1

Overview of Communication Systems

Success is not the absence of failure but the triumph of improved attempts.

In this Chapter

- ✓ A quick overview of nonelectrical telecommunication techniques highlighting their gross inadequacies for today's communication needs. You will find, however, that these ancient techniques are still indispensable in certain situations.
- ✓ A brief historical sketch of developments in modern (electrical) telecommunication from telegraphy to the Internet and 5G. Key developments in binary codes for data transmission, electronic components, transmission media, signal processing, and telecommunication services are summarised.
- ✓ A discussion of the elements of a communication system. Modern telecommunication systems may vary widely in complexity and applications, but they are all accurately represented by one block diagram. You will become conversant with the elements of this generic diagram and the roles and signal processing tasks that they perform. Information sources, information sinks, transmitters, and receivers are all briefly introduced.
- ✓ A detailed overview of the classification of communication systems. Every system is *simplex* or *duplex*, *analogue* or *digital*, *baseband* or *modulated*, circuit-switched or packet-switched. You will learn the features of each of these systems and find out just why we have undergone a transformation to *digital* and *IP* dominance.

COPYRIGHTED MATERIAL

1.1 Introduction

This first chapter provides a panoramic view of communication systems. It is intended to give a lucid and comprehensive introduction to the subject of *communication engineering*. We follow a nonmathematical approach and attempt to show you how each telecommunication concept fits into the overall picture. Armed with this knowledge, it is hoped that you will have enough inspiration and motivation to go on to the remaining chapters, where the principles and terminology are treated in more detail. If you do not have the time to study the entire book, it is advised that you work carefully through this chapter and the introductory sections of the remaining chapters. The material thus covered is suitable for a short course that presents a survey of modern telecommunications.

To drive home the fact that telecommunication generally means communication at a distance, we begin with a quick overview of various nonelectrical means of telecommunicating. The use of the word *telecommunication* is then narrowed to apply exclusively to the electrical techniques. After an important review of selected significant historical developments from telegraphy to 5G, we present a block diagram that adequately describes all

Communication Engineering Principles, Second Edition. Ifiok Otung.
© 2021 John Wiley & Sons Ltd. Published 2021 by John Wiley & Sons Ltd.
Companion Website: www.wiley.com/go/otung

communication systems. Each component and the processes that it contributes to the overall performance of the system are then briefly discussed in a nonmathematical way.

Different types and classifications of communication systems are discussed. Digital and analogue communication systems are compared, and analogue baseband systems are discussed in some detail. The features of digital baseband transmission are introduced. We show that modulated systems are essential to exploit the radio spectrum and the optical fibre medium and discuss the features of this class of systems. Our discussion includes brief references to some of the main modern applications of telecommunications, namely television systems, communication networks, telephony, satellite communications, optical fibre communication systems, and mobile communication systems. We also discuss and compare the different switching technologies of space switching, time switching, circuit switching, connection-oriented (CO) packet switching, and connectionless packet switching (CL) and explain why CL has become the dominant switching technology of the twenty-first century.

On completing this chapter, you will be well equipped to make informed decisions regarding the suitability of different types of system blocks (including transmission media) and various classes of communication systems for different applications. You will also have a clear overall picture of telecommunication and a good foundation on which you can build a more detailed knowledge of the subject. I hope that working through this chapter will inspire you towards a career in telecommunication or a very enjoyable further study of this exciting field.

1.2 Nonelectrical Telecommunication

This book is concerned with modern telecommunication or *communication over a distance*, which relies principally on electrical means to convey information from one point S, the sending end, to another point R, the receiving end. After this section, we will use *telecommunication*, *communication systems*, and other similar terms to refer exclusively to these modern technologies. However, before the advent in 1837 of telegraphy, the forerunner of modern telecommunications, many forms of nonelectrical 'telecommunications' existed. We will briefly introduce these nonelectrical communication systems and discuss their demerits, before turning our attention to various types and classifications of modern communication systems.

1.2.1 Verbal Nonelectrical Telecommunication

The most basic nonelectrical telecommunication was verbal, in which, for example, a town crier used a combination of a gong and their voice to broadcast information in an African village. Trumpets were also blown to summon to war, call a truce, or announce victory, depending on the distinctive sounds made. The transmitter in this case is the human vocal system and other suitable instrumental sound sources such as the gong, drum, and trumpet. The signal is a pressure wave called an *acoustic signal*, the transmission medium is air, and the receiver is every human ear within range of hearing. This form of 'telecommunication' is fraught with problems, which include the following.

- *Interference*: the ear also receives other *unwanted* sound signals in the environment, which serves as a common transmission medium. These sounds, which may be artificial or natural background noise, or the intelligible communication of other people, interfere with and hence corrupt the wanted signal. In the town crier example, an African woman would have to hush her chattering children, bleating goats, and barking dogs, or move away from the noisy zone in order to minimise interference with the town crier's message.
- *Nuisance*: the signal is received by those who do not want it. If you had to use this verbal means to 'telecommunicate' with a friend a few blocks away, your neighbours would not be pleased.
- *Huge losses*: the signal suffers a lot of *attenuation*, or reduction in the *amplitude* of the pressure wave as it is reflected at material boundaries and as it spreads out with distance over a wider wave front. For this reason, it is difficult to hear the town crier from within a closed room or from afar.

- *Limited range*: communication can only take place over small distances (i.e. small separations between S and R). To overcome the above losses, sound signals of high SPL (sound pressure level) must be used to extend communication range. However, this increases the nuisance effect and endangers the hearing of those close to the sound source. Besides, the human vocal system and other practical sound sources are severely limited in the SPL that they can generate.
- *Masking effect*: the wanted signal can be easily *masked* by louder signals in the medium. This makes hearing (i.e. *detection*) of the wanted signal impossible. The threshold of hearing increases with ambient noise. This means that the ear–brain system will be completely oblivious to the presence of one sound if there is another louder sound of about the same frequency.
- *Lack of privacy*: privacy is an important requirement that this means of communication is incapable of providing. Everyone (within range) hears your message and is potentially able to understand and make use of it. Therefore, this means of communication is only suitable for the broadcast of public information and would not be suitable for the communication of private, military, commercial, or classified information.
- *Delay*: even if we could overcome all the problems discussed above, *propagation delay*, the time it takes for the signal to travel from S to R, would still make this type of telecommunication unacceptable. Sound travels at a speed $v = 330$ m/s in standard air. Over very short distances, it seems as though we hear at the same instant as the sound is produced. However, imagine that S and R are separated by, say, $d = 2376$ km, a realistic international and in some cases national distance. The propagation delay is

$$t = \frac{d}{v} = \frac{2376000\ \text{m}}{330\ \text{m/s}} = 7200\ \text{s} = 2\ \text{h}$$

Thus, if you made an acoustic 'phone call' over this distance, the person would hear each utterance a staggering two hours after it was made. And if the person said, 'Pardon?' you would complete your call and hang up without knowing they wanted you to say it again. Real-time interactive communication would only be possible over short distances. This is in sharp contrast to the case of modern telecommunications using *electromagnetic waves*, which travel at the speed of light ($v \equiv c = 3 \times 10^8$ m/s). The propagation delay for the distance $d = 2376$ km is $t = 8$ ms. Barring other delays, you would hear each other practically instantaneously.

1.2.2 Visual Nonelectrical Telecommunication

Various methods of nonverbal telecommunication were developed as a vast improvement on the basic technique discussed above. Here, the information to be conveyed is coded in the form of visually observable events. The transmitter generates these events. Light signals convey the information through a transmission medium that is ideally a vacuum but is in practice air. The receiver is a human eye within visibility range along an unobstructed line-of-sight path. The problems of interference, nuisance, propagation delay, and masking by other signals are negligible compared to the verbal technique. Furthermore, the communication range is extended, especially in clear non-foggy weather. However, the receiver must have an unobstructed view of the signalling event and the major problem of lack of privacy remains. Visual nonelectrical telecommunication systems were identical in transmission medium and receiver, differing only in the type of transmitter. The most common techniques are now discussed.

1.2.2.1 Flags, Smoke, and Bonfires

The practice of raising a highly visible object, or performing complex coded motions with the object, has had universal application in warfare and ordinary social interactions over the centuries. Hand signals remain an indispensable means of communication in modern society. For example, hand signals are used by traffic wardens to control traffic. Hoisting a red flag or red light is still used to warn of danger, and raising a white flag in warfare communicates a message of surrender over distances that would be difficult to reach by verbal means.

An interesting record of a military application of this method of visual nonelectrical telecommunication around 1405 BC is given in the Bible. Hebrew soldiers divide into two parties to attack a city. One party waits in ambush outside the city gates while the other attacks the city and fakes defeat in order to draw the city guards away in pursuit. At a good distance from the city, the fleeing party raises a javelin from a vantage point, thus *signalling* to the ambush party to take the now defenceless city. This party takes the city and quickly starts a bonfire, which 'telecommunicates' their success to the retreating party and *signals* the beginning of a bidirectional attack on the guards, who are in shock at the sight of their burning city.

1.2.2.2 Heliography

Heliography was a system of communication developed in the nineteenth century in which rays of the sun were directed using movable mirrors on to distant points. Some links for heliograph communication were set up and the technique was reliably used, for example in 1880 by General Stewart to give a battle account over some distance. Reading a *heliogram* (a message sent by heliography) was said to cause eye fatigue.

A technique for transmitting information using variations in the pattern of light had been developed by the Greeks as far back as the second century BC. Different combinations and positions of torchlight signals were employed to represent the letters of the Greek alphabet. In this way messages could be coded and transmitted.

1.2.2.3 Semaphore

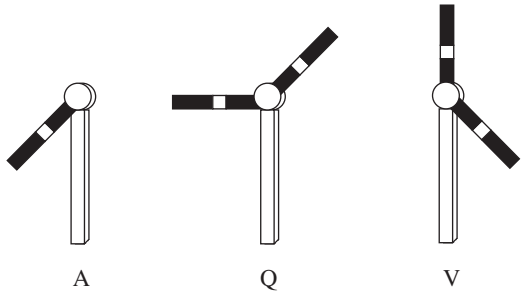
A semaphore consists of an upright post with one or more arms moving in a vertical plane. Beginning in the eighteenth century, different positions of the arms or flags were used to represent letters. Table 1.1 gives the conventional alphanumeric semaphore codes. The semaphore codes for letters A, Q, and V are illustrated in Figure 1.1. Note that some of the letters of the alphabet are shared with numerals 0–9. A numerical message is distinguished by being preceded with the code (0°, 45°). Code AR is used to indicate the end of signal and code *R* to acknowledge reception. In Table 1.1, the arm positions are given in degrees measured clockwise from vertical.

Table 1.1 Semaphore codes.

Symbol		Positions of semaphore flags	Symbol	Positions of semaphore flags
A	1	225°	N	(135°, 225°)
B	2	270°	O	(270°, 315°)
C	3	315°	P	(270°, 0°)
D	4	0°	Q	(270°, 45°)
E	5	45°	R	(270°, 90°)
F	6	90°	S	(270°, 135°)
G	7	135°	T	(315°, 0°)
H	8	(225°, 270°)	U	(315°, 45°)
I	9	(225°, 315°)	V	(0°, 135°)
J		(0°, 90°)	W	(45°, 90°)
K	0 (zero)	(0°, 225°)	X	(45°, 135°)
L		(45°, 225°)	Y	(90°, 315°)
M		(90°, 225°)	Z	(90°, 135°)

Angle is measured clockwise from vertical.

Figure 1.1 Semaphore codes for letters A, Q, and V.



Some governments built large semaphore towers on hilltops along strategic routes. For example, during the Napoleonic wars (1792–1815), a relay of semaphore towers was built in England from London to Portsmouth. By the year 1886, semaphore had been almost universally adopted for signalling on railways. Semaphore is still used today at sea by warships preferring radio silence and for mechanical railway signalling.

1.2.2.4 Demerits of Visual Nonelectrical Telecommunication

The major disadvantages of visual nonelectrical telecommunication include the following.

- *Low signalling speed*: the rates of signal generation at the transmitter and detection at the receiver are extremely low by modern standards. For example, if the torchlight or semaphore flag patterns are manipulated quickly enough to complete the representation of one letter every second, we have a maximum data rate of one character (or *byte*) per second. In modern systems, a *modem* that operates at 10 000 times this speed is considered slow. Note, however, that propagation delay is negligible, and the signal is received almost at the same instant as it is generated at the transmitter.
- *Manual operation*: a trained signaller is required at the transmitter to code the information and at the distant receiver to view and decode the signal patterns as they are transmitted. Automation of both transmission and reception is impractical.
- *Limited range*: the maximum separation that allows visibility between transmission and reception points may range from less than 50 m in thick fog to a few miles in obstruction-free terrain under very clear weather. However, this range is very low compared to distances involved in modern communications.
- *Lack of privacy*: even with the use of encrypted codes, the communication process is carried out in full view of all within the communication range.
- *Limited application*: a combination of the above factors means that visual nonelectrical telecommunication remains a system that can only accommodate a very few users (*low capacity*) for very limited applications. This is grossly inadequate to meet modern requirements of broadband communication.

1.3 Modern Telecommunication

Modern telecommunication began in 1837 with the invention of telegraphy. Sir Charles Wheatstone (1802–1875) and Sir William Fothergill Cooke (1806–1879) built the world's first commercial telegraph system in England in 1837. They laid the telegraph lines, consisting of six iron wires, along a 2.4 km railway track from Euston to Camden Town stations. Their invention was first publicly demonstrated on 24th July 1837 and described as a 'most extraordinary apparatus' in an advertisement. The receiver at the receiving station consisted of five galvanometer needles arranged in a row across the face of a grid of letters, as shown in Figure 1.2. Each needle (G1, G2, …, G5) could be deflected to the left or right by current flowing in the corresponding wire (W1, W2, …, W5), not shown in the diagram. A sixth wire, W6, was needed to provide a common return path for each of the five circuits. To

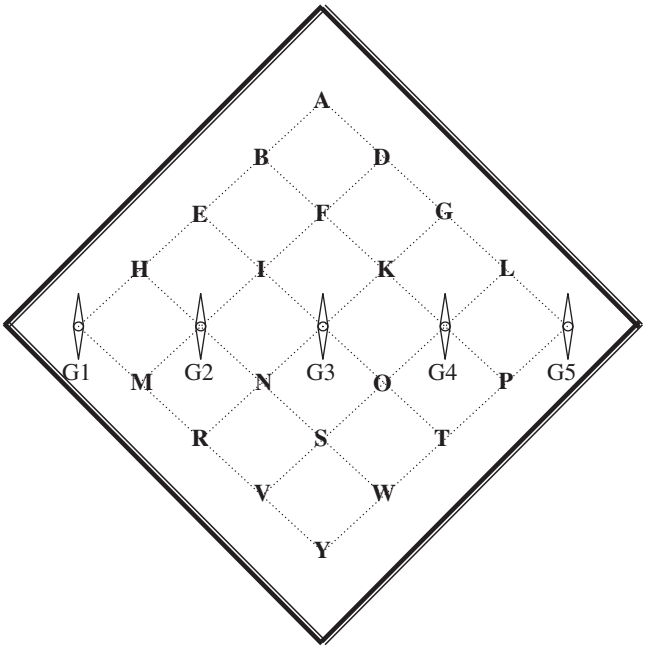


Figure 1.2 Wheatstone-Cooke 5-needle electric telegraph display grid.

transmit a letter, current flow was sent from the transmitter down two wires causing two of the needles to deflect in opposite directions and point to the right letter. For example, to transmit the letter A, current flow is established in wires W1 and W5 so that G1 deflects clockwise and G5 counterclockwise from vertical. But to transmit the letter Y, current flow is established in the opposite direction in the two wires so that G1 deflects counterclockwise and G5 clockwise from vertical.

