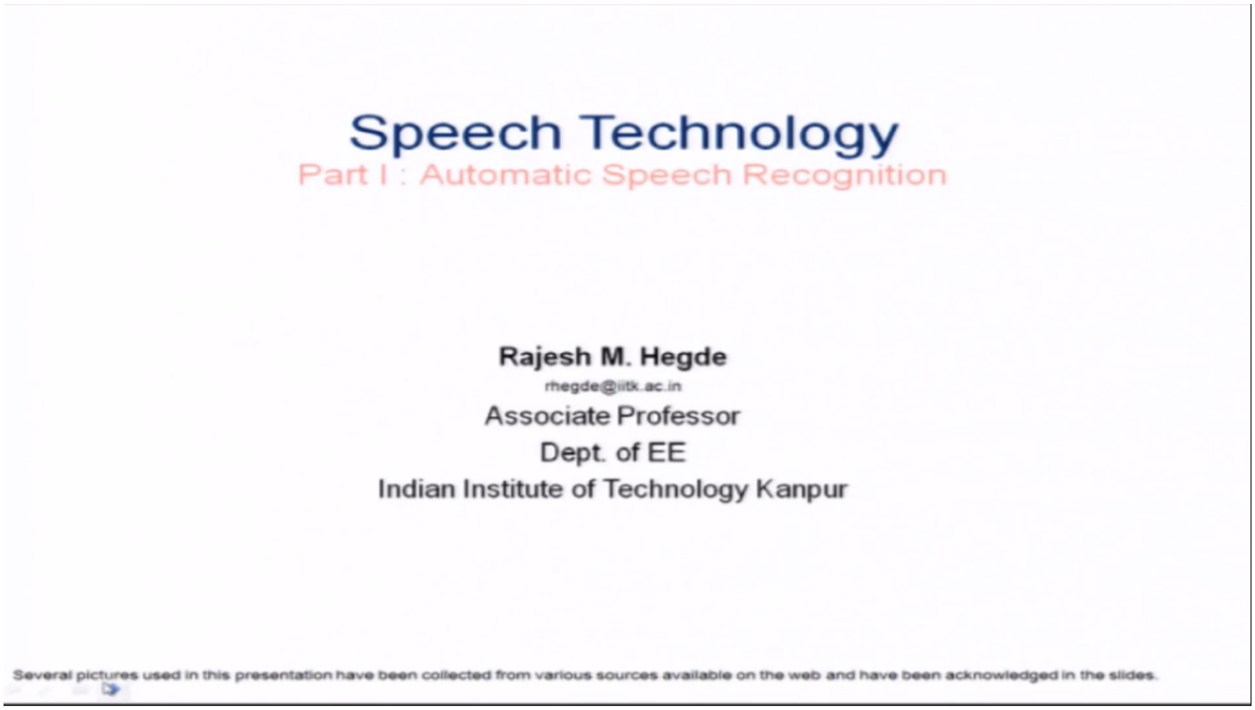


Speech Technology

Part I : Automatic Speech Recognition

Rajesh M. Hegde
rhegde@iitk.ac.in
Associate Professor
Dept. of EE
Indian Institute of Technology Kanpur

Several pictures used in this presentation have been collected from various sources available on the web and have been acknowledged in the slides.

