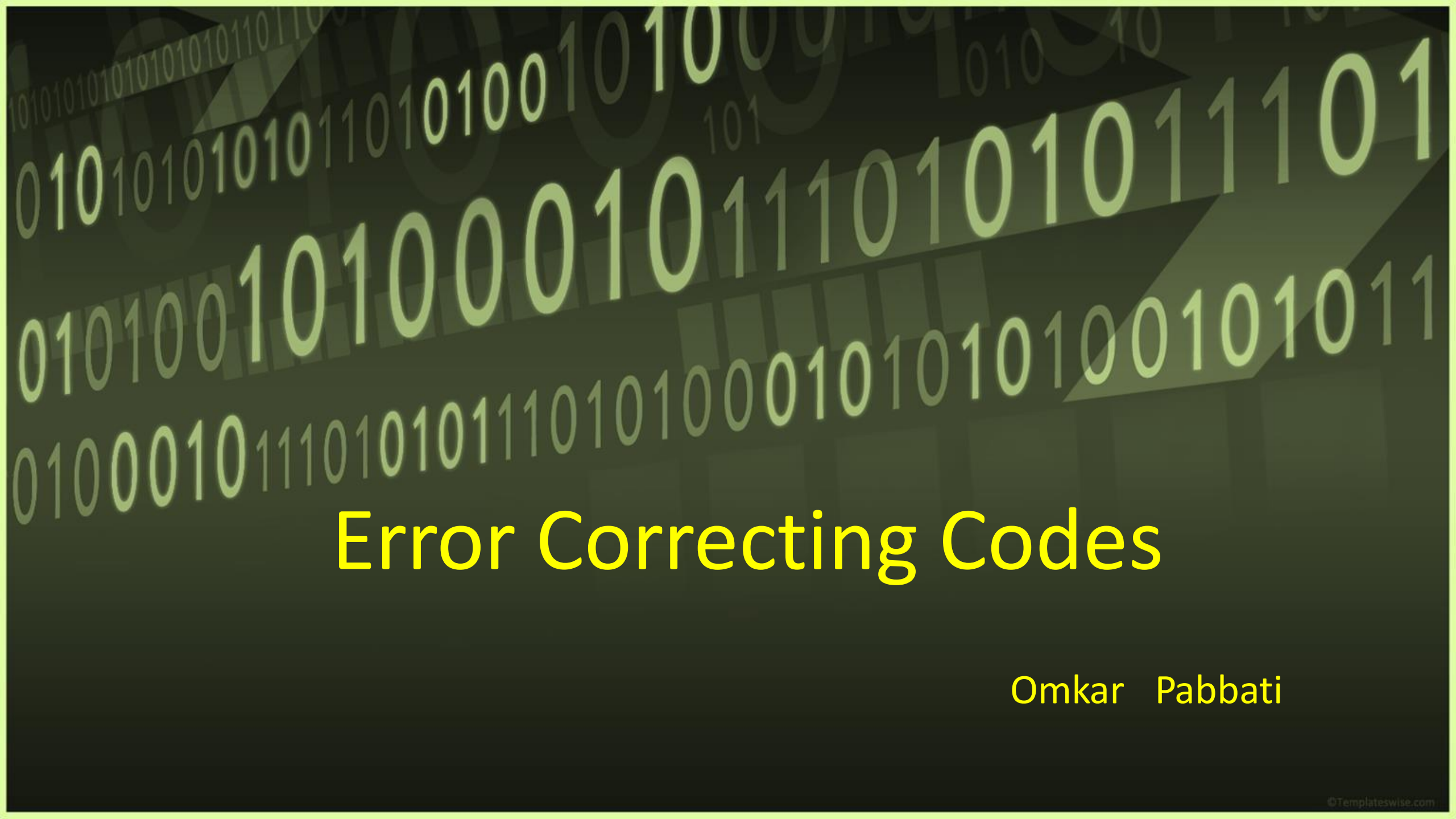


Error Correcting Codes

Omkar Pabbati

Introduction

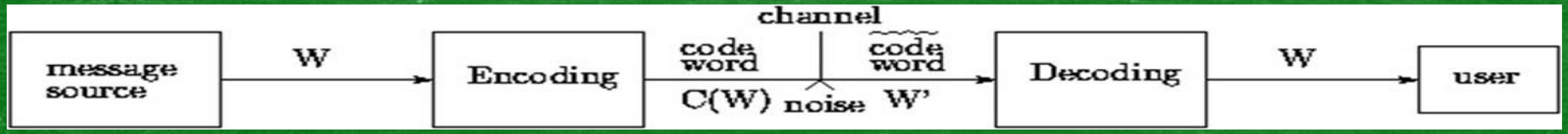
- Coding:
  - ✓ The process of assigning a code to something for classification or identification.
  - ✓ convert (the words of a message) into a code so as to convey a secret meaning.

