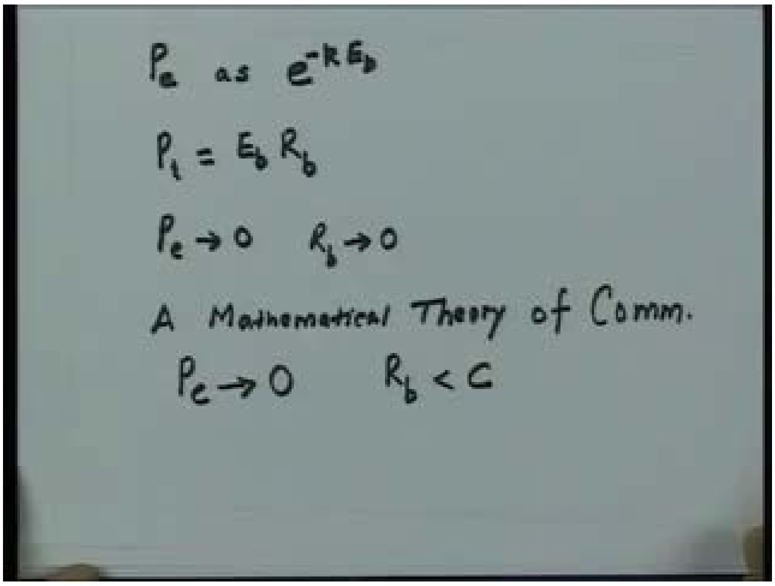
Information Theory and Coding
Prof. S. N. Merchant
Department of Electrical Engineering
Indian Institute of Technology, Bombay

Lecture - 01
Introduction to Information Theory and Coding

In all modes of communication, in general the communication is not error free. We may be able to reduce the accuracy in transmission in digital systems by reducing the probability of error that is Pe.

(Refer Slide Time: 01:06)



P_e as e^{-kE_b}

$P_t = E_b R_b$

$P_e \to 0 \quad R_b \to 0$

A Mathematical Theory of Comm.

$P_e \to 0 \quad R_b < C$

Let me say that probability of error is denoted by Pe, but it appears that, in general, it is not possible to have a communication which is completely error free if noise exists. Let us take example of some digital communication systems. It can be shown that in most of the digital communication system, Pe varies as e raise to minus k times some constant E B asymptotically. What it means is that if I can increase E B that is the energy per bit then I can reduce my probability of error Pe.

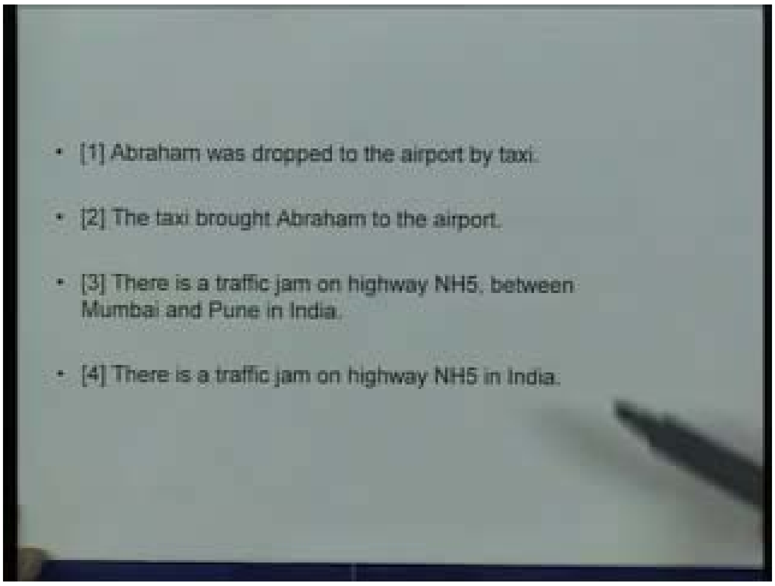
Now, the signal power, let me say is this P I is equal to energy into energy of the bit multiplied by the bit rate. What this implies is that if I want a very large value of E B then I should increase my P I for the given r B or if my P I is fixed then I should reduce my r B to get a large value of e B or I can do both. Now, looking at the physical limitations, increase of P I beyond a certain limit is not feasible. So, in this case if you want to have P e tending to 0 which is attained by increasing E B then for P I fixed, I have to reduce my R b. Hence to reduce Pe, I have to reduce R b.

What this implies is that in a communication channel where noise exists, it is not possible to have error free transmission or error free communication unless the value of R B is very low. Because to make Pe tending to 0 I have to make R B tending to 0 if I assume P I is constant. This is what all communication engineers thought until the publication of a paper by Shannon in 1948. Shannon was an engineer at Bell Systems and he published a classical paper in Bell System Technological journal in the year 1948. The title of the paper was A Mathematical Theory of Communication.

