Wireless Communications
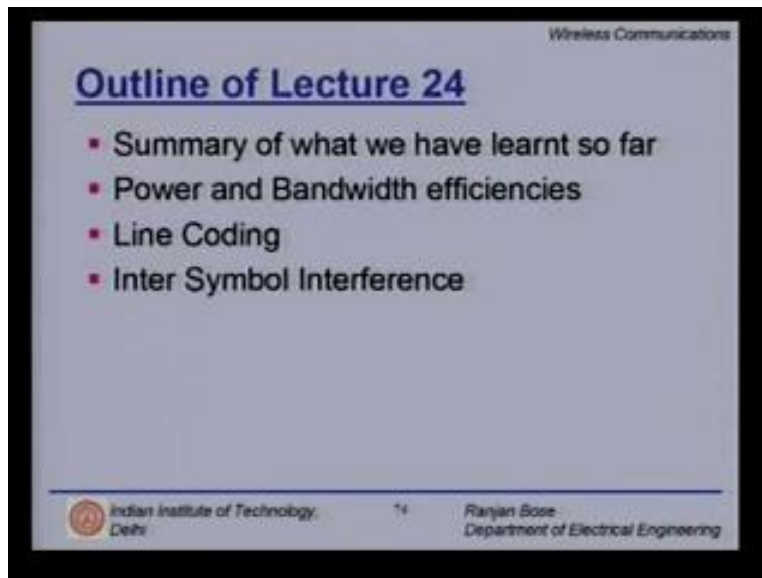**Dr. Ranjan Bose**
Department of Electrical Engineering
Indian Institute of Technology, Delhi
Lecture No. # 24
Modulation Techniques for Mobile Communications (Continued)

Welcome to the next lecture on modulation techniques for mobile communications, a brief outline for today's talk is as follows.
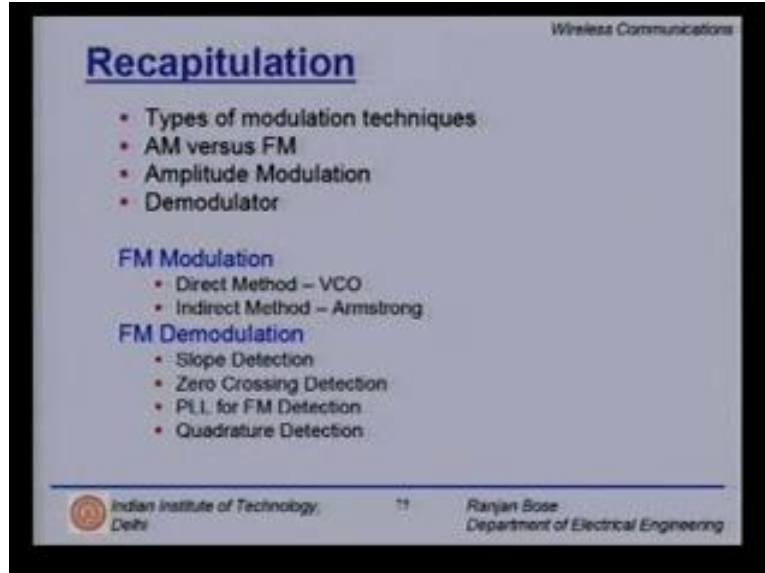
(Refer Slide Time: 00:01:28 min)

We will first summarize what we have learnt in the past couple of lectures followed by the power and bandwidth efficiencies. Then we will talk about line coding then we'll get into inter symbol interference and finally a peek into digital modulation techniques. That will be our topic for a couple of next lectures. So first let us summarize what we learnt last time.

In the previous classes we had a peek into the types of modulation techniques specifically for wireless communications. We looked at amplitude modulation versus frequency modulation specifically we dealt with AM how to generate AM and how to demodulate AM. In the last class we talked about FM or frequency modulation we looked at the direct method for generating FM as well as the indirect method.

(Refer Slide Time: 00:02:00 min)
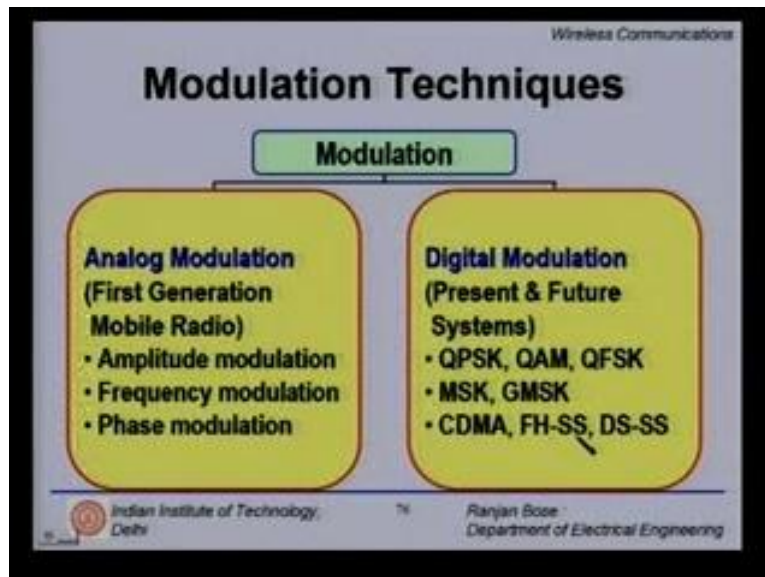
Then we considered the 4 popular demodulation techniques for FM which are the slope detection method, the zero crossing detection technique, the PLL or phase locked loop for FM detection and finally the most popular quadrature detection technique. FM we learnt was used in the first generation mobile communication standards but no longer employed for the second or beyond 2G or 2.5G generations.

