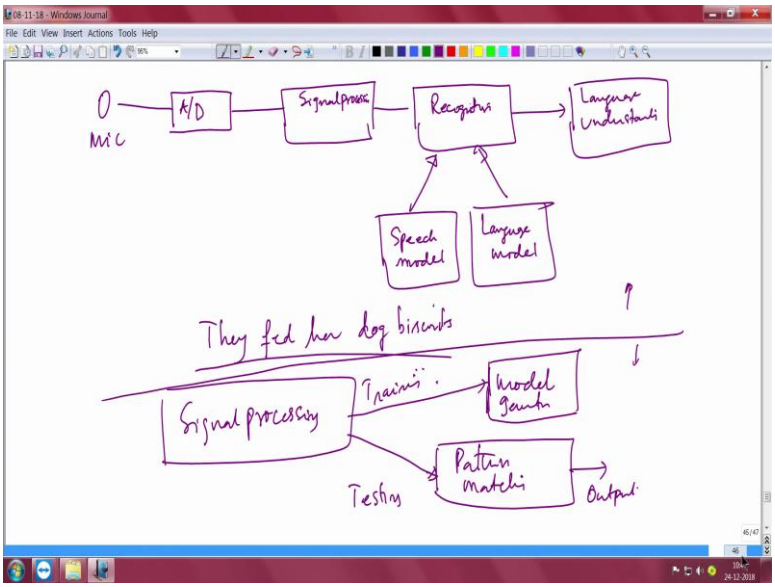
Advanced IOT Applications
Dr. T V Prabhakar
Department of Electronic Systems Engineering
Indian Institute of Science, Bangalore

Lecture - 26
Speech Recognition Part – 3

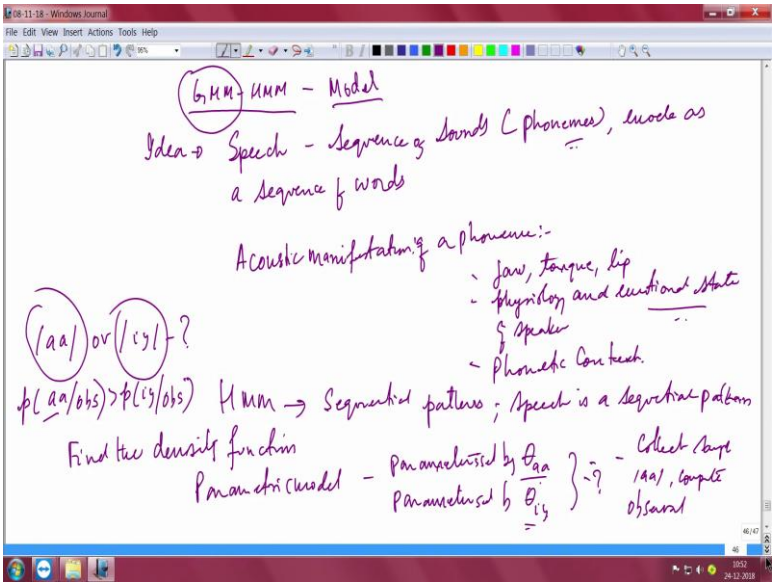
So, this signal processing block is a interesting block and lot of learning has to happen; training and testing has to happen in the signal processing block ah. So, it is good to know before you actually see a demonstration what actually happens inside the signal processing block. So, essentially when you talk about signal processing, you talk of these two terms. In any in any learning that you will come across in future whether it is a simple statistical models or you know neural network models or whatever you will come across these two terms: one is the training and the other is the testing. So, when you see training you are essentially creating the model. So, it is all about model generation ok.

(Refer Slide Time: 01:22)



And when you say testing you talk about pattern matching. So, this is the most important thing; we stopped this in the previous part. So, I am just speaking from here is indeed that you have signal processing divided into two parts: one is the model generation and the other is the which includes the training part. And, the other which includes the testing part which is essentially the pattern matching part.