What Shannon showed in this paper is that as long as the rate of transmission is less than a particular limit which he called as channel capacity then it is possible to have error free transmission… It is not necessary for R B to tend to 0 for Pe tending to 0. What he showed was that if I have my R B less than channel capacity c, then it is still possible for me to have error free transmission. The gist of Shannon's paper was that the disturbances which occur on a communication channel do not limit the accuracy of transmission what it limits is the rate of transmission of information.

Now, all this time I have been saying rate of transmission of information. What it appears that I should be able to quantify my information and this is exactly what information theory does? Information theory is a mathematical science which is characterized by a quantitative approach to the notion of information. The word information still quite deceptive, it is very difficult to universally quantify the word information. Let me try to explain this with an example. Let us take these four statements which I have here. These four statements which I have here:

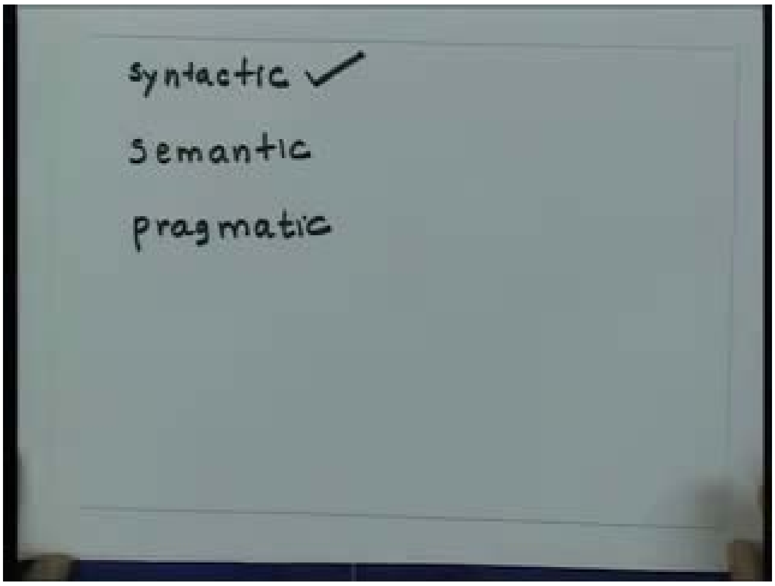
(Refer Slide Time: 06:50)



- [1] Abraham was dropped to the airport by taxi.
- [2] The taxi brought Abraham to the airport.
- [3] There is a traffic jam on highway NH5, between Mumbai and Pune in India.
- [4] There is a traffic jam on highway NH5 in India.

The first statement says a Abraham was dropped to the airport by taxi. The taxi brought Abraham to the airport. There is a traffic jam on highway NH5 between Mumbai and Pune in India. And the last statement says that there is a traffic jam on highway NH5 in India. If you look at these statements then I can say that there are three kinds of information. First is what is known as syntactic information.

(Refer Slide Time: 07:35)



syntactic ✓
semantic
pragmatic

Syntactic information is related to the symbols which I use to build up my message and to the interrelations, that is what I mean by syntactic. If you look at the statement number

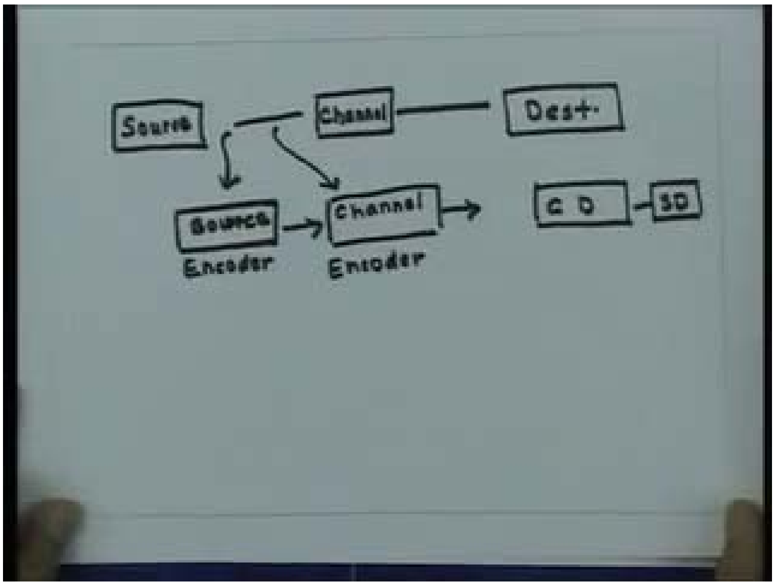
1 and 2, both the sentences in a colloquial meaning of information appear to be same but syntactically they are different. That is what I mean by syntactic information. Another form of information which you could have is what is known as semantic. Semantic information is related to the meaning of the message. It has a referential context. What I mean by semantic is let me take up again the previous example.