Speech Technology

Part 1 Automatic Speech Recognition

Rajesh M. Hegde

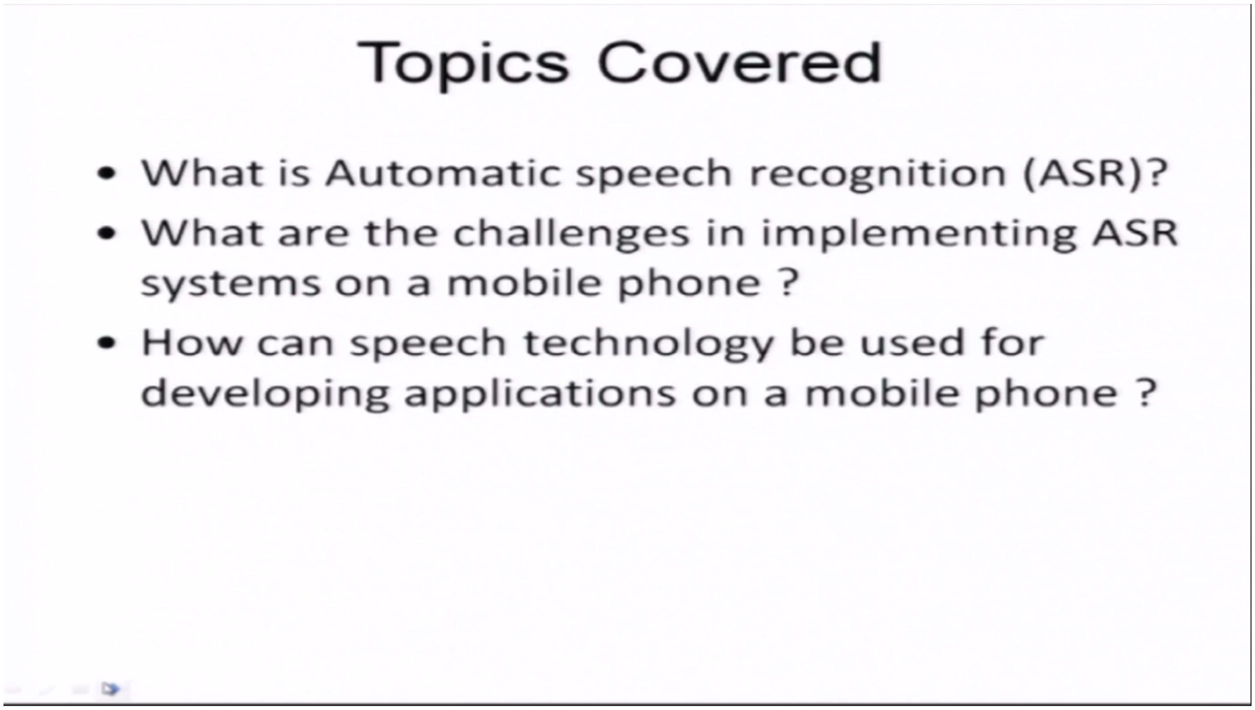
rhegde@iitk.ac.in

Associate Professor

Dept. of EE

Indian Institute of Technology Kanpur

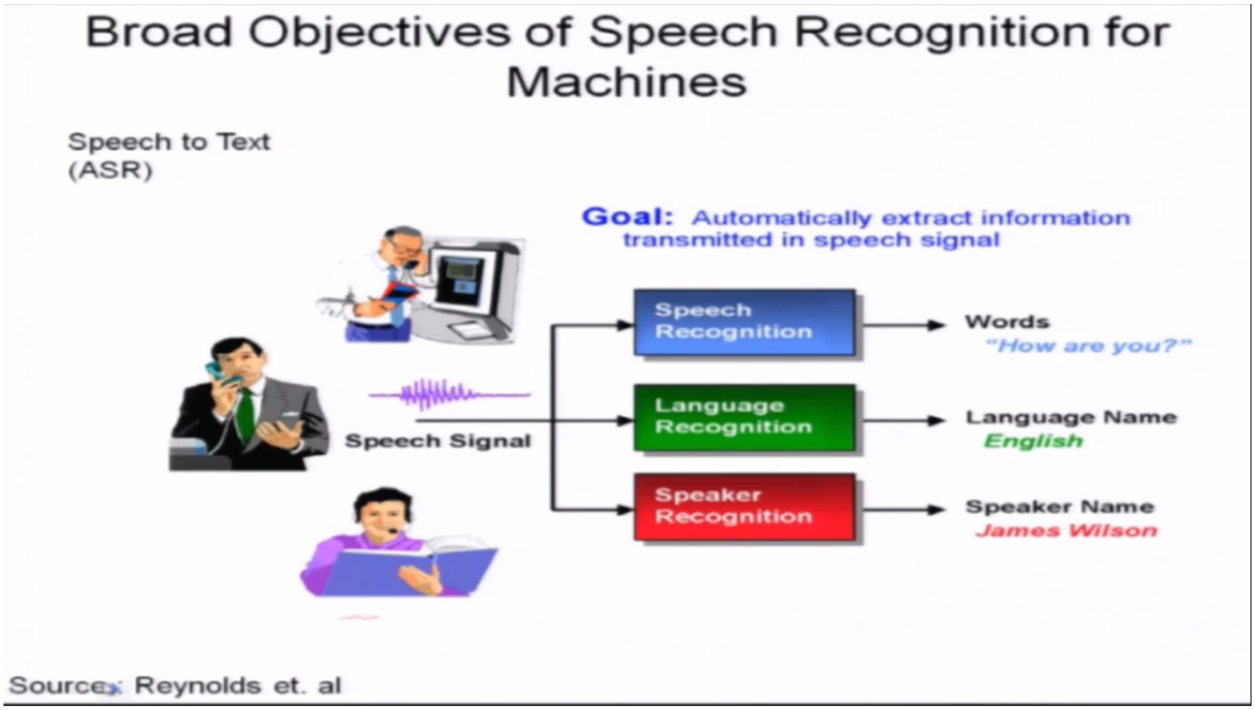
Hello everybody and welcome to this part on speech recognition and text-to-speech synthesis of this particular MOOC. What we will do in this particular set of lectures is I will be talking about speech technology. We will divide it into two parts. The first part which we are going to do right now is on automatic speech recognition.



Topics Covered

- What is Automatic speech recognition (ASR)?
- What are the challenges in implementing ASR systems on a mobile phone ?
- How can speech technology be used for developing applications on a mobile phone ?

So essentially what we'll cover in this block is we will discuss what is automatic speech recognition and what are the challenges in implementing automatic speech recognition systems on a cell phone because that is probably the most ubiquitous device which people use these days. We will then also discuss on how can speech technology be used for developing applications on a cell phone by taking some typical examples.



Broad Objectives of Speech Recognition for Machines

Speech to Text (ASR)

Goal: Automatically extract information transmitted in speech signal

Speech Signal

Speech Recognition → Words "How are you?"

Language Recognition → Language Name English

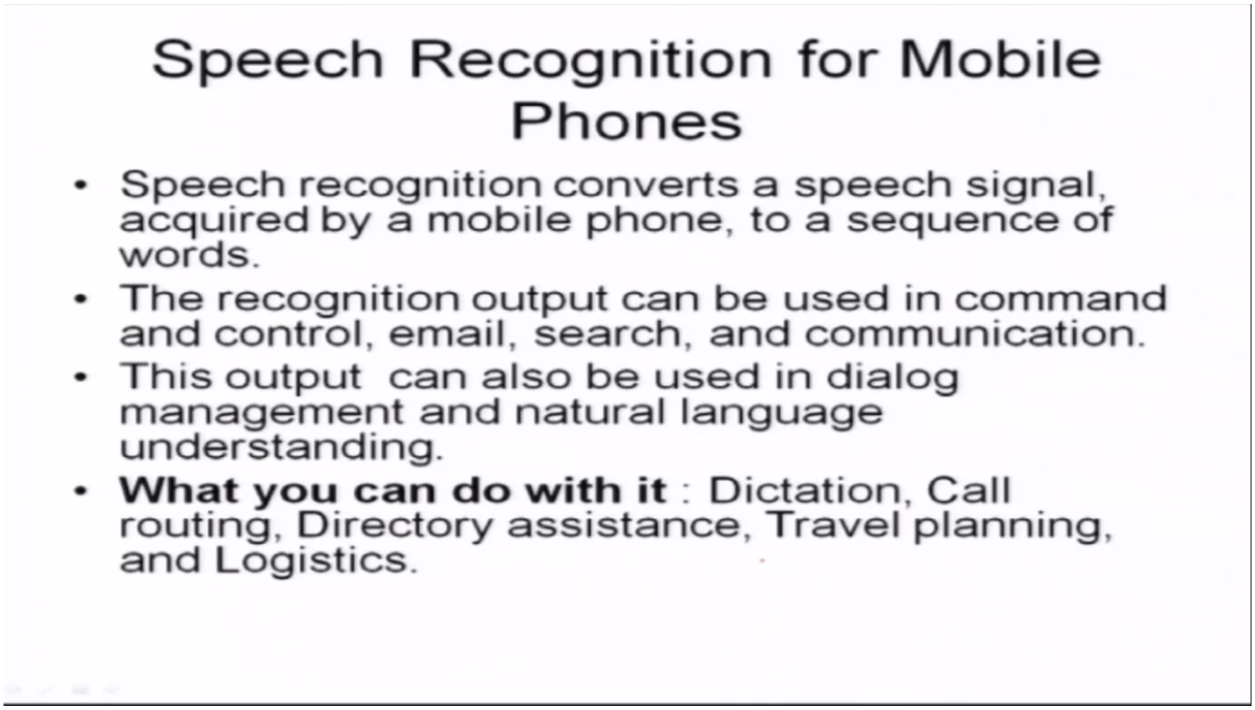
Speaker Recognition → Speaker Name James Wilson

Source: Reynolds et. al

Okay now let us see what are the broad objectives of speech recognition for machines. So essentially when we say speech recognition we are talking of speech recognition using or by machines. So what are the goal of automatic speech recognition? It is to automatically extract information transmitted in the speech signal. So you can also call speech recognition as a process by which the incoming speech signal is converted to text. Many people also refer to this as signal to symbol transformation. So you have a speech signal that can be recorded over a microphone which needs to be converted to text.