(Refer Slide Time: 01:53)



Now, whenever we talk what is the, what is actually happening we should know that a little bit in detail. Here I put on what exactly we are talking about in terms of this model; this is about the model this is the model. Speech essentially is nothing, but sequence of phonemes right. You have a sequence of phonemes, you want to encode them as sequence of words essentially you encode them in sequence of words. And, then you will be able to put these words with certain probabilities you will be able to estimate how these words sequences occur, sequence of words occur and then you will end up with a sentence.

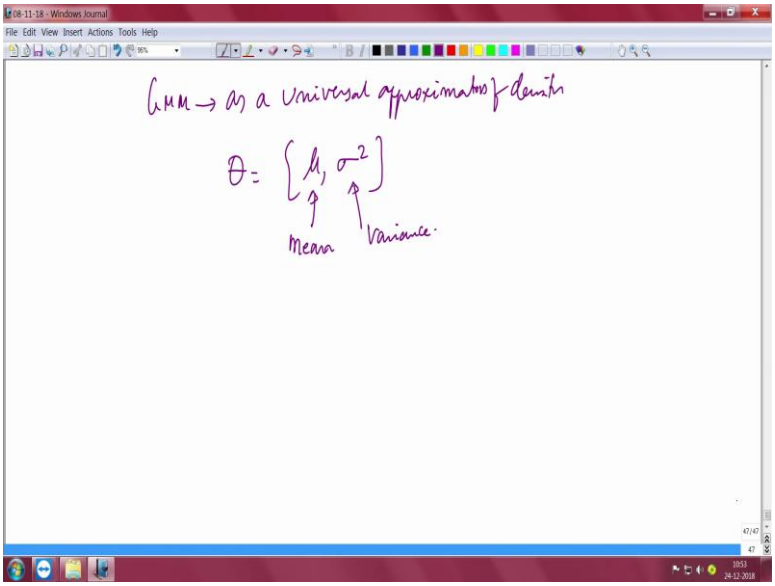
So, sentence comprising a words and each word comprising of several phonemes ok. Phonemes itself will have multiple states, you can have a start you can have a sort of a peak state and you can have a end state. All of that can also be visualized from the spectrogram that I showed you how you can have multiple states in within a phoneme. So, the acoustic manifestation of banana, we took that as an example. I can be talking in a very nice pleasant way and say, hi here is a banana I can say right I am when I am very happy. But if I am really you know aggressive and have some mood problem and all that I will say here is a banana right you actually mean the same word banana. So, the way you pronounce, the way you do things all depends on the position of your jaw, the position of your tongue, the position of your lip.

All these each one of us have different you know manifestations of our own physical systems. So, all of this including the emotional state of the speaker all of this has to be considered before you actually say anything about the word. So, here is the problem supposing you take I have mentioned to you about the phonemes right 24 consonants and 20 vowel based. I mentioned this already take one specific example of this phoneme aa aa iy. So, the question is when somebody said aa or iy what is the what did he or she say right. What who actually what is the actual utterance is it this or that, to for to decide that the computer system actually look at the probabilities.

The probability of aa given a set of observations if it is greater than the probability of iy given the setup observations you say that indeed the utterances this or the utterance is that ok. So, this is I am just looking at the phoneme level. Now, how do you get to this probability? You get to this probability by finding out the density function, probability density function ok. And this density functions are parameterized these are parameterized models, you say parameterize by theta aa and parameterize by i which is theta iy.

Now how to parameterize, what is this theta? Theta is nothing, but the fact that the density function is a mixture of Gaussian models is nothing, but a mixture of Gaussian. See the problem with Gaussian there is a limitation on what Gaussian can do and why we choose mixture Gaussian mixtures; the problem with the Gaussian is that if you take a simple Gaussian it is not very useful ok. So, essentially GMMs are very good for the; so, let me go to the next sheet.