(Refer Slide Time: 00:03:10 min)

We also learnt that modulation could be broadly classified into analog and digital. We have already had a revisit to the various popular analog modulation techniques, the amplitude modulation, the frequency modulation and the phase modulation.

The last two the frequency modulation and the phase modulation together can be grouped as angle modulation as we saw last time. The other technique is digital modulation which is used in the present mobile communication systems as well as in the future mobile communication systems. We will also consider the advantages of digital modulation techniques versus the analog modulation techniques. Some of the examples are QPSK, QAM, QFSK, MSK, GMSK used in GSM standards, CDMA and then some spread spectrum techniques used in WCDMA or CDMA 2000 standards. Here another emerging techniques for wireless communication for modulation is the pulse position modulation or a combination of PPM and PAM for ultra wideband communications also.

(Refer Slide Time: 00:04:47 min)

We had also compared the advantages and disadvantages of FM versus AM realized that FM signals are less susceptible to atmospheric noise because the information is stored as frequency variations rather than amplitude variations. We learnt that the atmospheric noise, effects the amplitude. AM signals on the other hand code the information in terms of the amplitude and hence are more susceptible to the variations in the atmospheric noise. AM signals occupy less bandwidth as opposed to FM but FM you can tradeoff bandwidth for the output SNR. The modulation index can be varied to obtain greater SNR about 6dB for each doubling in bandwidth. So you can give and take for FM there is a tradeoff whereas in AM we have no such flexibility. We also learnt that efficient class C amplifiers are good enough for FM but for AM we must use a class A or class AB amplifier.

3

(Refer Slide Time: 00:06:01 min)

Now let us start with digital modulations; in modern mobile communication digital modulation is used. Advancements in VLSI and DSP have made digital modulation techniques more cost efficient and more useful than analog transmission systems. They had been invented long time back but the advances in VLSI technology, the high speed chips, low power designs together have helped realize the M-ary modulation techniques. What are the specific advantages of digital modulation? Greater noise immunity as compared to analog, if you go into the digital domain you have much more flexibility in designing the noise margin. Easier multiplexing of information, voice, data and video; today all of this is coming as digital data beat voice, video, fax or any other multimedia application, everything is digital. So it is but natural to go in for digital modulation also.

It can accommodate digital transmission errors, source coding, encryption and equalization. In fact you can also combine modulation and coding together in a different part of digital communication called coded modulation techniques. Trellis coded modulation forms a part of that technique. Digital signal processors can implement digital modulators, demodulators completely in software. So these are some of the advantages why digital modulation techniques have become popular and are here to stay.

What are the factors that influence the choice of digital modulation? A desirable modulation scheme provides low bit error at low signal to noise ratio. Please remember modulation is not just a means to send your information over wireless channels. It is also helpful in overcoming the effects of noise. Efficient design of modulation techniques guarantee you immunity or at least better performance in terms of low SNRs.

4

(Refer Slide Time: 00:08:05 min)

Then digital modulation techniques perform well in multipath fading environments, even though digital modulation techniques are used popularly in other non-wireless applications. We are here concentrating on wireless channels. It has been found the digital modulation techniques perform well in fading environments, occupies a minimum bandwidth we will talk about the spectral efficiency today. Easy and cost effective implementation specially with the advent of DSPs but all of these effects cannot be achieved simultaneously so it allows for a tradeoff. So according to our requirement tradeoffs are made for selecting a digital modulation technique so you can use more bandwidth or less bandwidth tradeoff for bit error rates which are higher or lower. Let us now talk about the power and bandwidth efficiencies of modulations techniques.

(Refer Slide Time: 00:09:57 min)

Power efficiency is defined as the ability to preserve the fidelity of digital message over low level SNRs and is defined as eta subscript p, efficiency power as signal energy per bit divided by noise power spectral density. This is defined for a certain probability of error. So if you change the probability of error say from 10 raise to power minus 5 to the 10 raise to power of minus 7, the power efficiency will change because your numerator will change. In text books sometimes this is also termed as energy efficiency but we will maintain it as power efficiency.

The other part is bandwidth efficiency, it defines how efficiently the allocated bandwidth is utilized. So the whole game is to use the bandwidth efficiently so as to have the maximum possible immunity against noise. How do you measure that? In terms of bit error rate so given a certain amount of bandwidth, what is the minimum bit error rate you can guarantee and at the same time, your data rate should not go down too much. These are the 3 constraints that we have the data rate, the power requirement and the bandwidth. We want minimum bandwidth, maximum data rate and minimum power consumption for a modulation technique to be chosen.

