Speech Technology

Part II : Text to Speech Synthesis

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 II Automatic Speech Recognition

Rajesh M. Hegde

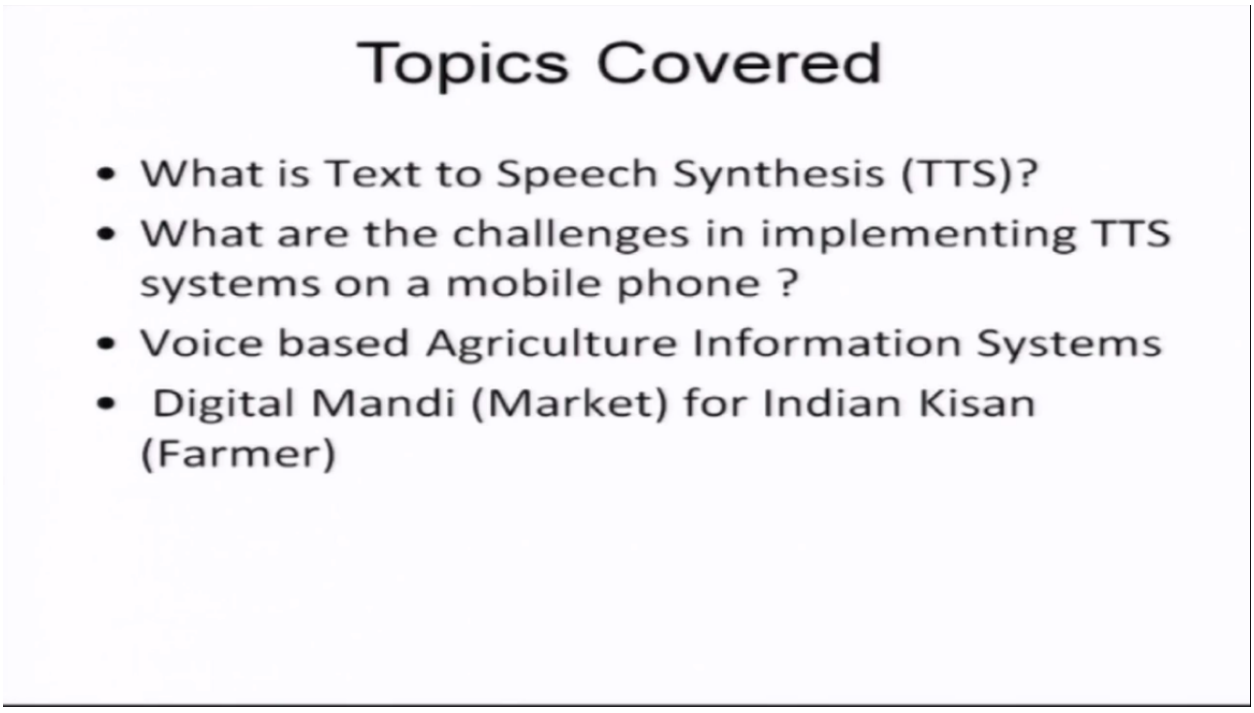
rhegde@iitk.ac.in

Associate Professor

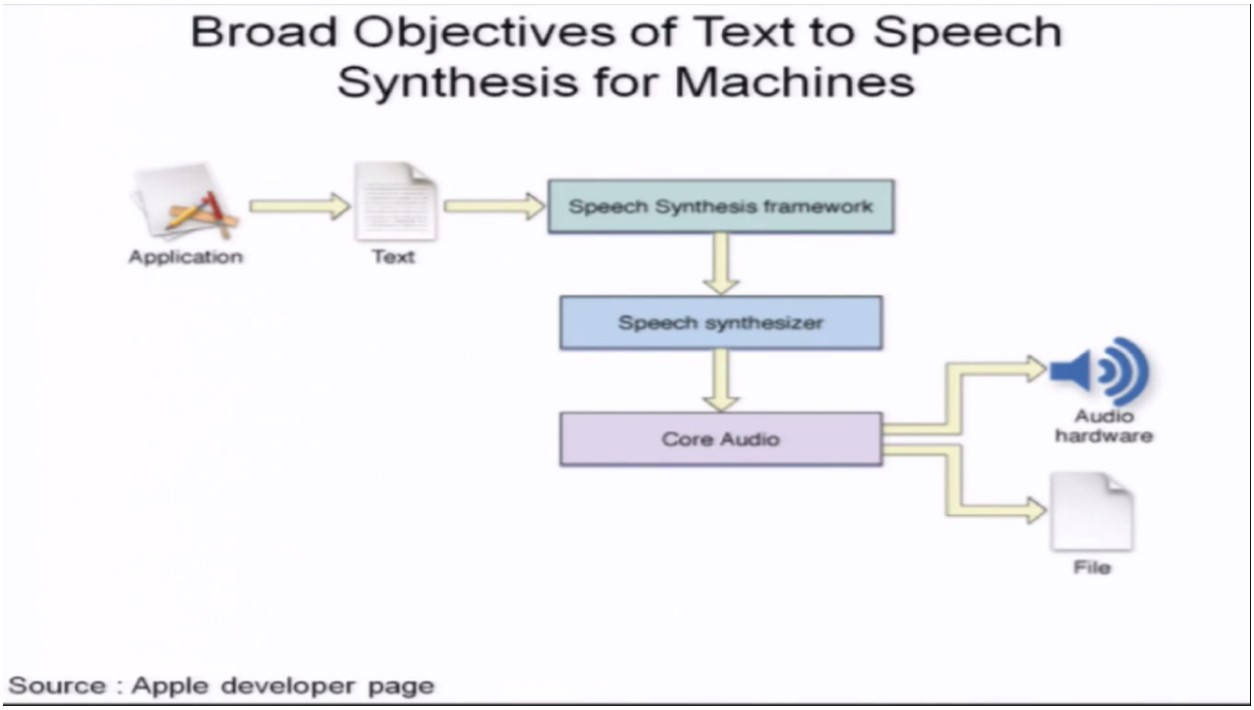
Dept. of EE

Indian Institute of Technology Kanpur

Welcome to this talk on speech technology for the MOOC. We are now looking at the second part of speech technology where we will focus on text to speech synthesis.



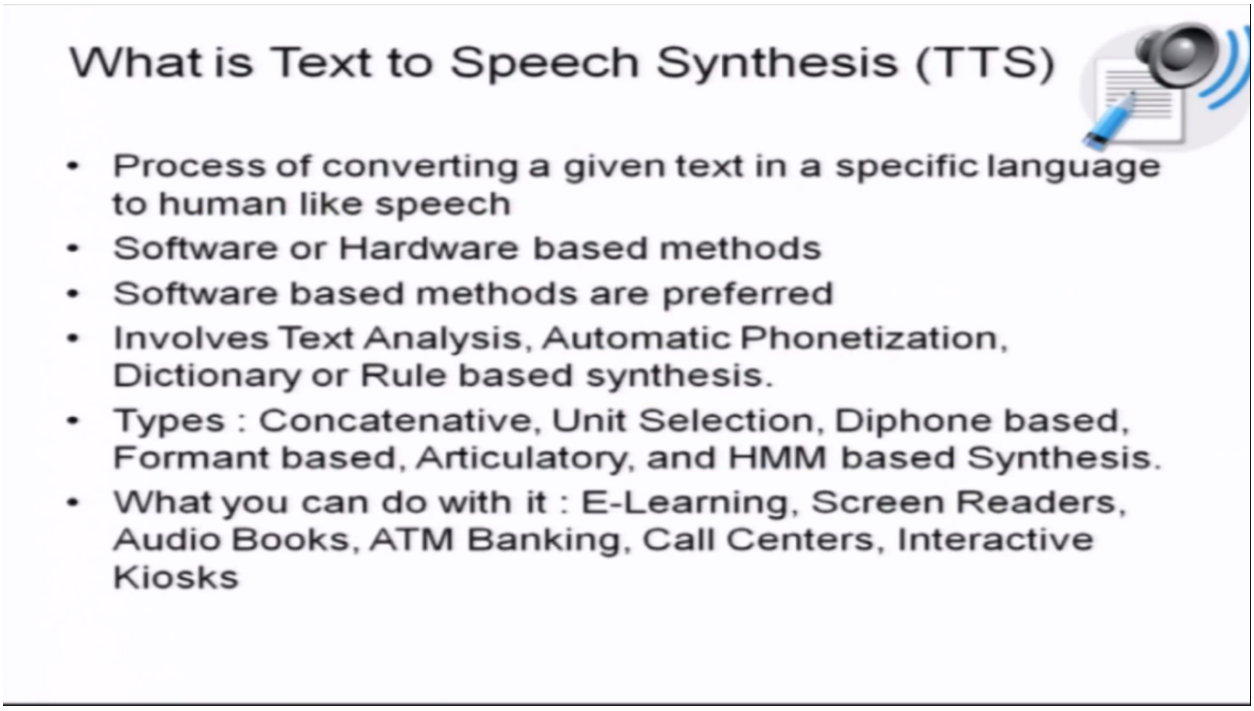
So what will we discuss or the topics covered in this lecture are as follows. We will discuss what is text to speech synthesis, what are the challenges in implementing text to speech synthesis on a cell phone, and we will also take a look at a couple of applications which are essentially in the agriculture domain. So we will talk about voice based agriculture information systems, a generic description of that and will take a particular example of what we call the digital market for the Indian farmer.