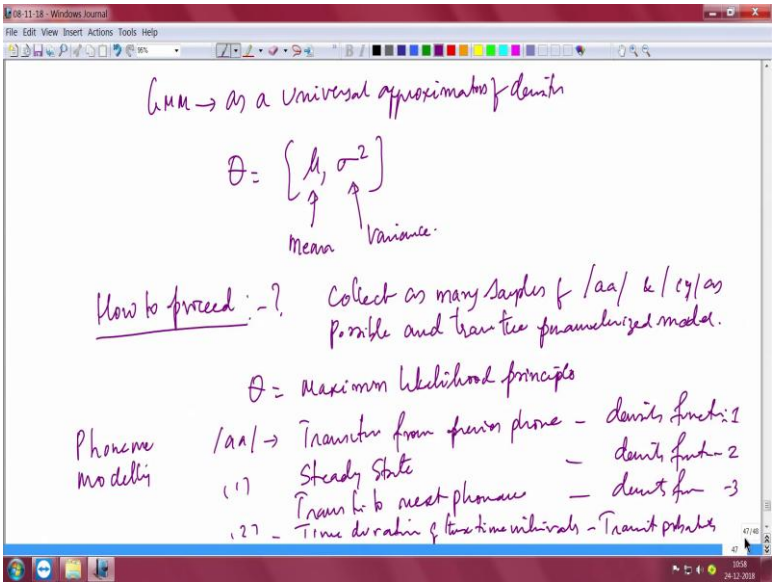
(Refer Slide Time: 06:04)



GMMs essentially our universal approximates. So, I will call them GMM as a universal as a universal approximators of densities of densities ok. If you look up Gaussian PDF you should be able to put down the Gaussian PDF there is a very well known expression. So, you can put down that essentially theta when you say when I wrote their theta I actually mean mu and sigma square ok, mu is the mu is the mean. And what is sigma square? Sigma square is the variants ok.

So, essentially you have to parameterize that is the; so, the whole model is a parameterized model. When you want to estimate these probabilities you have to go in for a for the you do it with the probability density function and you typically use Gaussian mixture models for the purpose of the estimation of these probabilities. So, how do you model this phoneme that is the next question. It is all good to say that a phoneme should be you know modeled as a as a let us say a density function and all of that and particularly as a probability density function of a Gaussian mixture model and all that. Why is this Gaussian mixture at all coming in the in the first place?

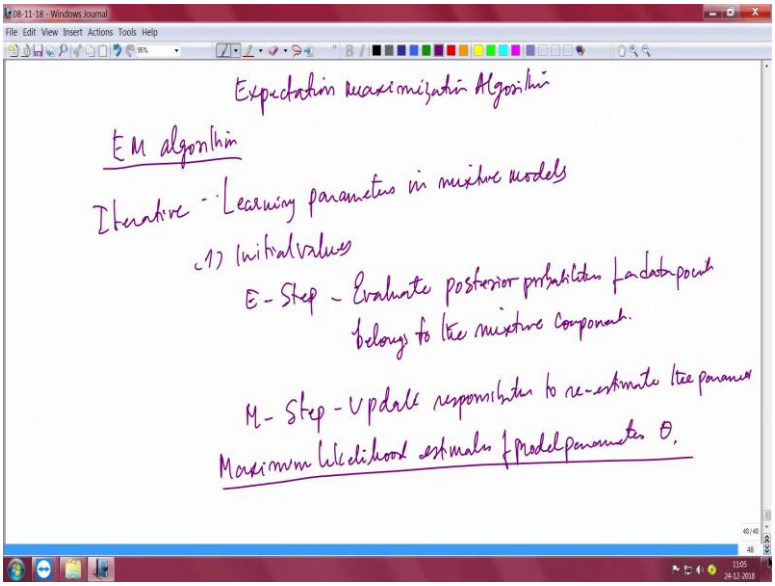
(Refer Slide Time: 07:55)



So, let me tell you let me explain with you with a very simple example, take this phoneme that we mentioned this aa right you take this phoneme aa right. If you take that you can model this phoneme in 3 phases you can be talking of entering that phoneme; that means, it is transiting from another phoneme somewhere. So, transition from a previous phoneme you can have one density function 1, you are remaining in that phoneme means it is in a steady state that is another density function 2, then you are transiting to the next to phoneme that can be another density function which is 3.