If you look at statement number 1 and 2, both are same semantically because both convey the same kind of information in a colloquial meaning of information. But if you look at statement number 3 and 4, they are not only different syntactically, but also they are different semantically. Statement number 3 gives me little more information than the statement number 4 by telling me that the national highway 5 is the highway existing between Pune and Mumbai. The third form of information which I have is what is known as pragmatic information. Pragmatic information is related to the effect and the usage of the message.

Again, I take the statement number all the statement number. Let us take the statement specifically number 3 and 4. Numbers 3 and 4, if I look pragmatically for a person who is staying outside India, this statement need not carry any message. This statement has some kind of effect or usage only for a resident in India. This is what I mean by the pragmatic effect of the message. Now, what we are going to dealing this course is the syntactic aspect of information. We will not be concerned with the semantic or the pragmatic effects of the messages. In a way I could say that in the syntactic form of information I will be concerned with the carriers of information that is the symbols, which I use for the message and not in the information itself.

A simple example could be like this, if I am interested in transmitting the message Abraham was dropped to the airport by taxi and another message the taxi brought Abraham to the airport. Both are same as far information is concerned in a colloquial meaning of information. But from the communication point of view the symbols which are using for transmitting message 1 and message 2 could be different and what we will be concerned is basically with these symbols. Let me take a simple example, suppose I have a source. This source wants to transmit some message to the destination .So I have a destination out here.

(Refer Slide Time: 12:08)

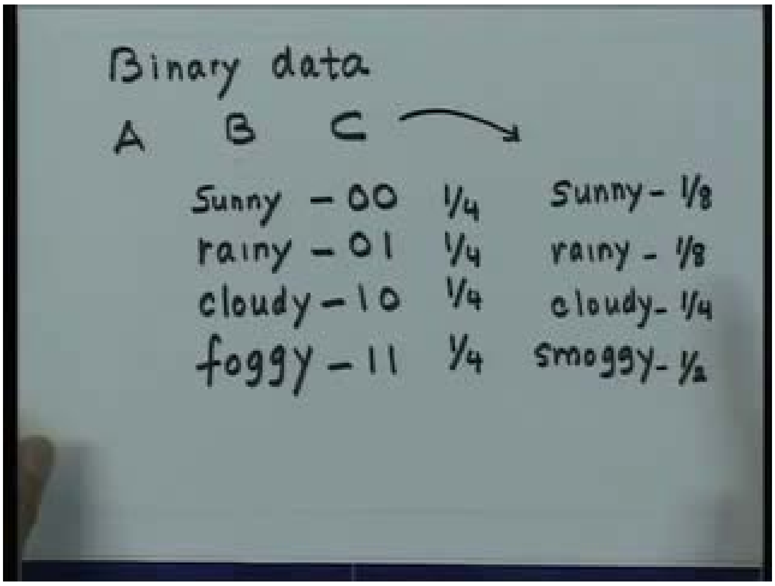


I have a source out here. What I do, using a communication system, is link up this source and destination with a channel. Now, the output of the source may not be really tuned to the way I should send it on a channel. So, what I have to do is convert this source output to a form which is suitable for the transmission on the channel. So, usually you will have 2 more blocks between source and channel. So, one block out here the first block which is falling the source would be what is known as source encoder. The job of the source encoder is basically to map the output of the source in a form which is suitable for a transmission on the channel.

When I am looking at the source encoder, I only look at the characteristics of the source. It is possible that the output of the source encoder may be ready for the transmission on the channel. But because of the characteristics of the channel it is required that you still modify the output of the source encoder and in a practical system what is going to happen is that you are going to have another block falling the source encoder and that would be basically the channel encoder. The job of the channel encoder is to take the input from the source encoder and get the output which is suitable for the transmission on the channel taking the channel characteristics into account.

Once the message has passed to the channel at the destination I am going to have the complementary block for both these block. For the channel encoder I will have what is known as a channel decoder followed by the source decoder. And the output of the source decoder goes to the destination. What Shannon did was with the help of his theory, he tried to find out the optimum source encoder and channel encoder for a given channel. This was possible because of the definition which he used for the information. Let me take another simple example to explain these phenomena.