Now the bandwidth efficiency eta subscript B is defined as R over B, R is the data rate B is the bandwidth and its units are bits per second per hertz. There is an upper limit an upper bound on this bandwidth efficiency that comes from the Shannon's theorem. So eta$_{B\ max}$ is C divided by B, C is the capacity of the channel, B is the bandwidth log to the base 2(1 plus signal to noise ratio). Please remember that the efficiency is proportional to the logarithm of SNR and inversely proportional to B but capacity itself will have a notion of bandwidth also. If you have to choose between signal to noise ratio and bandwidth, it is better to put in your resources for bandwidth than to increasing signal to noise ratio.

In fact this is the reason why CDMA systems are inherently more efficient, CDMA systems work at low SNR's, the spreading code spreads the initial signal thereby increasing the bandwidth and the signal to noise ratio is fairly low. So CDMA systems are by definition more efficient. We will talk about the spread spectrum techniques later in the lectures.

(Refer Slide Time: 00:13:33 min)

The definitions of bandwidth is important. How do you define bandwidth, because in the previous equation if seen that bandwidth prominently figures into the spectral efficiency equation. So it is important for us to understand how to define bandwidth. Why is the definition ambiguous? It is simple, whenever you send a signal, it is time limited because for every symbol duration you have to send one symbol in time but when you have a time limited signal, it cannot be frequency limited; bounded in time means unbounded in frequency. Clearly we cannot work with infinite bandwidth plugged into here. We have to have a useful practical definition of bandwidth so that we can carry on our calculations regarding bandwidth efficiency.

Clearly the definition of bandwidth can vary from application to application, from people to people. Let us consider the popular definitions of bandwidth which can be used. The definition of bandwidth are based on some measure on the power spectral density or PST of the signal. That's our starting point; for any random signal w(t), the power spectral density $P_w(f)$ is defined as absolute value $W_T(f)$ whole squared ensemble average divided by T where T tends to infinity, the bar denotes the ensemble average, small $w_T(t)$ is Fourier related with capital $W_T(f)$. Now the capital $W_T(f)$ has to be truncated at some point because clearly it is going to extend from minus infinity to plus infinity in frequency because my $w_T(t)$ is time limited. So what we do is $w_T(t)$ is defined as w(t) from minus T by 2 to T by 2 and it is zero elsewhere. So we are dealing with a truncated version of w(t).

(Refer Slide Time: 00:16:12 min)

Now the power spectral density or PSD of a bandpass signal is a PSD of its baseband complex envelope. If a bandpass signal s of t is given by real part of g (t) e raise to power j 2 pi f_c t where g (t) is the complex baseband envelope. Then the power spectral density of the bandpass signal is given by this equation basic analysis. Here P_g on the right hand side is the power spectral density of g (t). The absolute bandwidth of a signal is equal to the range of frequencies over which the signal has a non-zero power spectral density.

So this can be one of your definitions, the absolute bandwidth of a signal. Simply defined as range of frequencies over which the signal has a non-zero power spectral density, so of simple definition but not always a useful definition.

(Refer Slide Time: 00:17:26 min)
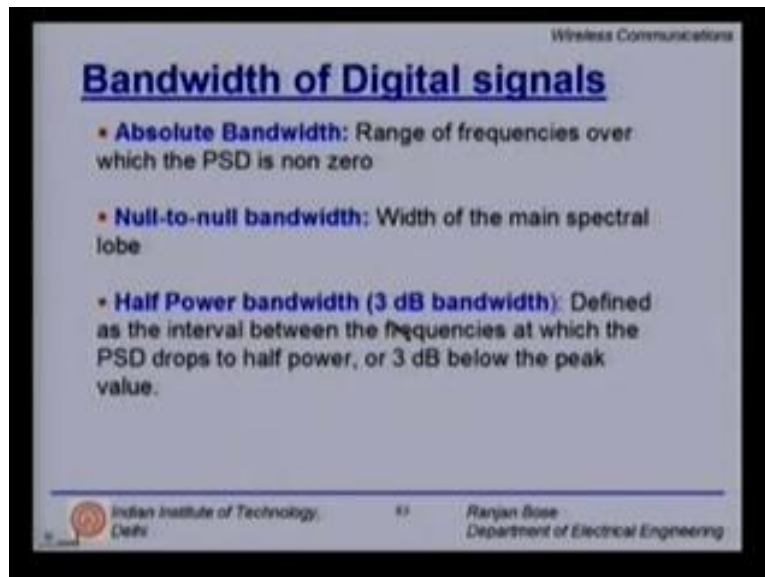
Let us continue our discussion on the bandwidth of digital signals. The absolute bandwidth is a range of frequencies over which the PSD is non-zero we have just seen, the other possibility is null to null bandwidth. It is the width of the main spectral lobe. So for example if you are dealing with a rectangular pulse in the time domain, in the frequency domain it will be sin function. Clearly the absolute bandwidth will be infinity but null to null bandwidth is a width of the main lobe. So if you make your pulse narrower your main lobe width will increase which is intuitive, if you shrink in the time domain you have to expand in the frequency domain.

