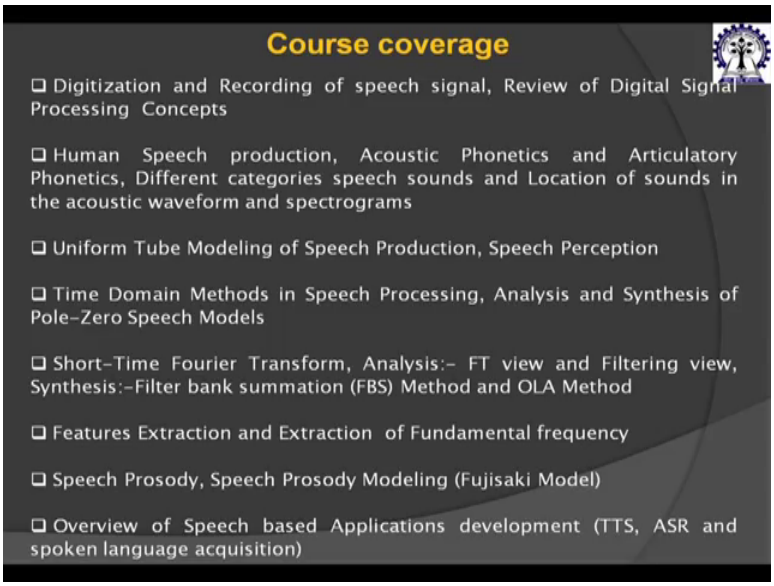
Digital Speech Processing
Prof. S. K. Das Mandal
Centre for Educational Technology
Indian Institute of Technology, Kharagpur

Lecture – 01
Introduction to Digital Speech Processing

So welcome, welcome all. So, this course is digital speech processing. So, I will take this course in 20 hours that means, that half 20 hours lectures. And this course mainly designed for speech part not that sub computing and that part. So, this course you do not expect that I will talk about that HMM or I can talk about that deep learning those things I will not cover in this course. This course mainly I covered the scientific or scientific aspects of the speech, and how those speech can be digitized, what kind of digital set up we should use, how do we locate the different kind of speech signal in that recording. So, all kinds of those things will be taught in this course. So, if you see the course coverage.

(Refer Slide Time: 01:17)

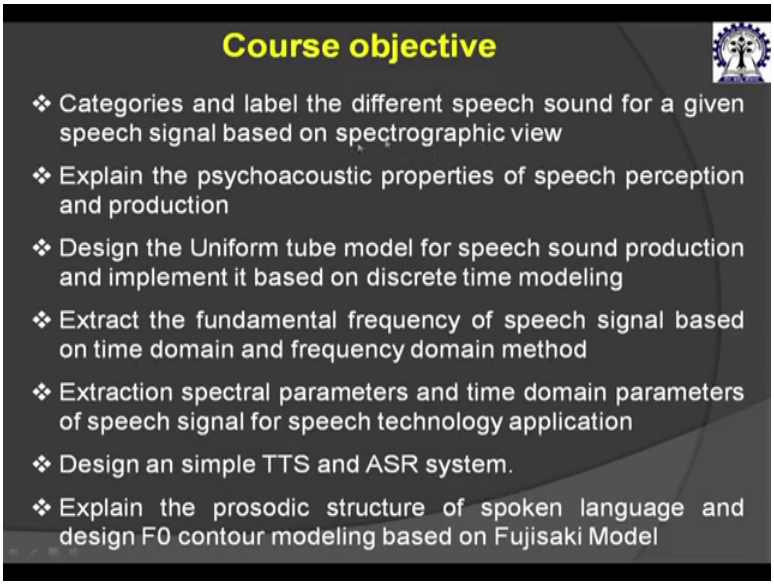


So, course coverage mainly covered if you see all are on the aspects of the speech signal aspects. Not that the aspects of soft computing and developing a system that kind of thins, but yes at the end of the course I should cover some of the portion or some of the important speech processing application like that TTS, ASR many people are talking about that TTS and ASR, but I am believing that there are not only the TTS ASR there is a other kinds of speeches speech application also, like that second language acquisition, action conversion all kinds of things. So, those will discuss I will discuss at the end of

the course that whatever the speech processing we have done entire course what is the main application of those kind of information in this course.