(Refer Slide Time: 15:53)



Binary data

A B C →

Sunny – 00	1/4	Sunny –	1/8
rainy – 01	1/4	rainy –	1/8
cloudy – 10	1/4	cloudy –	1/4
foggy – 11	1/4	smoggy –	1/2

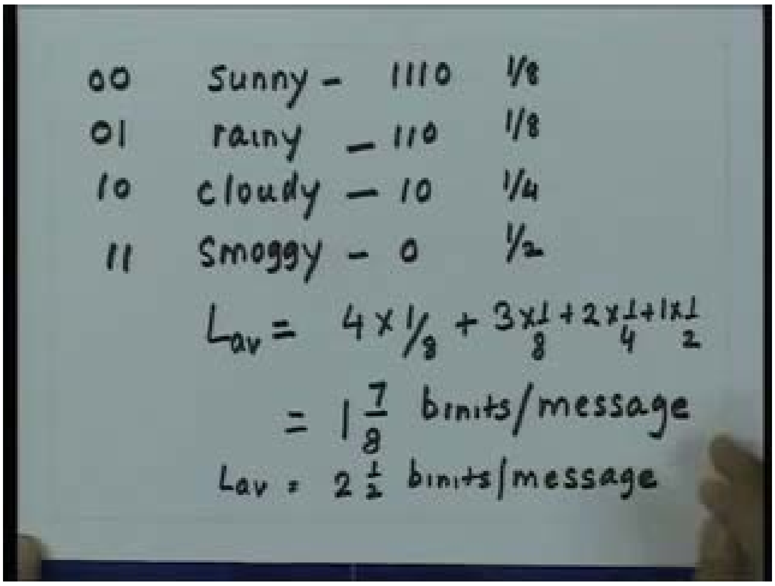
Let me assume that I am interested in the transmission of a simple binary data. I can obtain binary data by various means. The output of a computer is a binary data, the output of a fax machine is a binary data or what I could consider is that I have some kind of messages and using the source encoder I map it to a binary data. To understand this more clearly, let me take a simple example. Let me assume that I have 3 cities A, B and C. What I am interested is to develop a communication link between cities B and city A. The job of the communication link is to transmit the weather status of city B to city A at regular interval of times. So, let us assume that the weather in city B can be classified into four states.

Let us assume that the four states which the city B can have are sunny, rainy, cloudy and foggy. I am interested in transmitting the message that the weather at a particular instant of time in city B is either sunny, rainy, cloudy or foggy. One way of transmitting this message could be to convert these alphabets, s u n n y, into to ASCII codes and then transmit the binary data. More efficient way of transmitting would be to provide some kind of binary labels to these four messages. Let me assume that I provide the binary label 00 to the status sunny. I provide binary label 01 to the status rainy and similarly, 10 and 11.

Now, I also assume one more thing that at any given instant of time, the probability of the weather at this location B, could sunny, rainy, cloudy foggy are each one 4th.Iassume this. Now, my job is basically to design a communication link between B and A. A is interested in finding out what is the status of the weather. So, by transmitting 00, 01, 10, 11, any one of this, I can find out what is the status of the weather. Let me assume that I have another city C and I am also interested in finding out the weather status of the city C at location A. There is a small difference as far as city C is concerned. There are 4 states of the weather at the city C.

But these are sunny, rainy, cloudy and smoggy. A small difference, instead of foggy it could be smoggy. And, I also assume one more thing that the probability of this status of the weather at this location C are one-eighth or let us assume one-eighth, one-eighth, one-fourth and half. Now if I was asked to design a communication link between C and A. Again to know the status of the location of the weather at C at A, I could use the same labeling which I did previously 00, 01, 10 and 11. I could have done that.

(Refer Slide Time: 21:37)



00	Sunny -	1110	1/8
01	rainy -	110	1/8
10	cloudy -	10	1/4
11	Smoggy -	0	1/2

$$L_{av} = 4 \times 1/8 + 3 \times \frac{1}{8} + 2 \times \frac{1}{4} + 1 \times \frac{1}{2}$$

$$= 1\frac{7}{8} \text{ binits/message}$$

$$L_{av} = 2\frac{1}{2} \text{ binits/message}$$

But instead of that what I do is basically, for city C, I do something different. For, sunny, rainy, cloudy and smoggy, I assign the bits as 10,110,1110 and 0. The probabilities here are one-eighth, one-eighth, one-fourth and half. So, instead of trying to assign 00, 01, 10 and 11, what I assign is this. Now, what is the reason for assigning like this? Why cannot I assign the bits like this and why I am assigning like this? Let me try to answer this question in a different manner. What is the difference between a communication link between city B and A and that between C and A? As a communication engineer what I am supposed to design a communication link which has minimum cost of operation.

Now, criterion for the cost of operation could be decided by the number of bits which I transmit per message per second on average. That could be one of my criterion. If I take that as my criterion, then you can see that if I use this kind of labeling between my messages and binary bits, then every number of bits which I require to transmit binary bits per message would be given by L av. I am using 4 bits, so 4 multiplied by probability of sunny that is one-eighth, plus rainy is 3 multiplied by one-eighth, cloudy is 2 multiplied by one-fourth and smoggy is 1 multiplied by half. This comes out to be 1 seven-eighth binary digits which I call as binits per message.