Okay so we will start off just as we did in speech recognition by defining what are the broad objectives of text to speech synthesis for machines. So for that let us consider this diagram that we have here. So you noticed in speech recognition that it was a process of converting the speech signal to text or signal to symbol transformation. Now text to speech synthesis essentially is exactly the opposite of speech recognition. So given the text you would want to synthesize the speech signal.

So for that if you take a look at the diagram here. You have an application. Now the application could be anything for example you are pulling information from a database or you are probably typing in some text or you have a document that you want to read out. So that is essentially converted to text. So either the information is available in the form of text or it needs to be converted to text. Some info coming from the application needs to be converted to text.

Now this text is fed into the speech synthesis framework. Now the speech synthesis framework does some initial processing on the text for example things it does one of the simple examples would be smoothing on text and then it will plug it in into a speech synthesizer which is the heart of the text-to-speech synthesizer. Now that speech synthesizer output is piped to the core audio system of the cell phone or the machine now using which you generate the required audio. Now this the broad objective of text to speech synthesis for machines is therefore to convert text to speech.



Now let us look into a little bit detail as to what is text to speech synthesis. So the definition process of converting a given text in a specific language to human-like speech. I like the stress on human-like speech because typically the output of speech synthesizers even in the state-of-the-art applications is not human-like. It is not smooth. you et this synthetically sounding signal. Now the synthesis can be done using two methods. One is software base. The other is hardware base. Initially, when text-to-speech synthesis started they used to have this very fancy looking machines which used to use you know parts which would generate different sounds and therefore they would produce complex speech but software type of speech synthesis is the most preferred.

Now what are the broad components to text to speech synthesis? So it involves text analysis. The first step would obviously be text analysis. So given the input text you would want to analyze for what language it is, what is the construct of the sentence, what is the continuity of the sentence, where are the word boundaries, etcetera. Following text analysis you typically do what is called an automatic phonetization; what do you mean by automatic phonetization is given the text for example we took the example of a phrase called their car. We split their into three phonetic units. So obviously in text to speech synthesis also you would want to break down the given sentence into the most minimal parts which are phonetic units and therefore you need something that will do automatic phonetization breaking down the sentence into smallest phonetic units is automatic phonetization.