So, what I aspects from the every learner, that not that cover the course. From the learners prospective as a learner of the course of digital speech processing at the end of the whole course you should able to do following this course objective. What are the course objective? First is categorize and label the different speech sound for a given speech signal, based on the spectrographic view and time domain speech view.

(Refer Slide Time: 02:37)



So, not all those are written in the slides, I can you can I can say something on the in here also. So, if you see that categorize and label the different speech sound for a given speech signal based on the spectrographic view or time domain signal view. So, what I aspect from the earner? That, what is once I completed the course the learner should able to do? Or learner should able to categorize. Suppose I give you a speech signal and I told you to records a speech signal let us your name. You record your name you should able to record that that speech in the computer using the computer. Then using some you can say that open source software's there are many speech processing software's which are available, using those software you should able to view the spectrogram and time domain signal and label the different speech signal point.

So, for example, suppose I say the label that a consonant to vowel or tough consonant to vowel transition. So, those things you will learn here, how do I label which is tough

consonant what is look like in speech domain, and what is the transitory part what is content. So, those things you learn this course, and you should able to use that knowledge for a labelling the speech signal for a given or of a given speech signal whatever I say that you can. So, in examination also I will sometimes I will give you a spectrogram and told you that identify the words based on the manner of articulation of the speech.

So, I can expect that you should develop a skill not that theory, what is tough constant what is articulatory place of that velum that palate not all kind of things, but given a speech signal I should able to categorize and label the speech signal. Next one explain psychoacoustics and psycho acous or you can say the psychoacoustics properties of speech perception and speech production. That is not that much of skill oriented things, but yes you should know what is the how the speech is produced and how the speech is perceived. So, there is a lot of problem I will solve using the speech perception and speech production.

Next one is suppose you know that this is a speech production mechanism, that here the speech is produced using the vocal cords and there is a tube. You should have a experience with that if you see the flute. I can sing the flute and pressing the different hole and I can create the different kind of sounds. So, how a human being can produce different kinds of sound using this vocal tract. So, this is a tube and that is a vocal cords which is vibrate and create the sound, if you see the shehnai or if you see there is a membrane things in the beginning. And there is a long you can see that there is a long tube to produce the different kind of sound.

So, using the uniform tube model speech sound production and implementation of it is using the signal processing you should be able to do. So, I should explain how to model the this human vocal cords and how it should be implemented using the digital signal processing techniques. So, some part of digital signal processing I will cover, but that I am assuming that you should know the basic part of digital signal processing. So, using those principle I should able to implement that human vocal tract. So, that uniform tube model and I am mathematically model also.

Next one extract the fundamental frequency, or different kind of speech parameters. Why I required? If you see what is the propose of this course, if I say ask you the what is the

purpose of studying the digital speech processing. That is a subject then what is the purpose? If you see today in modern scenario or you can say 21st century there is a lot of research in speech domain. Why because if you think that human speech is the main communication media. Let us I take you the scenario, even a person does not know the or you can say the literate you can say that he does not know the script, he does not know the grammar, but he can speak effectively he can speak.