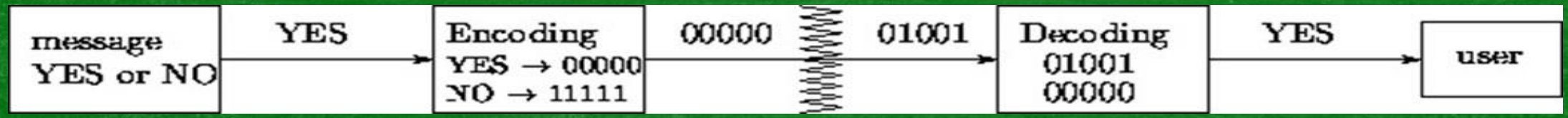
Error detection is the detection of errors caused by noise or other impairments during transmission from the transmitter to the receiver.

Error correction is the detection of errors and reconstruction of the original, error-free data.

example

- Error-correcting codes are used to correct messages when they are transmitted through noisy channels.





Definition of Code

Block code: all words are the same length.

A q-ary code C of length n is a set of n-character words over an alphabet of q elements.

Examples:

C_1 = {000, 111} binary code of length 3

C_2 = { 00000, 01100, 10110} binary code of length 5

C_3 = {0000, 0111, 0222, 1012, 1020, 1201, 2021, 2102, 2210} ternary code of length 4

Types of codes

- The codes are mainly classified as block codes and convolution codes

- Block codes: These codes consists of 'n' number of bits in one block or code word. This code word consists of 'k' message bits and (n-k) redundant bits. Such codes are called (n,k) block codes.

- Convolution Codes: the coding operation is discrete time convolution of input sequence with impulse response of the encoder . The convolution encoder attempts the message bits continuously and generates the encoder sequence continuously.

The codes can also be classified as linear or nonlinear codes.

- Linear Codes: if the two code words of the linear code are added by modulo-2 arithmetic, then it produces third codeword in the code.

  This is very important property of the codes, since other code words can be obtained by addition of existing code words.

- Nonlinear Code: Addition of the nonlinear code words does not necessarily produce third codeword.

Types of Error

- There are mainly two types of errors introduced during transmission of the data. Namely Random errors & Burst errors.

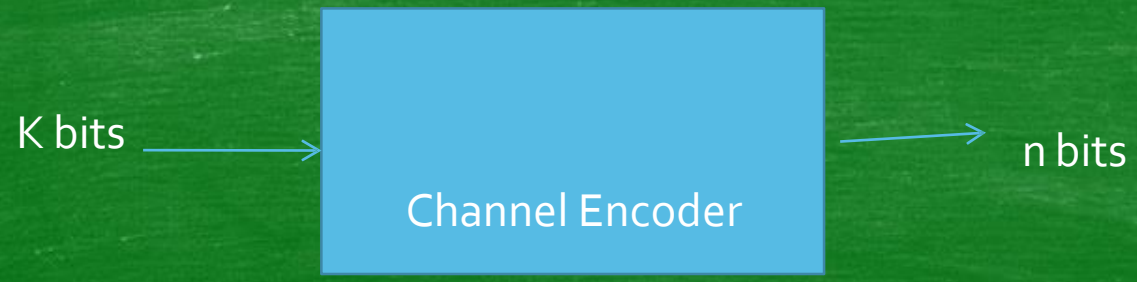
(i) Random Errors: These errors are created due to white Gaussian noise in the channel. The errors generated due to white Gaussian noise in the particular interval does not affect the performance of the system in subsequent intervals. In other words, these errors are totally uncorrelated. Hence they are also called random errors.

(ii) Burst error: These errors are generated due to impulsive noise in the channel. These impulsive (burst) noise are generated due to lightening and switching transitions. The noise bursts affect the several successive symbols. Such errors are called burst errors. The burst errors are dependent on each other in successive message intervals .

Note: Data Burst: Any relatively high-bandwidth transmission over a short period. Transmission that combines a very high data signaling rate with very short transmission times - i.e., the message is compressed.

Redundancy

- **Redundancy** is the number of bits used to transmit a message minus the number of bits of actual information in the message.