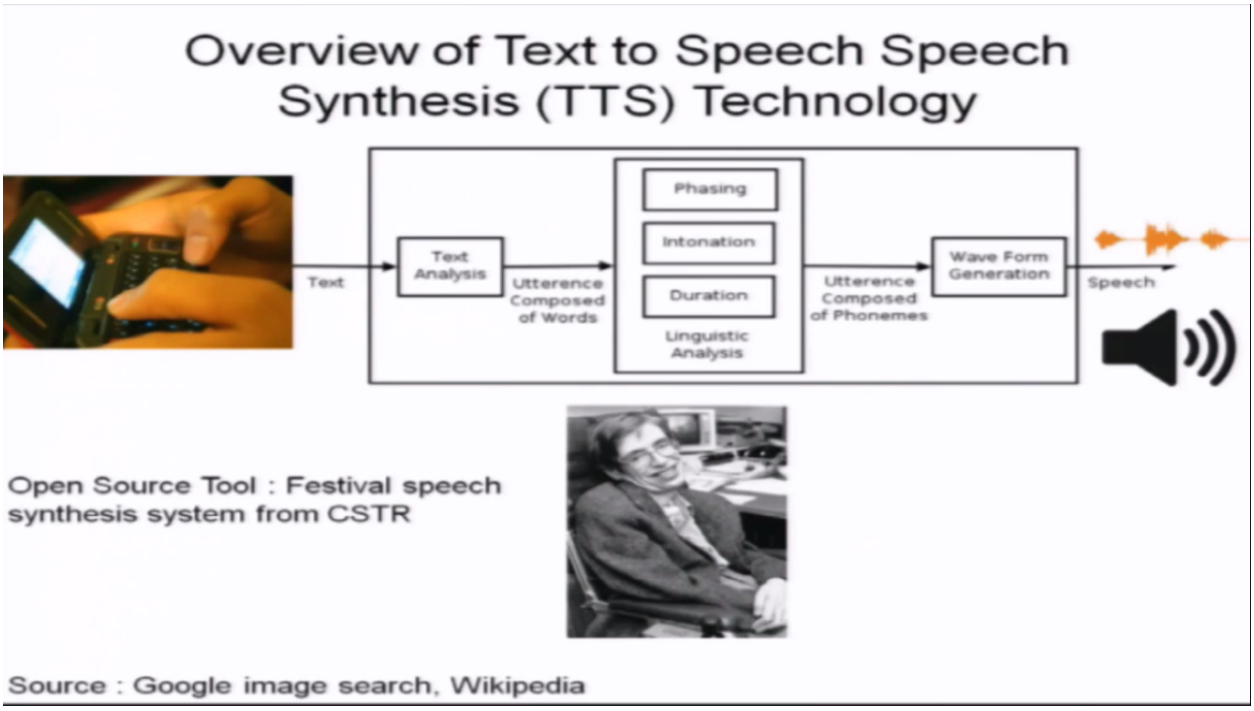
Now apart from breaking it down now when you synthesize a sentence you also need to know what is the underlying grammar or what is the underlying dictionary. And therefore that is a very important thing which we call as dictionary or rule-based synthesis. Now what are the types? What are the types of text-to-speech synthesis? So the first one is concatenative speech synthesis.

So what do you mean by concatenative is speech synthesis? As we discuss a sentence can be broken down into smaller phonetic units. So obviously we are looking at recording these small phonetic units and putting them all together in a sequence and playing them all. Now when I do this obviously I am going to get the same sentence that would have been produced by me but the only problem there is you have a very synthetically concatenated signal which needs some kind of smoothing. So that is the basic way.

You have more advanced methods like unit selection. You got diphone based synthesis. Now what do you mean by diphone based synthesis? We took a look at individual phonetic units. Now if you combine two of these phonetic unit you are actually creating a unit that is of a slightly larger duration. Now when units, sound units are of a larger duration they tend to produce smoother sounds then compared to sound units that are very smaller in duration. So combine two phones it becomes a diphone and a diphone based text to speech synthesis system is something that does concatenation on the diphones.

So you also have some technically more advanced schemes like formant based synthesis. Articulatory based synthesis and you also have something called Hidden Markov Model base synthesis. They are all more technically advanced schemes but they use some kind of signal processing to produce a better sound.

