Digital Voice and Picture Communication

Prof. S. Sengupta

Department of Electronics and Communication Engineering

Indian Institute of Technology, Kharagpur

Lecture - 01

Introduction

(Refer Slide Time: 00:00:52 min)

Welcome to the lecture series on Digital Voice and Picture Communication.

(Refer Slide Time: 00:01:03 min)

This subject has become a very important one in today’s technology. We all are familiar with the voice communication and picture communication independently. But, I mean today the world has moved to such a kind of technological growth that in order to effectively and efficiently disseminate the information pertaining to voice, picture and the other Medias which are there very efficiently so there we need to use the digital technology.

Now we know that the last century had two remarkable inventions. One is the telephony and the other is the television. Using the telephony one was able to transmit the voice signals from one place to the other; I mean overseas calls were made possible; I mean one could directly talk to each other that itself was a remarkable discovery in those times and the television added one more dimensionality in the sense that where one could have the picture transmission as well as with that the audio information is there in a synchronized form.

Of course telephone was more like a one to one communication a personalized communication, television was like a broadcast where the interactivity feature was not there; I mean the television stations are transmitting the picture information along with the audio and then we are receiving it in our television receivers; no scope as such for the interaction. I mean even broadcast was possible over the speech and audio also in the sense of the discovery of the radio, that really gave rise to the voice communication or the audio communication over long distances where we know that different modulation techniques that can be employed so that the audio signal can be actually can actually modulate a carrier and then a long distance transmission is possible.

We know about all these technology. But since 1980s when the computers were available in plenty and especially the processing powers started in getting available at our desktops the need was felt that how to process the digital information pertaining to voice and picture. Picture of course will include both still pictures in the sense that which is not changing with respect to time and then the video information which is changing with respect to time as well.

Voice, as we just understood, voice is the normal speech signal that we are using for verbal communication. But when we require any high fidelity audio like....... when we listen to any music, a very high quality music that would require a different technology altogether in order to process the audio information very efficiently. Now, when the digital technology came in that time we had to really go in for a different technology as such in order to efficiently digitize and then represent the information.

So in this course this is what we are going to study. So we are going to see the technological aspect of it and we will see that how efficiently the voice and picture information can be represented and transmitted from one place to the other. So this is what concerns us and we will be understanding the technological aspect of this in this course.

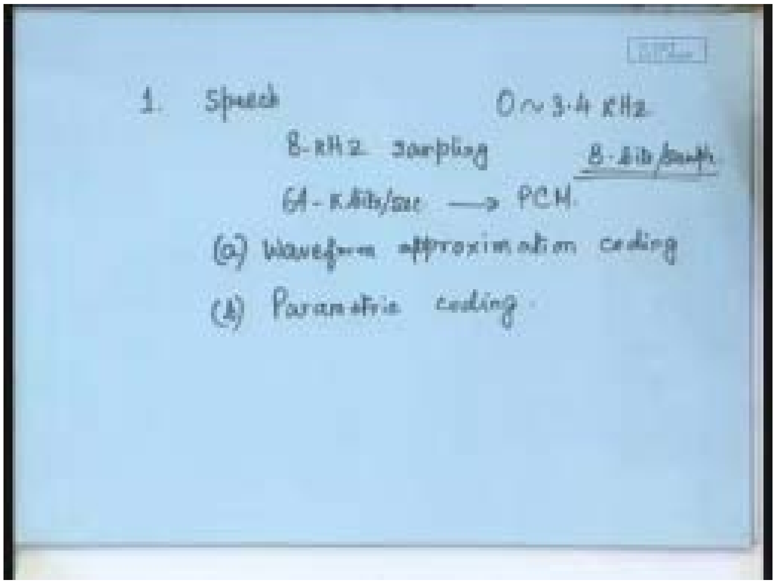
So this course as such will have some distinct modules. We will begin this course with the processing of speech information. So the first is the speech with which we will start and in order to efficiently represent the speech what we need to do is to efficiently compress the signal. We know one thing that the speech signal as such is normally restricted to a bandwidth within 3.4 kHz. 3.4 kHz is........ 0 to 3.4 kHz is a bandwidth that we require for speech. So, if we want digitize speech signal then we naturally have to sample the speech signal at a rate which is at least equal to or little higher than the Nyquist sampling rate. So given by that 3.4 kHz signal would require something more than 6.8 kHz of sampling rate and the sampling rate which is normally utilized in the digital speech processing is 8 kHz. So we require 8 kHz of sampling and if we go in for a regular pulse code modulation technique and we represent the speech samples, each speech sample if we represent by 8 bits per sample then that would require......... since the sampling rate itself is 8 kHz we will be requiring 8 multiplied by 8 that is a bit rate which is equal to 64 Kbps.so 64 Kbps per second will be the PCM or Pulse Code Modulation requirement considering 8 bits per sample.

Now 64 Kbps is something which.......... okay the least lines can provide you very easily a bandwidth equal to 64 Kbps but there are applications or there are situations when we need to compress the information even further. Although the bandwidth efficiency is increasing and the availability of bandwidth; today we have broadband availability very easily but yet you also consider that the number of users that has increased many folds. So as result of that no matter whatever effort you may make to increase your bandwidth at the same time the number of users are also increasing day by day and what you have to always remain concerned is that what is the amount of bandwidth that will be ultimately available for each of the individual signals or for each individual application and that is generally very restricted and it can be restricted due to many situations. Your channel capacity may be the limiting factor, the medium which you are using for transmission.

Naturally, for wire line transmission and for wireless transmission you are going to have different bandwidth requirement so going by that you have a requirement to reduce the bit rate further drastically. So the Pulse Code Modulation technique that we can think of as a very easy and simple representation of the speech samples, that needs to be augmented further.

So there were many speech coding techniques which were developed in the 80s, 90s and even lot of research is there till today also where efficient speech coding techniques are getting developed. Now, speech coding techniques are such we can divide into two broad categories. The first that we will be studying will be the waveform approximation coding and the other that we are going to learn after waveform approximation coding is what is called as the parametric coding.