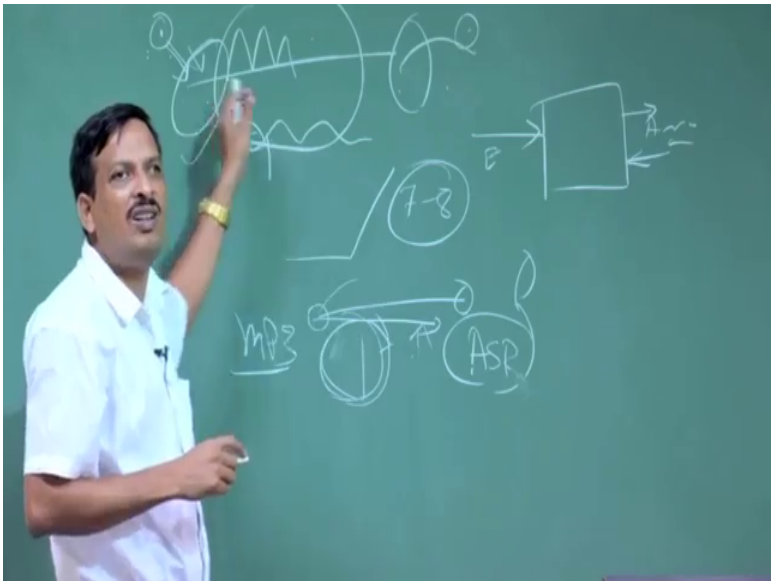
So, I can say the speak is the speech is the common natural mode of communication among the human being even speech is the sound is the communication among that any live things or you can that think about that also. So but if you see considering the human being speech is the most you can say that most easiest or you can every people use the speech mode communication or most natural communication medium. So, what scientist want? Scientist want can machine will act like a human being today there is a lot of artificial intelligence lot of soft computing let us of human intelligence we talk about.

Now scientist are trying to developed that can I developed an algorithm or can I developed an systems by which a human being can talk to a machine. Think about an application suppose you go to the railway stations to buy a ticket. So, instead of giving instruction to the machine by dialling 1 2 3 if you want this dial one dial 2 kind of things, you can replace with a chaos with a speech mode. I want to buy a ticket for say Kharagpur to Howrah. I just told to the computer give me the ticket and then computer ask how much money, the computer says at this much this amount of money give you the money done.

Lot of lot of this kind of continuous communication is required. Another if you think there is a lot other aspects speech communication. Think about the security biometry speech can be used for one of the biometry for the human being. That is why that speaker verification speak recognition human voice indentify all kinds of business are going on because speech carries the speaker biometry, because every speaker produce the speech to communicate the information during the production he impose some signature of that person also. So, that kind of things is going on. Similarly other like communication think about the pure communication thing forget about this kind of technology that human communication telephone I want to send a voice from one point to another point. Now how can I today if you see that? Today is that IBPS communication tie one that this channel cost is so high. So, I can I reduce the channel cost reduce the bandwidth. So, all

kind of things we want to do so that is will speech coding. How do I developed an efficient algorithm?

(Refer Slide Time: 10:19)



So, that with a minimum cost I can transfer a signal speech signal from here to here in a real time situation. So, I want some kind of algorithm or compression kind of things which can compress my speech and transfer that things. Think about that your CD, audio CD the music CD or you can say that vocalist that song CD. Earlier a CD can contain 7 to 10 song if it is stored in original format now think about MP3 a compression technique speech coding compression technique MP3, I can compress 160 song in single CD.

So, speech coding is also another aspects. So, while doing the speech this any kinds of application, the first thing is that I have to know the speech. What is the scientific aspects of the speech? After recording how it is behave like that. So, all kinds of things we have to know. So, the course aim is not to deal with that soft computing those things. Deal with that what is speech, how it is produced? Which features I should exploit to develop the speech base application? Which features scale is what kind of information? What is what do how the human being produce the speech? How the human being perceive the speech? All kinds of things will be covered.

So, in this course I will cover the extraction of different speech parameters details I will discuss during the speech parameter extraction what do you mean by parameter and what

kind of parameter you think is suitable. So, those kind of things we will discuss there is a signal processing algorithm we use those algorithm why we use this algorithm is very important. Do not read anything which is just this has written have read it I copied it and give the exam not this. Why this is important is very important. As a engineer or as a scientist you should know why I am doing this thing.

So, different kind of speech parameter extraction I will explain this parameter extraction algorithms are available in the book if you I have referred 2 books you know that you can go to those books the all kinds of algorithms are explained, but in the class what I explain that why we are doing these kind of parameter what is the advantage what is the disadvantage. Those kind of things I will cover in this class.

So, at the end of the course is suitable to write the algorithm for find out the different kind of speech parameter, that is my expectation. Then extraction of the spectral time domain parameters speech signal those things is then design of simple TTS. And ASR system I will cover and I will cover some part of the prosody modelling because today prosody modelling one of the important aspects in speech processing class.