- Data compression is a way to reduce or eliminate unwanted redundancy, while checksums are a way of adding desired redundancy for purposes of error detection when communicating over a noisy channel of limited capacity.

Hamming Distance

- The Hamming distance between two words over the same alphabet is the number of places where the symbols differ.

1. Example : d(100111, 001110) = 3
   - Look at 100111
     001110

Example: d (10101, 01100) = 3

d (*first, second, fifth*) = 3

- For a code , C, the minimum distance d(C) is defined by

$d(C) = \min\{d(c_1,c_2),| c_1, c_2 \in C, c_1 \neq c_2\}$

Example

Find the minimum Hamming distance of the coding scheme in Table

Solution ***We first find all Hamming distances.***

$d(000, 011) = 2$	$d(000, 101) = 2$	$d(000, 110) = 2$	$d(011, 101) = 2$
$d(011, 110) = 2$	$d(101, 110) = 2$		

The d_{min} in this case is 2.

Example

Find the minimum Hamming distance of the coding scheme in Table

Solution ***We first find all the Hamming distances.***

$d(00000, 01011) = 3$	$d(00000, 10101) = 3$	$d(00000, 11110) = 4$
$d(01011, 10101) = 4$	$d(01011, 11110) = 3$	$d(10101, 11110) = 3$

The d_{min} in this case is 3.

Detection and Correction

- A code C can detect up to s errors in any codeword if

$$d(C) \geq s + 1.$$

- A code C can correct up to t errors if

$$d(C) \geq 2t + 1$$

For a hamming distance of 5, how many errors can be detected? How many errors can be corrected?

Hamming Bound

Hamming bound for a Binary code is given by,

$$2^{n-k} \geq \sum_{j=0}^{t} \binom{n}{j}$$

Codes that attain the Hamming bound are called **perfect codes**.

Determine the hamming bound for a ternary code.

Coding Theory

- Coding theory - theory of error correcting codes - is one of the most interesting and applied part of mathematics and informatics.
- All real systems that work with digitally represented data, as CD players, TV, fax machines, internet, satellites, mobiles, require to use error correcting codes because all real channels are, to some extent, noisy.
- Coding theory results allow to create reliable systems out of unreliable systems to store and/or to transmit information.
- Coding theory is a practical method to achieve high data rate.

(Bandwidth, redundancy and security)

Introduction to information theory

- The Performance of the communication system is measured in terms of its error probability. An errorless transmission is possible when probability of error at the receiver approaches zero.
- The performance of the system depends upon available signal power, Channel noise and bandwidth. Based on this parameters it is possible to establish the condition for errorless transmission. These conditions referred as Shannon's theorems.
- The information theory is related to the concepts of statical properties of messages/sources, channels, noise interference etc. The information theory is used for mathematical modeling analysis of the communication systems.

Measure of Information

- Let us consider the communication system which transmits messages m1,m2,m3……, with probabilities of occurrence p1,p2,p3……., the amount of information transmitted through the message m_k with probability P_k is given by

- Amount of information $I_k=\log_2(1/P_k)$

- Unit is bits (binary digits)

Physical interpretation of Amount of Information

- We know Indus university declares large number results every semester. suppose that you received the following messages.

1. Result is declared today
2. Result is declared today
3. Engineering Result is declared today
4. Electronics communication of B.Tech results declared today.
5. Electronics communication of B.Tech 6^{th} semester results declared today.

➢ *(probability of occurrence is very small but amount of information received is great)*

Entropy (Average Information)

Definition: The average information per message is called entropy, it is represented in bits/message.

- Consider that we have M-different messages. Let these messages be $m_1, m_2, m_3 \ldots m_M$ and they have probabilities of occurrence as $p_1, p_2, p_3, \ldots p_M$. Suppose that a sequence of L messages is transmitted. Then if L is very very large, then it may say that,
- p_1L messages of m_1 are transmitted, p_2L messages of m_2 are transmitted, p_3L messages of m_3 are transmitted ..... p_ML messages of m_M are transmitted.

Hence the information due to message m_1 will be,

$$I_1 = \text{Log}_2(1/p_1).$$

- Since there are p_1L number of messages of m_1, the total information due to all messages of m_1 will be

$$I_1(\text{total}) = p_1L\log_2(1/p_1)$$

Contd…

- Similarly,

$$I_2(total)=p_2 L \log_2(1/p_2)$$

- Thus the total information carried due to the sequence of L messages will be,

$$I(total)= I_1(total)+ I_2(total)+.....+ I_M(total)$$

$$I(total)= p_1 L \log_2(1/p_1)+.....$$

Average information = Total information/Number of messages.