Now, this is not going to be sufficient you also need to know the time duration of these intervals and then the trans and therefore, you need to arrive at the transition probabilities. This whole system has to happen in real time as you are talking right. Therefore, estimating these parameters in real time is normally done using algorithms; you need another algorithm to do that and that algorithm which is very important thing in these the expectation maximization algorithm ok.

(Refer Slide Time: 09:21)



Let me put that down expectation Maximization algorithm EM algorithm, it is called EM algorithm ok. Now, it is nothing, but iterative scheme for learning it is a iterative scheme, for learning parameters this what you are interested in your have to learn the iteratively you have to learn iterative learning. So, algorithm let me ensure that I show you these learning parameters in mixture models in mixture models ok. So, you so, essentially this is a hit and miss kind of a system you start somewhere and then you end up with the model parameters.

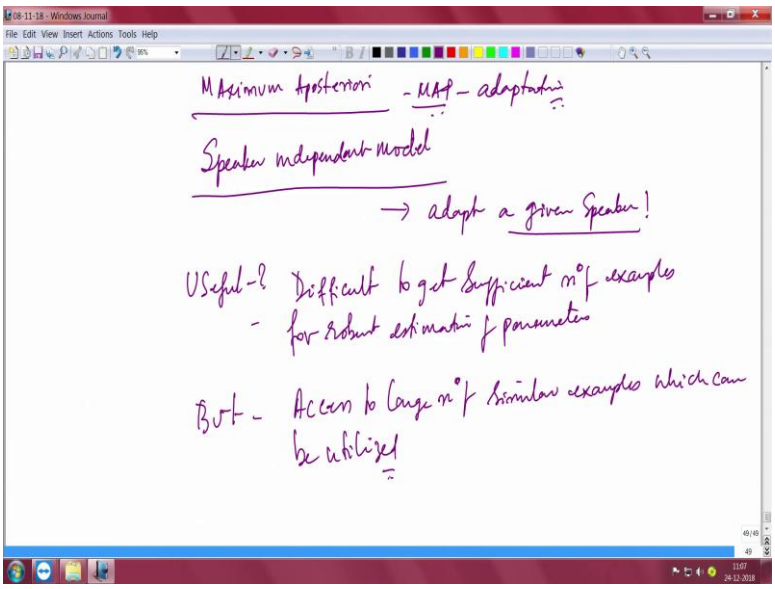
So, what do you do is first step in this is called expectation maximization; you choose some initial values you choose some initial values. And then you do what is known as a E step; this e step is nothing, but evaluate the posterior probabilities of data points. You evaluate posterior probabilities of a data point of some point of it I will just give us an example of some of the data point which belongs to that belongs to the mixture component compo component ok. Then what do you do? You do a M step that is why it is called EM, it use it then it what you do is since you have chosen some initial values you have to go and update.

You go back and do an update the to re-estimate, update responsibilities and call it responsibilities to re-estimate maybe parameters. You go on doing this step by step till you are able to match the exact PDF, the density function which looks very similar to the original. That is when you do when you spoke you got something your utterances related

to something which you only have seen you only observe that. And, somehow you are trying to fit that observation back into the model applying these parameters into the model into the GMM model and then getting back the original as close as possible to the original.

Then you say that the model has converged and you have actually found out the probability density function of the PDF of you have estimated the model parameters essentially which can generate the PDF essentially that is what it means. So, this theta so, all after you do all this EMs algorithms and all that you will end up with your objective. What was your objective? To get the model parameters of this the GMM parameters of the GMM essentially which uses maximum likelihood estimate methods to estimate the. So, you will use maximum likelihood estimates of model parameters of theta you will be doing that.