This five-needle telegraph system was simple enough to allow unskilled operation, but it required six wires to make the connection between transmitting and receiving stations and only 20 letters could be transmitted, which made it necessary to transmit some words with a modified spelling. It was soon replaced in 1838 by a cheaper two-needle telegraph, which was also eventually superseded in 1845 by a single-needle system. The single-needle telegraph coded each letter using a unique combination of movements of one needle to the left and right. By arranging for the needle to strike small metal pipes that emitted sounds of different pitches, each letter was heard as a unique melody at the receiver.

At about the same time that Wheatstone and Cooke developed their needle-deflecting telegraph, two Americans, Samuel Finley Breese Morse (1791–1872) and Alfred Vail (1807–1859), devised a more efficient telegraph system, which was eventually adopted all over the world. Morse and Vail demonstrated their system on 24th May 1844 by sending the message 'WHAT HATH GOD WROUGHT' over the first telegraph link in America from the Capitol at Washington to Mount Clare Depot in Baltimore, a distance of 64.5 km. The Morse–Vail telegraph consisted of a key or switch that was pressed to complete a circuit containing a battery connected by a single pair of wires to a distant sounder. The key and battery constituted the transmitter, the pair of wires the transmission medium, and the sounder the receiver. Transmission of textual information was achieved using 39 characters (26 uppercase letters of the alphabet, 10 numerals, and 3 punctuation marks: comma, full stop, and question mark).

In what follows, we introduce in chronological order the various coding schemes which were developed to represent characters and then we discuss some of the most significant developments in telecommunication since the advent of telegraphy.

1.3.1 Developments in Character Codes

1.3.1.1 Morse Code

Each character in the Morse–Vail telegraph was represented by a pattern of dots (•) and dashes (—) called *Morse code*, invented in 1838. The innovative and enduring principle behind the Morse code, which has sustained its application into the twenty-first century (e.g. in amateur radio), is that characters that occur frequently are represented using shorter codes, whereas less frequent characters are assigned longer codes. For example, the Morse code for letter E (the most frequent in the English alphabet) is a single dot •; the code for T (the next most frequent letter) is a single dash —; but the code for letter Q (one of the least frequent letters) is — — • —. Morse code has been extended over the years to include other characters and control signals. An international version is shown in Table A.1 of Appendix A.

In actual transmission using Morse code, a dot is a voltage pulse of duration one unit (say T) generated by pressing a key to close the transmitter circuit for this duration, a dash is a voltage pulse of duration $3T$, and dots and dashes within a Morse code for a character are separated by a silence interval of duration T. Characters are separated by a silence of duration $3T$, and words by $7T$. At the receiver the incoming message is identified by ear, using the sounds produced. For example, the letter E sounds like 'dit', the letter T sounds like 'dah', and the letter Q sounds like 'dahdahdidah'. The value of T used depends on the transmission speed (words per minute) at which the receiving operator can copy accurately. The spacing between characters and the spacing between words can be increased above $3T$ and $7T$, respectively, to accommodate a slow operator.

In the original Morse–Vail telegraph system the incoming message was recorded on a moving strip of paper by using an electromagnet to move a marker which sketched the code waveform comprising a sequence of short (dot) and long (dash) pulses. As an example, the statement (excluding quotes) 'THANK GOD' would be conveyed by the voltage pulse sequence shown in Figure 1.3. The system, however, developed into receiving by ear from the clicks of the receiver. This was more convenient and faster for a trained operator who could handle 40–50 words per minute. By 1914, automatic operation involving tape-reader machines and *multiplexing* (transmitting more than one message at a time over a single wire) allowed telegraph transmission speeds of 400 words per minute to be achieved.

The development of automated teleprinter communication technology in the 1920s marked the beginning of the end of manual *landline* (i.e. wire-connected) Morse–Vail telegraphy. Its commercial use was discontinued in 1932 by the British Post Office, and in the 1960s on railroads in America and Australia. Landline application has, however, continued on railroads in some developing countries and amongst a few die-hard enthusiasts in the US. Furthermore, Morse code continues to be used in *radio telegraphy* by radio amateurs worldwide and by a few ships at sea as a low-cost alternative to satellite communications.

1.3.1.2 Baudot Code

Efforts to improve the telegraph system led to the invention in 1871 of the first device for *time division multiplexing* (TDM) by the French engineer Jean-Maurice-Émile Baudot (1845–1903). The multiplexer consisted of a copper

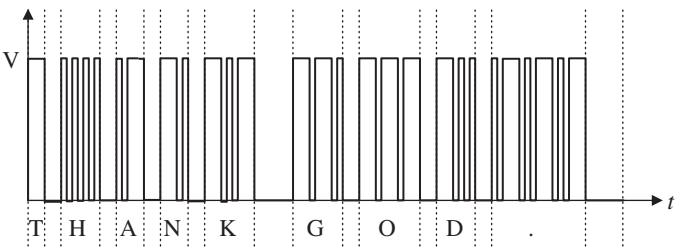


Figure 1.3 Morse code's sequence of voltage pulses for THANK GOD.

ring divided into several equal sectors (say N). Each sector had five contacts which could be opened or closed, giving 32 ($= 2^5$) possible combinations for coding characters. A brush arm rotated in a circle around the copper ring and sequentially picked up the code combinations from each sector. Connecting this arm to a transmission line allowed N messages to be simultaneously sent.

The existing *variable-width* Morse code was not suitable for the automatic transmission and reception operation of Baudot's multiplexer, so he devised a new fixed-length code which was patented in 1874. The *Baudot code* was modified in 1901 by Donald Murray (1865–1945) – a New Zealand born farmer, journalist, telegraph engineer, and entrepreneur. After further modification in 1932 by the International Telegraph and Telephone Consultative Committee (CCITT) – now International Telecommunications Union, Telecommunications Sector (ITU-T), the *Baudot–Murray* code became known as the International Telegraph Alphabet No. 2 (ITA-2).

Table A.2 of Appendix A lists the ITA-2 code. Each character is represented using five *symbols* drawn from a binary alphabet, namely a *mark* (voltage pulse) and a *space* (no voltage pulse), corresponding to *b*inary dig*its* or *bits* 1 and 0, respectively. With *fixed-width* codes like this, there is no need for the short pause (between characters) and long pause (between words), as in the Morse code. Rather a *codeword* 'space–space–mark–space–space' or 00100 is used to separate words. Characters need no separation between them, except that before actual transmission each character is framed by preceding the 5-bit character code with a start bit (usually a space) and terminating with a stop bit (usually a mark), giving seven transmitted bits per character.

Using 5-bit codes, only $2^5 = 32$ different characters can be represented. This is insufficient to cover 52 uppercase and lowercase letters of the alphabet, 10 numbers, and various punctuation and control characters. Raising the number of bits above five was not an option, owing to hardware constraints in the electromechanical technology employed at the time. To work around this 5-bit constraint, lowercase letters were not represented and two 'shift codes' were used. More specifically, a letter shift code 11111 is transmitted to indicate that all subsequent characters are uppercase letters until a figure shift code 11011 is encountered, which causes all following characters to be interpreted as numbers or punctuation marks. Baudot can therefore be credited with inventing the use of an *escape code* – still essential in modern computer systems – which is a code that changes the way the system interprets subsequent codes. Thus, 26 of the 32 codes were used to represent uppercase letters; 4 were used to designate blank, space, carriage return and line feed; and the remaining 2 were used for the letter and figure shift codes. Issuing a figure shift code allowed the 26 codes to be reused for representing numbers and punctuation marks, as shown in Table A.2.

The ITA-2 code was adopted in teleprinters and the telex (*tel*egraph *ex*change) network operating at a transmission speed of 50 symbols per second. By the middle of the twentieth century it had replaced Morse as the most widely used telegraph code and was still used in telex networks operating at speeds $\leq$300 symbols/second until the turn of the century. In other data communication systems Baudot code was replaced by American Standard Code for Information Interchange (ASCII) and *extended binary coded decimal interchange code* (EBCDIC) codes, which could represent a larger number of characters. The memory of Baudot, however, lives on today, transmission speed in symbols/second being called *baud* (Bd) in his honour.

1.3.1.3 Hollerith Code

In 1881, Herman Hollerith (1860–1929), a young American statistician working for the US Census Bureau, devised a means of recording census data on punched card for automatic machine reading. The punched card was later standardised to 186 mm × 82 mm in size, with a clipped top left-hand corner and 12 rows and 80 columns. A column was used to represent one character by punching holes in a selection of the 12 row positions available. One card could hold 80 characters. Recorded data were read electronically by moving the punched card between brass pins in a 'tabulating machine'. At positions with holes, the pins on opposite sides of the card made contact, thereby completing an electrical circuit and registering a value.

The pattern of holes for each character was defined by the *Hollerith code*. With 12 binary digits – i.e. 12 hole-or-no-hole positions – a total of $2^{12} = 4096$ characters can be represented. However, only 69 characters (52

upper and lowercase letters, 10 numerals, and 7 punctuation marks and symbols) were defined. The use of 12 rows (rather than, say, 7) allowed each character to be uniquely represented using only a few holes. Bear in mind that the holes were initially manually punched. Besides, many holes in one row would have made the card more liable to tear. Thus, all 10 numerals and 2 letters were represented with just one hole, the remaining letters were represented using two holes, and two or more holes were used to code the less frequent punctuation marks and symbols.

Hollerith's punch card technology was first employed in 1887 for calculating mortality statistics, but it gained popularity when it was used for the 1890 US census. Thanks to the new technology, it took just six weeks to analyse the 1890 census data more thoroughly than the previous 1880 census, which took seven years of toil by hand. To market his highly successful invention Hollerith founded the Tabulating Machine Company in 1896, which in 1911 merged with two other machine manufacturers to form the Computer-Tabulating-Recording Company. In 1924, the company was renamed International Business Machines Corporation (IBM) and it dominated the world of computing until the 1970s.

The Hollerith code continued to be used for data representation until the 1960s, when IBM developed a new character code, EBCDIC, for its mainframe computers. But punched cards remained the primary means of data input to computers until the early 1970s, and the technology continued to be used in some establishments up to the early 1980s. Many today regard Herman Hollerith as the father of *information processing*.

1.3.1.4 EBCDIC Code

The *extended binary coded decimal interchange code* (EBCDIC – pronounced ebb-sea-dick) is a proprietary 8-bit character code developed in the 1960s by IBM for use on its computing machines. It provides $2^8 = 256$ codewords for the representation of 256 different characters including a wide range of control characters. Table A.3 gives a list of EBCDIC codewords. Note the following features.

- The first 64 codewords (hexadecimal 00 → 3F) specify control characters.
- The codewords for uppercase and corresponding lowercase letters differ in only one bit position (bit 7). For example, the codewords for letters 'm' and 'M' are 1$\underline{0}$010100 and 1$\underline{1}$010100, respectively.
- The 62 alpha-numeric characters (52 uppercase and lowercase letters and 10 numbers) occupy 7 noncontiguous blocks in the code table. Their codewords are restricted to those whose lowest four bits (i.e. first hexadecimal digit) have decimal values in the range 0–9. It is for this reason that the term *binary-coded-decimal* (BCD) features in the name of the coding scheme. A drawback of this noncontiguous arrangement of letters is that it makes computer programs for manipulating textual information more complex.

The EBCDIC character code has always been perceived as an IBM ploy for locking in its customers. Other computer manufacturers and the rest of the world adopted the ASCII code in the 1960s. The rapid growth in personal computers (PCs) starting in the late 1970s brought IBM's domination of the computer industry to an end. EBCDIC's use has therefore since become confined to a vanishing minority of historical computer systems. It is worth emphasising that EBCDIC is inherently incompatible with ASCII. A PC requires code conversion to be able to communicate with an IBM mainframe computer that uses EBCDIC, and a PC software program that manipulates letters of the alphabet by the relative decimal values of their ASCII codewords would not run properly on an EBCDIC system.

1.3.1.5 ASCII Code

The ASCII code was first developed in 1963, and updated in 1968 as ANSI X3.4, by the American National Standards Institute (ANSI) – formerly the American Standards Association. ASCII uses seven bits, which allows representation of $2^7 = 128$ characters covering 26 lowercase letters, 26 uppercase letters, 10 numbers, punctuation marks and mathematical symbols, and a wide range of control characters.

Table A.4 of Appendix A gives a complete listing of 7-bit ASCII codes. To use this table, note that the number in the top left-hand corner of each character entry is the decimal value of the code. The first and second columns

of the table give, respectively, the first hexadecimal (hex) digit and the least significant four bits $b_4b_3b_2b_1$ of the ASCII code for the character in the corresponding row. Similarly, the first and second rows give the second hex digit and the most significant three bits $b_7b_6b_5$ of the code. Thus, the ASCII code for the character M is decimal 77, hex 4D, or binary 1001101. We may write these, respectively, as 77_{10}, $4D_{16}$ or 1001101_2.

The following features of the ASCII coding scheme may be observed in Table A.4.

- The three most significant bits indicate the type of character. For example, the C0 control characters are in the first two columns $b_7b_6b_5 = 000$ and 001; uppercase letters of the alphabet are in the fifth and sixth columns $b_7b_6b_5 = 100$ and 101; and lowercase letters are in the seventh and eighth columns $b_7b_6b_5 = 110$ and 111.
- The allocation of codewords to numbers (0–9) and letters of the alphabet (A–Z and a–z) follows a binary progression. For example, the codewords for the numbers 0, 1, and 2 are respectively 0110000, 0110001, and 0110010, the last four bits being the binary equivalent of each decimal number. The codewords for the letters R, S, and T are 1010010, 1010011, and 1010100, respectively. Mathematical operations can therefore be performed on the codewords that represent numbers and alphabetisation can be achieved through binary mathematical operations on the codewords for letters. For this reason, ASCII codes are described as *computable* codes.
- For ease of generation on a keyboard, lowercase and uppercase letters differ only at the sixth bit position (b_6). For example, the ASCII codes for letter 'A' and 'a' are 1*0*0000 and 1*1*0000, respectively. This bit position (b_6) is changed on the keyboard by holding down the shift key.