Now what you can do with text to speech synthesis? There are several things you can do with it. One of them is E-Learning. So when I say E-Learning text to speech synthesis is being widely used in learning languages these days. For example if I want to learn a new language I know how the text looks like and I would like to listen to the text. Now text to speech synthesis allow me to interactively learn a new language. It essentially also tells me how a particular word is produced. It's pretty interesting in that sense that you do not know how I know I do produce a word or a sentence I speak a sentence but I am not doing it in the way a native speaker would do. So if I have a text-to-speech synthesizer it essentially trains me to speak the language in the way the native speakers do. Fine. This is one of the applications. So you have screen readers. I think most of you are used to something called Amazon Kindle these days. So you have a text-to-speech engine incorporated into Kindle which instead of reading you can probably listen. Now this will help people who are you know specially-abled and not able to read things. You have audio books. You have ATM banking where the banking machine will tell you what you did. You have got call centers and you have got interactive Kiosks where you can actually information. We will take an example of a campus directory information system. So for example you are in a university campus and you want to get information on reaching a particular place. So you ask where am I and machine will probably tell you at which location you are and if you want to go to a particular location name the location or speak the location it will give you the routing to that particular location. Fine.



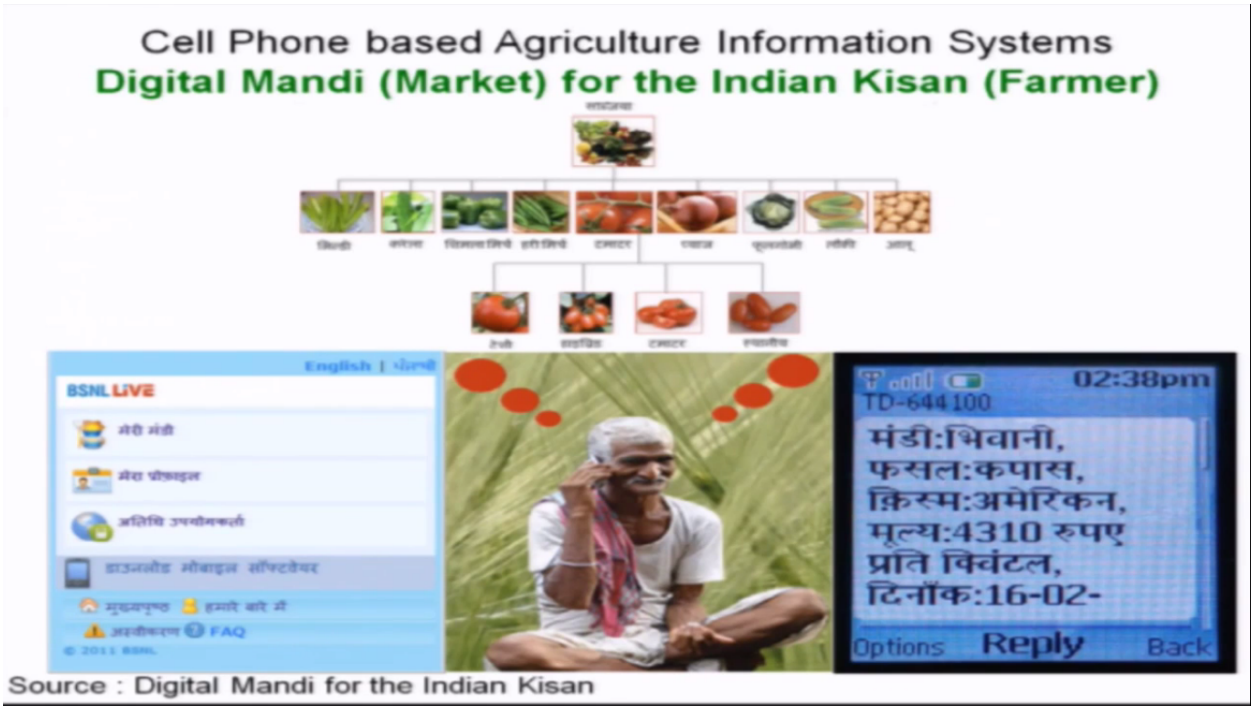
So let us look a little bit more detail not very technical but a broad detailed diagram of text-to-speech synthesis. So if you can take a look at this figure. Let us take this very simple example where someone is keying text into a cell phone. Right. So I want to key in text into a cell phone and I want to listen to that. So I key in text. The first block if you can see is text analysis. So what the text analysis does is it analyzes the text for various things and it actually creates an utterance composed of words. One of the simple ways of looking at it is given a sentence, it breaks it down into words. Now the words are broken down into phonetic units.