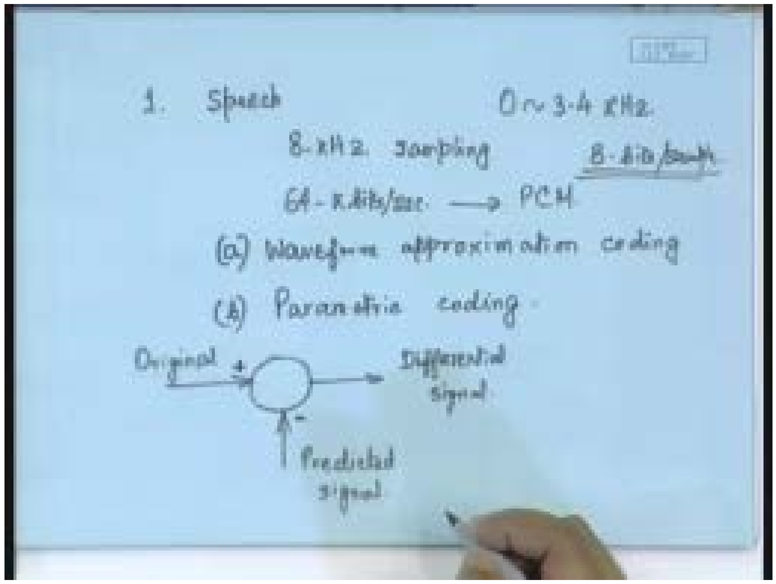
(Refer Slide Time: 11:42)



Now as that name implies the waveform approximation coding means that whenever you have any waveform bit speech or the signal communication can come from any input and it is a time varying signal. So we consider any type of a signal that is varying with respect to time and that can be approximated using some waveform standard waveforms coding techniques. So waveform approximation coding, well, you have a variety of such coding techniques which are available.

Now PCM, as I told you, is the basic representation but if instead of PCM for example we do some kind of a prediction of the signal like whatever is coming as an incoming signal if we are able to predict what it is going to be, how to predict? We simply use the previous samples. So using the previous samples if we can generate a predicted signal so we have a predicted signal and then we have the original signal so you take the difference between these two so take this with a plus sign, take this with a minus sign so what you have as the difference is the differential signal or what you can also call as the prediction error.

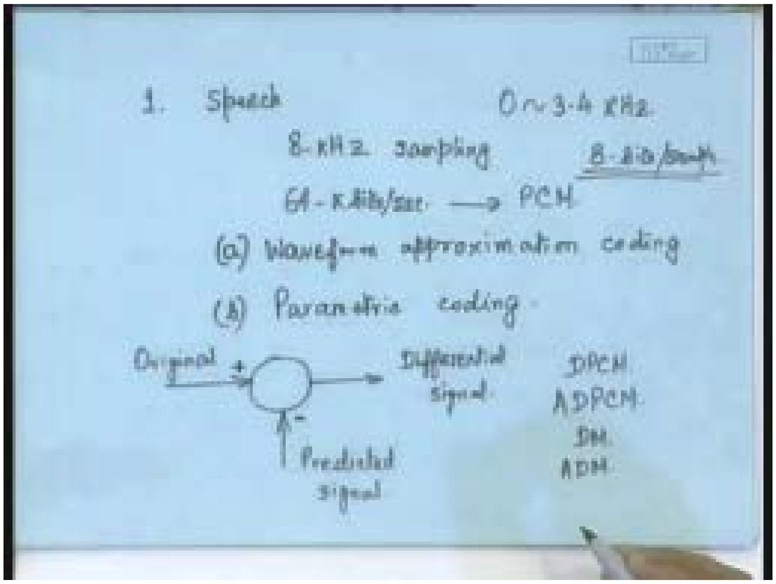
(Refer Slide Time: 13:29)



So this differential signal what is available to you, you can then use a Pulse Code Modulation technique on this differential signal. So you are going to call that technique as the Differential Pulse Code Modulation or in short form you are going to call this as DPCM. Now this is one such technique. There are many other techniques which we will come across while discussing about the waveform approximation coding. Some of them are very efficient ones. We will be talking about the regular differential pulse code modulations scheme. We will see how to make that process adoptive and we will talk about the Adoptive Differential Pulse Code Modulation scheme or the ADPCM.

We will also learn about the Delta Modulation which in short form we call as DM where sample by sample we just quantize it to one of the two levels. Again the differential again this Delta Modulation one can use with an adoptive control over its step size so one can have an Adoptive Delta Modulation or called as ADM. So there are many such waveform coding approximation coding techniques which are available.

(Refer Slide Time: 14:44)



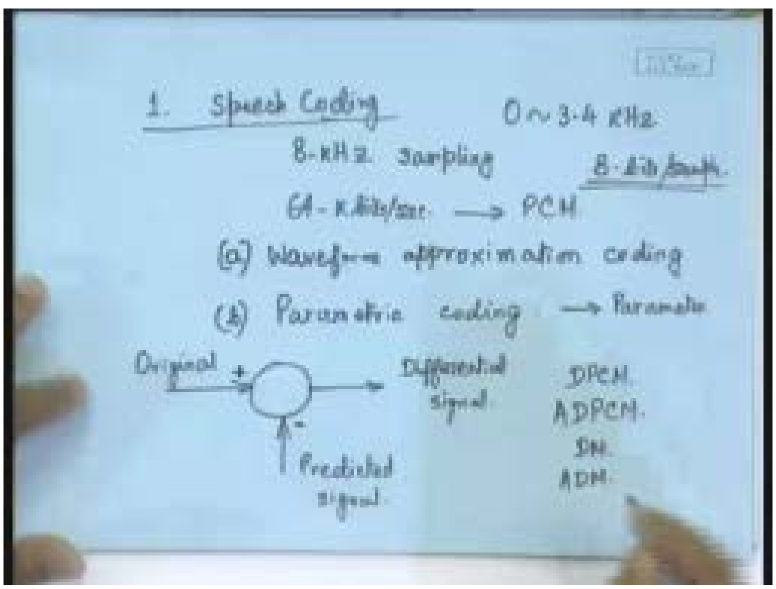
Now I was telling that all these things can be applied to speech signals even for other time varying signals also. So there is nothing special about the speech that we can talk of as far as this waveform approximation coding techniques are concerned.

You can apply for speech just like the way you can apply for any other varieties of signals be it biomedical signal, be it any one dimensional signal coming from other sensors you can very easily use that; there is nothing particular to speech that you are using. but whereas the next technique the next group of techniques which we are going to learn that is to say in parametric coding, as the name implies that parametric word has come from parameters.