Instead of this, if I had used this kind of labeling, you can find out the number of bits required for this. It will turn out to be more than 1 by seven-eighth or another way of saying would be. Let me assume that I used this kind of assignment for my previous link between B and A. If you try to do that way you will find out that every number of bits required would be 2 and half binits per message. Let us take the communication link between C and A. I have got the value of 1 seven-eighth. If I had used this kind of a mapping, I would have got the value which is higher than this or if I use generally this 00, 01, 10 and 11 then I would have got 2 binits per message. So, I save 1 by 8 binits per message that come roughly about 6 percent.

Now, there is a question that arises, by this kind of a mapping, what I have done is that I have been able to save 1 by 8 binits per message for the transmission between C and A. The question that arises is, is it possible for me to get a mapping which is better than this? So, fortunately this case it is possible. If it is possible, then how low can I go? Is it possible for me to answer that question? That is the second question I have. The third question is that if it is possible for me to go then how do I synthesize this kind of a mapping? In information theory and coding, jargon, I would say this mapping from the message to the binary bits is what is known as formation of a code.

So what I am asking is that if it is possible for me to go even below 1 seven-eighth binits per message. Then how do I design that code? That is the third question I have. And finally, a very important question that comes to my mind is why is this different? Why is there a difference in the number of binits per message required for transmission of the weather status from B to A is different from that of C to A? What is the reason? If you look at these both the examples, this is the example which I have for location C and this is the example which I have for location B.

In both the cases, it is important to realize that when I did this kind of a mapping, I had probability of this message somewhere behind my mind in designing of this code. Since the probability of occurrence of all this were equal what I did I assigned 2 bits to each whereas, if you look at this case what a small trick I played was that I knew that the probability of occurrence of smoggy is on a higher side. So, I assigned less number of bits to declare the status of smoggy. So, this is a clever kind of coding which I have done. I did it in a heuristic manner.

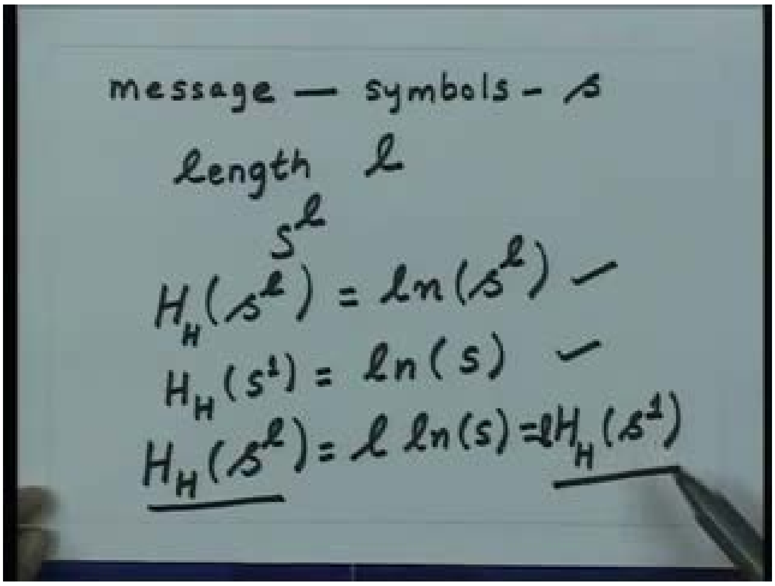
But I could exploit why I could get this reduction was that I could exploit the characteristics of this messages. In this case though the messages, as from communication point of view, I have to transmit 4 messages here. In the earlier case also I had to transmit 4 messages except the difference between foggy and smoggy. Except for that difference, there is no difference. But since we are interested in syntactic aspect of information, from a communication engineer point of view, it makes no difference for me what 4 messages I have. But still I am getting a reduction in the latter case is because of the statistics of the messages which I have here and the statistics of the messages, which I have in the earlier case. This is the basic difference which I have exploited..

What this example shows is that since on the average I require less binits per message for transmitting the information about the weather of city A to city A and from B to A. It means that during my transmission from C to A, I have less amount of information. Is it possible for us to quantify information in a manner which reflects this observation? Since there is a less amount of information, for transmission from C to A, it appears that I require less number of binits per message. How do I quantify this? Let us have a look at it.

From a previous example, which we discussed, it is clear that the definition of information is concerned with the probability of occurrence of the various messages. With this intuition, let us go ahead and define a measure for information. The founder of information theory was Shannon. Shannon, in his paper a ‘Mathematical Theory of Communication’ which he published in 1948, set the foundation for the present day modern information theory. What he did was, he defined a measure of information and showed that this measure of information had properties which very well correlated with the physical models of communication and using this definition he could answer many of the questions which we had posed earlier.