Okay now if you take a look at this figure broadly you can classify speech recognition into three parts. One is given the input speech signal. So you see the speech signal which is described in this particular form. So you have a incoming speech signal which is input to a module called speech recognition and let me assume that I am saying how are you. The objective of this particular module is to make sure that it outputs the words how are you. So essentially the machine recognizes what you are speaking.

So you'll also have some elide applications like language recognition where one speaks in a particular language and the system tells in which language the speaker was speaking or you could have something called speaker recognition where you are essentially trying to find out who was speaking from a set of speakers. This gave you a broad objective of what is speech recognition and different possibilities there.

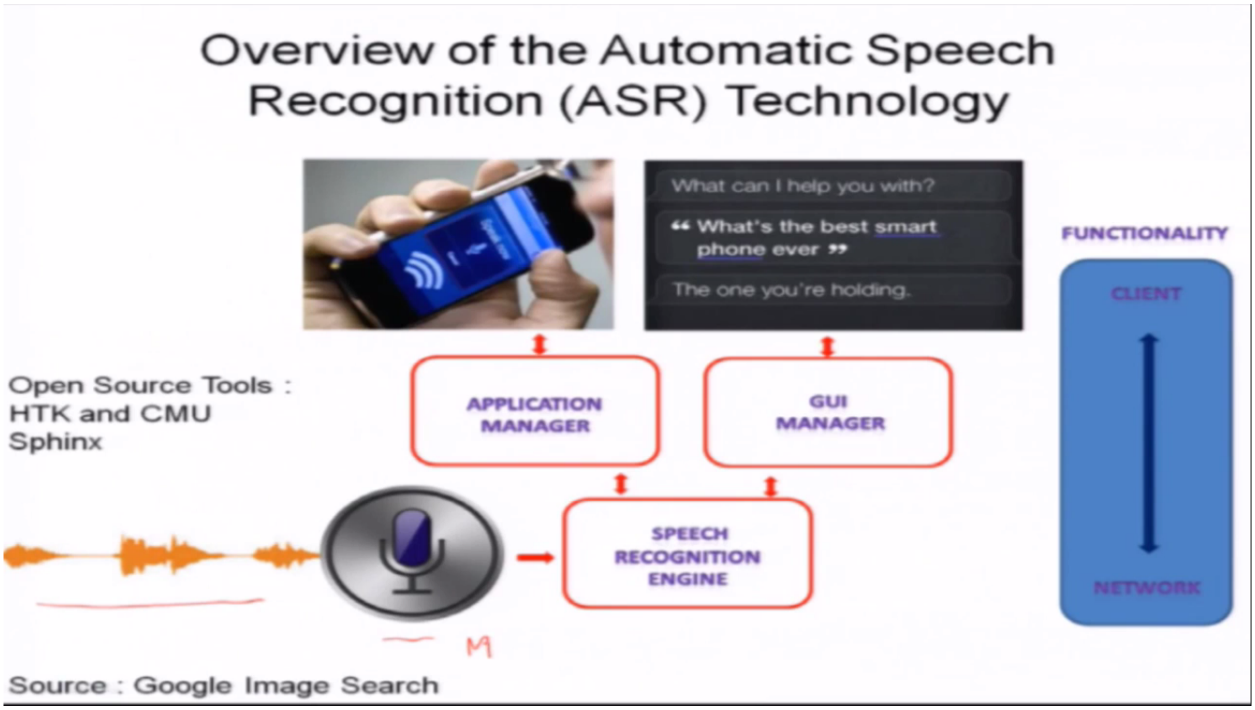


Speech Recognition for Mobile Phones

- Speech recognition converts a speech signal, acquired by a mobile phone, to a sequence of words.
- The recognition output can be used in command and control, email, search, and communication.
- This output can also be used in dialog management and natural language understanding.
- **What you can do with it** : Dictation, Call routing, Directory assistance, Travel planning, and Logistics.

And now let us get into a little bit of nitty-gritty on how we do speech recognition on a mobile phone. So the definition of speech recognition it converts a speech signal acquired by a mobile phone to a sequence of words. So the recognition output can be used in many places. Some of these applications are in command and control. When I say command and control what it means is essentially I'm trying to switch on a device or off device. So I will say switch on. I'll say switch off. So these are command and control applications. So I could probably dictate my email and that's one other application or I could do search via voice typically you have these applications on Google these days or I could use speech as a form of communication between machine to machine.

Now the output can also be used in dialogue management and more complicated things like natural language understanding. Right. So what can you do with it? There are several applications. Some of the broad applications are you can dictate using speech. You can do call routing. You can do directory assistance which means say I want to search for a particular phone number I speak the name of the person I get back the telephone number of that person. I could do travel planning booking tickets using voice or I could do some kind of complex logistics using voice.



Overview of the Automatic Speech Recognition (ASR) Technology

What can I help you with?

“ What's the best smart phone ever ”