Another popular definition is the half power bandwidth also known as a 3 dB bandwidth. It is defined as an interval between the frequencies at which the power spectral density drops to half power or minus 3 dB below the peak value. In electrical engineering the half power beam width is usually used, half power bandwidth is one of the most commonly used definitions.

FCC also has a definition of bandwidth. Occupied bandwidth is the band which leaves 0.5% of the signal above the upper band limit and 0.5% below the lower band limit. In other words 99% of the signal power is contained within the occupied bandwidth. You have left off 0.5 on the left side, 0.5 on the extreme right side whatever you have left with this 99%.

8

(Refer Slide Time: 00:18:57 min)

Another definition could be, everywhere outside the specified band the PSD is below a certain stated level; not necessarily 3 dB. So you can have different definitions of bandwidth. [Conversation between professor and the student (Refer Slide Time: 19:57)] This 0.5% is fixed or it's varying. This 0.5% is fixed by FCC definitions but for example if you talk about the another definition, if you look at the popular definition of UWB or ultra-white band definition they don't adhere to the 3 dB definition that adhere to something as 10 dB of the peak limit. So you can define everywhere outside the specified band where the power spectral density is below 10 dB of the peak value. It depends how strict you want to make your definition, the stricter you make the more frequency band you encompass. Typically if you are going by this other definition, you really want to be strict 45 to 60 dB of attenuation could be specified.

(Refer Slide Time: 00:20:46 min)

9

Now let's change gears and move into a slightly different but related domain called line coding. We are making case for introducing digital modulation techniques. Line coding is not digital modulation but it is important to have carrier at the receiver. We should not loose synchronization between the transmitter and receiver. All the good digital modulation techniques would first determine the board rate, the carrier frequency if not known and then do a demodulation. Line coding will ensure that you have enough flips between zero's and one's or enough changes so that the synchronization can be maintained. Line coding is also important for wireless communication. Line codes are used to provide particular spectral characteristic of a pulse train. We will understand this definition of line codes by some examples which you will take just now. So you can understand line coding by simply defining it in terms of certain properties in the frequency domain. For example I do not want any power present in the DC value.

Some of the commonly used line codes are as follows; one is the return to zero or RZ. RZ implies that the pulse returns to zero within every bit period, non-return to zero or NRZ is more popular than the RZ, do not return to zero in a bit duration. Then a more popular code, the Manchester code which is the special type of NRZ, it uses two pulses to represent each binary symbol. So no matter what happens whether you get a string of one's or a string of zero's, you will have transitions. Remember if you had a simple positive pulse to represent a one and a negative pulse to represent a zero and you did have a long chain of one's, there will be no transition. RZ and NRZ do take care of this problem to some extent, Manchester code will ensure no matter how long is your chain of one's or zero's, you will always have transition. Hence you can recover back the carrier or you can recover back the synchronization at the receiver. You can have either unipolar line coding that is you work between zero and V volts or bipolar minus V to plus V volts that's your choice.

(Refer Slide Time: 00:23:54 min)

10
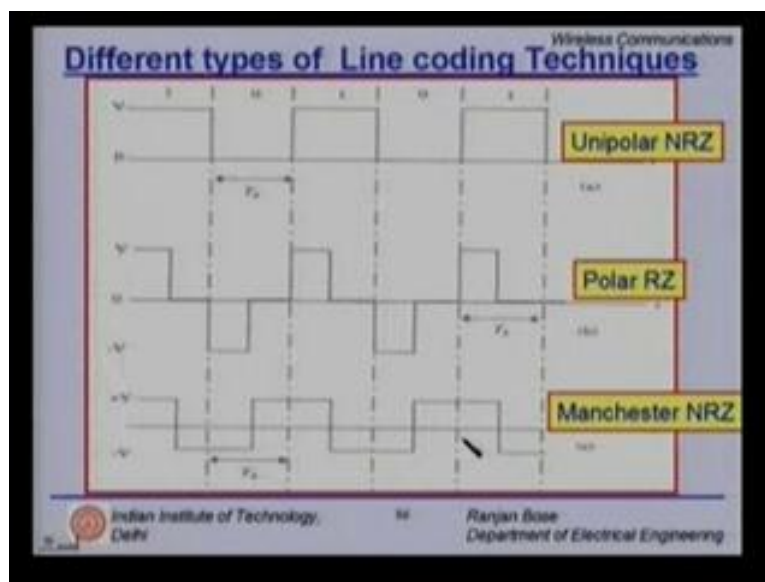
So here in the time domain is a representation of the three popular line coding techniques. Here the top is a unipolar NRZ, non return to zero, here is a polar. So if we flip between V and minus V, return to zero. So within every bit interval I do return to zero thus the whole idea and here Manchester NRZ is we ensure that regardless of what input strings comes in I will always have transitions.