ASCII was an instant success and became a de facto international standard. However, as you can see from Table A.4, the code was inadequate even to cover international characters such as é, ü, ô, etc. based on the Latin alphabet. It was therefore necessary to internationalise ASCII. This task was undertaken by the International Organization for Standardization (ISO) leading to an international standard ISO 646 in 1972 and a subsequent revision in 1991. The ASCII code table shown in Table A.4 was designated the international reference version (IRV) and identified as ISO-646-IRV, which is synonymous with the US ASCII version ISO-646-US. Provision was then made for national versions to be created by substituting other graphic characters in place of the least needed 10 characters, namely @ [\ ] ^ ‘ { | } ~. For example, the German version (ISO-646-DE) is identical to Table A.4 except that these 10 characters are replaced by the characters § Ä Ö Ü ^ ‘ ä ö ü ß, respectively. Similarly, the French version (ISO-646-FR) has the characters à ° ç § ^ ‘ é ù è ¨, respectively; the Italian version (ISO-646-IT) has the characters § ° ç é ^ù à ò è ì, respectively; and so on. Furthermore, allowance was made for versions to be created in which the IRV was altered in two other character entries by substituting the pound sign £ for the number sign # and/or substituting the currency sign ¤ for the dollar sign $.

The process of internationalisation of the 7-bit ASCII code was extended to include languages that are not based on the Latin alphabet, such as Arabic, Greek, and Hebrew. As a result, up to 180 ASCII-based character code versions were eventually registered. The main drawback of a multiplicity of versions is that the character represented by a given code number (say decimal 91) is not unique but depends on the ASCII version of the source system. This poses a nuisance in information interchange between computers using dissimilar versions. National variants of ISO-646 are now obsolete having been replaced by less problematic schemes.

1.3.1.6 ISO 8859 Code

A more interoperable character code scheme was developed by the European Computer Manufacturers Association (ECMA) in the mid-1980s by extending the codeword from seven to eight bits. This allowed up to 256 characters to be represented. The new scheme was endorsed by ISO as ISO-8859. To maintain compatibility with the well-established ISO-646-IRV the 8-bit code table was designed as shown in Figure 1.4, with the left half (most significant bit, or MSB, = 0) identical to the ISO-646-IRV table and the right half (MSB = 1) providing an extra 32 control characters and 96 graphic characters. A US-ASCII character with code $b_7b_6b_5b_4b_3b_2b_1$ was therefore assigned an 8-bit code $0b_7b_6b_5b_4b_3b_2b_1$ under the new scheme.

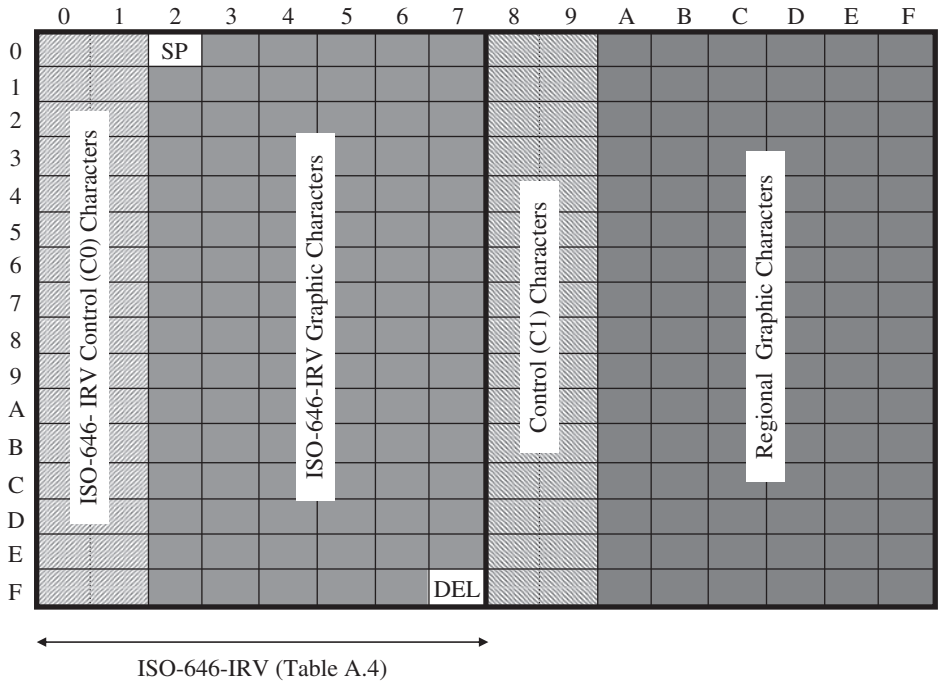


Figure 1.4 Layout of ISO 8859 (8-bit) character code table. The column label is the second hex digit (or most significant 4 bits) and the row label is the first hex digit.

Since there are more than 96 foreign graphic characters, it was necessary to develop multiple character sets, one set for a specified geographical region. This gave rise to more than 10 sets of 8-bit codes, which differed only in the area labelled *Regional Graphic Characters* in Figure 1.4. Escape sequences were defined for switching between character sets – analogous to changing typeheads on a typewriter. Thus, there was ISO-8859-1 or Latin Alphabet No. 1 for the needs of Western Europe. This is listed in Table A.5 of Appendix A, where the number in the top left-hand corner of each cell is the hex code of the character. So, for example, the 8-bit code for the ligature æ is hex E6 or binary 11100110 and the pound sign £ has code 10100011 (equivalent to decimal 163). Other character sets in the scheme included ISO-8859-2, which covered Eastern European languages (Albanian, Hungarian, Romanian, and Slavic), ISO-8859-3 to -8, which respectively covered Southern Europe, Northern Europe, Cyrillic (i.e. Bulgarian, Macedonian, Russian, Serbian, Ukrainian), Arabic, Greek, and Hebrew, and so on.

The 8-bit ISO-8859 scheme was an improvement on the 7-bit ASCII. It reduced the number of character sets involved in global information exchange to a smaller number of regional variants. However, it still involved switching between character sets and did not cover all the world's languages, most notably the East Asian languages.

1.3.1.7 Unicode

In 1987, three engineers from Apple Computer and Xerox – Joe Becker, Lee Collins, and Mark Davis – started a project to develop the ultimate character code: one that would be universal, uniform, and unique. Becker coined the term *Unicode* to describe this scheme, which aimed to cover all of the world's languages, assign a unique character to each code number or bit sequence, employ fixed-width codes (i.e. same number of bits in every codeword), and wherever possible maintain some compatibility with existing standards (e.g. ISO-646-IRV and ISO-8859-1). This was a mammoth task the success of which required the support of influential global computer companies such as Microsoft Corporation and Sun Microsystems.

The group did gain support and grew into an incorporated consortium in 1991. And in 1992 Unicode merged with and subsumed IS0/IEC 10646, a multilingual character code independently developed by ISO and at that time incompatible with Unicode. Version 13.0 of Unicode was released on 10th March 2020 as ISO/IEC 10646:2020, containing a staggering 143 859 characters covering almost all modern and historical writing systems in the world. Unicode has ample room for all graphic characters, format characters, control characters, and user-defined characters as well as future growth. The graphic characters include a wide range of symbols (technical, optical character recognition, Braille pattern, dingbats, geometric shapes, emoji, etc.) as well as ideographic characters used in *C*hina, *J*apan, *K*orea (CJK), Taiwan, Vietnam, and Singapore. Duplicate encoding is avoided by assigning a single code to equivalent characters irrespective of usage or language of occurrence. There are ongoing efforts to identify and define additional characters for inclusion in future versions of the standard, subject to a policy of not altering the codes already assigned to characters in previous versions, and not encoding logos, graphics, and font variants, nor 'characters' deemed idiosyncratic, novel, private-use, or rarely exchanged.

To fully cater for all the world's written languages, Unicode makes provision for 1 114 112 characters using integers or *code points* in the range $0 \rightarrow 10FFFF_{16}$. A code point is referred to by its numeric hex value prefixed by U+. For compatibility with US-ASCII and ISO-8859-1, the first 256 code points (i.e. 0000 → 00FF) are assigned as in Tables A.4 and A.5 of Appendix A. Each *encoded character* has a unique official Unicode name. For example, the character G is assigned code point U + 0047 and named *LATIN CAPITAL LETTER G*, and the character Ç, assigned code point U + 00C7, is named *LATIN CAPITAL LETTER C WITH CEDILLA*.

It is convenient to think of the 1 114 112 Unicode code points as divided into a total of 17 planes, each containing 2^{16} code points. Almost all common-use characters of all the world's languages are contained in the 65 536 code points (0000 → FFFF) available in the basic multilingual plane (BMP) or Plane 0. These Unicode characters can be represented in computer systems using fixed-width 16-bit codewords, unlike the seven bits of ASCII and the eight bits of ISO-8859. However, to cover all the characters in all 17 planes using unique (non-overlapping) codewords, the Unicode Standard specifies the following three distinct encoding forms, called *Unicode Transformation Formats* (UTF):

- *UTF-32*: each character is expressed using one 32-bit code unit which has the same value as the code point for the character. This is a fixed-width 32-bit character format, which quadruples the size of text files compared to the 8-bit ISO-8859. But, importantly, all characters of the world are uniquely covered using a single code unit access. UTF-32 is the preferred encoding form on some Unix platforms.
- *UTF-16*: most common-use characters of all the modern scripts of the world lie in the BMP (U + 0000 → U + FFFF). These characters can be represented by a single 16-bit code unit. This was the original design of the Unicode Standard, to be a fixed-width 16-bit character code. However, as the Standard evolved it became clear that not all characters could be fitted into code points ≤ U + FFFF. Thus, the code points' range was extended to include the supplementary planes (1 → 16). An area of the BMP that was unallocated up until the extension was then set aside for surrogates, and a pair of 16-bit code units (called surrogate pair) in this area was used to represent (i.e. point to) a code point in a supplementary plane. Note that neither of the code units of a surrogate pair is used to represent a character in the BMP. Thus UTF-16 is a variable-width encoding form that uniquely codes characters by using one 16-bit code unit for each of the more frequent BMP characters, and two 16-bit code units for each of the much less frequent characters of the supplementary planes. A code unit conversion is required to convert a surrogate pair to the Unicode code point (> U + FFFF) and hence the supplementary character represented by the pair. UTF-16 typically requires half the memory size of UTF-32 and is the preferred encoding form in most implementations of Unicode other than on Unix platforms.
- *UTF-8*: this is the preferred encoding form for the Internet. It was prescribed to provide transparency of Unicode implementation in ASCII-based systems. Each Unicode code point in the range U + 0000 → U + 007F is represented as a single byte in the range 00 → 7F having the same value as the corresponding ASCII character. Then, using 8-bit code units or bytes of value >7F, the remaining code points are represented by two bytes (for U + 0080 → U + 07FF), three bytes (for U + 0800 → U + FFFF), and four bytes (for U + 10 000 → U + 10FFFF).

Thus UTF-8 is also a variable-width encoding form that uses codewords ranging in width from one byte for the first 128 (ASCII) characters to four bytes for the supplementary Unicode characters. The UTF-8 form works automatically as follows: if a byte – let's call it Byte 1 – starts with bit 0 then it represents a character; or else if Byte 1 starts with bits 110 then the character is represented by the concatenation of Byte 1 and the next byte (i.e. Byte 2); or else if Byte 1 starts with bits 1110 then the character is represented by the concatenation of Byte 1 and the next two bytes (i.e. Byte 2 and Byte 3); or else if Byte 1 starts with bits 11110 then the character is represented by the concatenation of Byte 1 and the next three bytes (i.e. Byte 2, Byte 3, and Byte 4). Thus, a 1-byte codeword has prefix 0; a 2-byte codeword has prefix 110; a 3-byte codeword has prefix 1110; and a 4-byte codeword has prefix 11110. Furthermore, all Bytes 2, 3, and 4 have prefix 10, and this distinctively identifies them as *not* Byte 1, which prevents the error that would otherwise have been possible if one, for example, started reading a multibyte codeword from other than its first byte. Therefore, picking any byte at random inside a UTF-8 file, one can correctly read the character as follows: if the byte's prefix is 0 then that's the entire 1-byte character; or else the byte is part of a multibyte codeword formed by concatenating byte(s) to its left and/or right until the longest possible multibyte ($\leq$4 bytes) codeword is formed that has a first byte starting with 11 and a last byte starting with 10.

Unicode has been described as the ultimate character code for the twenty-first century. It has gained worldwide acceptance, including in East Asia, where it has the capacity to fully cater for all the ideographic characters in JIS (Japan Industrial Standards) and other East Asian standards (e.g. Chinese GB 2312-1980, Korean KS C 5601-1992, and Taiwanese Big-5 and CNS 11643-1992). Unicode may not be as efficient in handling a language as its regional character code set (e.g. ISO-8859-8 for Hebrew), but Unicode's multilingual capability is a compelling advantage in an era of globalisation.

1.3.2 Developments in Services

Data transmission by telegraphy was soon followed by voice transmission with the invention of the telephone by Alexander Graham Bell in 1876. Initially, voice communication was only possible over short distances between two people using permanently linked telephone sets. The novelty of switching made it possible for two subscriber terminals connected to the same switching office or local exchange to be manually linked by an operator. In 1897, an undertaker, A. B. Strowger, developed a step-by-step electromechanical switch that made automatic switching possible. The technology was thus in place for the communication of voice and data over wire line connections, although the speed and quality of transmission left much to be desired by today's standards.

A wide range of digital communication services, too many to discuss here, have been developed since the 1830s and especially during the last five decades. Here we briefly introduce a few of the most significant services in roughly chronological order. Some services, such as the telex and facsimile, were largely limited to business applications, whereas others, such as the telegram and email, made significant penetration into both home and business.

1.3.2.1 Telegram

Digital communication services began with the *telegram* in the late 1830s when members of the public sent messages on the Wheatstone–Cooke telegraph from Paddington railway station to Slough at one shilling per message – a large sum in those days. With the widespread use of the Morse–Vail telegraph all over the world, the telegram became a well-established communication service starting from 1845.

The military used the telegram to communicate intelligence, deploy troops, and disseminate battle news; railway companies used it to regulate trains, and this significantly extended railway capacity; industry used it for a wide range of business communication (e.g. transmitting racing results, stock prices, urgent memos, etc.); and the police quickly adopted it in their fight against crime. A much-publicised early application was the arrest of John Tawell

for the murder of his mistress on 1st January 1845. Tawell fled the crime scene and got on a train departing Slough to London where he hoped to make good his escape. However, his failure to reckon with the power of the new telecommunication invention proved his undoing. A telegram giving a description of the murder suspect beat him to the London station and the police were waiting when his train pulled in. Tawell was arrested, later tried, found guilty of murder, and hanged.

The telegram was also a great success with the general public. Messages – charged according to length – were transmitted via telegraph, received and printed on paper at the destination telegraph station, and delivered by hand to the recipient's address. By 1874, the United Kingdom had over 650 000 miles of telegraph wire, and more than 20 000 towns and villages were part of the UK network. However, by the second half of the twentieth century, the telegram service was forced into steady decline due to developments in other communication services – both analogue (e.g. telephony) and digital. The UK inland telegram service was eventually discontinued in 1982.