= I(total)/L.

Average information is represented by entropy of the source. and it is denoted by H.

$$H= I(total)/L.$$

$$H= p_1 \log_2(1/p_1)+ p_2 \log_2(1/p_2)+.....+ p_M L \log_2(1/p_M).$$

Properties of entropy

- Entropy is zero if the event is sure or it is impossible .i.e. H=0 if Pk=0 or 1.
- When Pk=1 /M for all the M symbols , then the symbols are equally likely. For such source entropy is given H= $\log_2 M$.
- Upper bound on entropy is given as, Hmax= $\log_2 M$.
- The entropy is maximum when both the messages are equally likely.

Information rate

Information rate (R) is represented in average number of bits of information per second.

- Information rate is represented by R and as given as, R=rH

Here R is the information rate..

H is the Entropy or average information.

Information rate R It is calculated as follows:

- R=(r in message/second) X (H in information bits/message)

  R=Information bits/second

- We know that the signal is sampled at nyquist rate . Nyquist rate for B Hz band limited signal is,

  Nyquist rate=2B samples/sec.

- Since every sample generates one message signal per second,

  r=2B messages/sec.

Eg. The signal band limited to 4 kHz

Find the sampling rate and determine the information rate with entropy of 1.9043

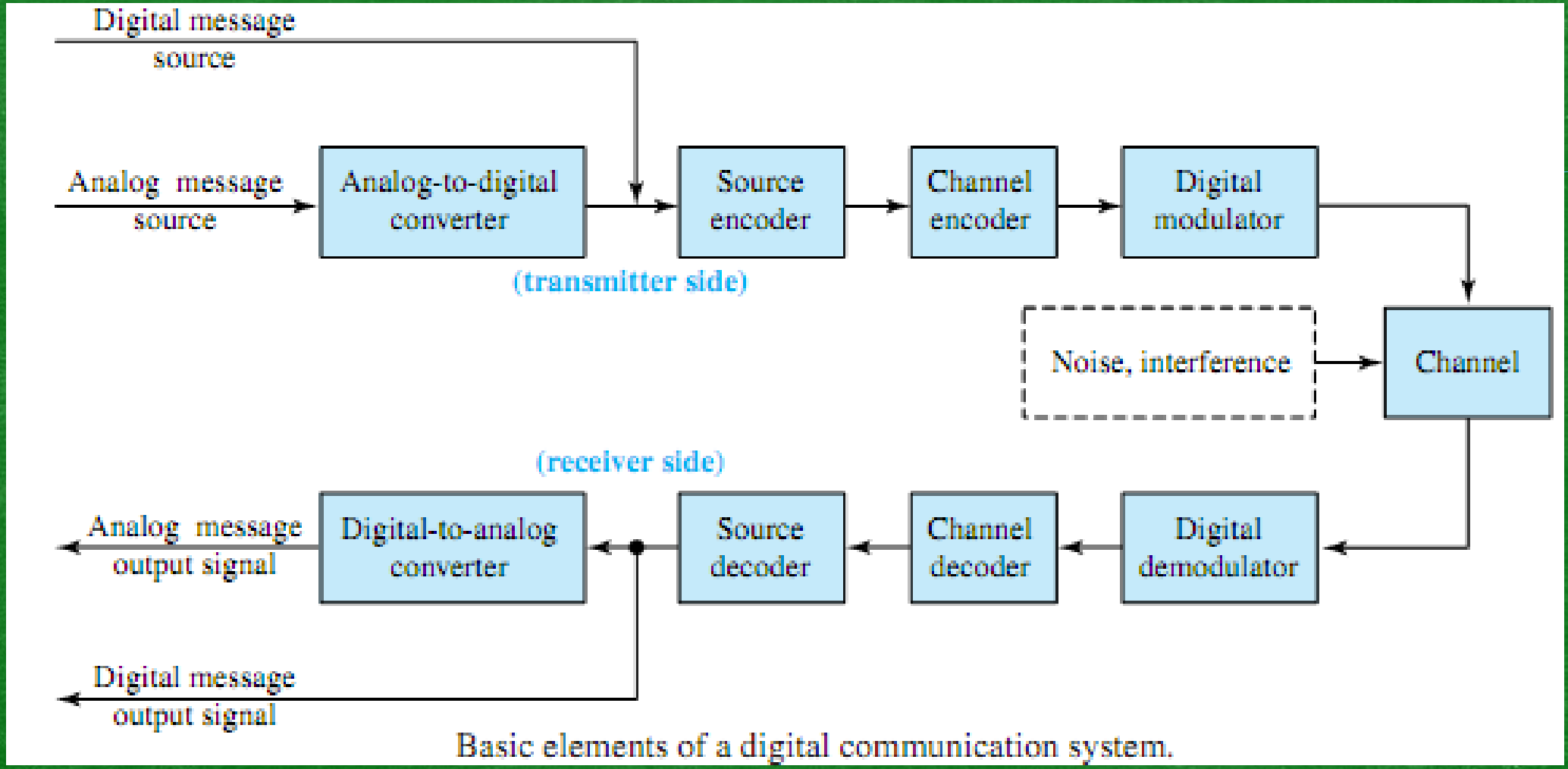
Ans: 15.2346 bits/sec.