Before Shannon, in 1924, we had Niquist who also talked about information, not directly but indirectly. In his paper, he discussed how to transmit messages. For that matter, he used the word characters over a telegraph channel with maximum speed but without distortion, he did not talk about information as such. Then, in 1928, Hartley was the first person who defined information measure. Let us look into the simplest definition of information measure which was first provided by Hartley. He assumed that message can be constructed with the help of symbols. These symbols can have s possibilities.

(Refer Slide Time: 33:13)



message — symbols - s

length l

s^l

$H_H(s^l) = \ln(s^l)$ ✓

$H_H(s^1) = \ln(s)$ ✓

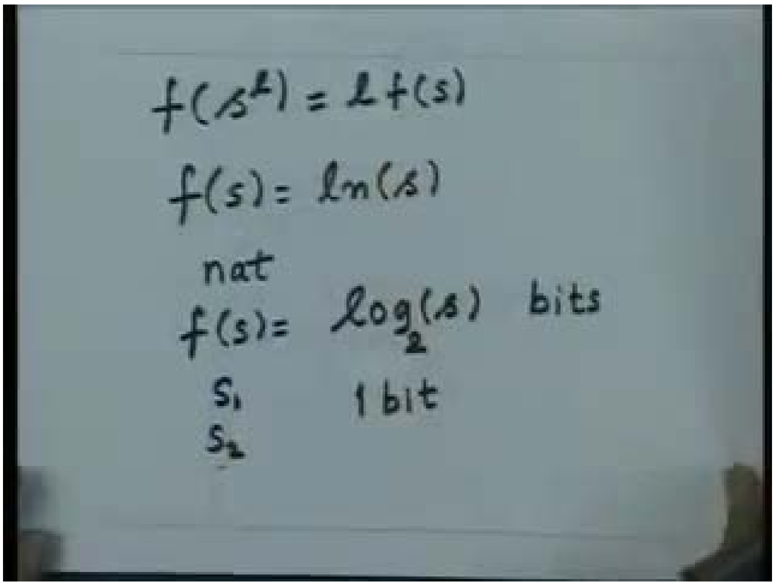
$\underline{H_H(s^l)} = l \ln(s) = \underline{l H_H(s^1)}$

So, I have a message. Message is constructed out of symbols. Each of these symbols which I have can take s values. So, each of the symbols which forms the message can have s possibilities. What this means that now if I am looking for messages of length l symbols. So, if have messages of length l then I can form s raise to l distinct messages. If I assume this, then he defined the measure for information as for a particular message of length l consisting of l symbols. Each of the symbols have s possibilities then the information associated with this message is defined as h is the information associated with the message of length l. And let us use a suffix as h which denotes the Hartley's definition is defined as log of this value.

This definition has an interesting implication. If I take a message of length one, then by this definition it means I have the information contained in that message of length 1 equal to log of s. Now, if you look at this relationship and this relationship, I can easily write h of s l equal to l of log of s is equal to this. And this is the relationship which I have between 2 messages of different lengths. This is a message of length of l and this a message of length 1. What it says that by this definition the information contained in a message of length l is l times the information contained in a message of length 1. This definition is intuitively very satisfying.

We expect that if I have a message of length l then I should have information contained in that l times the information contained in a message of length 1. So, Hartley's definition is very convincing. Now, the reason for choosing the function log is not very difficult to explain. What I want intuitively is that information contained a message which is of length l should be l times the mess information contained in message of length 1.

(Refer Slide Time: 37:40)