1.3.2.2 Telex

The *telex*, started in 1932 by the British Post Office, was another major digital communication service to be introduced. A teleprinter (or teletypewriter) in the premises of one subscriber was connected to another subscriber's teleprinter via the *public switched telephone network* (PSTN) that had evolved since the advent of analogue telephony in 1876. A text message, typed in at the keyboard of the sending teleprinter, was represented in 5-bit Baudot (ITA-2) code, and transmitted down the line using *on* (mark) – *off* (space) – *keying* (OOK) of a 1500 Hz carrier voltage at a speed of 50 Bd. The message was received and automatically printed on paper at the destination teleprinter.

Telex quickly grew into a global service particularly popular with government departments and a wide range of institutions and business organisations. Documents sent by telex had legal status. However, although the technology behind telex was steadily improved over the years – from manual to automatic switching, from 50 Bd to 300 Bd transmission speed, and from mechanical teleprinters to PCs – demand for the service began to decline steadily in favour of *facsimile* in the 1980s and *email* in the 1990s.

1.3.2.3 Facsimile

Facsimile service, or *fax* for short, allows paper documents – handwritten, printed, drawings, photographs, etc. – to be electronically duplicated over a distance. In 1980, the CCITT specified the first digital transmission standard for fax, called Group 3 (G3) standard. The connection between the sending and receiving fax machines is in most cases via the PSTN, and sometimes by radio. The G3 fax standard became the most widely used and could transmit an A4 page in less than one minute, typically 15–30 seconds. A later G4 standard required a digital-grade telephone line and could transmit an A4 page in less than five seconds with a resolution of 400 *lines per inch* (lpi).

The G3 standard has a resolution of 200 lpi, which means that the paper is divided into a rectangular grid of picture elements (*pixels*), with 200 pixels per inch along the vertical and horizontal directions, amounting to 62 pixels per square millimetre. Starting from the top left-hand corner, the (A4) paper is scanned from left to right, one grid line at a time until the bottom right-hand corner of the paper is reached. For black-and-white-only reproduction, each pixel is coded as either white or black using one bit (0 or 1). For better-quality reproduction in G3 Fax, up to five bits per pixel may be used to represent up to $2^5 = 32$ shades of grey. The resulting bit stream is compressed at the transmitter and decompressed at the receiver in order to increase the effective transmission bit rate. At the receiver, each pixel on a blank (A4) paper is printed black, white, or a shade of grey according to the value of the bit(s) for the pixel location. In this way the transmitted picture pattern is reproduced. Early fax machines used a drum scanner to scan the paper directly, but later machines formed an image of the paper onto a matrix of *charge-coupled devices* (CCDs), which build up a charge proportional to incident light intensity.

Fax had its origin in chemical telegraphy invented in 1842 by Alexander Bain (1810–1877) and involved damp electrolytic paper. Fax did not, however, come into any serious use until 1902 when a German, Dr Arthur Korn (1870–1945), developed a suitably sensitive photoelectric cell, which allowed a mechanism to be devised for converting a photographic negative of the picture into an analogue electrical signal. In 1924, AT&T (American

Telephone and Telegraph Corporation) demonstrated *telephotography*, as fax was then known, by transmitting pictures over a telephone line from Cleveland to New York.

By the end of the 1920s, pictures of distant events for publication in newspapers were being routinely sent by fax. In 1935, the service began to come into more widespread use amongst businesses when the fax machine became more affordable. This came about through eliminating the need for a photographic negative and producing the electrical signal using light reflected directly off the picture being scanned.

At the turn of the century, the popularity of fax for business communication began to be challenged by email, which is particularly convenient for sending documents that are in electronic (word-processed) form as it can be done without the sender leaving their desk or feeding any paper into a machine. And by first scanning the documents into electronic form and sending them as file attachments, email may even be used for transmitting hard copies and documents that must be supported by a signature and/or headed paper.

1.3.2.4 The Digital Era

Digital communication services have grown at an astounding rate during the last five decades. Computer networking started in 1971 with the Advanced Research Projects Agency Network (ARPANET). By the end of the 1990s, this simple project had evolved into an *Internet revolution* having enormous global impact on every aspect of business and social life, and spurning completely new jobs, business methods, work patterns, social pastimes, vocabulary, crimes, etc. *Social media*, *search engines*, *e-commerce*, *video*/*music streaming*, *e-learning*, and *online transactions* are just a few of the countless services made possible by the Internet, which is now a global network of computers, devices, and (in short) things.

A cellular telephone concept, which was first demonstrated by Motorola in 1972, quickly caught on and led to a global *mobile communication revolution*, driven by steady improvements in successive generations of nonbackward-compatible terrestrial wireless communication technologies from 1G in 1979 to 5G in 2019. The major areas of improvements included changing (i) from *analogue* technology in 1G (launched in 1979) to *digital* technology in 2G (launched in 1991); (ii) from circuit switching to packet-switched *high-speed data* network of data rate ≥ 144 kb/s in 3G (launched in 2001); (iii) to an all-IP (Internet Protocol) *mobile broadband* network of peak data rate ≥ 100 Mb/s in 4G (launched in 2011); (iv) to a high-mobility (≤ 500 km/h), low-latency (≥ 1 ms), *ultra-broadband* network of peak data rate ≥ 10 Gb/s in 5G (launched in 2019).

At the same time as mobile communication technology was evolving, there was enormous progress in the development of the transceiver device or user terminal (enabling access to the network), culminating by the year 2000 in the *smartphone*, which is unarguably the most versatile and consequential pocket-sized device or gadget in humankind's entire history. Think of it! The smartphone is a telephone handset, data terminal, video terminal, video camera, still camera, calculator, audio recorder and player, clock and alarm, route finder (i.e. navigational aid), personal digital assistant (e.g. notebook, diary, calendar, reminder), gateway to the Internet, and much more; yet it is not just portable (i.e. untethered to one location) but pocket-sized. The smartphone has become a must-have item for work, leisure, and travel, and its global penetration has even surpassed the all-time great success stories of the car and washing machine. For example, on 29th February 2020, the website http://live-counter.com [1] estimated that there were 1.318 billion cars in the world. This figure should be compared to an estimate of 3.5 billion smartphone users worldwide in 2020 by http://statista.com [2] and a report in http://gartner.com [3] of total global smartphone sales of 1.524838 billion units in the year 2019 alone.

In 1973, the US department of defence began the development of a *global positioning system* (GPS) comprising a space segment of 24 satellites, a control segment of several earth stations, and a user segment. The system eventually became operational in 1995. GPS was designed primarily for the US military to provide estimates of position, velocity, and time of a GPS receiver and hence of the person or unit bearing the receiver. However, civil applications of the service have since grown to the extent that satellite navigation (enabled primarily by the US GPS) is now indispensable to aeronautical and maritime transportation, mining, surveying, and countless other activities; and is also an important tool for route determination (called satnav) by drivers and the general population around

the world. In addition to the US GPS, other less dominant *global navigation satellite systems* (GNSS) have also been launched, including Europe's Galileo (which went live in 2016), Russia's GLONASS (the development of which started in 1976 but full global coverage was only attained in 2011), and China's BeiDou (which started limited regional coverage in 2000 and attained full global coverage in 2018).

By the early 1970s, advances in integrated circuit technology had made digital transmission a cost-effective way of providing telephony, which was hitherto an exclusively analogue communication service. *Pulse code modulation* (PCM), a signal processing technique devised by British engineer Alec Reeves (1902–1971) back in 1937, was employed to digitise the analogue speech signal by converting it into a sequence of numbers. Digital exchanges and TDM began effectively to replace analogue technology, i.e. analogue exchanges and *frequency division multiplexing* (FDM) – in the PSTN during the 1970s, although the first TDM telephony system had been installed as far back as 1962 by Bell Laboratories in the USA. The momentum of digitalisation of the entire transmission system increased in the 1990s, until by the turn of the century the telecommunication networks of many countries had become nearly 100% digital. Globally, in 2020, the only remaining analogue portion of the telecommunication network is the *local loop* connecting, via copper wire pair, a subscriber's landline telephone handset to the local exchange, or a street cabinet beyond which transmission is entirely digital.

Digital audio broadcasting began in September 1995 with some field trials by the British Broadcasting Corporation (BBC). Although analogue *amplitude modulation* (AM) and *frequency modulation* (FM) radio broadcast will continue into the foreseeable future, digital radio now provides the best in broadcast sound quality, especially in the presence of multipath distortion. Terrestrial digital television broadcast was first launched in the UK in 1998 and provides a more efficient utilisation of the radio spectrum and a wider service choice than its analogue counterpart. Transition to terrestrial digital television broadcasting and a total shutdown of analogue television broadcasting (a process known as *digital switchover*) has been completed in many countries around the world. For example, analogue TV broadcasting was fully terminated in the UK on 28th November 2013 and in China on 14th May 2016. Also, TV broadcasts via satellite have been migrated from analogue transmission using FM to digital transmission using a selection of bandwidth-efficient digital modulation techniques. It is therefore likely that analogue television broadcasting, first started by the BBC in 1936, will cease to exist within this decade.

By the turn of the century, supported by a variety of reliable mass data storage media, the digital era had fully arrived, and the convergence of communication, broadcasting, computing, and entertainment had become a reality. This convergence is best epitomised at the device level by the smartphone and at the network level by the Internet. Both support multimedia streaming, audio and video (IP) telephony, mobile computing, and gaming.

In view of the rapid pace of development and diverse range of services, a detailed study of all telecommunication applications would be an extremely challenging and inadvisable task. Fortunately, the underlying principles do not change and the best way to acquire the necessary expertise in this exciting field is to first become grounded in its principles. It is for this reason that this book will focus on a thorough yet accessible treatment of the principles of communication engineering, drawing on modern examples and applications as necessary to ensure a complete understanding of the concepts presented as well as a better appreciation of the current state of the art.

1.3.3 Developments in Transmission Media

Much of the growth in telecommunications would have been impossible without significant developments in transmission media which provide suitable paths for signals to travel from source to destination. The first telegraph circuit built by Wheatstone and Cooke between Euston and Camden stations consisted of iron wires insulated with cotton and laid in iron pipes buried beside the railway line. The insulation and hence the line failed whenever the wires became wet. This forced Wheatstone and Cooke to abandon the idea of buried cables and opt rather for suspending the wire from telegraph posts using glass insulators. From this very humble beginning, transmission medium technology developed to encompass in chronological order copper cables, and radio and optical fibre,

all of which have been used for short- to long-distance links. We will briefly review the historical development of these three media. But note that other transmission media suitable only for very short links ranging from a few centimetres to a few metres have also been developed over the years. These include microstrip lines in printed circuit boards, metallic waveguides used, for example, to connect an outdoor antenna to an indoor receiver unit, and infrared radiation used, for example, in the remote control of gadgets.

1.3.3.1 Copper Cable

The telegraph poles that crisscrossed America and Europe in the early days of telegraphy carried iron or steel wires, one wire for each circuit with the ground providing the return path. It was known back then that copper was a much better electrical conductor than iron. In fact, the first ever undersea telegraph cable, laid between Dover (UK) and Calais (France) in 1851, consisted of four copper conductors insulated with a coating of gutta-percha (the natural latex of the native Malaysian *Palaquium* tree brought to the UK in 1843) and protected by iron sheathing. Copper was also the conducting material for the first successful trans-Atlantic telegraph cable laid in 1866 (after three previous failed attempts in 1857, 1858, and 1865) between Valentia Island (Ireland) and Newfoundland (USA), a distance of 1852 miles. Copper, however, is a soft metal and the annealing technique necessary to make it strong enough to support its own weight from pole to pole had not yet been discovered. This situation changed in 1877, when Thomas Doolittle (1839–1921) invented a suitable copper wire manufacturing process.

As a result, in the 1880s copper wire replaced iron as the standard transmission medium, and the single wire circuit (with ground return path) was replaced by a two-wire (i.e. wire pair) circuit, which significantly reduced noise. Further improvement was obtained through the introduction of twisted-pair cables in which the two wires comprising one circuit are independently insulated and twisted around each other. Multipair cable was introduced, consisting of a lead tube containing up to 100 copper wires, each insulated with paraffin-impregnated cotton or gutta-percha.

By 1900, to avoid the unsightly congestion of aerial wires, short-distance cables began to be laid underground in conduits made initially of creosoted wood and then of vitrified clay. The first long-distance underground cable was laid in 1912 and carried analogue telephone signals between Philadelphia and Washington DC. Long-distance underground wire pairs were replaced by coaxial cables in the 1940s. Underground cable cores bearing the twisted wire pairs of the PSTN *local loop* are a significant global investment of the twentieth century which could not be simply discarded in the digital era. In the 1990s, technological advances (especially in very large scale integration (VLSI) circuits to perform the required complex signal processing) allowed the copper lines of the local loop to be used as a vital medium for digital transmission known as the *digital subscriber loop* (DSL). From 2003, there was a rapid global uptake of DSL-enabled broadband connection to the home.

AT&T invented the coaxial cable and installed the first experimental coaxial cable system between New York and Philadelphia in 1936, followed in 1941 during World War II by the first commercial installation between Minneapolis and Stevens Point. A coaxial cable (or *coax* for short) consists of an insulation-covered, tube-shaped outer conductor (sometimes of braided design) with a second conductor fixed into the centre of the tube and separated from it by other insulation. It provides better protection from interference and larger bandwidths than twisted-pair configurations. The early L1 coaxial cable installations of the 1940s could carry 480 telephone calls, whereas the L5 systems of the 1970s were vastly superior and could carry 132 000 simultaneous calls. The laying of the first transatlantic coaxial cable (TAT 1) spanning 2240 miles between Oban in Scotland and Clarenville in Newfoundland was completed in 1956. With an initial capacity of 36 telephone circuits, TAT 1 remained in service for 22 years, although several other TAT coax cables were laid in the ensuing years to provide more capacity. The last transatlantic coaxial cable to be laid was TAT 7 in 1978 with an initial capacity of 4000 telephone circuits. In the 1980s, coaxial cables were superseded by optical fibre and most long-distance coaxial cable installations have now been retired.

As the above evolution of the transmission medium from iron wire to copper coax took place, improvements were also made in insulation material, from cotton through gutta-percha and paper to finally (synthetic) plastic.

1.3.3.2 Radio

James Clerk Maxwell (1831–1879), an outstanding university professor by age 25 and widely regarded as the most brilliant British mathematician and physicist of the nineteenth century, made a theoretical prediction of the existence of radio waves in 1864. His theory is embodied in four equations, known as *Maxwell's equations*, which when combined yield two wave equations – one for a magnetic field and the other for an electric field – propagating at the speed of light. Simply put, Maxwell's theory stipulates that a changing electric field generates a changing magnetic field in the surrounding region, which in turn generates a changing electric field in the surrounding region, and so on. In this way a coupled electric and magnetic field travels out in space at the speed of light ($\approx 3 \times 10^8$ m/s), and the resulting wave is known as an *electromagnetic wave*.