If you see there is lot of work is going on speech prosody there is a lot of segmental speech work is done if you see the TTS most of the TTS are in even if in their language I have developed one TTS in Indian language using different method I will explain that also. But it is not as like natural as human being. So, what we are missing is the speech prosody. So, today if you heard any speech scientist any speech lab if you go there lot of people are working on speech prosody. How what do you mean by speech prosody, and there is a different application also speech prosody, suppose I give you one example that I have seen that this is a very important may be important research problem also. Think about that I have since I have dealing I am I am part of an center for education technology. So, I have seen many lecturing video which have recorded and even if foreign lecture video also I have seen which are recorded, but if you find there is a difficulty of understanding of the speech of different language speaker. Even all speakers are saying in the same let us English language. Suppose a Japanese people is giving a lecture on English language, and a Chinese people also giving a lecture on English language and I am sitting there, but I am not understanding fully his English or his or her English. Because it is come with a Japanese accent all kind of because if it is not the first language of is English is not first language of that Japanese speaker.

Suppose I am seeing a seeing a lectures of a our north Indian people or our south Indian people their English is little bit of different from the Bengali. Even my English may not be 100 percent understand or 100 percent you can say that intelligibility of that English is not that good. If you suppose you are a speaker of a American English when you listen this English you said this intelligibility of the speech is not that good. Now think can I make a device, I am saying in Bengali English and it convert to let us American English American accented English.

So, intelligibility of the speech is increased. Think about I have developed a systems and I am was the system in here. So, when you download the lectures and when you listen the English, accent of the English is converted as per your preference. So, that kind of I am not saying that language transformation somebody is giving a lecture in English I am listening in Bengali, I am not saying speech to speech conversion. I am saying simple action conversion that has I am speaking in the Bengali accented speech some American speaker listening it in a American accented speech. Same things an American speaker giving a lectures in American accented language I want to listen in Bengali accented language.

So, those kind of tremendous kind of applications second language acquisition also there is a some application of speech kind of things. So, even if speech research I will I will told the that the ASR lot people are doing research in ASR automatic speech recognition. And if you heard about that there is a HMM model hidden markov tool (Refer Time: 16:46) model, or I people are saying that, now people are saying that this model s not sufficient for those language which resource is very less resource constant language. So, what kind of alternative speech technology I can developed which is used the speech science? So, that I can think about new kind of model like that exploitation of speech prosody in ASR is a much more you can say the serious research which is going on by different group of people are doing it.

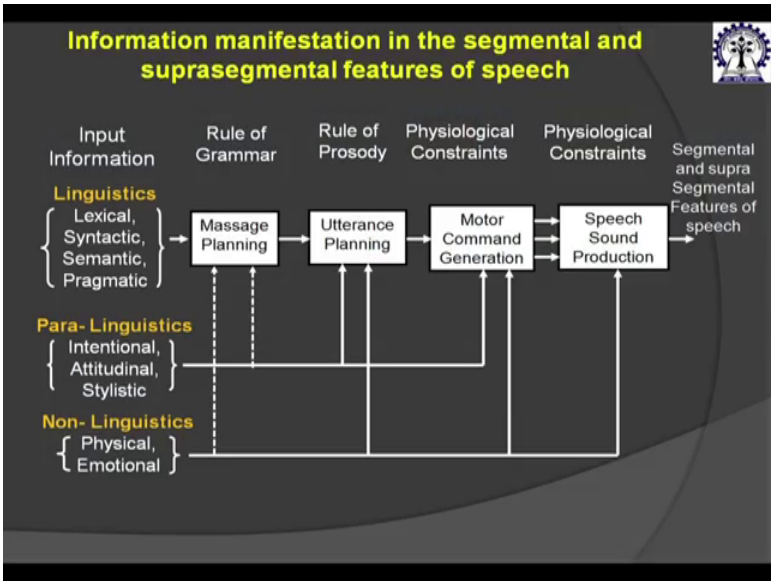
So, those kind of lot of speech applications are you can say the think about in your mind and you have to develop some expertise or some skill on which itself So that you can think about what kind of algorithm or technology or soft computing algorithm o have to sue to do these kind of things. So, this part I will this the, so if you see this my course outcome or course objective have never written the soft computing part. So, I am not covering on that part. So, that part may be cover in the other subject.

