here and I want you to think about this assumption all along the sound that you heard was clean right. All along we never even spoke about any background noise as though that word banana was so clear to you that is not the case you are going to hear this word banana in very noisy environments, the system has to decode with lot of noise associated. Speech is associated with a lot of background noise therefore, this story of word splitting it into phonemes each phoneme having feature vectors.

And putting the feature vectors together, applying this language model, acoustic model and the I language model is all done under highly noisy conditions. And how good you build your decoder, how good you build your word creation from solution a to solution b essentially talks about; how would you are able to recover that word under as low SNR's as possible Signal to Noise Ratios as possible.