(Refer Slide Time: 00:24:41 min)
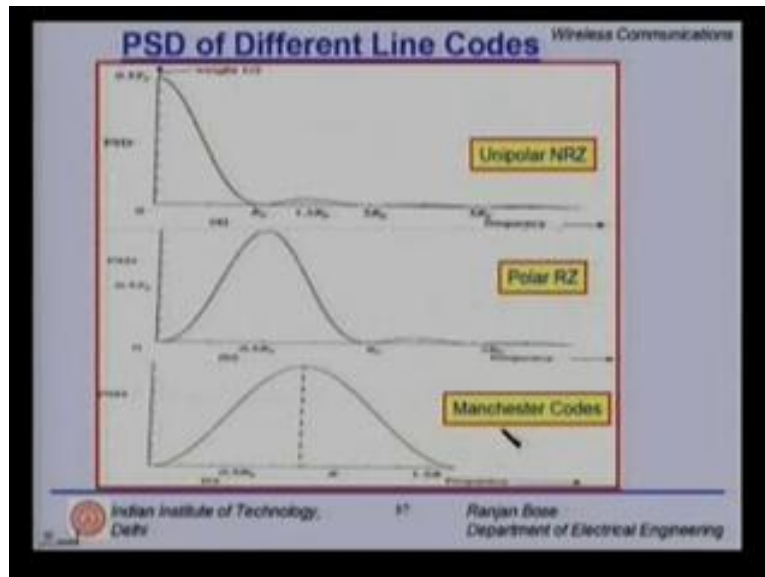
Now let's look at from the power spectral density perspective for different line codes. Here you have the unipolar NRZ, we have a high energy at the DC a close to the DC level this is not desirable. Here is the polar RZ, we shift the peak away from zero and then they are the Manchester codes which look like this in the frequency domain and there are many other techniques for line codes.

(Refer Slide Time: 00:25:19 min)

11

What are some of the important properties that we are looking at? How do you find out whether a line code is good? First of all the transmission bandwidth should be as small as possible. So here if we go back, you have to see which one has a larger transmission bandwidth that's more expensive. Power efficiency for a given bandwidth and a specified detection error probability should be as small as possible. Error detection and error correction capability, it should be possible to detect and preferably correct detection errors. Favorable power spectral density desirable to have a zero power spectral density at zero frequency. Adequate timing content that's what we are trying to recover, it should be possible to extract exact timing or clock information. So it's important for the clock recovery to maintain synchronization between the transmitter and receiver. Transparency should be possible to transmit a digital signal correctly regardless of the patterns of one's and zero's. This is last but not the least.

(Refer Slide Time: 00:26:44 min)

Let us also look at slightly different but again a related issue of pulse shaping and the related problem that we have of intersymbol interference. We are still in the process of building grounds for introducing digital modulation techniques. Now first let us understand the problem of intersymbol interference. First let us take a bandlimited channels, almost all channels are bandlimited. Some are more bandlimited than the others, fiber optic channels are less band limited they have a lot of bandwidth, twisted copper or telephone lines have much smaller bandwidth but no matter what you have in general you are dealing with a bandlimited channel unless it is an ideal channel.

Now what happens is whenever there is a signal input to a channel, it is convolved with the impulse response of the channel. Convolution by definition is a spreading operation, unless you are convolving with a delta function which corresponds to channel of infinite bandwidth. So any channel which does not have infinite bandwidth will expand the output. The input will get expanded at the output, talking in a very basic term.

Consequently two symbols which have been sent one after the other might interfere in time at the receiver. Let us look at an example. Suppose I have an ideal pulse being transmitted through this band limited channel. So what we see is a distorted spread out version of the same pulse. It also attenuated to some extent. Now if you are interested in sending another pulse, what we will get out from the band limited channel is another distorted spread out version of the pulse. Here clearly I had taken enough precaution not to overlap the pulses so that I can distinguish my one's and zero's but here they have overlapped, thanks to the band limited channel and this is called intersymbol interference, this is a problem.

At the receiver we will receive an addition of these two and here it will be much more distortion than what is seen here and this is being shown only for two pulses, it could be for several pulses. We have to overcome this effect, it puts serious limitations on a data rate. So there is thankfully a technique proposed by Nyquist that overcomes this effect. When rectangular pulses or in general pulses of any shape passes through a bandlimited filters it spreads. Why? Convolution is a spreading process I can do nothing about it. This inter symbol interference or ISI leads to errors in the transmission of symbols but there are pulse shaping techniques which can overcome or reduce the effect of intersymbol interference. So let us quickly revisit some of the good pulse shaping techniques and before that what is of the general philosophy behind pulse shaping techniques. Why should it work, how can we still recover what was sent even when we have a bad case like this.

(Refer Slide Time: 00:31:03 min)

So the Nyquist criteria for ISI cancellation. ISI can be nullified if overall communication system is defined so that at every sampling instant, the receiver the response due to all of the symbols except the current symbol is equal to zero. What does it mean? That let the symbols interfere I do not care but at the time when I am going to sample except my desired symbol, everybody else should be going through a zero point. Let there be a finite value after that or before that but at that instant of sampling I should have zero contributions from other symbols except the one that I want to sample.