Now once these are done there are several things that need to be done. The first one or the most important part is something called linguistic analysis. You do analysis on the language in which the text was piped to the system. So the three main things in linguistic analysis are phasing, intonation and duration. Duration is essentially trying to find out that particular unit is how long in a particular language. The second one is duration of course and the third one is called phasing. Now the phasing is how are these words or units let us take units in a particular word. How are these units phased so that the speech that you produce looks natural in that particular language.

Now following these analysis you have utterance composed of phonemes or phonetic units which you have discussed. Now given the phonemes you pick the waveforms corresponding to that phonemes, concatenate them and what you get at the output is this synthesized speech signal. So for those of you are very technically savvy you can look at several open source tools. So you have got something called the festival speech synthesis system. It's kind of you know a software which you can download install on your machine and pipe in your text and see how it sounds. So it has this basic structure where you can actually tell which we want a male or female speaker speaking. You want uniphone or diphone synthesis etcetera. One of the classic examples if you take a look at the picture you know who this person is and he uses text to speech synthesis.

One of the classic examples where TTS is being used is this. I'm sure you'll be able to search and find out who this person if you are not sure.

Now what are the popular commercial applications? I explained this already Siri and Google Voice. Now how does text or speech synthesis play a role here for example if I am querying a restaurant using Siri. Now of course the system pulls out the restaurant but it then also speaks out to you and says that you have to do this. You get essentially speech based information in the sense that probably you are driving a car and you don't want to see take a look into the map it will reproduce that as a speech signal you could probably listen to it and drive to that particular location. So is the case with Google Voice.



Now what are the applications of speech synthesis in some let me take an example of a particular application that we have developed at IIT Kanpur which uses text-to-speech synthesis. I like to stick to only agricultural information systems. So let us take this particular example of a cell phone based agriculture information system. So what we do in this system is essentially the farmer is able to access the crop prices in his native language.

Now he's not going to call into the system in this particular application but he is going to register one time using his cell phone say for example let us take a look at this slide. So he uses his cell phone, registers for say three crops and three markets that he wishes to have information using this kind of interface. So you have a profile created for the farmer. You can take a look at the rightmost screenshot. A profile is created from him. He says he wants information about three crops of three markets on these days and this application has got two modes of transmission; one is SMS. The other is voice calls. Now what we are interested in the context of this topic is essentially the voice call part. So here we use the text-to-speech synthesis system.

So how do you use it? You can see these three SMSs that we generated. These are typical examples for those of you do not understand the language this SMS are they are in a language in an Indian language Hindi. So let me read out the first SMS to the leftmost side. So what the SMS is essentially is on February 16th in a particular market called Hodel the price of a particular crop ladies finger in this case was 2750 per quintal. So essentially if it was in English we do have systems in English but this is for a particular language. Now the leftmost SMS that you see will be transmitted to the farmer. But if he is not able to read the SMS he also gets a voice call which essentially reads out this particular text as speech.

Now for this purpose we use text to speech synthesis and most of the technology that I have described earlier is used in producing the speech signal from the text that I've shown here.

So broadly what we have done in these lectures on speech technology is that we have discussed what is automatic speech recognition and what is text-to-speech synthesis. We have also discussed how this technology can be used in developing applications on the cell phone by taking two examples of socially relevant applications where a farmer can access agricultural commodity prices.

So I hope this lecture has helped you thank you and if you have any questions you feel free to interact on the MOOC. Thank you.