Source coding

- Objective: an objective of source coding is to minimize the average bit rate required for representation of the source by reducing the redundancy of the information source.
- Source encoding attempts to compress the data from a source in order to transmit it more efficiently.
- This practice is found every day on the Internet where the common Zip data compression is used to reduce the network load and make files smaller.
- The aim of source coding is *"to take the source data and make it smaller".*
- Eg,. Huffman codes, golay code etc.

Channel coding

- channel encoding, adds extra data bits to make the transmission of data more robust to disturbances present on the transmission channel. The user may not be aware of many applications using channel coding.
- eg,. Reed-solomon codes, turbo codes, cyclic codes, convolution codes etc.
- A typical music CD uses the Reed-Solomon code to correct for scratches and dust. In this application the transmission channel is the CD itself.
- Cell phones also use coding techniques to correct for the fading and noise of high frequency radio transmission.



Basic elements of a digital communication system.

The process of efficiently converting the output of either an analog or a digital source into a sequence of binary digits is called source encoding or data compression.

The process of adding patterns of redundancy into the transmission path in order to lower the error rate

Terms related to source coding process

- ✓ Codeword length : The length of a codeword is the number of binary digits in the code words.
- ✓ Average codeword length:

$$L = \sum_{i=1}^{n} P(x_i) L_i$$

- ✓ Code efficiency
- ✓ Code redundancy
- ✓ Code variance

Code efficiency

- Code efficiency (η)= H/L
- Code redundancy (γ)= 1- η
- Code variance: variance is the measure of variability in code word lengths.

$$\sigma^2 = \sum_{k=0}^{M-1} P_k \left(n_k - \bar{L} \right)$$

$\bar{L}$ Is the length of the code word

Note: Variance should be as small as possible

Huffman Coding – a type of prefix code

Basic idea : Assign to each symbol a sequence of bits roughly equal in length to the amount of information conveyed by the symbol.

Huffman encoding algorithm:

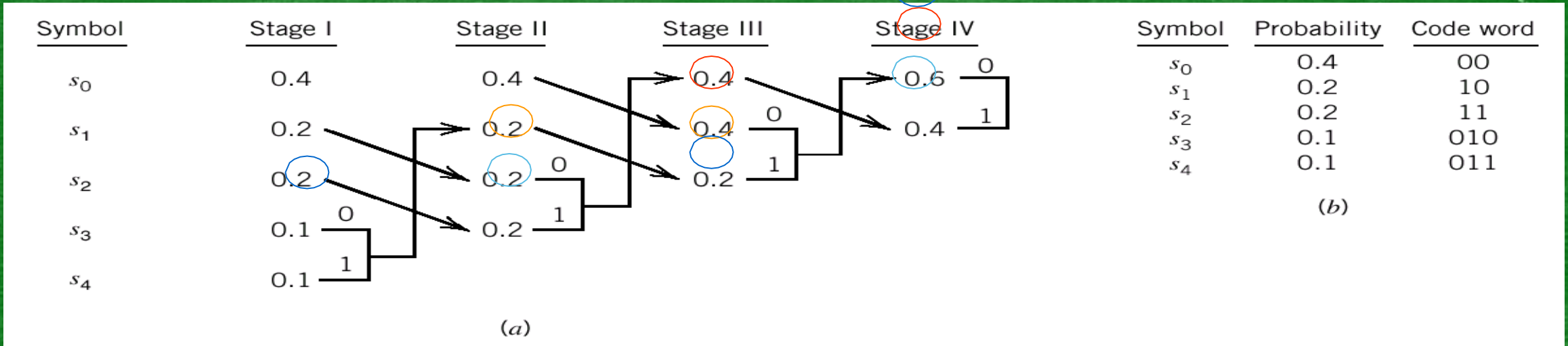
Step 1: The source symbols are listed in order of decreasing probability. The two source symbols of lowest probability are assigned a 0 and 1.