Now, since t is introduction class let us talk about that how human being produce speech. So, there is a suppose I am giving a lectures, how I am producing the speech. What or what kind of activity is going on in my body to production of this kind of speech? So, if you see the slides you can close look in the slides also, the slides will be shared to you that that I happening once I want to speak some message formulation is happening in our mind. And so, what is happening? If I want to speak a sentence what I want to say that is created that is that is come from our mind, and that is called message planning.

Once the message planning done then it is goes to the language model, what kind of language code you should use if it is Bengali then the words will be selected if it is English some others words will be selected. And based on the linguistic that is called linguistic in person linguistic coding it will come to neuro muscular action So that coding has to be executed by a human vocal cord. So, different kind of muscle has to be activity is involved.

So, muscle command that is neuro muscular action command will be generated. And that command will goes to the this whole function body. And this has a 2 part one is called source and one is called muscular action or acoustics system, and to produce the different kind of speech. So, message planning then the linguistics part. So, if I explain this one in the block diagram basis.

(Refer Slide Time: 19:45)

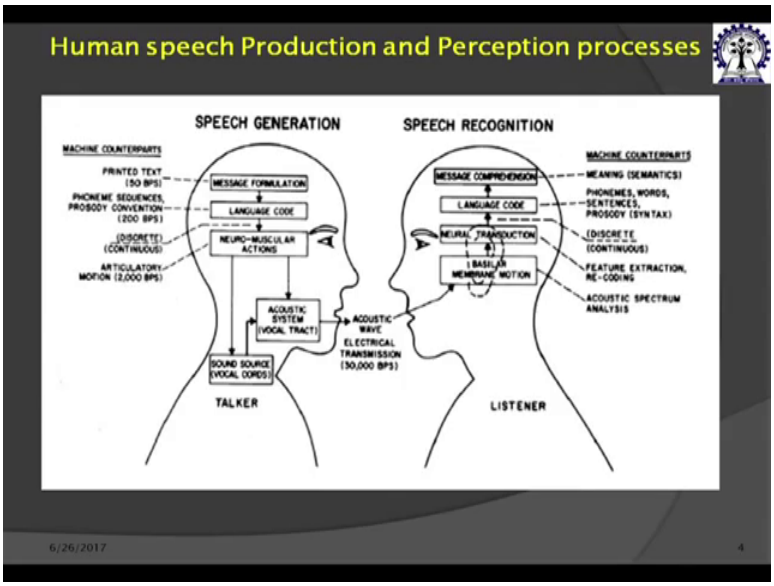


So, message planning that is the rule of grammar, what kind of word I should say which is well, this is unknowingly without grammar I can produce this speech. So, grammar is not important to generate the message planning. The person who does not know the language he can also speak in that same language. So, this is although it is it is rule of grammar it is automatic. So, grammar is not primary criteria. Grammar is discovered by us to explain the phenomena of what message planning is going on.

So, speech message planning which is lexical syntactic, semantic and pragmatic. Then come to the rule of prosody utterance planning. How I produce the speech I can produce this very excited manner, I can produce the very low manner, I can may be the very sad. So, those kind of prosody planning will come here utterance planning then motor command will be generated and speech production system is produce the speech.

If you see in message planning lexical syntactic, semantic, pragmatic then if you see the utterance planning paralinguistic, intentional, attitude, stylistic every human being has a different kind of speech styles. Then non linguistic parameters also there which is come under motor and utterance planning also. If you see the physic physiological problem also there somebody has a very thick vocal cords and very you can say the short kind of things. So, he can produce the very voice which is very the fundamental frequency is very low. And somebody has a different kind of style of speaking somebody has a stammering problem. So, all are come from here motor command generation tract planning, and then the speech is produced.