The significance of Maxwell's theory for telecommunications is that if we can somehow generate an electromagnetic wave having one of its parameters – amplitude, frequency, or phase – varied in sync with the variations of an information-bearing voltage signal then we can transmit information from one point to another at the speed of light and without the need for a cable connection. This is *radio* or *wireless communication*. For example, to implement *wireless telegraphy* we can switch the wave on or off to signal a mark or space, respectively, and at some distance recover the sequence of marks and spaces by detecting the presence and absence of the wave during each signalling interval.

A German physicist called Heinrich Hertz (1857–1894) provided an experimental verification of Maxwell's wave theory in 1888. Using a wire connected to an induction coil to generate the electromagnetic waves and a small loop of wire with a spark gap to detect them, he measured their wavelength and showed that the waves were reflected and refracted in a similar way to light.

The most significant contribution to the establishment of radio communications at the turn of the twentieth century was made by Guglielmo Marconi (1874–1937), who was born and grew up in Bologna, Italy but moved to Britain in 1896 in search of financial support for his radio experiments. In 1897, he sent signals by radio over eight miles from Lavernock Point to Brean Down, made the first radio transmission from ship to shore while on a visit to Italy, and formed the Wireless Telegraph & Signal Company. In 1899, he put his invention to work for the international yacht races being held off New York harbour at that time. He followed the yachts around the course and reported the races for two newspapers using wireless telegraph from yacht to shore. This event helped bring wireless telegraphy to prominence in the US. In December 1901, Marconi achieved the first transatlantic radio transmission covering 2800 km from Poldhu in Cornwall (UK) to Signal Hill in Newfoundland (Canada); and by 1907 he had established a transatlantic wireless telegraph service.

Radio communication started in 1897 with point-to-fixed-point wireless telegraphy using Morse code – a digital transmission that was simple enough to implement using the technology of the time by switching the radio signal on and off. It was not long, however, before the broadcast and mobility potential of radio became obvious. In 1904, Marconi started a service of Morse-coded news broadcasts to subscribing transoceanic ships, although any Morse code literate person with the right equipment could receive the news for free. The exploitation of radio communication grew very rapidly during the twentieth century to a point today where modern society is inextricably dependent on radio communication. Some of the most notable early developments that made this possible are summarised below.

Reginald Fessenden (1866–1932) invented amplitude modulation (AM) in 1906, which for the first time allowed the amplitude of a radio carrier signal to be continuously varied according to variations of an analogue (voice) signal. This paved the way for audio broadcasts and radio telephony once the design of vacuum tube transmitters and receivers was perfected in the mid-1910s. The first AM radio broadcast was the transmission of the US presidential election returns from the Westinghouse station at Pittsburgh in November 1920. Commercial radio broadcasts started in 1922 and stations proliferated in the US after it was finally figured out that the service could be financed through advertising. In the UK, the BBC was established and started its broadcasts in the autumn of that year, paid for through licence fees – a practice which has continued till today. The first transatlantic radio telephone service

was established in 1927 between London and New York using a 60 kHz radio carrier. But with a capacity of only one call at a time the service was extremely expensive, attracting a tariff of £15 for a three-minute call.

Edwin Howard Armstrong (1890–1954) invented the frequency modulation (FM) technique in 1934 and FM radio broadcasts started in 1939. The FM technique varies the frequency of the radio carrier in proportion to the variations of the voice signal and yields a superior received signal quality compared to AM. In 1940, the Connecticut State police put into operation the first two-way FM mobile radio system designed by Daniel Noble (1901–1980), a professor of electrical engineering in the University of Connecticut.

In 1936, radio broadcast of television was started by the BBC in the UK. The first (black-and-white) television sets cost as much as a small car. Television had been invented in the 1920s, with significant contribution from John Logie Baird (1888–1946) – who invented mechanical television and successfully demonstrated a working prototype on the 26th of January 1926 to members of the Royal Institution in London. However, it was Philo Farnsworth (1906–1971) who demonstrated the first all-electronic television system in 1928.

In the 1940s, radio began to be used to provide high-capacity, long-distance transmission links for telephone and television signals. Called *microwave radio relay*, the system consisted of a relay of line-of-sight radio transceivers arranged on towers and had the advantage of lower construction and maintenance costs compared to coaxial cable transmission systems. It provided an alternative to cable and by the 1970s carried most of the voice and TV traffic; but it was surpassed in the 1980s by optical fibre.

In 1945, Arthur C. Clarke, famous as a science fiction writer, suggested in an article in the British radio magazine *Wireless World* that a long-distance radio link could be established using a satellite located in an equatorial orbit with a period of 24 hours. In this orbit known as *geostationary orbit* (GEO), the satellite would appear stationary when viewed from any point on the earth's surface. A signal transmitted from an earth station to the satellite could be retransmitted by the satellite back to earth and received at all points on the visible earth. This would make possible broad-area coverage and long-distance radio communication that was more reliable than the only method available at the time via HF radio transmission. This vision of satellite communications became a reality 20 years later when the first commercial satellite, *Early Bird* (later renamed *INTELSAT 1*), was launched into GEO and commenced operations on the 28th of June 1965. The first spacecraft (*Sputnik I*) had been launched by the USSR (now Russia) in 1957, but it was the satellite *Telstar I*, built by the Bell Laboratories and launched in 1962 into a 158-minute low earth orbit (LEO), that relayed the first television signals between USA and Europe. Demand for satellite communication has grown steadily since its inception. In the 1970s and 1980s, satellite systems were used for international and domestic telephony and television distribution. The arrival of superior optical fibre transmission systems in the 1980s shifted telephone traffic from satellite to optical fibre links. A bold attempt in the 1990s to provide global mobile communication using a constellation of satellites in LEO or medium earth orbit (MEO) ended in commercial failure due in part to the unforeseen runaway success of terrestrial cellular mobile communications. But satellite communication is now well established as the leading means of television programme distribution directly to homes, and of communication at sea and in remote, undeveloped or disaster areas. The GPS service is also a satellite communication application. Satellite communication is also used to provide broadband Internet access for subscribers in remote areas or in regions without adequate terrestrial communication infrastructure.

1.3.3.3 Optical Fibre

The use of optical fibre as a transmission medium was first proposed in 1966 by Charles K. Kao of Standard Telecommunications Laboratories, in the UK. An optical fibre is a dielectric waveguide about 125 μm in diameter made from high-purity silica glass. It consists of a glass core surrounded by a glass cladding of lower refractive index. A plastic sheath covers the cladding to give mechanical protection. An optical cable contains several fibres used in pairs – one for each transmission direction. Light waves propagate within the inner core over long distances

through total internal reflection at the core-cladding interface. This light can be switched on and off to signal bits 1 and 0, respectively, and thereby convey information.

At the time of Kao's proposal the best available fibre material had a loss of about 1000 dB/km, but Kao was convinced that the high loss was due to absorption by impurities in the glass, and that losses ≤20 dB/km could be achieved to make long-distance optical fibre transmission links possible. The first breakthrough in reducing fibre attenuation came in 1970, when Robert Maurer of Corning Glass Works used fused silica to make a fibre having an attenuation of 16 dB/km at a wavelength of 633 nm. Since then, improved fabrication technology and a move to higher wavelength regions – called *windows* – offering lower optical losses have resulted in dramatically reduced attenuation. There are four such windows at 850, 1310, 1550, and 1625 nm. The first-generation fibre systems were installed in 1977 and operated at 850 nm with an attenuation of 3 dB/km. Most operational systems today use the second window at 1310 nm with attenuation 0.5 dB/km and the third window at 1550 nm with attenuation 0.2 dB/km. The technique of Raman amplification enables use of the fourth window at 1625 nm (of attenuation around 0.3 dB/km) especially in ultra-dense wavelength division multiplexing (WDM) systems. The attenuation figures quoted here refer to intrinsic losses only. This is the loss inherent in the fibre material due to *Rayleigh scattering* and absorption by impurities. Optical fibre is also subject to external sources of attenuation known as *extrinsic losses* (e.g. due to sharp bending, coupling, and splicing), which may sometimes account for the bulk of the total attenuation.

Optical fibre has numerous advantages over radio and copper lines as a transmission medium, but the most significant are its low attenuation (stated above) and large bandwidth. This small attenuation allows a wider spacing of repeaters in optical fibre transmission links than is possible in copper lines, leading to reduced costs. For example, the last (and best) transatlantic copper cable TAT-7 had 662 repeaters compared to only 109 repeaters in the first (and poorest) transatlantic optical fibre cable TAT-8.

Optical fibre conveys an optical carrier of very high frequency – about 200 000 GHz. So, assuming a transmission bandwidth of about 5% of the supported carrier frequency, we see that optical fibre offers a potential bandwidth of about 10 000 GHz. The technology exists today – using soliton laser, erbium-doped fibre amplifier (EDFA) and dense wavelength division multiplexing (DWDM) – to allow long-distance transmission at several terabits per second on optical fibre without regeneration.

Starting from the 1980s, optical fibre has replaced coaxial cables in all long-distance and undersea transmission lines, and in the inter-exchange and metropolitan area networks. The first international submarine optical fibre system was installed in 1986 between the UK and Belgium. In 1988, the first transatlantic optical fibre system, TAT-8, was installed between Europe and the USA with a capacity of 560 Mb/s. Several more transatlantic optical fibre cables have been installed since then, up to the latest TAT-14, which became operational in 2001 with a capacity of 3.15 Tb/s.

Two transmission media frontiers have withstood the optical fibre revolution started in the final years of the twentieth century. First, the local loop – the so-called *last mile* of the PSTN – continues under the dominance of copper wire pairs. The reason for this is largely economic. Fibre in the local loop makes ultra-broadband connection to the home possible, but it involves significant extra costs for cable installation and new terminal equipment, costs which (unlike those of the long-distance segment) cannot be spread over several subscribers. Furthermore, the power supply required to operate the subscriber unit is currently sent from the local exchange along the same copper wire pair that carries the message signal, whereas fibre being an insulator imposes the burden of an alternative power supply arrangement. But these problems are not insurmountable and optical fibre is now beginning to replace copper in the local loop and a twenty-first century broadband network is emerging as a natural successor to the PSTN legacy of the last century.

The other frontier, the only one that is impregnable to optical fibre, is the connection or access provided by radio for mobile communication units on land, sea, and air, and in remote, undeveloped or disaster-stricken areas. The mobility and rapid deployment capabilities of radio in such applications are unique qualities that cannot be delivered by optical fibre.

1.4 Communication System Elements

Modern communication systems vary widely in their applications and complexity, but they are all accurately represented by the block diagram shown in Figure 1.5. A variety of information sources feed a transmitter with the message signal to be transmitted. The transmitter transforms the message signal into a form that is compatible with the type of communication system and that is suitable for passage through the transmission medium with an acceptably small distortion. The output of the transmitter, known as the *transmitted signal*, is placed into the transmission medium, which conveys it to a receiver located at the intended destination. The received signal at the output of the transmission medium and input of the receiver is a distorted version of the transmitted signal. Noise, distortion, and reduction in strength have been introduced by the medium. The receiver has the task of removing (as far as possible) the transmission impairments and undoing each operation performed by the transmitter. It then delivers an exact or close copy of the original message to a user or information sink. The receiver is also selected to match the characteristics of the transmission medium and to be compatible with the type of communication system.

In what follows, we give a brief overview of information sources and sinks. A nonmathematical introduction to the processing tasks of the transmitter and receiver is also presented, which lays a good foundation for a more detailed treatment in the rest of the book. However, an introduction to transmission media is postponed until the first section of Chapter 5, which is devoted to a detailed treatment of the topic.

1.4.1 Information Source

The information source or input device acts as an interface between the communication system and the outside world and provides the message signal that is processed by the transmitter. There are four main classes, namely audio, video, and data input devices, and sensors.

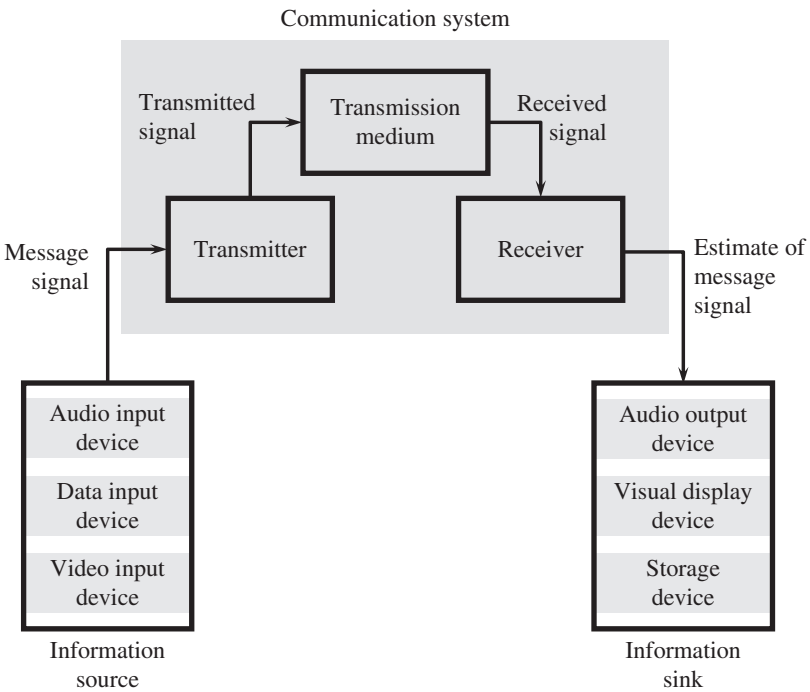


Figure 1.5 Block diagram of a communication system showing its major elements.

1.4.1.1 Audio Input Devices

The *microphone* is the most basic audio input device. It converts *acoustic* (sound) pressure wave into an electrical signal (current or voltage) of similar variations, which serves as the message signal input of the communication system. A voltage supply is often required for the operation of the microphone, or for the amplification of the weak electrical signal it generates. Other audio input devices may include a music keyboard, musical instruments digital interface (MIDI), audio players, and other digital music instruments. There are different types of microphones based on the principle of their operation.

Dynamic microphone: the dynamic or electromagnetic microphone is based on Faraday's law of induction, which stipulates that a changing magnetic field will induce voltage in a conductor lying within the field. In the *moving coil* dynamic microphone shown in Figure 1.6, a wire coil enclosing a fixed permanent magnet is physically attached to a diaphragm. Pressure waves from a sound source fall on the diaphragm, causing it to vibrate along with the coil. The motion of the coil in the field of the permanent magnet induces an electromotive force (electrical output) in the coil, which varies in sync with the sound pressure waves. A *moving magnet* dynamic microphone is also possible in which the coil is fixed and the magnet is attached to the diaphragm. In this case, sound waves moving the diaphragm also move the magnet, causing a change in the magnetic field in which the coil lies and hence inducing electromotive force in the coil. A type of dynamic microphone with much better fidelity at high frequencies is the *ribbon microphone*. Here a light (low-inertia) aluminium ribbon is suspended edgewise between the poles of a magnet. Electromotive force is induced in the ribbon when it is moved in the magnetic field by air flowing past it from sound waves.