So what we do is that the speech signal is represented in terms of certain parameters. so if it is possible to extract a set of parameters from the speech which can characterize the speech information completely or at least we can say reasonably then what one can do is that instead of sending the waveform approximated signal or its differential form or whatever instead of that you only send the parameters and using those parameters if at the decoder end or at the receiver end one is able to extract the parameters and reproduce in the same way as that of the techniques which was used to extract the parameters you just do a reversal of that. So using those parameters you synthesize the speech signal. And if you can do that then that requires much less information to the much less number of bits to represent that information as compared to that what you would be doing normally with the waveform approximation.

So definitely if 64 Kbps is the reference that we use for the pulse code modulation one can come down to a bit rate of 16 Kbps or let say 9.6 Kbps some kind of rates like that one can have using efficient waveform approximation coding techniques. But if you want to reduce your bit rate further in a very drastic way then you have to go in for something like a parametric coding technique there because it is parameters can be more efficiently represented.

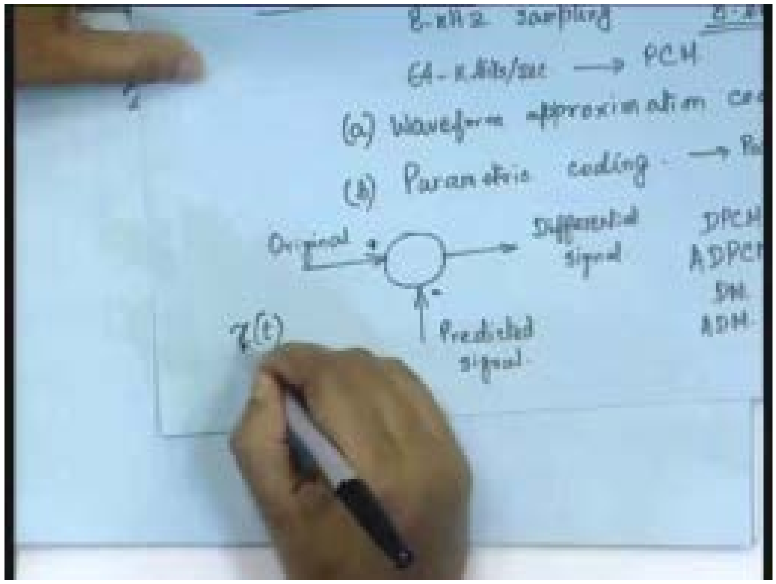
(Refer Slide Time: 17:40)



So we will be studying, I mean under the speech coding techniques we will be studying both these aspects that is the waveform approximation as well as parametric.

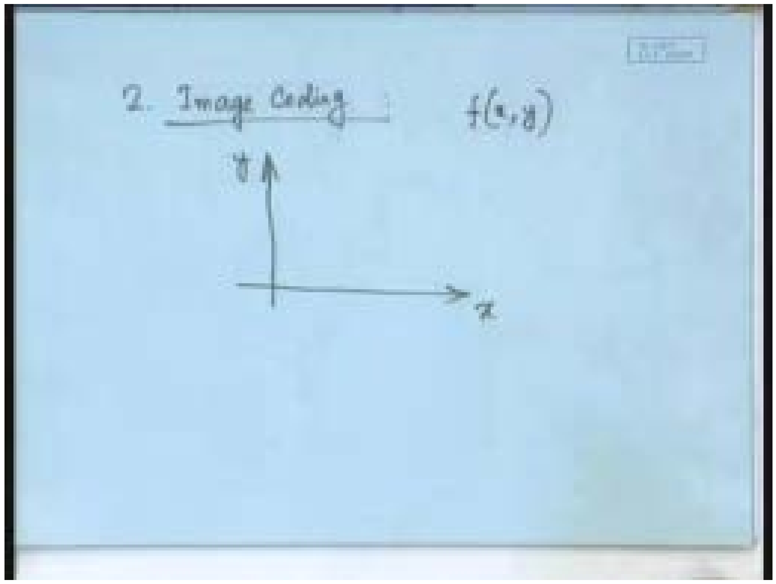
Next we come to the image coding technique. In image coding actually image is a signal in the sense that speech is a one dimensional signal, there the variation of the speech signal is only with respect to time whereas in case of images it is a two dimensional signal whereas speech signal we can call as x(t) where t is the time with respect to which the signal x is changing.

(Refer Slide Time: 18:29)



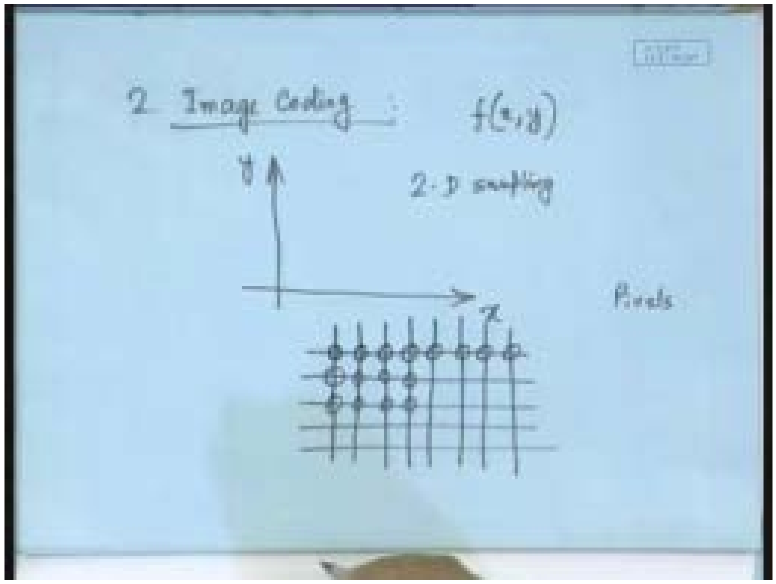
For the images we have to take the signal as a two dimension so we represent the image intensity by a function f and we call this as a function of x and y where x and y are nothing but the spatial direction. So x we can take as the horizontal direction and y we can take as the vertical direction. So using this x and y coordinates you can specify that which particular point in the entire image that you are looking for.