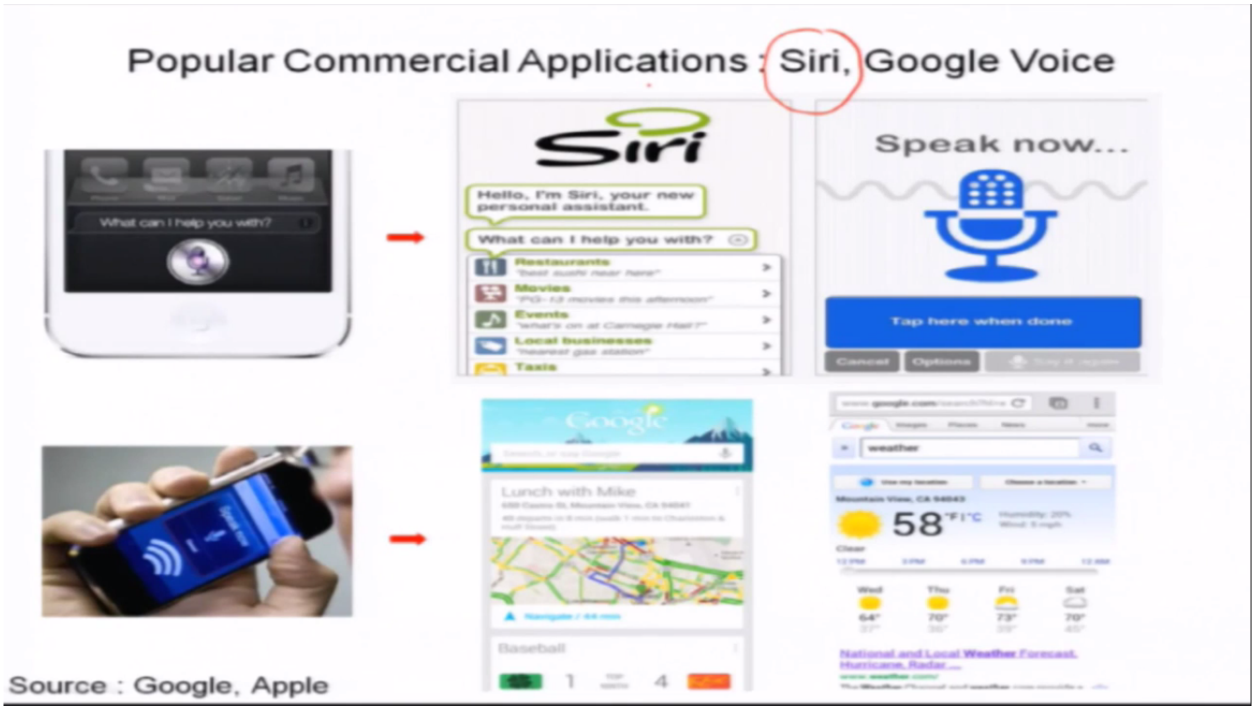
The one you're holding.

Open Source Tools : HTK and CMU Sphinx

Source : Google Image Search

Okay. Now this block diagram gives a very broad overview of the speech recognition technology. So let me see if I can start from one end. So you essentially have the speech signal which I am trying to mark out now which is input into a microphone. So let me call this as a microphone. Now this voice signal is essentially fed into something called the speech recognition engine. Now the speech recognition engine essentially recognizes speech or it converts incoming speech signal into text. Now this is obviously connected on both sides to an application manager and also to a graphical user interface manager. An application manager is the manager that essentially pipes the data coming in from the interface to the device. So the device in our case is supposed to be the cell phone. Okay now what are the examples you can have. Say for example you are talking to a phone and asking what can I help you with. So you probably say what is the best smartphone ever and then it says the one you are holding. So there are two aspects to this; what can I help you with is recognized by the cell phone and once it finds out that you are holding the best cell phone it tells you that the one you are holding. So essentially this involves both ASR as well as text to speech synthesis which we will call as TTS.

So essentially the other aspect that one needs to notice when you are doing recognition on a cell phone these are very resource constrained devices the computing power is less when compared to a desktop and therefore typically the functionality goes as a client server or as client network mode of running.

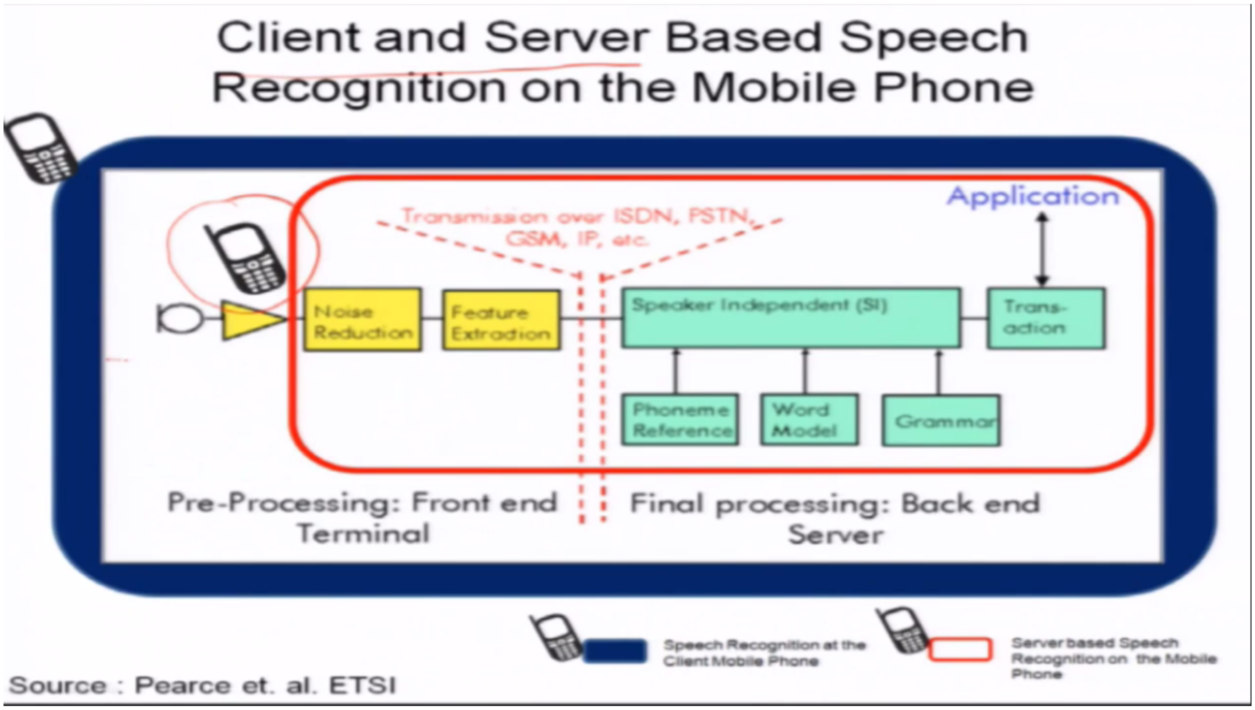


Popular Commercial Applications : Siri, Google Voice

Source : Google, Apple

Okay now what are popular commercial applications of speech recognition if you are interested in? You have this very popular application called Siri on Apple iOS. Now Siri actually helps you with a lot of things. So some of these things are captured in this picture. So you can search for restaurants. You can search for movies. You can search for events. You can search for local businesses. You can hire a taxi. There is a lot of things you can do probably if you have an iOS device you should take a look at Siri and try accessing information using speech.