(Refer Slide Time: 14:04)



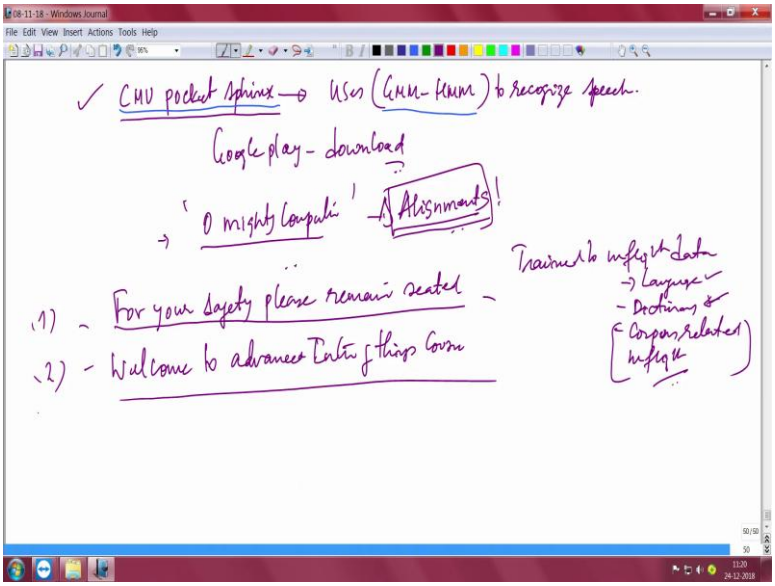
So, essentially you may also end up in a sometimes when you see the demo you will also end up with a certain way of doing it using maximum a posteriori it is called map a posteriori a post teriori ok. It is also called map adaptation ok, I will give you a feel for this then we will go and see what all you can do with this. See it is like this right you have speaker you have speaker independent, you have speaker independent model ok. You have a speaker independent model, you want to adapt this speaker independent model to a given speaker adapted to a given speaker this is the key.

So, they use this technique of map maximum a posteriori adaptation in order to do that. Now question is - where is it useful, why is it useful, why are we discussing this because the demonstrations that you see can actually you can do map adaptation ok. So, for example, why is it useful? It is useful because let us say it was difficult for you it was difficult for one to get sufficient number of examples for robust estimation of parameters ok. But, what you were what do you have with you is you have access, but I will say, but you have access to large number of similar; this is important similar examples which can be utilized. So, many acoustic application speech to text applications actually you can do map adaptation also ok.

So, you can see all of this involves intricate math and it mostly uses the Bayesian techniques estimations. So, I strongly urge that I have not covered too much of math here I want you to go and read a little bit of background. So, that you will be able to understand what exactly is happening; you could connect easily once you know the math a little bit better all it. So but; however, this is already sufficient for you, I hope you got a feel of how a phoneme essentially is modeled before it arrives at it. There is phoneme, there are states in the phoneme, there are the essentially you are looking at a parameter estimations of that phoneme. We mentioned about the 3 states which I wrote it here.

So, transition from previous phoneme then you can have a steady state, you have a transition to the next phoneme. All 3 of them can be modeled as density functions, you have multiple density functions 1, 2 and 3 in this case. You also have time durations and all of it falling into the GMM HMM model and the whole exercise is to estimate the model parameters. Estimation of model parameters is done using expectation maximization algorithm and you arrive at the closest match to what was spoken whatever was uttered. So, that is a goal. Now, let us move on to see some demonstrations of what is possible and how you can actually build your own application to do this.

(Refer Slide Time: 18:55)



Let us start with a small demonstration Payal and Rahul, I have put this demonstration together and have done a wonderful job to give you a feel of the problem. What you can do is if you have an android phone go to Google play ok. And download app called CMU pocket sphinx; it is written here go and download CMU pocket sphinx. This essentially users all that we spoke about in the, it uses the GMM HMM model to recognize the speech. It takes the phoneme, splits it into 3 parts I will just recap quickly into 3 parts. The transition, steady and then going on to the next stage right; all the 3 parts it is done it uses the transition probabilities builts the model, you had to estimate the model parameters all of that it does.