13

Nyquist derived the condition for impulse response for an overall communication system so that there is zero ISI. So the overall system has to be optimized which includes the transmitter the channel and the receiver, all taken together and this is the h effective. This impulse response must have some properties and what is the property? That h effective nT is equal to constant for n is equal to zero and is equal to zero for n not equal to zero. So I am sampling at these intervals and it should be zero for all other sampling instances other than its own sampling instance, T_s is the symbol period, n is an integer, k is a non-zero constant as simple as this.

(Refer Slide Time: 00:33:03 min)
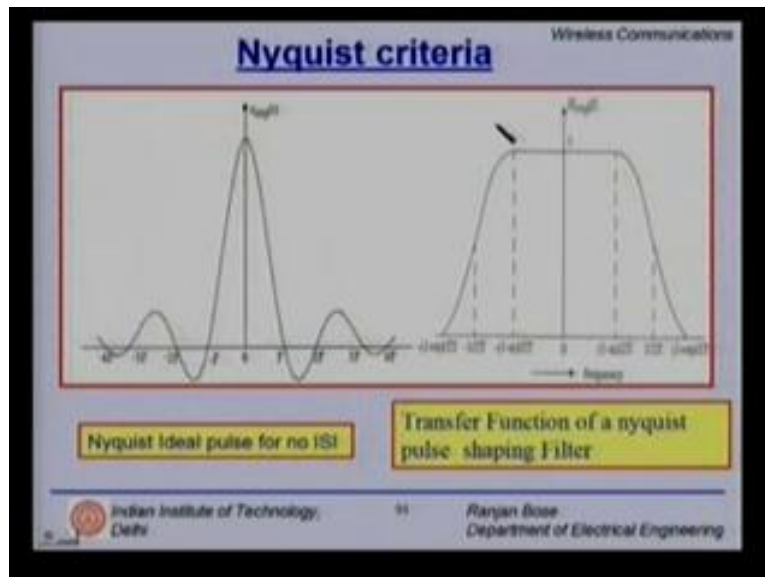
So let us look at a simple example consider the sync pulse in the time domain so the x axis is time and here this pulse satisfies the requirement, at time T is equal to zero it has a finite value but at T_s, $2T_s$, $3T_s$ and so and so forth it goes through a zero and similarly before that. So only at its own sampling interval is it non-zero; rest all its zero. Hence at those instances it will not contribute to intersymbol interference. So this is an ideal Nyquist pulse for no intersymbol interference. How does it look in the frequency domain? You can have the transfer function of a Nyquist pulse shaping filter as follows. So Nyquist criteria can also be understood in the frequency domain.

Let's talk about the mathematical analysis of Nyquist criteria. So what is your h effective? It is delta t convolved with p (t) your pulse, the impulse response of the channel and the impulse response of the receiver. This should have the property that it is non-zero only at its own sampling duration interval and it is zero for all other sampling instances, p (t) is the pulse shape of the symbol, h_t as I said is the channel impulse response, h_r is the receiver impulse response and the star denotes the convolution operation.

14

(Refer Slide Time: 00:34:12 min)

The desired transfer function H effective can be obtained as H effective is equal to 1 over f_s the rectangular function rect f over f_s and in general Nyquist showed that the transfer function of a filter for zero ISI is given by H effective is equal to 1 over f_s rect f over f_s convolved with z of f. What is z of f? Is any arbitrary even function which is zero outside the passband of the rect filter, this also satisfies zero ISI. So you don't have to worry about only using the sync function which is much more difficult to generate and realize in real life. You can have a host of other pulses which will help you realizes zero ISI condition, h effective should have some other properties; it should have a fast decay with a small magnitude near the sample values because they could be jitters in the sampling interval.

You may sample not exactly at the sampling instance but a little before after that. So I should not take in a little bit too much of the interference in terms of the inter symbol interference. It should have fast decay and very small values near the zero crossings. For an ideal channel $h_c t$ is delta and should be possible to realize a closely approximate shaping filters at both receivers and transmitters to produce the desired H effect. Even if it is not, you can come up with good pulse shapes which will overcome or minimize the effect of intersymbol interference.

Let us look at an example, one of the good examples is the raised cosine roll off filter. It is the most popular pulse shaping filter used in mobile communications. It is a Nyquist filter, it satisfies the Nyquist condition. So the raised cosine roll off filter is given by the following equation, the impulse response $h_{RC}(t)$ is equal to, this is a sin function multiplied by cos pi alpha t over T_s divided by 1 minus 4 alpha t over 2 T_s whole squared. Here T_s is the sampling interval, alpha is a roll off factor; as the value of alpha or the roll off factor increases the bandwidth of the filter also increases.

15

(Refer Slide Time: 00:36:47 min)

So this is a neat way, choosing a values of alpha will shrink or expand your filter in the frequency domain that will have repercussions on the bandwidth requirement of your applications but for all values of alpha, it keeps on meeting the Nyquist criteria. As the value of roll off factor increases, the time side lobe levels decrease. The symbol rate $R_s$ that can be passed through a bandpass filter is given by 1 over T_s approximately is equal to 2B over 1 plus alpha. So if you have to have a larger value of R of s, you must decrease alpha for the same bandwidth, so it's a design parameter. You can choose values of alpha and achieve a certain symbol rate and hence a certain data rate without violating the Nyquist criteria, beta B here is the absolute bandwidth.