So essentially what Siri does is you will probably tap a button and then you will probably say something into Siri and it will give you the information that you need. Now one other typical application that very popular application is Google Voice. Now the advantage of Google Voice is that it is available on any iOS and you could do lot of things using the power of Google. So essentially you are connecting to the power of Google using your voice. So one of the applications could be probably locating an address as it is shown here or you would want to find weather information using Google Voice. So probably I'd say I want the weather in my city. City name and whether that would give me the weather information. Well these are some of the popular applications.



Client and Server Based Speech Recognition on the Mobile Phone

Application

Transmission over ISDN, PSTN, GSM, IP, etc.

Noise Reduction

Feature Extraction

Speaker Independent (SI)

Trans-action

Phoneme Reference

Word Model

Grammar

Pre-Processing: Front end Terminal

Final processing: Back end Server

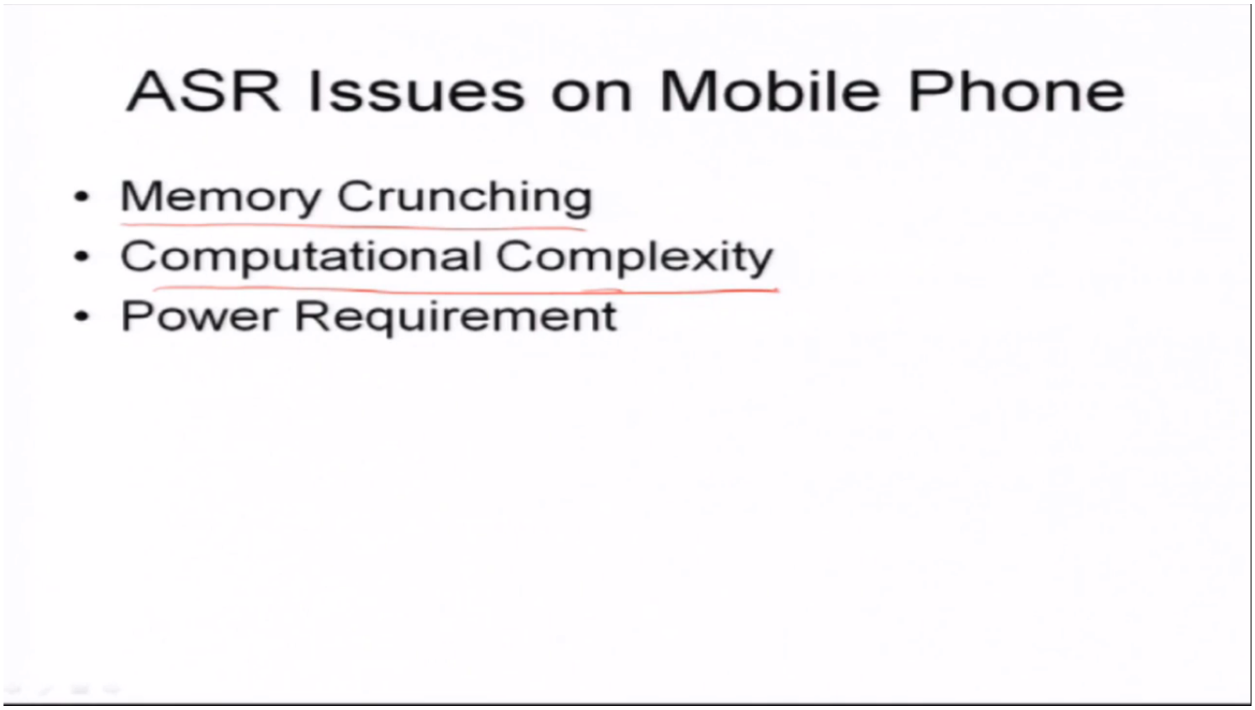
🟦 Speech Recognition at the Client Mobile Phone

□ Server based Speech Recognition on the Mobile Phone

Source : Pearce et. al. ETSI

So as I said earlier when we look at a cell phone the recognition process technically is not done on the device because the device is resource contained. So what you do is you follow a client-server based approach where the client is your cell phone. So you speak into a microphone of your cell phone. The first part actually does some kind of noise reduction which is very important because you often speak in noisy environments. Then there is a process of feature extraction which is slightly technical but it actually gets some meaningful numbers from the speech signal. Now that is piped into the speech recognition engine. Now typically that is speaker independent because you would want the speech recognition engine to be independent of any speaker who speaks. So what are the inputs to this? The inputs are the phoneme reference, the word models and the grammar. So it's something like you make the machine understand what are the units of speech, what is the grammar that it needs to know to recognize speech.

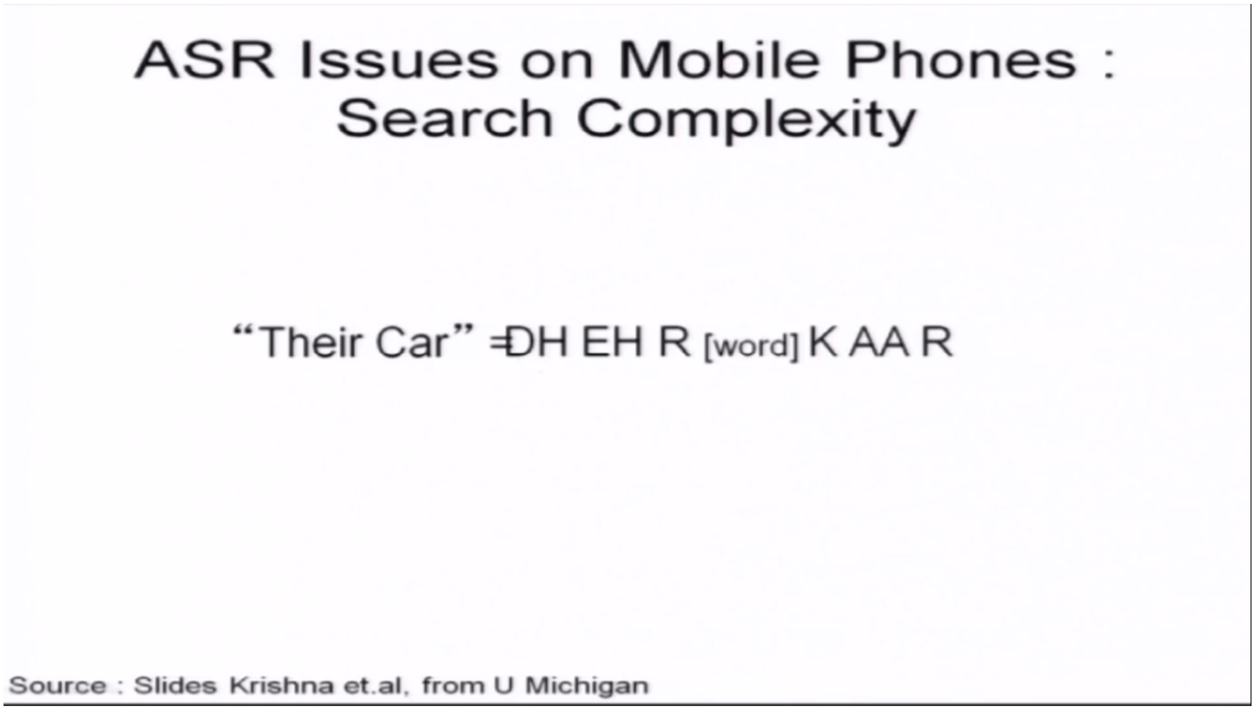
Now the transmission is typically done over ISDN, GSM, IP there are different ways of transmission. Now once the recognition is done so you are obviously looking at some kind of transaction which is coupled to an application. For example if you are doing a banking transaction you would say something it would be coupled to that particular application. So broadly so therefore you can divide speech recognition into the front end and you can the final processing is called as the back end. Fine.