$f(s^l) = l f(s)$

$f(s) = \ln(s)$

nat

$f(s) = \log_2(s)$ bits

s_1 1 bit

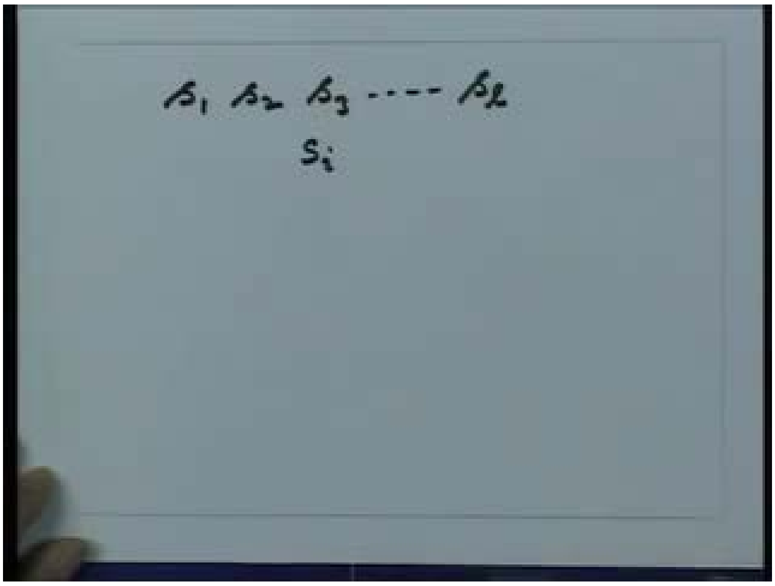
s_2

Then what I should do is that I should have chosen my function in such a way that f of s l should be equal to l times f of s. Now, it is not very difficult to show that the only function which satisfies this requirement is a log function. So, we define f of s as log of s where the logarithm is to the base e. This is your natural logarithm. So, the unit of information which I get using the Hartley's measure, is known as nat. Probably nat means natural logarithm. Is it possible for us to use a base other than the base e? The answer is yes. Instead of trying to define a natural logarithm, I could define the information in terms of logarithm to the base 2. So, I can have my definition of f s as log of s but this time it is to the base 2. In this case, the units of information will be bits. The choice of the bits is not very important but in a practical for our course we will be using the base 2.

So, the information will be measured in terms of bits. The choice of 2 also has a physical implication. Let us take an example. I have two messages, s 1 and s 2.So, in this case, if I have two messages s 1 and s 2. If I go by this definition the information contained in this 2 messages s 1 and s 2 of the same length let us assume the length of s 1 is 1 and the length of s 2 is 1. And each of this message can take either 1 or 0, then the information contained in this message would be equal to 1 bit. Now, if you have 2 messages then to identify this message require physically 1 bit and it very well correlates with the information theoretical approach too. So that is the reason for choosing the base 2.

There is a small problem with the definition of Hartley’s approach to the measure of information. The difficulty is, in this I have assumed that the symbols which form a message are all equi probable. The second assumption which I have made is that there is no relationship between the symbols in this message. What Shannon did was he went beyond the results of Niquist and Hartley. He defined an information measure which takes care of this fact that I could have a message consisting of l symbols. But it is possible that each of these symbols need not occur with equal probability.

(Refer Slide Time: 41:56)



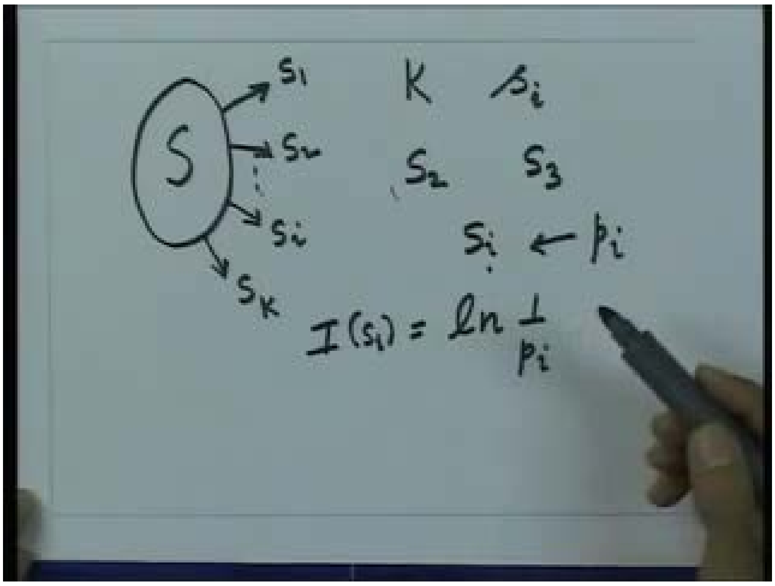
$s_1\ s_2\ s_3 \cdots s_l$

S_i

That is the first thing what it means that if I have a message composed of some symbols s 1 s 2 s 3like this up to s l and each of this symbols I take any of this s possibility. It is not necessary for me to have this s possibilities for the symbols s I to be equi probable. Shannon said it is not necessary. That is the first contribution he made. Second contribution that he made was that it is also not essential for me assume that s 1 s 2 s 3 and s l occur independently. In most of the communication systems and the sources, which are linked in the communication system, you will find the way the symbols are being generated in a message are inter dependent.

What it means is that occurrence of s 2 could be dependent on the occurrence of s 1.Occurrence of s 3 could be dependent on the first symbols s 1 and s 2. If I take that into consideration then how do I define the information contained in this complete message and Shannon developed the theory for defining the information for such messages? That is his major contribution. Let us start with the Shannon's definition of information to understand this.