Piezoelectric microphone: when a piezoelectric material is subjected to mechanical strain, voltage is induced between its ends. The operation of a *piezoelectric microphone* is based on this important property. If one end of such a material is fixed and the other end is attached to a diaphragm, sound waves falling on the diaphragm will cause a strain in the material. This induces a voltage that varies according to the strain and hence according to the sound pressure.

Carbon microphone: in a *variable-resistance* or *carbon microphone*, carbon granules are packed in a chamber formed by one fixed and one movable electrode. The movable electrode is attached to a diaphragm. Sound waves falling on the diaphragm vary the packing density of the granules and hence the resistance of the chamber. The variable resistance R is converted to a varying current I by connecting a constant voltage V across the chamber since, by Ohm's law, $I = V/R$.

Electret microphone: an *electret microphone* is of variable capacitance design. A capacitor is formed using two metallic plates, one fixed and the other movable. The movable plate is a thin metallic layer diaphragm. The fixed plate is a metal plate covered over with a film of Teflon-like plastic that has a permanent electric charge Q, hence the name *electret*. Air pockets are trapped between the electret and the fixed electrode due to irregularities on the surface of the electret. When sound waves fall on the movable diaphragm, the size of the air pockets and hence the capacitance C of the capacitor is varied. This produces a variable voltage V according to the relation $V = Q/C$. An alternative variable capacitance design is the *condenser microphone* in which the charge Q is maintained by an external voltage source.

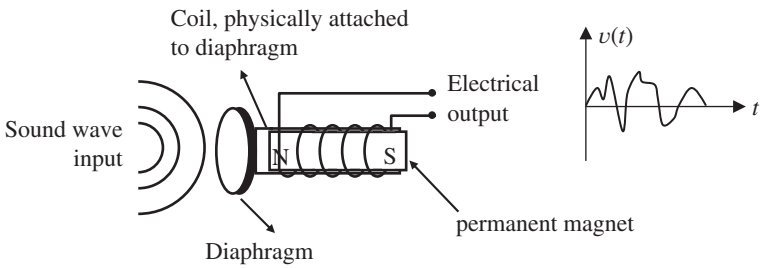


Figure 1.6 Moving coil dynamic microphone.

1.4.1.2 Video Input Devices

A video input device presents a video message signal to the communication system. The video signal is a variable voltage that originates from a visual signal, which consists of light intensity (or luminance) and colour (or chrominance) information. The most basic video input devices are the *video camera* (e.g. digital camera, analogue camcorder, and webcam) for dynamic three-dimensional (3D) images and the *scanner* (e.g. image scanner, 3D scanner, barcode reader, and fingerprint scanner) for still images. Secondary input devices include those that read pre-recorded video signals such as the video cassette recorder (VCR), which reads analogue video signals stored on magnetic tape in two-reel cassettes, and digital video recorders.

1.4.1.3 Data Input Devices

There is a wide variety of data input devices, which generate data serving as the message signal input of a communication system. Many of these devices work in conjunction with a computer system or other suitable *data terminal equipment* (DTE). A few examples of data input devices are given below.

- A *keyboard* generates the Unicode or ASCII code corresponding to the key pressed.
- Several devices are used to convert finger and hand movement into computer commands, which may involve data generation under the control of a computer program. Examples of such devices include the *mouse*, *trackball* (an upside-down mouse), *joystick*, *touchscreen*, *touchpad*, *pointing stick*, and *virtual reality systems*.
- There are different types of devices that read stored data and present these to the communication system for processing and/or transmission. A *barcode reader* or *laser scanner* scans a barcode, which encodes information (e.g. the price of an item) in black lines of varying width. A *magnetic-ink character reader* reads information written on a document using magnetic ink characters. A *magnetic strip reader* reads data stored magnetically on magnetic strips, such as those found on credit cards, security access cards, etc. *Optical character readers* and *optical mark readers* are used (in conjunction with software) to read printed characters and to detect and decode marks on paper. *Punched card readers* read information stored as an array of small holes and spaces on paper. Other data-reading input devices include the *optical disk drive* for reading data from an optical disk, the *disk drive* for reading data from a floppy or hard disk, and the *tape reader* for reading data from a magnetic tape.
- A *digitising tablet* is used to convert graphics information (drawn on the tablet) into computer data. The drawing is displayed on a computer screen and the associated data may be processed by a communication system.
- A *smart pen* is used to capture handwritten or other graphics information, which may then be converted into data for processing in a communication system.

1.4.1.4 Sensors

Sensors measure physical quantities such as temperature, pressure, mass, etc., and convert the measurement into an electrical signal that serves as the message signal input for the communication system. Sensors are used in telemetry systems to obtain information from remote or inaccessible locations, e.g. monitoring the status of equipment on a spacecraft or detecting the accumulation of ice on the wings of an aircraft. They are also used in numerous systems including automatic data-logging systems, security systems, safety systems, traffic control systems, manufacturing systems, and process control systems. For example, a tipping-bucket rain gauge generates a voltage pulse each time a standard-size cup tips after filling with rainwater. The irregular sequence of voltage pulses serves as the message signal, which may be transmitted and analysed to obtain the rainfall rate. A security system may use a sensor to detect movement. An alarm (or message) signal is then generated and transmitted to a console and used to initiate appropriate action, such as the ringing of a bell.

In the most general sense, every input device can be described as a sensor of some sort. The camera senses reflected light, the microphone senses sound pressure, the keyboard senses key press, the mouse senses hand movements, etc. However, the classification of input devices presented here is useful in identifying the type of information provided by the device to the communication system, whether audio, video, data, or a more general measurement of a physical quantity.

1.4.2 Information Sink

An information sink is the ultimate destination of a transmitted signal. It may serve as an interface between human users and the communication system, making the transmitted information understandable through sight or sound. It may also serve as a repository, storing the transmitted information for later processing, re-transmission, or display. There are therefore three types of information sink, namely audio output devices, visual display devices, and storage devices.

1.4.2.1 Audio Output Device

Generation of sound: the loudspeaker is an audio output device that converts the electrical signal output of a communication system into mechanical vibrations at the same frequencies. Figure 1.7 shows the basic operation of a moving coil loudspeaker. A cone is physically attached to a wire coil, which encloses a permanent magnet lying along its axis. The electrical signal is fed as current into the coil. This current sets up a magnetic field that varies proportionately with the current and has maximum strength along the coil axis. The interaction of the field of this electromagnet with the field of the permanent magnet results in a force that varies in sync with the field strength of the electromagnet and hence with the input current. This varying force causes the coil along with the attached cone to vibrate, thereby producing sound waves that follow the input current variations. The vibrations also generate sound waves travelling in the backward direction, which are reflected by the loudspeaker casing, and may interfere in a destructive manner with the forward waves. To prevent this, the casing is padded with sound-absorbent material. It is usual to have separate loudspeakers, a *woofer* for the low frequencies and a *tweeter* (with a smaller cone) for the high frequencies.

Sound pressure level (SPL): a sound signal or sound wave is an oscillation in pressure, stress, particle displacement, or particle velocity in an elastic medium such as air. When this sound wave is received and processed by the ear–brain mechanism, the sensation of hearing is produced, provided two conditions are met: (i) the oscillation or vibration frequency must be within the audible range; and (ii) the SPL produced by the vibrations must be high enough, i.e. above the threshold of hearing.

The ear only perceives sound signals of frequency in the range 20 Hz to 20 kHz. This is the theoretical audible range of frequencies. However, in practice most people have a narrower hearing range, with the top end decreasing with age after about age 30 years. Sound pressure may be expressed in dyne/cm^2, microbar, or Newton/m^2, where

$$1 \text{ dyne/cm}^2 = 1 \text{ microbar} = 0.1 \text{ N/m}^2 \tag{1.1}$$

The threshold of hearing for an average person below age 30 when the sound signal is at a frequency of 1 kHz is 0.0002 dyne/cm^2. That is, sound of 1 kHz in frequency will be heard (by this standard listener) only if the vibrations

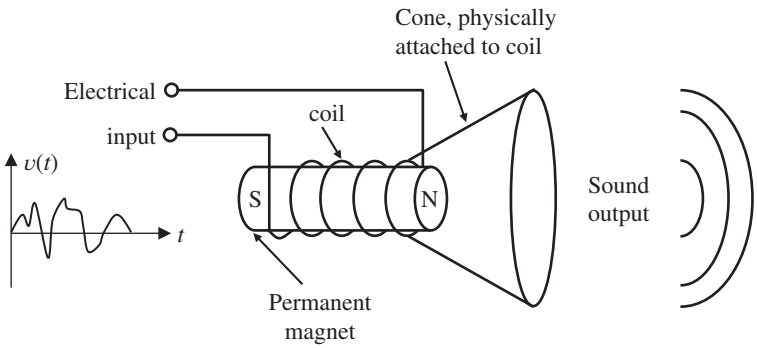


Figure 1.7 Audio output device: a loudspeaker.

produce (in the ears) a sound pressure of at least 0.0002 dyne/cm^2. This value has been adopted as a reference P_{REF} for expressing SPL in decibel (dB). Thus, the SPL of a sound of pressure P dyne/cm^2 is expressed in dB above P_{REF} as

$$\text{SPL} = 20\log_{10}\left(\frac{P}{P_{REF}}\right) \quad \text{dB},$$
$$\text{where,} \quad P_{REF} = 0.0002 \text{ dyne/cm}^2 \tag{1.2}$$

The SPL of ambient noise in a quiet home is about 26 dB, that of formal conversation is about 62 dB, whereas that of jet planes at take-off is about 195 dB. The threshold of hearing increases irreversibly with age above 30 and is raised to a higher level in the presence of ambient noise. The second effect is transient and is referred to as *noise masking*. The *dynamic range* of the ear is typically 120 dB, bounded at the bottom by the threshold of hearing and at the top by the threshold of pain. Sounds of SPL < 0 dB are inaudible, whereas sounds of SPL ≥ 120 dB will cause pain and may result in an immediate and irreversible loss of hearing.

Frequency response of the ear: whereas SPL is an objective measure of the amplitude of sound vibrations, another closely related term, *loudness*, is not. Loudness is the response of the human ear to the amplitude of sound waves. It is a subjective attribute that allows us to place a given sound at a point on a scale from soft to loud. The loudness level of a sound is expressed in *phon*, which is numerically equal to the SPL (in dB) of a 1 kHz reference tone. The sensitivity or response of the human ear to sound varies markedly with frequency. A 1 kHz sound at an SPL of 130 dB is deafeningly loud, but a 22 kHz sound at the same SPL is not loud at all. In fact, you would not hear the latter (because it is outside your audible range), although both vibrations produce the same high level of sound pressure.

The ear acts like a band pass filter, shutting out all frequencies outside the audible band. Within the audible band itself, the response of the ear is far from uniform. Figure 1.8 is the ISO-226 standard curves of the SPL to maintain constant loudness at various frequencies. It shows, for example, that to produce sound of loudness 40 phon, one needs SPL of 40 dB at 1 kHz and a higher SPL of 77 dB at 30 Hz. We observe also that the ear's frequency response is more variable when dealing with softer sound, but reasonably flat with loud sounds (SPL > 100 dB).

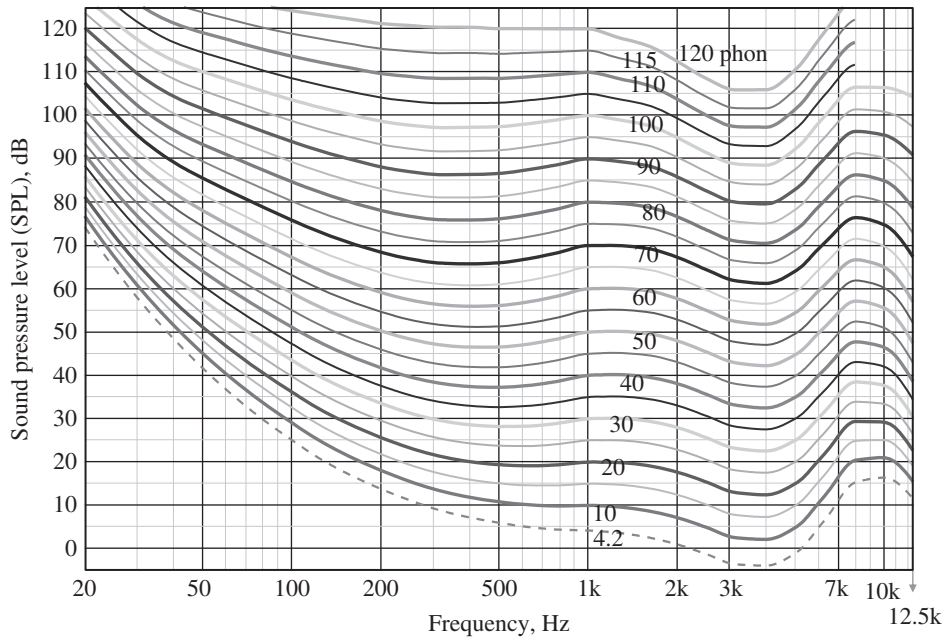


Figure 1.8 Contours of equal loudness (phon) as a function of frequency and sound pressure level.

1.4.2.2 Visual Display Devices

Three broad classes of visual display devices are common as information sinks in communication systems. Cathode-ray tube (CRT) and flat panel displays present what is usually called a *soft copy* of the received message signal, whereas printers produce a *hard copy* of the message.

Cathode-ray tube (CRT): cathode ray tubes display information (pictures, text, graphics, etc.) by using an accelerated beam of electrons to excite electrons in phosphor atoms, which emit visible energy as they return to their unexcited state. The basic parts of a CRT are shown in Figure 1.9. The CRT consists of an electron gun, a focusing system, a deflection system, and a phosphor-coated screen. Electrons are boiled off a hot cathode surface, a process called *thermionic emission*. The number of electrons emitted by the gun and hence the brightness of the display is controlled by a negative voltage on the control grid. A focusing system is needed to prevent beam spreading due to mutual repulsion among the electrons and to focus the electron beam to a small point on the screen. The focusing system may be an electrostatic lens formed by the electric field of a positively charged metal cylinder, or a magnetic lens formed by the magnetic field of a current-carrying coil mounted around the outside of the CRT envelope.

The deflection system is used to direct the beam to any spot on the screen. Figure 1.9 shows an electrostatic deflection system. Magnetic deflection may be obtained using current in a pair of coils mounted on the top and bottom of the neck of the CRT envelope for horizontal deflection and another pair of current-carrying coils mounted left and right for vertical deflection.

The phosphor atoms on the screen at the spot impacted by the accelerated electron beam will glow for a short time. The duration or persistence of the glow and its colour depends on the type of phosphor. To maintain an image without flicker, the entire screen is refreshed at a regular rate by repeatedly moving the electron beam along rows of the screen from top to bottom, while simultaneously varying the beam intensity to reproduce the desired image.