ASR Issues on Mobile Phone

- Memory Crunching
- Computational Complexity
- Power Requirement

So what are the issues for speech recognition. Not going too much into technical details but what are the broad issues one would face if you were to do speech recognition on the cell phone. See the problem mainly is that of memory crunching. As I said earlier the cell phones have got limited memory. They're small devices and therefore you do not have a lot of memory to do the processing. Of course the problem of computational complexity is always there and therefore you need more computational power on the cell phone which it obviously does not have and then comes the requirement of power. Cellphones have got batteries which do not last for a long time and therefore there is obviously a problem of power requirement.

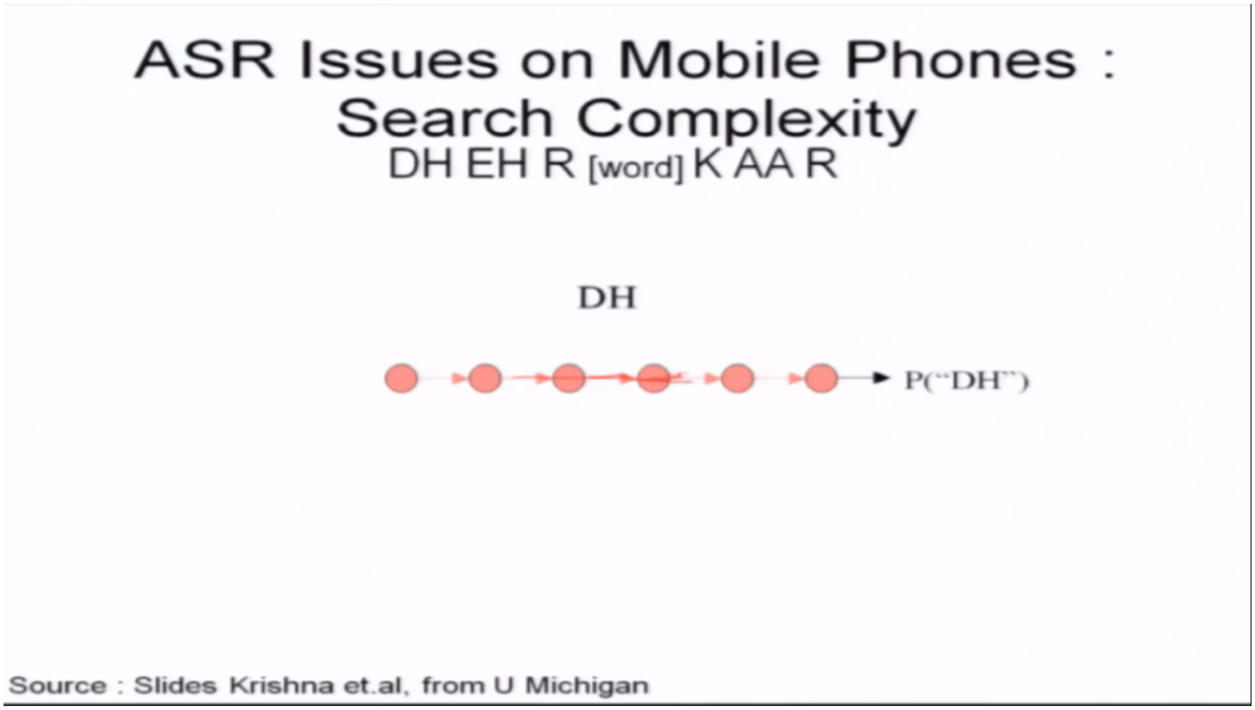


ASR Issues on Mobile Phones : Search Complexity

“Their Car” ⇒DH EH R [word] K AA R

Source : Slides Krishna et.al, from U Michigan

Now let us take a simple example to illustrate what are the issues involved in speech recognition. Now let us say I want to recognize a phrase their car. So what will do is let us say they split up into these smaller units Dh eh and r. So what we are trying to do is we want to recognize their car. So what does the recognition essentially do in such case? Let us take a very naive example but a very good slide which is available from University of Michigan. So I want to recognize these basic phonetic units which will form obviously the word their car which has got two words; their and car. Okay. Now these are actually something called as we build what are called as statistical models. They are called hidden Markov models let us not get into too much of detail into that but essentially you are trying to build a probabilistic model which will find from the given speech signal what is the probability that the first unit was the. So that's what we are trying to do here.