(Refer Slide Time: 00:39:10 min)
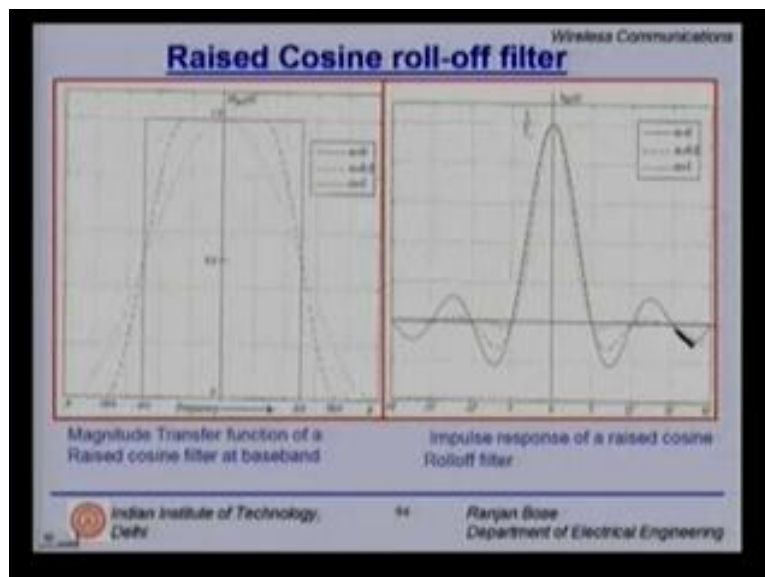
How does the raised cosine filter look in the frequency domain? So here the x axis is the frequency and for alpha is equal to zero, it is the rect function and as you increase the alpha take it to 0.5 it rolls off and then as you increase the roll off factor as a name suggest, the roll off is gradual. Sharp for alpha is equal to zero, gradually increasing for alpha is equal to 0.5; roll off factor for alpha is equal to 1 is much more gradual, I am increasing my bandwidth. How does it look in the time domain? The rect clearly corresponds to the sin function with roll off factor 0.5 it deviates from the sin function here and what you must see is that these side lobes have decreased in value and as you increase it further, the side loops are further reduced. What does it mean, I am concentrating more and more energy in a lower main lobe. My pulse is actually becoming narrower, my frequency is increasing. All of these pulses meet the Nyquist criteria, note at zero T_s, $2T_s$, $3T_s$ they all pass through the zero.

(Refer Slide Time: 00:41:01 min)

Now let us look at another popular pulse shaping filter which does not exactly satisfy the Nyquist criteria but comes fairly close. So this is a non Nyquist filter, it is used in conjunction with minimum phase shift keying MSK modulation. You can see here why pulse shapes and modulation techniques go hand in hand. The Gaussian low pass filter has a transfer function given by H_Gf equal to exponential minus alpha squared f squared. Alpha is defined as under root ln 2 over root 2 B where B is a 3 dB bandwidth of the baseband Gaussian shaping filter. The impulse response of this Gaussian filter is given by this following equation. So Gaussian filter followed by minimum shift keying MSK together is known as GMSK and is popularly used in GSM forms. Gaussian filters are used when cost and power efficiency are major factors and again low bit error rate are required.

17

(Refer Slide Time: 00:42:36 min)
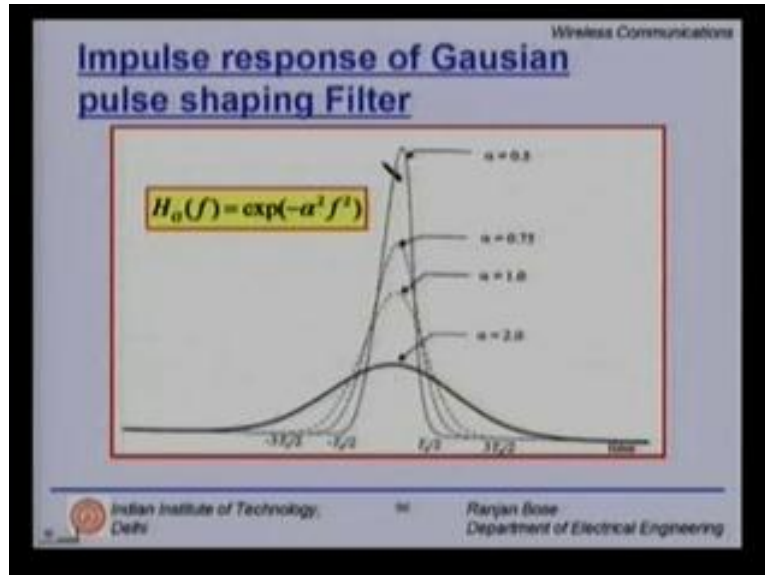
How does the impulse response of a Gaussian pulse shaping filter look like? Here we have plotted for different values of alpha, x axis represents time; here for alpha is equal to 0.5 it is a sharpest peak it's a narrow pulse and then as you increase your values of alpha, the pulse spreads. In the frequency domain I'll have a narrower and narrower filter. So again alpha is a design parameter depending upon, how strict you have to be with your filter design, how expensive is a bandwidth you would like to use different values of alpha. These are the pulses that you will actually transmit. Here $H_G f$ is exponential minus alpha squared f squared corresponding to the different values of alpha.