For a colour display, each spot has three phosphor dots that glow with a red, green, and blue colour. Three electron guns are used, one for each colour dot. Any desired colour can be produced at each spot by (additive) mixing (of) these three primary colours in the right intensity. For example, a yellow colour is obtained by exciting the green and red phosphor dots equally without exciting the blue phosphor dot – that is, the electron gun aligned with the blue phosphor dot is switched off.

Despite its excellent display quality, the weight, bulk, and power consumption of a CRT make it unsuitable for portable devices (e.g. pocket calculators) and the large displays required for example in high-definition television (HDTV). After the turn of the century, CRTs began to be superseded by flat panel displays.

Flat panel displays: flat panel displays have smaller volume, weight, and power consumption than a CRT. They are now the preferred technology for use in a wide range of applications requiring monitors, e.g. TV, computers,

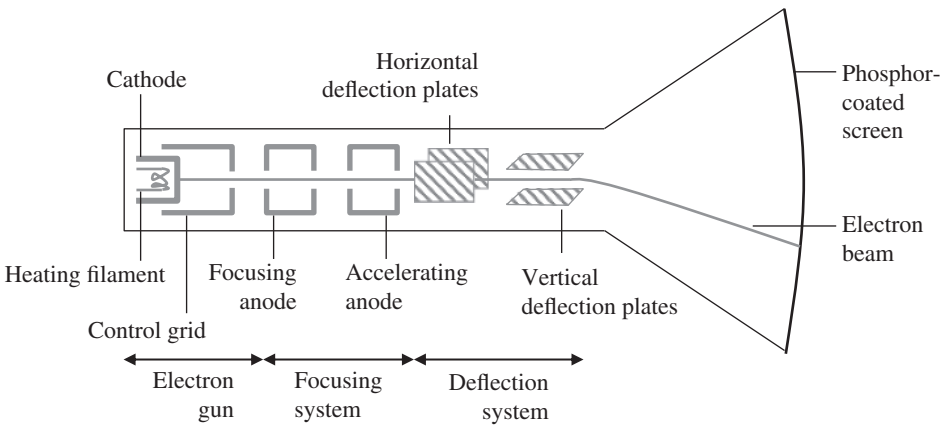


Figure 1.9 Basic components of a CRT.

mobile telephone units, etc. There are two main classes of flat panel displays. One class, called *emissive display*, works by converting electrical energy into light. The other class, called *non-emissive display*, uses an arrangement that modulates (i.e. blocks or passes to varying degrees) light from an external or internal source, thereby creating a graphics pattern.

Examples of an emissive flat-panel display include a *light-emitting diode* (LED) display, *plasma-panel* (or gas discharge) display, and *thin-film electroluminescent* display.

In a LED display, a matrix of diodes is arranged to form the screen spots or picture elements (called *pixels*) and a picture is displayed by applying a voltage to light the diode at each required location. Recent developments have led to widespread use of *organic light-emitting diode* (OLED) displays in which the electroluminescent substance is an organic compound that emits light in response to electric current. Sony produced the world's first TV set using an OLED display screen in 2007. In addition to TV and computer screens, OLED displays are now used in a variety of portable devices, including the smartphone.

In plasma-panel and electroluminescent displays, the volume between two glass panels is filled with a suitable substance – a mixture of neon and some other gases for the former, and a manganese-doped phosphor for the latter. One panel carries a series of vertical conducting ribbons, and the other carries a series of horizontal ribbons. The intersection between a vertical ribbon and a horizontal ribbon defines a display pixel. To light a pixel, firing voltages are applied to the pair of ribbons that intersect at that point. This breaks down the gas at the pixel to form glowing plasma, or, in the case of the electroluminescent display, causes the manganese atoms at the pixel to absorb electrical energy, which makes them glow. The firing is refreshed at a regular rate (e.g. 60 Hz) using the picture definition stored in a refresh buffer.

A *liquid crystal display* is non-emissive. In this case, the volume between two glass plates is filled with a liquid crystal compound (LCC), a substance that has a crystalline arrangement of molecules but flows like a liquid. One of the glass plates contains rows of transparent conductors and a light polariser. The other plate contains columns of transparent conductors and a light polariser at right angles to the polariser in the other plate. The intersection of a row conductor with a column conductor defines a pixel. A pixel is turned on when the LCC molecules at its location are unaligned and is turned off when these molecules are forced to align by the application of a voltage. This happens because the unaligned molecules at the pixel twist the polarised light, causing it to pass through both polarisers. On the other hand, if a voltage is applied to two conductors it aligns the LCC molecules at their intersection and prevents the twisting of the light so that it cannot pass through both polarisers, thereby becoming blocked.

Printers: impact printers produce hard copies by pressing the face of formed characters against an inked ribbon onto the paper. The characters are usually formed using a dot-matrix print head, which has a rectangular array of protruding pins, by retracting appropriate pins. The quality of the print depends on the dot size and on the number of dots per inch (or lines per inch) that can be displayed.

Various nonimpact printing techniques have been devised. Laser printers create a charge distribution on a rotating drum using a laser beam. Ink-like toner particles are then emitted from a toner cartridge onto the drum. These particles adhere only to the charged parts of the drum. The particles are then attracted from the drum to the paper and pressed into the paper by heated rollers. Ink-jet printers use an electric field to deflect an electrically charged ink stream to produce dot-matrix patterns on paper. Electrostatic printers charge the paper negatively at the right spots. These spots then attract a positively charged toner to create the required picture or text. Electrothermal printers apply heat in a dot-matrix print head to create the required pattern on a heat-sensitive paper.

Colour printing is obtained by mixing the right quantity of the colour pigments cyan, magenta and yellow. For example, equal mixture of magenta and cyan gives blue and equal mixture of all three pigments gives black. Laser colour printers deposit the three pigments on separate passes to produce a range of colour patterns, while ink-jet colour printers shoot the three colours simultaneously.

Finally, we should mention that *3D printing* has become commonplace in recent years. If viewed as an additive manufacturing process then a 3D printer is not a display device. However, if seen as a process to produce a physical

3D object from information stored on a computer then a 3D printer qualifies as a display device. A 3D printer produces a physical object through layer upon layer deposition of material. Information stored on the computer specifies the model to be produced, including the material to be deposited at each *voxel* (i.e. 3D pixel).

1.4.2.3 Storage Devices

Many communication services are designed to store the received message signal for visual or audio display later. In email, the text message is stored at the recipient's email server until the recipient accesses the server to obtain a soft copy (screen display) or hard copy (printout) of the message. Voicemail service provides for the storage of digitised speech message, which is played back later to the authorised recipient at their request. In these services, the information sink (in Figure 1.5) is an appropriate storage medium.

The desirable characteristics of a storage medium include large storage capacity, high access and data-reading speeds, portability, durability, low cost, data integrity, and reusability. Four types of storage media are most used as information sinks in communication systems.

Magnetic tape: this is made of a thin coating of a film of magnetic metal or ferromagnetic particles (embedded in a binding material) on a plastic strip of width 6–50 mm and length up to 1.5 km. The overall thickness of the tape (magnetic coating and substrate strip) is about 15–60 μm. The tape is contained in a two-reel *cassette* and is read from and written to in a tape drive or deck. The drive winds the tape from one reel to the other, causing it to move past a read/write head. Magnetic tapes offer large storage capacity and come in a variety of incompatible formats. The leading brand, known as Linear Tape-Open (LTO) Ultrium, offers a capacity of up to 30 terabytes (TB) when data are compressed. Other advantageous features of the magnetic tape include that they are reusable and easy to edit, compact, and portable. However, they have several demerits, which limit their use mostly to data archiving and hard disk backup. They are easily damaged through breaking or deterioration if exposed to heat or humidity. Recorded information degrades with continuous use, due to tape wear. Information can also be ruined by strong magnetic interference. Furthermore, access to data on the tape is slow and sequential.

Magnetic disk: magnetic disks may be fixed or removable. A flat-disk substrate is coated with a thin film of magnetic material. The disk is either a floppy disk, with a flexible substrate, or a hard disk, also known as a *hard disk drive* (HDD), with a rigid substrate. The surface of the disk is divided into several evenly spaced concentric tracks and each track is divided into sectors. Data are stored in these sectors. During a read/write operation, the disk is rotated at constant speed and the read/write head is moved in a radial direction to position it at the desired track. This radial movement is referred to as *seeking*. Floppy disks are magnetically coated on both sides and contained within a plastic with a spring-loaded door that is forced open when the disk is inserted in its drive. An HDD, including the disk and read/write head, is hermetically sealed to prevent contamination by dust particles and the head glides on a thin film of air without making actual contact with the disk. A PC hard drive usually consists of several disks (with a head for each disk) mounted on the same rotating axle. In this case, the set of tracks at the same radius on all the disks is called a *cylinder*. HDDs have two standardised form factors. The 3.5 in. disk has dimensions (length × width × height) = 146.99 mm × 101.85 mm × 26.11 mm, whereas the 2.5 in. disk has standard length 100 mm and width 69.85 mm but they have been produced in a variety of thicknesses ranging from 5 to 19 mm. HDD storage capacities have grown significantly over the years and is currently as high as 14 TB and 5 TB, respectively, for the 3.5 in. and 2.5 in. drives.

Magnetic disks have several advantages. Access time is less than 28 ms in high-speed drives. Random data access is possible using an index or directory (stored at a specific location on the disk) that lists the physical location of all data objects. Magnetic disks are reusable, cheap, and easy to write to, read from, or edit. The removable hard disks and floppy disks are also portable. However, the magnetic disks can fail, data on the disk becoming unreadable if certain critical sectors are damaged or otherwise corrupted.

Floppy disk drives are much slower than hard drives and have very limited storage. Floppy disks came in two standard sizes of 5.25 in. (with an initial storage capacity of 360 kB, later improved to 1.2 MB) and 3.5 in. (with a storage capacity of 1.44 MB). They were extremely popular and ubiquitous as a portable storage medium for

computer files in the 1980s and 1990s, but at the turn of the century they were superseded by more rugged and higher-capacity portable storage media such as the compact disk and the USB memory stick, and from around 2006 computers were no longer manufactured with floppy disk drives. To find a higher-capacity and compatible replacement for the floppy disk, efforts were made in the 1990s to introduce the 3.5 in. *floptical disk*, which used an optical tracking mechanism to improve head positioning and with a data storage capacity of up to 120 MB. These efforts ultimately failed in the market due primarily to the success of the optical disk.

Optical disk: optical disks include CD-R, magneto-optical, DVD-R, and DVD-RAM discussed below. The compact disk (CD) is based on a 12 cm plastic disk that stores data in the variation of the reflectivity of densely packed spots on the disk surface. The disk is read by bouncing a laser beam off the surface. Both the *compact disk recordable* (CD-R) and the *digital versatile disk recordable* (DVD-R) are write-once, read-many. On the other hand, the *magneto-optical disk* and the *DVD random-access memory* (DVD-RAM) can be erased and rewritten. CDs and DVDs have numerous advantages. They provide a very large storage capacity, 650 MB for a CD and up to 17 GB for a double-sided, dual-layer DVD. Data stored on a CD-R cannot be altered later, which ensures the integrity of archived information. Optical disks are very durable as the reading process is without physical contact with the disk and so does not cause any mechanical wear. Stored information can be read countless times with practically no degradation. There is ample provision for error correction and the error rate is below 10^{-15}. The disk is cheap, portable, compact, and not easily damaged, and the recording technique is immune to electric and magnetic fields. The main disadvantages of optical disks are that most of them, except magneto-optical disks and DVD-RAM, are not reusable. Once data have been written, they cannot be changed. Secondly, unlike magnetic disks that use the same head for reading and writing, a special disk writer is required to write information onto an optical disk. Thirdly, CD access times are about 10–20 times longer than that of magnetic hard disks.

Solid state disk (SSD): an SSD, also called a *flash drive*, stores data in non-volatile semiconductor memory cells made from metal-oxide-semiconductor field-effect-transistor (MOSFET) and fabricated as integrated circuit (IC) chips. Unlike magnetic hard disks, they have no moving parts, run silently, and have much quicker access times. The 2.5 in. SSD has storage capacity which may be as high as 4 TB. Since around 2006, SSDs have grown in popularity as the preferred storage medium in high-end notebooks, laptops, and PCs.

1.4.3 Transmitter

The transmitter is essentially a signal processor. Its primary role is to transform the message signal into a form that best suits the transmission medium, complies with any regulations governing the communication service, and meets system design objectives. How the transmitter does this depends on the type of information source, transmission medium, and communication system. For example, if the transmission medium is radio then the signal processing tasks of the transmitter would necessarily include modulation and radiation. The transmitter would process the signal in such a way as to ensure that it is transmitted in the authorised frequency band and at the permitted transmitted power level. If the main design objective is to minimise the cost of receivers then a simple AM would be preferred to a more complex digital signal processing technology.

Subject to the above considerations, the signal processing performed by the transmitter may include one or more of the following processes. A detailed discussion of most of these processes can be found in later chapters of the book.

Source coding: also called *source encoding*, this deals with the efficient representation of the message signal. Several processes may be involved, depending on the type of communication system. The following are examples of source coding.

- *Lowpass filtering* to limit the bandwidth of the message signal.
- *Multiplexing* to combine several message signals for simultaneous transmission.

- *Message formatting* to represent the message in digital format. This includes character coding of text and graphics using a suitable character code, such as Unicode, and analogue to digital conversion (ADC) of an analogue message signal, such as voice and video.
- *Data compaction* to minimise the number of symbols or bits that represents the message. Examples of data compaction include variable-length codes (e.g. Morse code and Huffman code), which assign shorter codewords to more frequent source symbols, and the Lempel–Ziv algorithm, which uses a codebook built from the message.
- *Lossy data compression* to reduce the number of bits used for representing the message by removing some details considered insignificant because a human observer will not perceive their absence. Examples of lossy data compression include predictive coding (e.g. differential PCM and delta modulation), low bit rate speech coding, transform coding, motion compensation, subsampling, colour table, and truncation.

Encryption: the message – in this context referred to as *plaintext* – is processed at the transmitter and converted into a *ciphertext*, which is disguised in some way to ensure security. The process may be accomplished by scrambling the bits of the plaintext using an *encryption key*. Only authorised receivers will have the correct *decryption key* with which to *decipher* or *decrypt* the ciphertext back to its original plaintext. This process is illustrated in Figure 1.10.

Channel coding: also called *channel encoding*, this process is necessary to ensure that the message signal is compatible with the transmission medium (or channel). It attempts to protect the message signal against channel distortion and transmission errors. It may add some redundancy to the transmitted symbols. Examples of channel coding include pre-emphasis/de-emphasis in analogue systems (e.g. FM), line coding in digital baseband systems, and provision for error detection and correction at the receiver. In fact, the process of carrier modulation introduced next is also a form of channel coding, which is required for some transmission media such as radio.