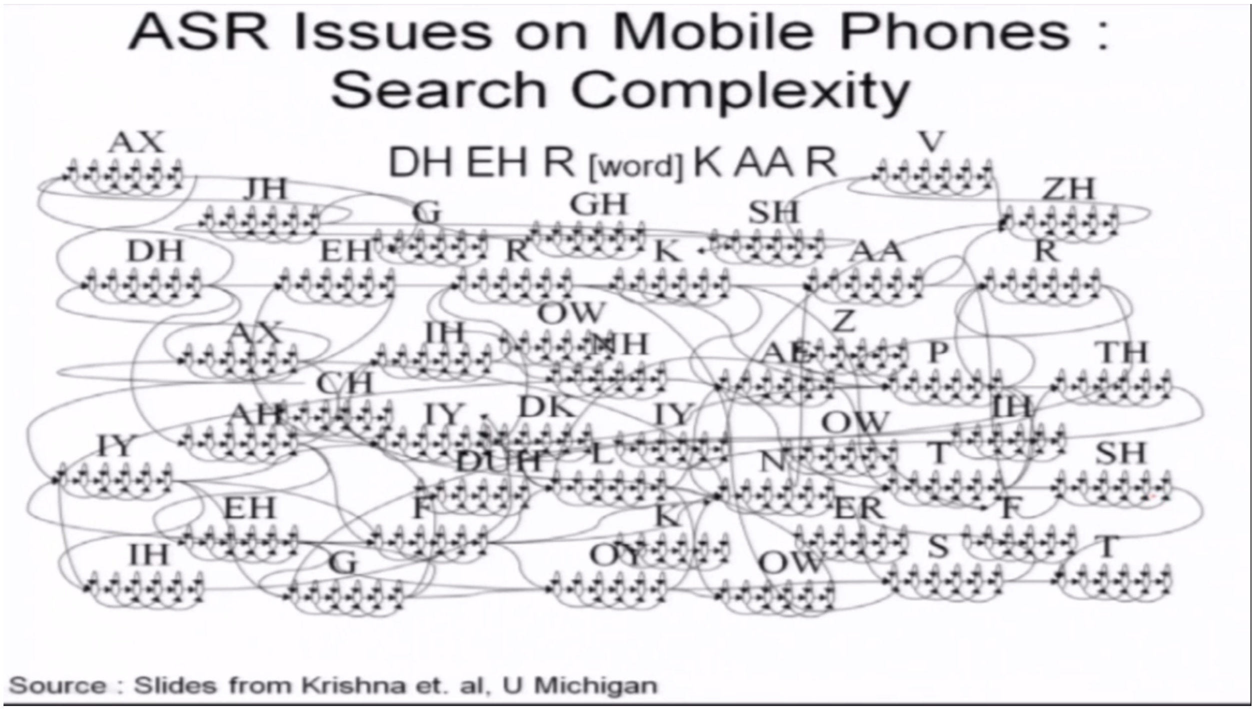


ASR Issues on Mobile Phones : Search Complexity

DH EH R [word] K AA R

Source : Slides Krishna et.al, from U Michigan

So now this is your hidden Markov model. This is a simple animation which we can run through. So you traverse through this model. So this is your Dh. Okay now you also coupled Eh and R so these are two other phonetic unit models. So you cascade them. So you have their. Okay. Now the problem with speech recognition is that it could be recognized as their or it could be recognized as the ear as you can see from the slides. So it's always often a possibility that it could be recognized as their or the ear because they are very close. So it depends on how well your system is whether it recognizes as their or the ear.

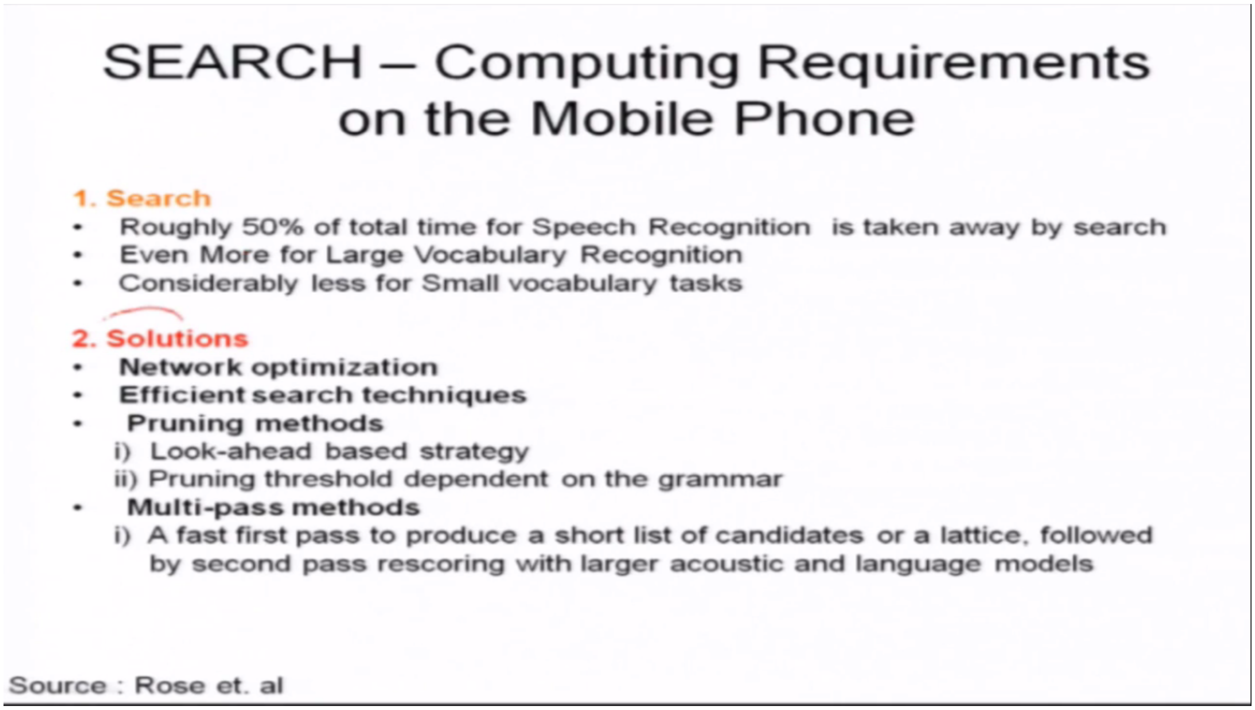


ASR Issues on Mobile Phones : Search Complexity

DH EH R [word] K AA R

Source : Slides from Krishna et. al, U Michigan

Now the complexity is not over here because we consider only two possible applications. Now on top of this you have got the next word car which obviously has got some possibilities like cap or cat. Now this is a very small subset of the search problem we have taken. Now what happens is there are numerous possibilities for their. It could be their, the ear, or now it could go on and on and become very complex because in practice you have got several possibilities for the and car. So you essentially have to run through all the possibilities and come up with an engine that gives you the best possibility and the correct one the topmost one is supposed to be the output of the machine. So what this essentially tells us is that the search is pretty complex when you do speech recognition.