(Refer Slide Time: 19:16)



Now, here it is digital, I mean just like the way the speech also in order to process in the computer we require sampling and after sampling we require a digitization and then it is the processing of the digitized speech samples that we are considering. In a very similar way images also are to be digitized and the image signal which will be coming from the camera that has to be represented in a two dimensional form by proper sampling the image information on a grid. Grid means something like this that supposing the grid spacing is this say (Refer Slide Time: 20:10) these are the spacing of the grids that we use horizontally and these are the spacing of the grids which is in the vertical direction; we show the grids like this. So it is an array of picture elements that we form, so we call those arrays of picture elements as pixels and those pixels are nothing but the intensity as you find at the grid location. So, at all these grid intersections the intensity that you have are nothing but the samples of the image. So it is a two dimensional sampling that you have.

(Refer Slide Time: 21:00)

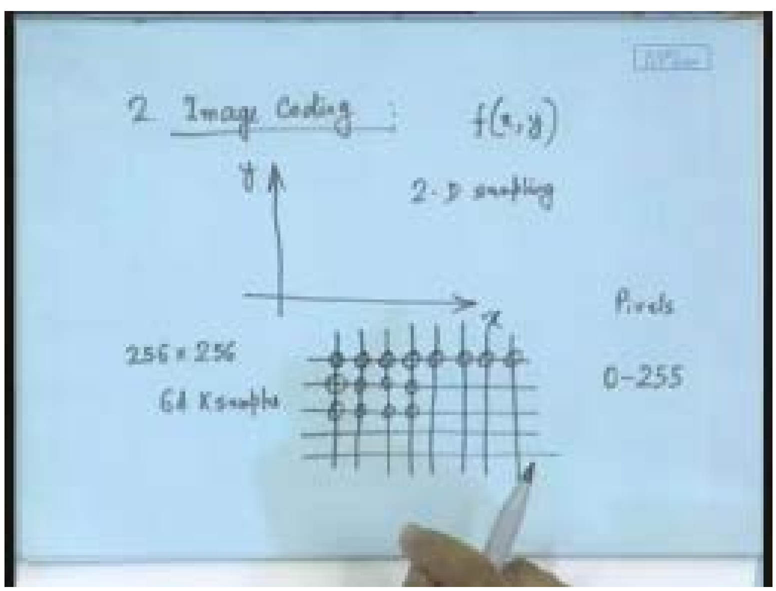


And as you can very well understand that closer your grid spacings are better will be your resolution with which you are representing the pixels in an image. And then again if the question is asked that how to decide on the spacings of these samples or rather to say the grid spacings for that matter also we have to go in for the Nyquist rate.

Just like the way for time varying signals we have a maximum bandwidth which is available or we rather band restrict the signals before digitization there for the speech it was 3.4 kHz in this case the bandwidth or the frequency content that we have going to talk of is the spatial frequency. So here it is a variation with respect to space and again it will be having two components of variations: one is along the x and the other is along the y. So, depending upon your spatial frequency content you have to decide your sampling frequency. In this case the sampling frequency also will have to be in terms of the distance or inter-sampling distance. So if your frequency content is more obviously you need to sample in a much finer way, you need to have more samples within one unit length. So this is where we will be ultimately finding that the image signals will require lot of information to represent.

Even a single image just you imagine.............. let us take a very typical size say we take a grid of 256 by 256 image samples which means to say that there are 64 kilo samples in an image or you call it as picture. Now each of these samples can be represented by 8 bits of representation for intensity. So, if we have 8 bits for intensity then we have the intensity levels between 0 and 255 where 0 will be the darkest intensity or it is the minimum intensity, the blackest one and 255 will correspond to the maximum brightness. So it is the extreme white pixel that will be having intensity of 255 and in between its intensity variation will take place within the range of the 0 and 255. So something like mid-range 127 or 128 intensity equal to 127 128 you can say as a gray that is something which is in between the black and the white and if it is slightly above somewhat above 128 then you can say that it is a gray image but closer to a white. If it is 64 you say it is a gray image but closer to black. So we call everything as the gray scale. But if you have to transmit the image information from one place to the other just think over that how much of that how many bits you require. 156 by 256 example 64 kilo samples and each sample requiring 8 bits per sample so 8 bits per sample which means to say that you have got 512 kilo bits of information for every image.

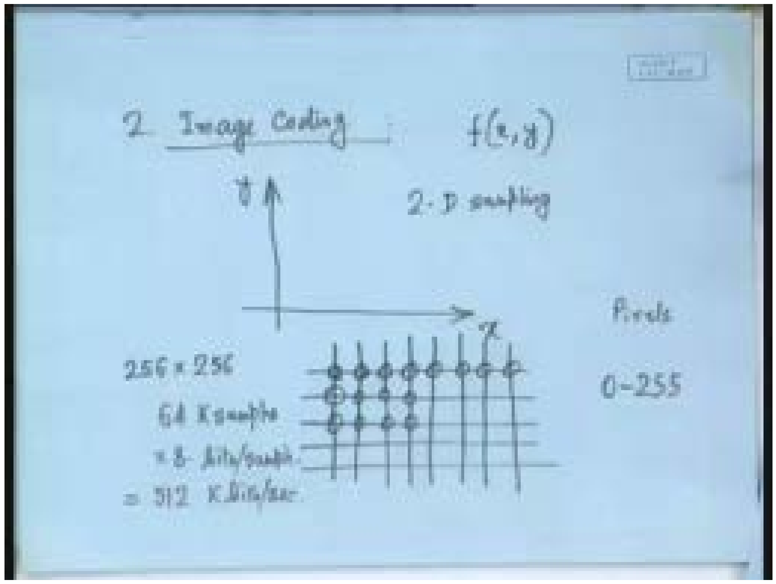
(Refer Slide Time: 23:43)



Now you can say that let me take my own time, if the bandwidth is restricted so what I will be transmitting my image slower. Well, all the time you cannot do it and you cannot do it for all applications. Like say, for example, supposing a medical conference is going on, some doctors they have met over a video conferencing and they are located at different geographical locations on the globe. One doctor may be in India, one doctor may be in UK, another doctor may be Australia, somebody may be in USA so we have them distributed all over the world. Now, whenever we are transmitting an image to them naturally some decision making is to be made based on some medical images may be that the ultrasonography image or may be the x-ray image whatever. So naturally we have to transmit the image in a very efficient way so that it can be encoded without appreciable loss of time. So naturally the computation time is at a premium that is one fact and the second is that if you want to transmit the image within a very small time then you also require more number of bits to transmit so you require more efficient bandwidth. But in order to do that you have to apply some very efficient coding techniques or very efficient way of compressing the information.