(Refer Slide Time: 21:36)



So, once the speech is produced, this is radiated either from the mouth or from the nose. The once I have produced the sound using this vocal cord, if you see there is a velum inside it details I will discuss. And either the velum will be closed or opened if velum is closed some part of the sound will come to the nasal cavity. And some or if the oral cavity is completely closed. So, it will come to the nasal cavity, if it is nasal cavity is completely closed it will come to the oral cavity.

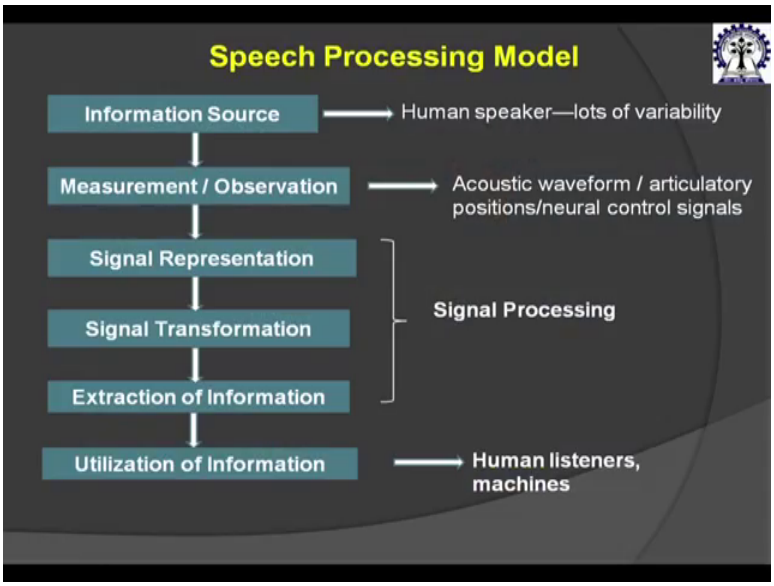
So, it is radiated so acoustics which is generated in here it is radiated from our mouth r nose cavity. And propagated in a acoustics wave. That is why is called speech acoustics is important. So, speech is produced by a human being. So, the acoustics wave is travel in the medium. You know that acoustics wave cannot travel without the medium. So, acoustics wave come in the medium and transmit it. Now once the acoustics wave is transited a listeners who is present to listen that voice the acoustics wave strike is in ear system. So, hearing system is there. So, the acoustics wave is there. So now, from there the acoustics wave has to be converted to the again neural signal. Then again neural signal has to fire the language code, and language code from the language code brain decipher the intended meaning or intended message of the speaker. So, listeners is again try to find out what the speaker want to communicate. Now if you see this is this is the process, but this process you can say that there is a lot of optimisation is possible. Means that human being has a tradition that in biological system we always please conserve the

energy. So, suppose I am speaking to some students, and once I his face gesture or I once I realise he is tune with the same topic, lot of message planning done which is very short.

Or you can say that the planning complete message or complete linguistically complete message is not required to transmit to the listeners. So, it may not be a linguistically correct sentence, or linguistically correct words or whole words I may even I am not speaken spoken, but listeners understand. So once the listeners understand speaker will say that I my purpose is complete. So, do not want to produce the whole sentence.

So, this kind of vulnerability that the how this optimisation is how this can be tackled in speech science, that is also important aspects. So, this is the speech production and perception mechanism of the human being. The details I will cover in the following lectures.

(Refer Slide Time: 24:40)



Now, if think about the engineering aspects so a human being, so if I see the engineering aspects that person speaking in the speech or you can say the I have generated the acoustics wave from my mouth. Now suppose this acoustic wave I want to transmit form this point to this point and these 2 points are far away. While this acoustics signal cannot reach here have you understand suppose I am speaking in this room the persons who are sitting in this room can be audible. But suppose you are sitting in your home you want to listen the same acoustics wave. So, I have to take the help of technology. So, what kind of technology? So, I have spoken the speaker is or radiated the speech in acoustics wave.

Now I have to use the communication technology which is electrical communication technology.