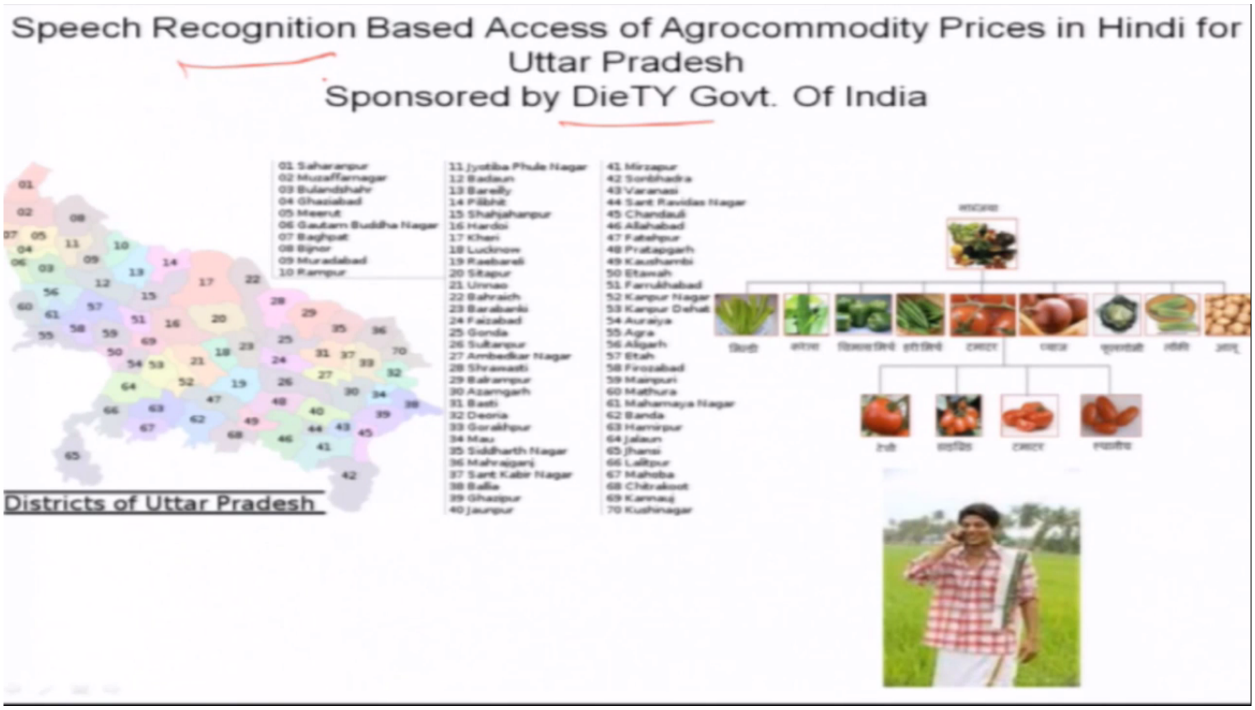
SEARCH – Computing Requirements on the Mobile Phone

1. Search
- Roughly 50% of total time for Speech Recognition is taken away by search
- Even More for Large Vocabulary Recognition
- Considerably less for Small vocabulary tasks

2. Solutions
- **Network optimization**
- **Efficient search techniques**
- **Pruning methods**
  - i) Look-ahead based strategy
  - ii) Pruning threshold dependent on the grammar
- **Multi-pass methods**
  - i) A fast first pass to produce a short list of candidates or a lattice, followed by second pass rescoring with larger acoustic and language models

Source : Rose et. al

Okay so search it takes roughly 50% of the total time. It depends on the vocabulary. You have a larger vocabulary of word the search is going to take more time. The smaller the vocabulary you are doing well on time efficiency. So there are several solutions. I will not get into the technical details but for those of you are technically savvy you can probably take a look at this. You can come up with efficient search techniques, network optimization some kind of pruning methods or you could use some kind of dynamic grammar which we call as multi pass method. So what I suggest is this can be taken a look for people who are technically savvy otherwise broadly you need to understand point one.



Speech Recognition Based Access of Agrocommodity Prices in Hindi for Uttar Pradesh
Sponsored by DieTY Govt. Of India

Districts of Uttar Pradesh

01 Saharanpur
02 Muzaffarnagar
03 Bulandshahr
04 Ghaziabad
05 Meerut
06 Gautam Buddha Nagar
07 Baghpat
08 Bijnor
09 Muradabad
10 Rampur
11 Jyotiba Phule Nagar
12 Badaun
13 Bareilly
14 Pilibhit
15 Shahjahanpur
16 Hardoi
17 Kheri
18 Lucknow
19 Raebareli
20 Sitapur
21 Unnao
22 Bahraich
23 Barabanki
24 Faizabad
25 Gonda
26 Sultanpur
27 Ambedkar Nagar
28 Shravasti
29 Balrampur
30 Azamgarh
31 Basti
32 Deoria
33 Gorakhpur
34 Mau
35 Siddharth Nagar
36 Mahrajganj
37 Sant Kabir Nagar
38 Ballia
39 Ghazipur
40 Jaunpur
41 Mirzapur
42 Sonbhadra
43 Varanasi
44 Sant Ravidas Nagar
45 Chandauli
46 Allahabad
47 Fatehpur
48 Pratapgarh
49 Kaushambi
50 Etawah
51 Farrukhabad
52 Kanpur Nagar
53 Kanpur Dehat
54 Auraiya
55 Agra
56 Aligarh
57 Etah
58 Firozabad
59 Mainpuri
60 Mathura
61 Mahamaya Nagar
62 Banda
63 Hamirpur
64 Jalaun
65 Jhansi
66 Lalitpur
67 Mahoba
68 Chitrakoot
69 Kannauj
70 Kushinagar

Okay. Now what are the applications of speech recognition? I'll take one examples of a system that is built at IIT Kanpur. This is essentially funded by the Government of India. We are trying to do speech recognition based access of agriculture commodity prices. Now if you take a look at the picture you see that we have done this for a state called Uttar Pradesh in India which has got several states. You can see the listing of the districts that I think we have only around 70 districts here. So essentially what this system does is a farmer can call into a particular telephone number and the system prompts him to say the name of the district. He says the names of the district and the crop price he is interested and the system essentially gives him back the information that he requests for. So essentially speech recognition can be used in the development of several socially relevant applications apart from the usual fancy application like Siri and Google Voice.

So thank you. This is the end of the first module and what we have discussed here is the basics of automatic speech recognition and a possible application which we have implemented here.

The next part we will cover text to speech synthesis. Thank you.