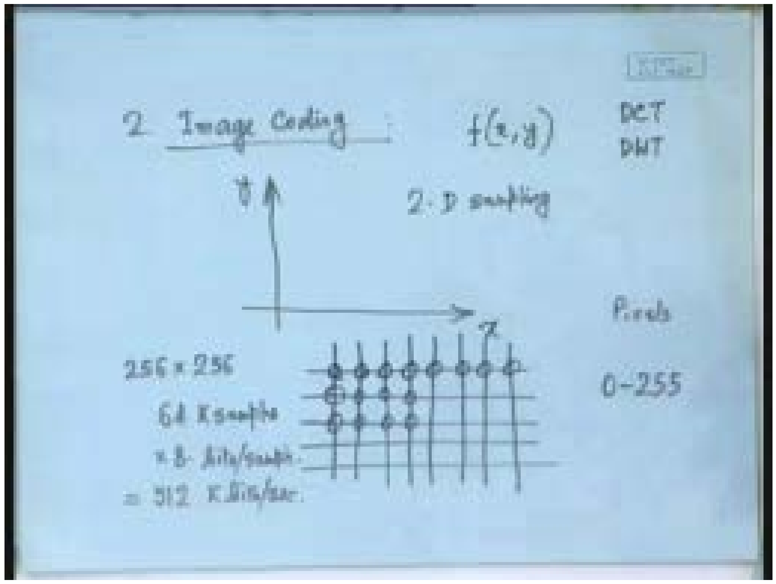
Actually for the speech signal or for image signal we luckily have lot of redundancy which is present. Now think that why for the speech we were talking about a differential encoding. Because successive samples of speech happen to be somewhat correlated to each other. If we pick up a sample at some instant the very next sample is not going to be very drastically different from the present sample. So instead of representing the absolute magnitude of that sample we say that we take the differential why because the differential is going to have much lower dynamic range so we require less number of bits to represent that. It is the same with the images. Successive pixels are going to have more or less very similar intensity. So there also you can adopt various techniques in order to reduce the information content.

(Refer Slide Time: 28:17)



First of all you can use the same techniques like the differential pulse code modulation even for images also. like say for example, one sample predicting the very next sample and then we represent the error content in that or otherwise we exploit the spatial redundancy in a different manner in the sense that we transform the image from the space domain representation to a different domain where the representation is more compact, more energy efficient and we choose that kind of a scheme for image coding. We will be learning all these in this particular course.

(Refer Slide Time: 29:28)



So under the image coding we will be learning about the different computation techniques and especially we will talk about the discrete cosine transform which is a very popular technique and has been adopted in the standards in the image coding standards that has evolved internationally. So DCT or Discrete Cosine Transform we will be studying. And also, we will be studying about the discreet wavelength transforms because wavelength is another very efficient way of representing the images where the image can be analyzed into a multi-band component and then each individual band can be efficiently coded and this gives rise this gives us a flexibility that what is called as the space frequency localization. So we can know that which particular spatial location of the image has got what frequency content and that can be very efficiently represented using the discrete wavelength transform.

After the image coding we will be coming to the topic of video coding. Now the difference between image and video is just that in the case of image the variation was only in space. But in the case of video the variation is not only with respect to the x and the y directions that we consider for the images but also with respect to time. So here we have actually three dimensional signals in the sense that the image intensity f will not only be a function of x and y but it will also be a function of time t. So, in the case of video signals what we have to take is that, there will be a sequence which we capture from a video camera and the video camera normally captures the sequence at a very fast rate; fast rate in the sense that we are habituated to capture the video at 30 frames per second.

Why the rate of thirty frames per second?
It is because in video or in continuous motion case; the same situation we also have while showing the movies or even for transmitting the television signals also. Since we have to give to the viewers a feeling of continuous motion there should not be any jerkiness that is present in the playback of the information that we do so that is why the rate at which a video frame is to be replenished should be as good as possible.

Now we can say that why 30, why not 60 or why not better than that. Well, ultimately it is the human detector that we have to consider because ultimately your eyes will be able to detect. So, if your eyes can have the feeling of continuous motion at 30 frames per second then why not restrict to 30 frames per second. Of course people have managed with frame rates of 10 frames per second or say 15 frames per second or I have rate that in some applications there is a severe bandwidth restriction; people might have worked with 7.5 frames per second also.

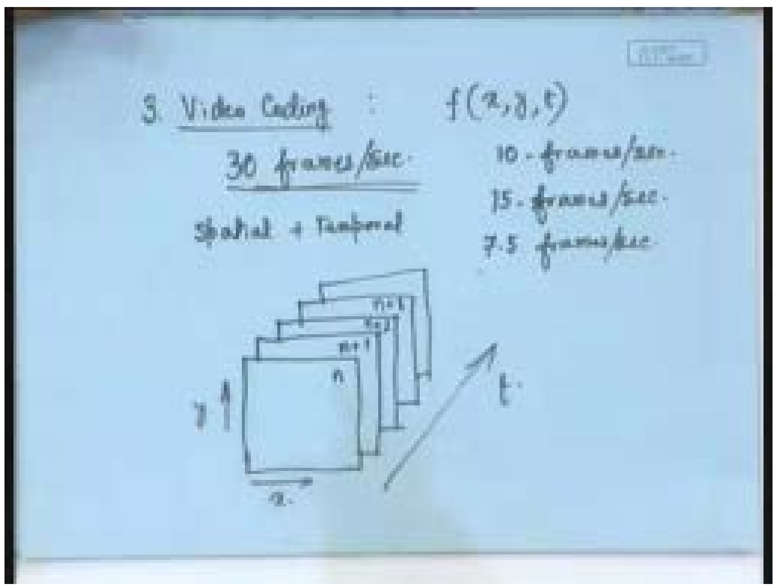
Now, if your frame rate is coming down like that 10 frames per second or even below 7.5 frames per second obviously you are going to have some amount of jerkiness in the motion content of the scene. but nevertheless 7.5 frames per second if you take, then your information content is also much less because in 30 frames per seconds within one second you have to transmit the content pertaining to 30 frames whereas here it is just one fourth of that, 7.5 frames per second of course definitely at a loss of perceptual quality.