(Refer Slide Time: 00:43:36 min)

Now let us take a brief look at the geometric representation of modulation signals because the whole effort so far has been to build a background. If the r total M possible signals, the modulating signal set S can be represented as S is equal to S_1 (t), S_2 (t) up to S_m (t). So what we have trying to do is we have n possible signals to be sent and for binary case M is just 2. So you can send 1 or 0 by representing it using S_1 or S_2, for higher level M-ary modulation scheme M greater than or equal to 2 the signal space may be represented as a vector space. Any point in the vector space can be represented as follows, S_i (t) is summation, S_{ij} phi$_i$ (t), a phi$_i$ represent the orthonormal axis. So it is an end dimensional space so j is equal to 1 through n; for in the case n is equal to 2 you can represent it on a paper x and y axis. So phi$_1$ and phi$_2$ will be two axis, they are orthonormal they could represent sin omega t and cos omega t or any other two orthonormal functions, S_{ij} is the projection.

Here phi js are the basis signal such that phi$_i$ integrated with phi$_j$, i not equal to j will give you zero, definition of orthogonal. Each basis signal is normalized to have unit energy that is integration from minus infinity to infinity phi$_i$ square (T) dt is one. So this represents a general framework over which we can define any modulation technique. We have to define orthonormal basis function and then in this vector space I can take points which will represent my various modulating signals.

(Refer Slide Time: 00:46:16 min)

What are the types of the digital modulation techniques? We can broadly classify them into linear, non-linear and spread spectrum. Now in linear is bandwidth efficient, useful for accommodating more users in a limited spectrum. The popular methods are QPSK, OQPSK, pi by 4 QPSK. Some of the non-linear techniques are FSK, GMSK we just looked at the Gaussian filter for the minimum shift key and MFSK. So basically frequency shift keying's and then the spread spectrum, inefficient clearly for single users but it is very popular for multi users scenarios.

19

Here for linear techniques the amplitude of the transmitted signal S (t) that is linearly with the message signal m (t), here the amplitude of the carrier is constant and here the transmission bandwidth is much greater than the minimum required signal bandwidth. So you have lower power but much larger bandwidth and it is closest to the Nyquist theorem which tells you, let you better put your resources in increasing a bandwidth and not in increasing the power. In subsequent lectures we will consider the various linear, non-linear and spread spectrum modulation techniques for mobile wireless communication systems. So let us summarize today's lecture; we first looked at the power and bandwidth efficiencies followed by the line coding.

(Refer Slide Time: 00:48:12 min)
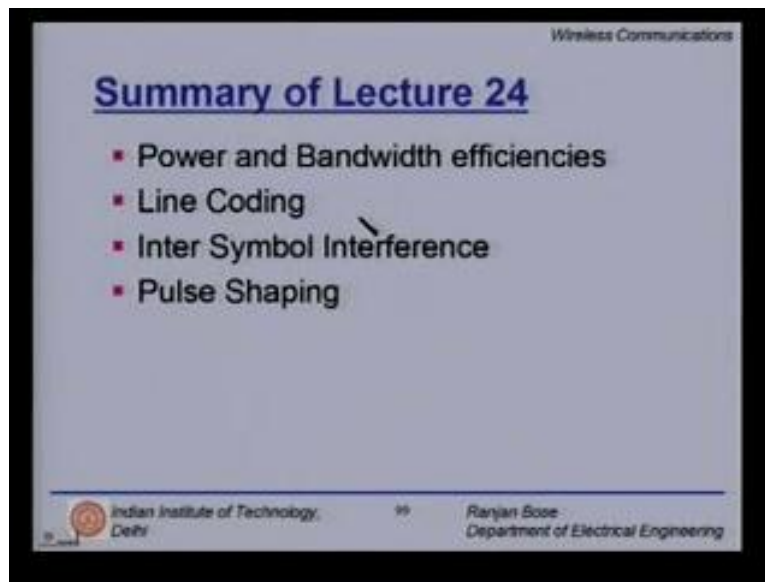
We looked at different kinds of line coding techniques they return to zero, they non-return to zero and the Manchester code. Then we talked about the problem of intersymbol interference, in mobile channels ISI is inevitable and then we looked at the Nyquist criterion to overcome the effects of intersymbol interference. We looked at some examples of pulse shaping filters which can be used for mitigating the intersymbol interference. In the subsequent lectures we will talk about different kinds of digital modulation techniques currently being used in current mobile communication systems as well as future mobile communication systems. We will conclude our lecture here.