(Refer Slide Time: 43:44)

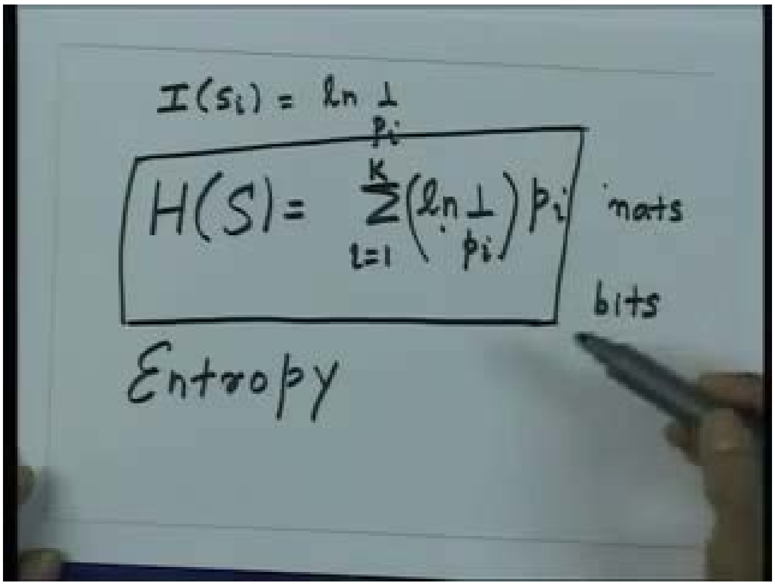


Let me assume that I have a source s. This source s emits different messages. Let us assume the messages which this source emits can take the value from s 1s 2. So the symbols which emit from the source s can be either s 1 s 2up to s k. That means I have k possibilities for a particular symbol s i. If I take any particular symbol, let us assume if I take the occurrence of s 2. When s 2 occurs is it possible for me to say how much information I gain when s 2 occurs? In the definition of my measure for information, we should be taking a fact into consideration that if the probability of occurrence of s 2 is very low and when s 2 occurs then the information, which I gain on the occurrence of s 2 should be high compared to say s 3, which occurs with high probability.

This the intuitive reasoning which goes behind in the definition of information measure. So, what it means that I should define a function which is inversely proportional to the probability of occurrence of s i. Now, I could have chosen any function which is inversely proportional to s i. Now, I can choose a function which is logarithmic. So, what I say that when s I occurs the information which I get from s I is I define as log of 1 by P I where P i is nothing but the probability of occurrence of s i.

Now, when I define like this, this is what is known as self-information for the symbol s i. If I define like this, this correlates very well with my intuitive reasoning for the information measure which says that if the probability of occurrence is low I should get more information. If the probability of occurrence is high I should get no information. This is what happens next is basically is that if you look that if the probability of occurrence can be always p I will always be lying between the values0 and 1. So, what it means is that if I have the value near 1 my probability of occurrence this will become 0. So, if I know that the symbol is always going to occur then the information contained in that symbol is obviously 0 because I know the symbol is always going to occur I do not gain anything out of that.

(Refer Slide Time: 47:34)



$$I(s_i) = \ln \frac{1}{p_i}$$

$$\boxed{H(S) = \sum_{i=1}^{K} \left(\ln \frac{1}{p_i}\right) p_i}$$

nats

bits

Entropy

So this is what I have done for the definition of self-information for a single symbol s I but now if I am interested since this output of this source is random in nature I do not know what symbols will emit at a particular instant of time. If I do not know what symbols are going to emit at a particular instant of time, then if I am still interested in finding out what is the average information contained in the source s, then I can use the probability theory to do that .It is very simple to find out the average information contained in source which emits s symbols from s 1 to s k.

What I will do is I will define that average information in terms of h s that is nothing hut equal to summation of log of 1 by p i multiplied by p is this over i equal to 1 to k. What it means is that, this is the average information which I get depending on the bases which I have. If I take the base to be natural log then I get in nats. If I choose the base to be 2, then I get what is this bits. So, to this is the first step which we have derived. This is the first important result which we have derived in information theory. What it says is that the information which I get from a source emitting symbols from s 1 to s k is given by this equation. In literature, this average information for source is known as entropy. When I derived this, I made a small assumption that the symbols are independent. If the symbols are not independent then this formula has to be modified a bit.

Next time, we will study the properties of the entropy which we have defined as the measure for information. It is very important to realize that though this definition seems to be reasonable and the relationship which we derive from this definition are internally consistent in the framework of our definition, it is not justification for defining this way but the beauty of this definition is that with the help of this definition and the framework which we develop based on this definition will help us to answer the question which we had posed earlier. As a communication engineer, I am interested to transmit and store information as compactly as possible.

This is the first thing which I want to do. Second thing what I want to find out is what is the amount of the information that I can transmit over a given channel. Are there any limitations? As I discussed in my earlier in the earlier part of my lecture today that Shannon showed random disturbance on a channel does not set limit to the accuracy of transmission. But it sets a limit to the transmission rate of information. And if the transmission rate of information is less than the channel capacity then it is still possible to have error free communication without making the bit rate tending to 0. This is the major contribution of Shannon and based on this we will go ahead.