(Refer Slide Time: 33:28)

So there is always as we will be learning through this course also that there is always a kind of tradeoff that exists between the representation of the information content and the quality that we are going to deliver. We may be ultimately designing a very efficient scheme in terms of bandwidth utilization but that may not be appealing perceptually so then it is of no use. So we have to always make a good tradeoff that without severe perceptual degradation how efficiently you can encode. That is the challenge; that is the challenge with which all these techniques are developed.

Now, I was telling you about the image coding aspects that whereas in the case image coding we have the spatial variations, variations with respect to the x and the y. in the case of video there is also a temporal variation. So there is spatial as well as temporal variation and in both these dimensions the spatial dimensions as well as in temporal dimension we have lot of redundancy. Spatial we can definitely understand that given one frame the pixel intensities are not going to change drastically between the neighboring pixels. Unless it is at some age of some object boundary between some object and background then there some drastic change in the intensity is possible. But as long as you are moving within smooth areas of the image some uniformly ((neat ....35:24)) object that is not going to change the intensity drastically from one neighbor to the other. And same is in the case with temporal aspects also that if we capture a frame now and if we are maintaining 30 frames per second as the frame refreshment rate or the rate at which the frames arrive. So the very next frame we are representing like this then the very next frame we just keep it here (Refer Slide Time: 36:00) so this is the stack of frame so this is our time dimension whereas here the variation is with respect to x and y so we have the space variation within one frame, time variation across the frame. So if we call this frame as the nth frame then we will be calling this as the next frame or n plus 1th frame, this as n plus 2th frame, this as n plus 3th frame and so on.

(Refer Slide Time: 36:26)

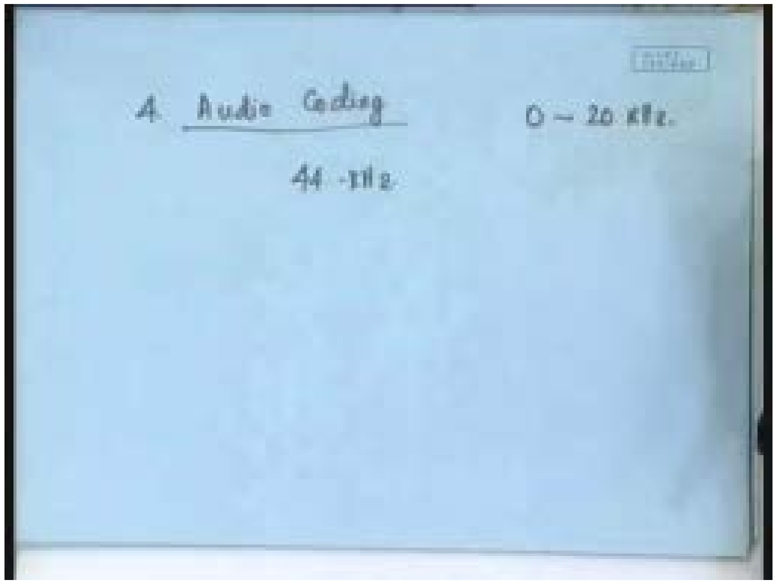


Now within successive frames we have at 30 frames per second we have a gap of only one thirtieth of a second. So in one thirtieth of a second or rather to say 30 milliseconds how much change can be there? After all you are taking the pictures mostly of real world objects. Real world objects are not going to move that drastically. Sometimes they may but most of the times you can say that in one thirtieth of a second there is not a significant movement, there is marginal movement. So at least we can use some kind of a differential technique over time which means to say that even if we apply any transformation techniques to consider the variation in space very efficiently we can use a differential technique to consider the variation against time which results in some kind of a hybrid encoding scheme that is to say having a differential pulse code modulation temporarily and having transformation techniques like the DCT or the DWT to encode the error that results out of that prediction.

So what you can do is that from the nth frame you can predict the n plus 1th frame, take the difference between load predicted and the original n plus 1th frame, take the error signal of that; error of the differential signal of that and then that differential signal will be having some variation with respect to space which you can exploit very well. The correlation that exists along the spatial dimensions can be exploited using the techniques like the DCT, wavelengths and all others. So we will be learning about the video coding techniques also in details and subsequent to that in this course we will also come to the question of how to efficiently encode the audio.

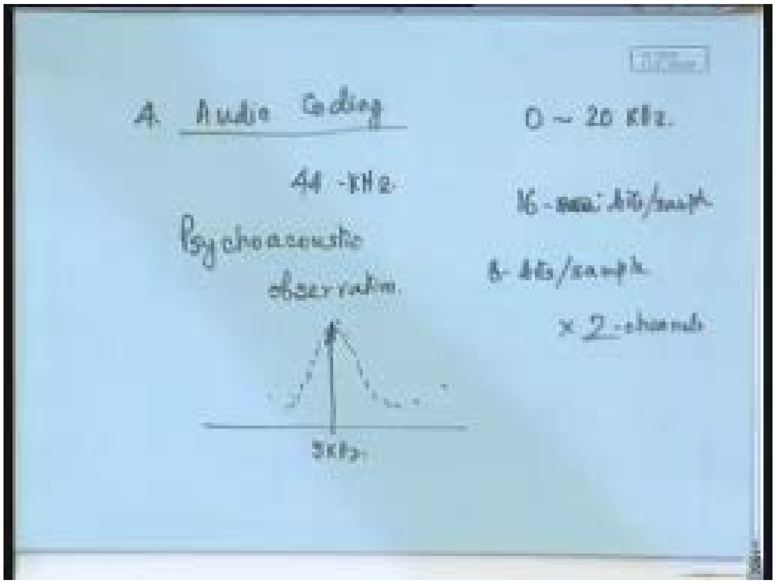
Now, again we are making a distinction between speech and audio. Speech is something that is a simple voice quality signal which is restricted to 3.4 kHz. Now if you get a channel of bandwidth which can just accommodate the speech samples obviously the same bandwidth will not be able to cater for having any high quality audio say music say some high fragility musical instrument is going on and then we will not be able to use the channel; I mean the channel which can be used for speech will not be good for music. So here for audio we have to make use of the entire audio frequency spectrum which we consider as 0 to 20 kHz people say 20 hertz to 30 kHz. So your signal may be all the way up to 20 kHz so in order to sample this signal again you have to fulfill the Nyquist rate so you have to sample it at a rate which is higher than 20 kHz and usual practice is to use 44 kHz as the sampling rate.