Step 2: These two source symbols are regarded as being combined into a new source symbol with probability equal to the sum of the two original probabilities. The probability of the new symbol is placed in the list in accordance with its value.

The procedure is repeated until we are left with a final list of symbols of only two for which a 0 and 1 are assigned.

The code for each source symbol is found by working backward and tracing the sequence of 0s and 1s assigned to that symbol as well as its successors.

Huffman Coding – Example



Symbol	Probability	Code word
s_0	0.4	00
s_1	0.2	10
s_2	0.2	11
s_3	0.1	010
s_4	0.1	011

(b)

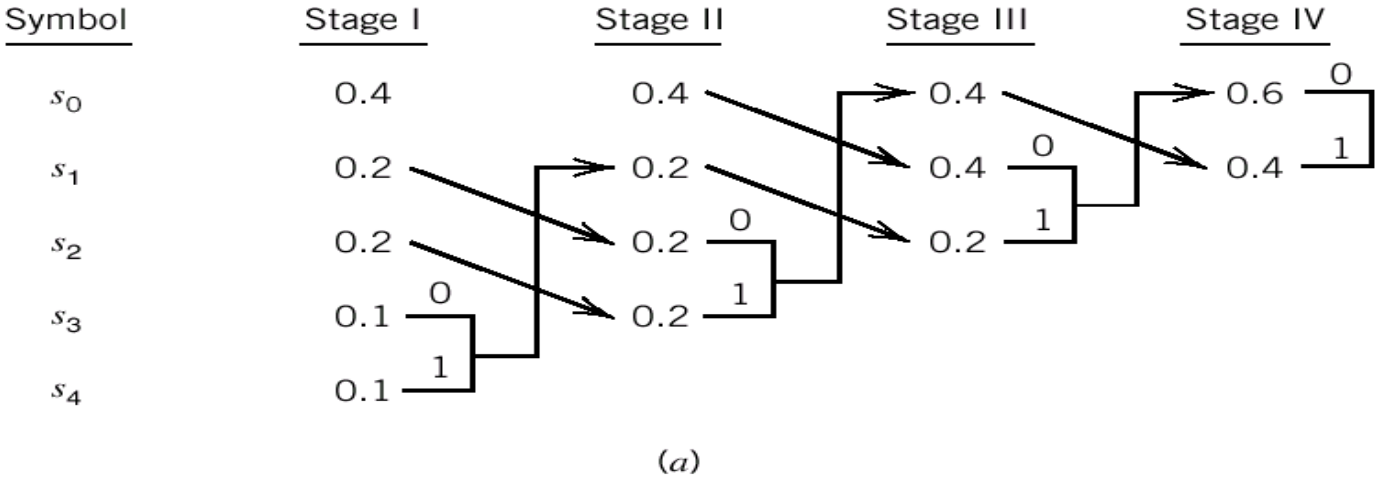
Step 1: The source symbols are listed in order of decreasing probability. The two source symbols of lowest probability are assigned a 0 and 1.

Step 2: These two source symbols are regarded as being combined into a new source symbol with probability equal to the sum of the two original probabilities.

The probability of the new symbol is placed in the list in accordance with its value.

The procedure is repeated until we are left with a final list of symbols of only two for which a 0 and 1 are assigned.
The code for each source symbol is found by working backward and tracing the sequence of 0s and 1s assigned to that symbol as well as its successors.

Huffman Coding – Average Code Length



(a)

Symbol	Probability	Code word
s_0	0.4	00
s_1	0.2	10
s_2	0.2	11
s_3	0.1	010
s_4	0.1	011

(b)

$$L = \sum_{i=1}^{n} P(x_i) L_i$$

= 0.4(2) + 0.2(2) + 0.2(2) + 0.1(3) + 0.1(3)

= 2.2

Huffman Coding – Exercise

Symbol	S_0	S_1	S_2	S_3	S_4	S_5	S_6
Probability	0.25	0.25	0.125	0.125	0.125	0.0625	0.0625

Compute the Huffman code by placing the probability of the combined symbol as high as possible.

What is the average code-word length?