So, you can start by recognizing speech in that Google play model by uttering the word 'O mighty computer' you should use this word. See everything in speech to text processing is all about alignments. This can be a nightmare for you it is all about alignments. If you do not align properly because, speech is a time it is a time series right you will be taking it the system will be taking samples of 10 milliseconds and start to decode the phonemes if the words from. So, if you say a sentence words and from within the word the different phonemes which make up the words. So, all of this will have to do you have to start somewhere some index has to be there.

So, that is essentially related to the alignment. So, this alignment time alignment is an important thing once you; so, this 'O mighty computer' is like you start and then you say

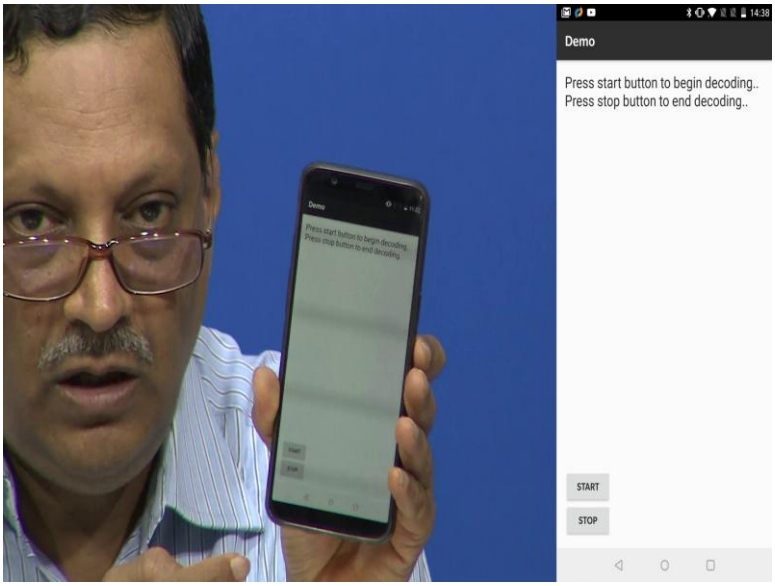
whatever you have to it has is expected to decode. So, please note in any implementation this alignment can be a very important problem for you. So, if you do that you will see something do not worry about it; what we have done in the lab is we have created as very simple example for the course is related to airplanes and passengers sitting in an airplane. For example, if there is a cabin announcement by the flight attendant for your safety please remain seated, supposing it is said like that particularly when there is turbulence the fight attendance keep announcing that you should remain seated in your seat.

That is something that you hear and you may want to sort of converted into some sort of textual form; I mentioned to you right in the beginning that speech to text has various application particularly for persons with different abilities. So, that is the reason why this example is extremely important. What we have done is we have trained the language model essentially a set of sentences which can which essentially its sentences whose occurrences are based on probabilities of words right that is essentially the language model. You are talking about utterances occurrences of words and their probabilities of in the sequence in which they can occur essentially related to probability essentially related to the language model.

Then dictionary is essentially all those words which comprise of the sentences that you can form. So, you have a dictionary model, you have the language model and all of that from a corpus of everything related to flight only flight corpus is included here. So, you create all the acoustic models, you create all the language models specific to the in flight system. Let us see if you do such let us say fine adjustment then what would happen if you use words with are which are specific to airplanes. What would happen if you use words which are away from an airplane scenario?