(Refer Slide Time: 40:22)



Thus, with 44 kHz of sampling rate you have to sample the audio signal. Audio signal is like speech it is a time varying signal only but the rate at which it is sampled is 44 instead of 8 kHz which is for speech. So definitely the audio information content is going to be much higher and again for audio we require a stereo quality in many of the cases like for efficient recording of audio or for efficient playback; in order to ensure that it is more appealing to our ears we go in for stereo channels which means to say that we have 16 samples or 16 bits we take per sample where we consider 8 bits per sample for each of the channels. So it is 8 bits per sample for each channel into two channels. So obviously it leads to 16 bits per sample and 16 bits per sample with 44 kHz is definitely a bandwidth which is considerably high and we need to adopt good efficient techniques for the audio coding also and in the case of audio coding actually some characteristics of audio are very efficient to utilize. So instead of going in for the standard waveform based coding techniques one considers some psychoacoustic behavior.

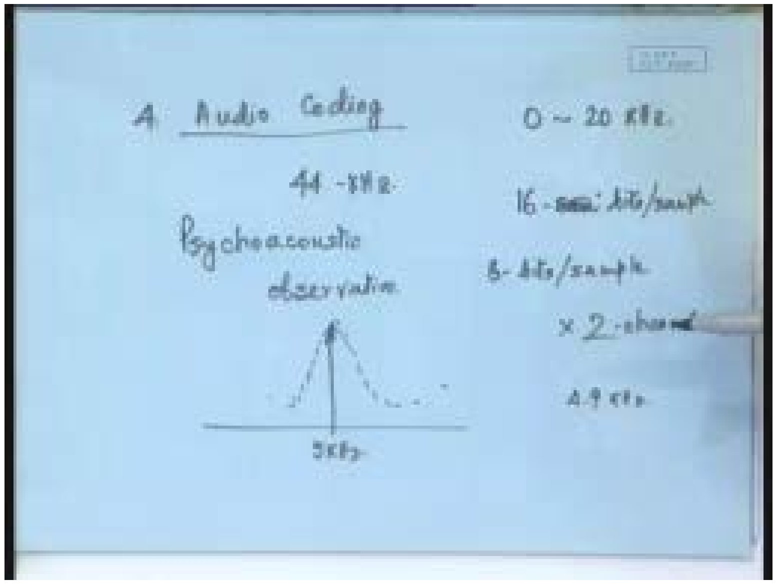
(Refer Slide Time: 43:09)



The psychoacoustic behavior has been very widely studied and based on the psychoacoustic observations it has been found that if we have the presence of a very strong signal at some frequency; let us say that at 5 kHz we have the presence of a very strong signal. In that case it is seen that this very strong signal tends to mask off the other audio signals which are there in the vicinity of this frequency. So there is some kind of a masking characteristic that beyond this 5 kHz or close to this 5 kHz there is some kind of a masking effect which gradually tappers off.

So if we have another audio component at 1 kHz that may not be masked off. Or if we have something at 10 kHz that is also far off from this 5 kHz component. But if we have something which is 4.9 kHz; definitely the 4.9 kHz is very close to the 5 kHz strong audio signal which is present and this 4.9 may be masked off or 5.1 may be masked off. So what we do is that we make a complete tonal analysis and then find out that what is the amount of what you say..... what is the masking effect at a particular frequency; when there are strong tonal components which are present what is the overall masking effect so that we allocate the bits considering those masking effects. This also we will be studying and there definitely we have to consider the psychoacoustic phenomenon into consideration which has been actually incorporated in the audio coding standards.

(Refer Slide Time: 44:04)



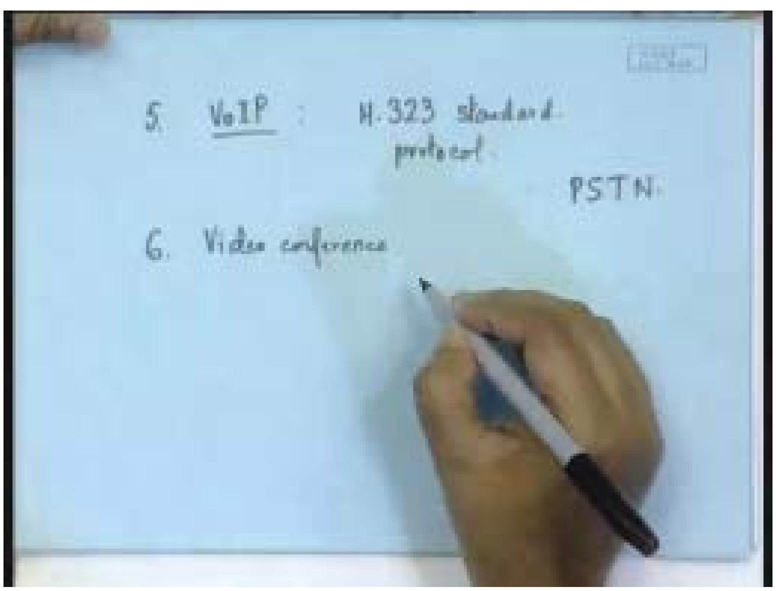
And after going through the audio coding which is the fourth topic fourth major module in our course content the next topic that we will be considering is the actual communication of the signal. What we mean here is that how the speech and the speech/audio/video these can be exchanged how this information can be exchanged from one place to the other. There we considered the networking aspect that if we have to send the information from one place to the other or if we have a kind of a conferencing like one terminal say multiple terminals have entered into a kind of an audio conferencing or in a video conferencing there the exchange of information has to continuously take place so then what is the requirement of such networks; what are the protocols that come in.

Therefore, our fifth topic in this course will be the voice over IP voice over internet protocol and in that we will be stressing on the networking and especially there we have to study the signaling requirements which are given by the H dot 323 standard. So we will be covering the voice over IP with respect to the standard protocols H dot 323. H dot 323 basically gives you the protocols for voice over IP.