So, somehow I have to convert this speech acoustical wave to an electrical signal, and transmit the electrical signal and this side I have to convert the electrical signal to the acoustics wave. So, that is why conversion from acoustic wave to the electrical signal is microphone, and conversion from the electrical signal to the acoustics wave is the loud speaker.

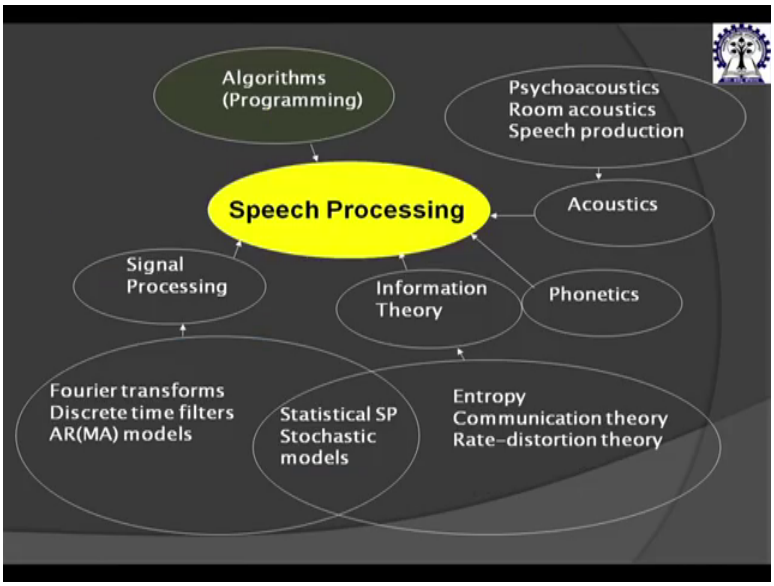
So, this is the mechanism I can I can say that I can transmit the speech from one point to another point. So now, I can deal with this electrical signal, whatever the coding technology all kinds of things I can do with this. That is why I say it is digital speech processing, I have not saying that the speech processing in analogue domain. I said digital speech processing, once I done I take this electrical signal in digital domain what kind of processing we should use.

So, that we can develop different kind of technology. So, if you see these slides the source of information is the speaker human speaker then measurement of observation this acoustics wave or a form has to be converted to the or you can say the acoustics wave from I have taken and observation has been done. Then the signal representation has to be done signal processing, then again human listeners I have to model the human listeners. So, that now if you see that technology I want to replace this speaker.

So, I say the message planning to acoustics wave generation that has to be done by a machine. So, I can say it is text or idea to acoustics wave. Here I want a machine which can converted that the signal which is coming here, or acoustics wave to understanding of the speaker's message. So, this kind it is not ASR, I do not say it is automatic speech not only recognition here understanding also is a part. Here also this figure modelling is not that he just read a text, once we talk we do not read the text, we understand the text and generate that text using the acoustics wave.

So, what all things are there we I will discuss on that things. So, that kind of technology. So, either technology should be here, technology should be here, or technology should be here. So, all kinds of speech technology we have developed to develop. Those kind of technology we should know what is speech and how it produced ok.

(Refer Slide Time: 28:23)



Now, if you see some speech processing is involved different you can say the dimensions or you can say the different discipline to work in speech processing. So, if you say the language communication. So, different kind of dimension. So, speech processing again involves algorithm it is psychoacoustics room acoustics speech production. So, acoustics part. Then information theory part, phonetics part signal processing part, statistical signal statistical signal processing entropy all kind of multi disciplinary actions has been involved in developing the speech technology ok.

So, I will some of the discipline I will cover not all the discipline I will cover. So, some of the discipline I will cover the mainly focused on acoustics part of the speech, and phonetics acoustics and phonetics and signal processing part of the speech. So, thank you. So, next lectures I will discuss how to record the speech. Because before you go to anything on speech processing, you should know how do I record my speech. And how do I see it in my (Refer Time: 29:29) computers. So, that whatever I say you can play with it. Because I say that do not it is not a theory type of class that I read this class and read this definition I will give the exam, not this. Developed and skill after seeing the speech I can able to say this probably the worst signal and it is coming from this region. So, those kind of expertise I want ok.

Thank you.