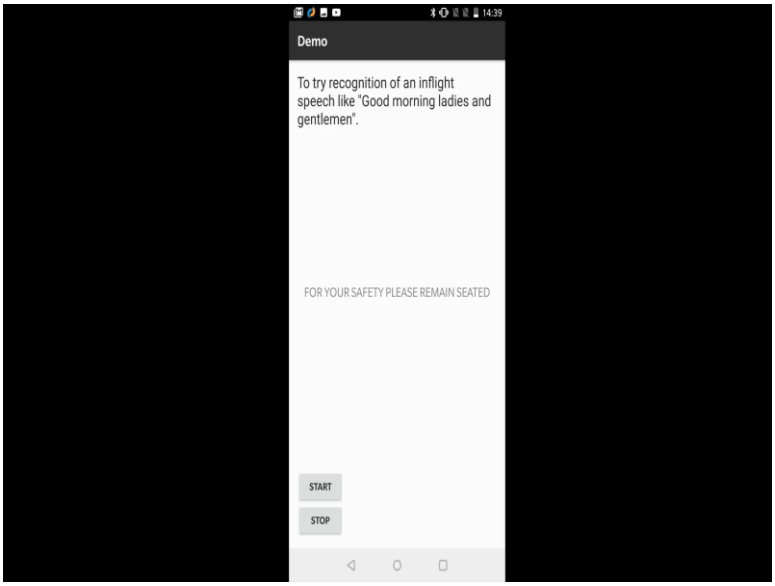
So, we will try 1 and followed it up by 2 ok. So, if we have trained this system to work for 1 anything related to 2 should not be decoded properly because, the language model, the dictionary and the corpus that was used to train the models were all specific to the in flight data ok.

(Refer Slide Time: 23:42)



So, let us turn our attention to this mobile phone app downloaded from Google play called the CMM pocket springs modified by our project associates in the lab to support 2 buttons: one is called start, the other is called stop. So, I hope you can see that here ok. Now I will turn it around, I will press start and I will utter these words and let us see what it decodes; for your safety please remain seated.

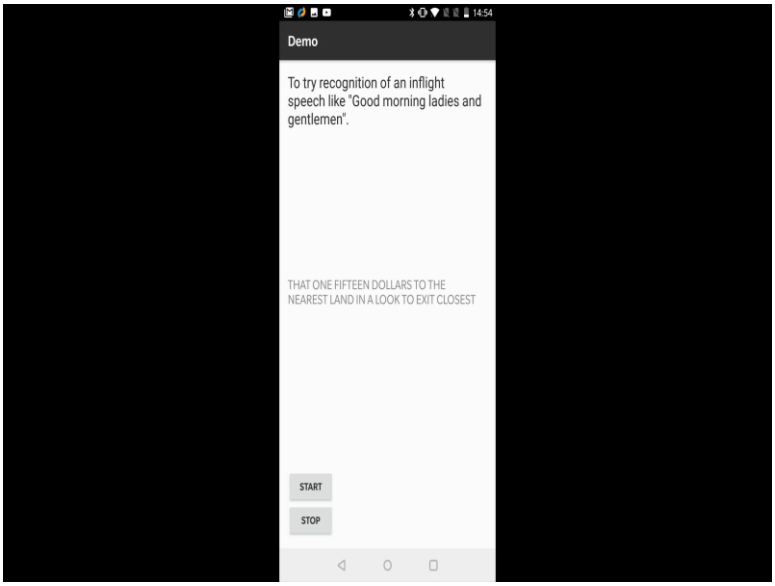
(Refer Slide Time: 24:20)



Look what it has decoded ok. So, I will read out for you for your safety please remain seated, that is what it has decoded is exactly what we had trained this complete model

for. Let us proceed now and try to see whether it will decode welcome to advance internet of things course, welcome to advanced internet of things course.

(Refer Slide Time: 24:50)



So, you see it has decoded its completely garbage. So, you can see that it is not really tuned to decode, it is not trained the training and the testing that has been done for pattern recognition you do testing and for training you get hold of the model parameters right; both have not actually happened for in flight data. So, this you can now try with another app you tune it for something else then you will be able to proceed. And then so, this is an example of how you can use yourself; you yourself can do this in your by sitting in front of your computers with an internet and a android phone you can try this on your by all by yourself.