Now what has happened is that slowly the networking has encroached; I mean already the networking has encroached our day to day live to such an extent that we will be having a mixture of the traditional communication devices and the network communication devices; like voice over IP we use for the telephoning purpose where the voice communication will go through the internet backbone rather than going in through the normal telephone lines and it will be completely digitally processed; from the source it is digitally processed and then it is sent over the packets and the packets are delivered to the destination whereas in the case whereas people also still use the traditional telephones for voice communications. So you have a combination of the voice over IP telephones and also the Public Switched Telephone Network or PSTN terminals are also available. So there needs to be some kind of an interface between PSTN and VoIP so that there is no barrier as such.

A VoIP user should be able talk to a PSTN user very easily just like the way a PSTN user also can initiate a call and talk to a VoIP user in a very easy and in a transparent manner no matter wherever the users are; whether they are in the VoIP domain or PSTN domain a very easy access should be there. So this aspect we will be talking of when we talk of the voice over IP and then we will come to the last topic in our subject which is the video conference.

(Refer Slide Time: 48:29)



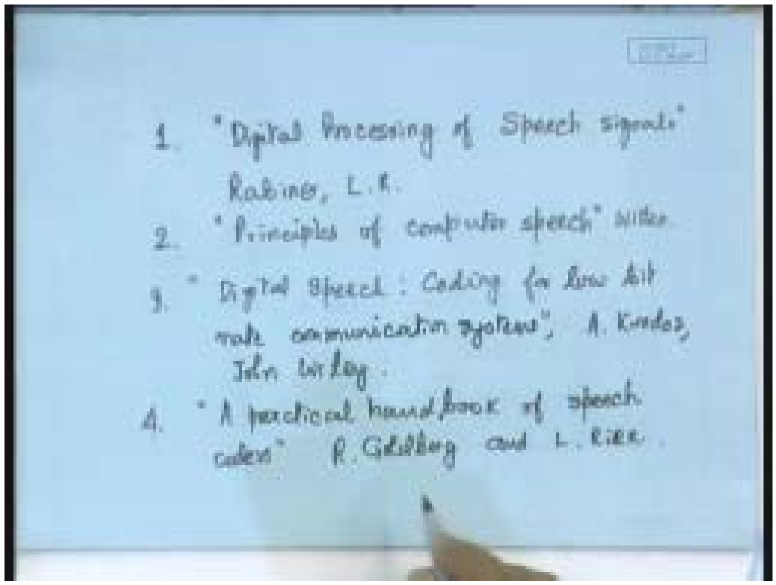
So voice over IP is for the voice transmission mostly. But of course the VoIP techniques can be used for the video signaling as well. but voice But video conferencing, actually conference techniques also is going through a change in era. When video conferencing first came in then it was made possible over dedicated communication channels; it was possible to achieve video conferencing through the ISDN the Integrated Services Digital Network and slowly now the situation is changing with the easy availability of internet backbone.

Now the video conferencing is also becoming popular over the IP based network. So there are some efficient protocols to support that which is the Session Initiation Protocol or what is called as the SIP protocol which also we will be studying under the video conferencing. So we will study the ISDN based video conferencing also; we will talk about the IP based video conferencing especially the aspects of SIP protocol and also in our last lecture possibly we will try to cover about the wireless video conferencing also which will be the future in the coming days. So this is the broad outline of our course that we will be having in a nutshell; we will be having speech coding techniques, we will be having we will be studying about the image coding, we will be studying about the video coding, voice over IP and no audio coding in between voice over IP, video conferencing. So that easily takes 40 lectures for us in which we will be able to give you some comprehensive coverage of the technology that is in use for the digital voice and picture communication.

So we will actually be starting with the speech coding from the next class onwards. So may be that today I can give you some idea about the reference materials that would help you to some extent. So for speech there are some many good books which are available. So I can recommend some of these books for you so that your understanding can be better if you listen to these lectures along with reading some reference books.

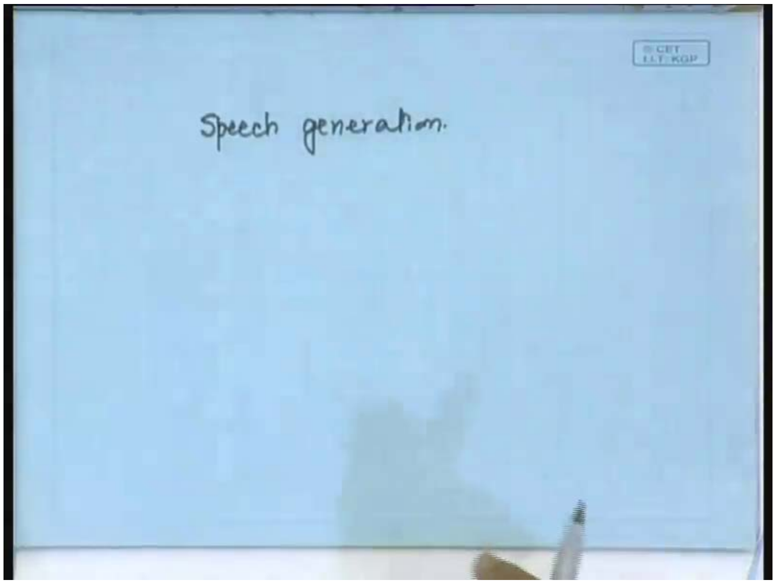
So, one very popular book is Digital Processing of Speech Signals by Rabiner. So this is the title and this is by Rabiner L. R. And another book you can refer to is Principles of Computer Speech, this is by Witten. Then you can also refer to a more recent book; it is Digital Speech: Coding for low bit rate communication systems; this is by Kondoz. This is actually a John Wiley publication book and another interesting book is a Practical Handbook of Speech Coders by R. Goldberg and L. Rick.

(Refer Slide Time: 53:46)



So we will begin with the topic from the next class. So first I will present the speech processing or rather to say the speech generation models.

(Refer Slide Time: 00:54:10 min)



So for this we need to study some physiological aspects of speech generation process because only if we understand the speech generation process properly then we will be able to efficiently represent the parameters which are associated with speech. So we will start with the speech generation and then after the speech generation and finding out the speech generation models for that we will go further into the speech coding aspects, thank you.