Channel coding is, however, often defined very narrowly as just *error control coding*, a simple example of which is *single parity check* in which the message bits are taken k bits at a time and one extra bit (called *parity bit*) is added to produce a codeword of $k+1$ data bits having an even number of binary 1 (for even parity check) or an odd number of binary 1 (for odd parity check). With this scheme, the receiver can detect the occurrence of a single bit error or any odd number of bit errors in a codeword. However, if two or more even numbers of bit errors occur in a codeword, this will go undetected. Nevertheless, this simple scheme does lead to a significant reduction in the number of undetected errors. For example, if a system incurs on average one character error per transmitted page

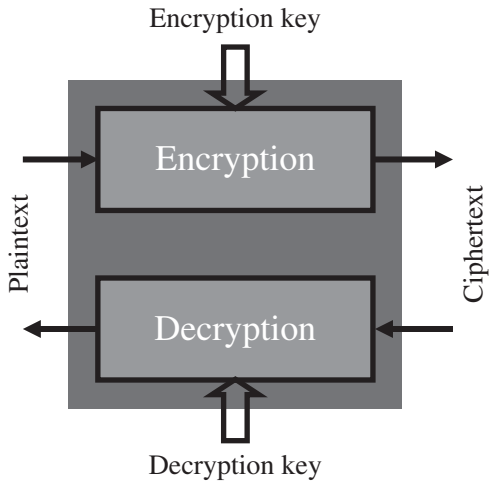


Figure 1.10 Encryption at the transmitter and decryption at the receiver.

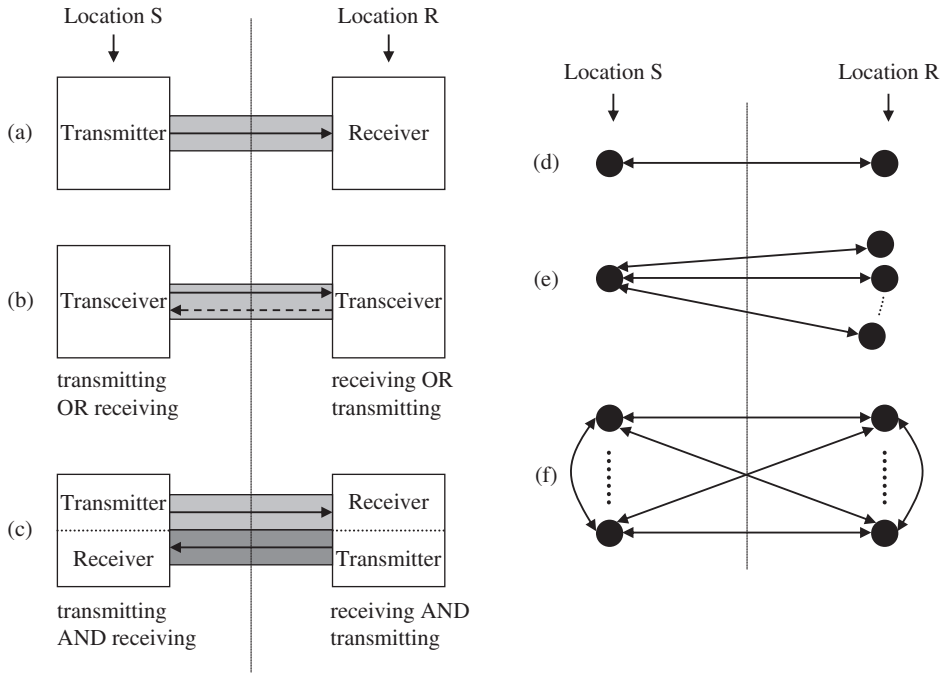


Figure 1.11 Communication systems: (a) Simplex; (b) Half-duplex; (c) Full duplex; (d) Point-to-point; (e) Point-to-multipoint; (f) Multipoint-to-multipoint.

of text, the introduction of single parity check with $k = 7$ leads to the number of undetected errors being reduced to one character in about 2500 pages of text on average.

A large variety of error control coding schemes have been devised over the years with capabilities that far exceed those of the single parity check, allowing not only the detection of a wide pattern of bit errors but also the reliable correction of one or more bit errors in a codeword.

Carrier modulation: this process translates the message signal frequencies from baseband into a frequency band that is suitable for and allows efficient exploitation of the transmission medium.

Spread spectrum modulation: this involves the transmitted signal being deliberately spread out over a wide frequency band. This technique was previously designed for military communications to protect against frequency-selective fading, interference, and intentional jamming; but it is now also employed as a multiplexing and multiple access technique in non-military communications.

Radiation: in a wireless communication system, the message signal must be processed and radiated as electromagnetic waves in the desired direction. The radiating element required depends on the type of wireless system. In general, *antennas* of various designs are used for radio communication systems and *light-emitting diodes* or *laser diodes* are employed in infrared and optical communication systems.

1.4.4 Receiver

A receiver is required at the other end of the transmission medium (i.e. destination of the transmitted signal) to process the received signal to obtain a close estimate of the message signal and deliver this to the information sink. Note that the receiver's input comes from the original message signal after it has been processed by two

blocks of the communication system, purposely in the transmitter and undesirably in the transmission medium. The receiver undoes the processing of the transmitter in a series of steps performed in reverse order to that of the transmitter. This may include one or more of the following tasks.

- Radio *reception* using a receive-antenna to convert the received electromagnetic waves back to a voltage signal.
- Spread spectrum *demodulation* to remove the carrier spreading.
- Carrier demodulation to translate the signal back to baseband frequencies.
- *Clock extraction* and *synchronisation* to recover the clock signal (if any) used at the transmitter in order to use the same timing intervals for operations at the transmitter and the receiver.
- Channel *decoding* to correct errors, remove redundancy, etc.
- *Decryption* to recover the plain and undisguised symbol sequence.
- Source decoding to recover the original message signal from the symbol sequence. This may involve *demultiplexing* – which breaks up a composite (multiplexed) signal into the components belonging to each of the multiple users – digital-to-analogue conversion (DAC), lowpass filtering, etc.

1.5 Classification of Communication Systems

There are four major ways of classifying the wide variety of communication systems in operation today. If we consider the direction of transmission then the communication system will be either *simplex* or *duplex*. If we consider the type of signal transmitted by the system then we have either an *analogue communication system* or a *digital communication system*. Comparing the frequency content of the transmitted signal to that of the original signal leads to a classification of the system as either a *baseband communication system* or a *modulated (or bandpass) communication system*. Finally, in terms of how a transmission path is established to convey the signal from source to destination, we may have either a *circuit-switched* or *packet-switched system*.

1.5.1 Simplex Versus Duplex Communication Systems

If information flows in only one direction, the communication system is referred to as *simplex* (SX). Typically, information originates from one transmitter and has one (or more) receiver as its destination. These receivers do not have the capability of responding through the *same* communication link.

By the above explanation, interactive digital television broadcast is a simplex communication system. Although a customer can respond (via a telephone link), information transmissions from the television house to the customer and from the customer to the television house are carried out on two different links. *Radar*, a system that measures electromagnetic signals reflected off objects (such as ship, aircraft, missile, storm, hydrometeor, etc.) in order to determine their range and velocity, is also a simplex system. In this case, although the signal may be reflected by the object back towards the transmitter S, giving an appearance of bi-directional information flow, there is strictly speaking only one transmitter, which is co-located with one receiver. The reflecting object is neither a transmitter nor a receiver.

Other examples of simplex systems include audio broadcast (AM radio, FM radio, music services, and digital radio), television broadcast (satellite, cable, and terrestrial), paging services, telemetry, and remote control.

When information can flow in both directions on the same link, the communication system is referred to as *duplex*. The communication equipment at both locations S and R is equipped with transmission and reception capabilities and is therefore referred to as a *transceiver* – a name that is a portmanteau of the words *transmitter* and *receiver*. The system may be designed in such a way that simultaneous communication in both directions is possible. This is a *full duplex* (FDX) system and requires a separate channel (e.g. a different band of frequencies, a different time slot, or a different wire pair) being allocated for each direction of communication. If, on the other hand, information can only flow in one direction at a time then the system is referred to as *half-duplex* (HDX).

The most common example of a FDX system is public telephony (both fixed and mobile). Broadband communication via DSL in the PSTN is also FDX, as is computer interconnection in local area networks (LANs). An example of an HDX system is the *walkie-talkie* used for wireless voice communication between two locations. Transmission in either direction uses the same radio frequency band. The handset at each location can be switched between transmit and receive modes, so that at any given time one location transmits while the other receives.

In both simplex and duplex systems, if communication takes place between only two transceivers then the system may be further described as a *point-to-point* communication system. If there is one transmitter or transceiver communicating with several receivers or transceivers, we have a *point-to-multipoint* system. If there are many intercommunicating transceivers (as in a LAN, or a video conference system linking more than two locations) then we have what is called a *multipoint-to-multipoint* communication system. In this last case, information flow between two transceivers or *nodes* is essentially bi-directional and therefore a simplex multipoint-to-multipoint system is not possible. The Internet is a multipoint-to-multipoint communication system. A radio or television broadcast system is a good example of a point-to-multipoint simplex system. See Figure 1.11.

An important limitation of simplex systems is that there is no return path for the receiver to automatically request re-transmission in the event of an error. Thus, there are two options to deal with the problem of errors: (i) ignore them in noncritical systems or (ii) include sufficient redundancy in the transmitted signal so that, in the event of an error, the receiver can make a reliable guess of the transmitted information. The guess may occasionally be wrong. This system of error correction is called *forward error correction* (FEC). Duplex systems are not under this limitation and can also use the more reliable technique of error correction called *automatic repeat request* (ARQ), in which the receiver automatically requests retransmission once an error is detected.

1.5.2 Analogue Versus Digital Communication Systems

An important classification of a communication system, from the point of view of the technology involved in its implementation, is based on the type of information signals conveyed, whether *analogue* or *digital*. Analogue communication systems convey analogue signals, while digital systems convey digital signals. However, note that a digital communication system may employ an analogue waveform called a *carrier signal* to convey the digital signal, as in the use of modems for data transmission over public telephone networks, which were originally designed for analogue speech signals. We will defer an introduction to analogue, digital, and other types of signals until the next chapter and focus here on discussing the advantages of digital over analogue communication, which are so great that even analogue signals such as speech and video are transmitted digitally by first converting them to a digital format at the transmitter. This *analogue-to-digital conversion* (ADC) at the transmitter is followed at the receiver by the reverse process of *digital-to-analogue conversion* (DAC).

The advantages of digital communication over its analogue counterpart include:

- *Low cost*: inexpensive digital circuits may be used to implement the signal processing tasks required in digital communication. With continuing advances in semiconductor technology, the cost of VLSI circuits will drop even further, making digital communication systems cheaper than their analogue counterparts despite the simplicity of the latter. Cost is an important factor that determines whether a new technology is successfully assimilated into society. The digital revolution of the 1990s, which we referred to earlier, was driven to a large extent by the falling costs of digital circuits of increasing computational power.
- *Privacy and security*: increased reliance on telecommunication systems for private, business, and military communications and the sale of entertainment and information services calls for *secrecy*, *authenticity*, and *integrity*. The first requirement ensures that the information is received only by an authorised user, whereas the last two requirements assure the receiver that there has not been any impersonation of the sender and that the information has not been deliberately or accidentally altered in transit. Digital communication permits data encryption to be easily implemented on the information bit stream in order to satisfy these requirements.

- *Dynamic range*: the dynamic range of a communication system refers to the amplitude ratio between the strongest and weakest signals that the system can process with an acceptably low level of impairment. Signals of a wider range of values (from very small to very large) than is possible with analogue systems can be accurately represented and transmitted with negligible distortion in digital communication systems. The dynamic range may be increased as much as desired by increasing the number of bits used to represent each sample of the analogue signal during the ADC process. The penalty, of course, is increased bandwidth requirements.
- *Noise immunity*: the number of errors in the received data, or the bit error ratio (BER), may be very small even in the presence of a significant amount of noise in the received signal. Although the precise value of the received signal will be changed by additive noise, the change will only rarely be large enough to force the signal value beyond the range that represents the transmitted bits.
- *Regeneration*: digital communication systems allow the possibility of regenerating (at sufficiently closely spaced regenerative repeaters) clean new symbols or pulses, which are free from all impairment effects and are (ideally) an exact replica of the original transmission. Thus, unlike analogue systems, noise does not accumulate from repeater to repeater and no further signal distortion occurs beyond that which was introduced at the analogue-to-digital conversion stage. Digital signals may also be stored in various storage media (e.g. optical or magnetic disks) and processed or re-transmitted later without loss of fidelity.
- *Error correction*: it is possible to detect and even correct errors in the received data using various coding techniques, which generally insert some redundancy in the transmitted data.
- *Flexibility*: the signal processing tasks of a digital communication system may be readily reconfigured simply by changing the software program, without any need to change the hardware. Modification of system functions can therefore be implemented more cheaply and speedily.
- *Integrated services*: voice, video, and data can all be represented in a common bit stream format and transmitted simultaneously in a common communication system. Multimedia communication, the Internet, and a host of other modern communication services are only feasible through digital technology.

Digital communication also has several disadvantages. However, in most cases, these disadvantages are under the system designer's control, and their effects may be reduced as much as desired by making a suitable trade-off. It should therefore be noted that the following disadvantages are far outweighed by the advantages discussed above, and this accounts for the transformation of global telecommunication into an all-digital network at the turn of the century.

- *Large bandwidth*: digital communication systems may require more bandwidth than analogue systems. For example, a 4 kHz bandwidth is adequate for analogue speech transmission, whereas digital speech transmission using standard (64 kb/s) PCM requires a minimum bandwidth of 32 kHz. The spectrum of all transmission media, especially radio, is limited. Transmission techniques that minimise the required bandwidth are therefore preferred in order to increase the number of services, users, and bit rate per user that can be accommodated. Various low bit rate speech coding and data compression techniques have been devised to reduce the bandwidth requirements of digital audio and video transmission and storage at the price of a somewhat reduced signal quality.
- *Complexity*: digital communication systems generally perform more complex processing operations on the input signal and require more sophisticated circuitry. Synchronisation usually must be maintained between receiver and transmitter. However, advances in semiconductor technology make circuit complexity a less significant disadvantage. Most of the signal processing tasks may be performed in a single highly reliable and affordable VLSI unit, which can be easily replaced in the unlikely event of a malfunction. Furthermore, some digital transmission techniques, such as the asynchronous transfer mode (ATM) – discussed in Chapter 13 – make synchronisation a less challenging issue.
- *Quantisation distortion*: analogue signals such as speech must be converted to digital form prior to transmission or processing in a digital communication system. This conversion introduces an irreversible quantisation distortion. However, this distortion may be made as small as the system designer wishes by increasing the number of quantisation levels. The price for this improvement is increased bandwidth requirements.

1.5.3 Baseband Versus Modulated Communication Systems

Communication systems may also be classified as either *baseband* or *modulated*, depending on how the information signal is prepared for the transmission medium. When the transmission medium can pass signals of frequency around 0 Hz and upwards, it is described as a *lowpass medium*. The original information signal (called the *baseband signal* since it usually contains frequencies near 0 Hz) may be conveyed to the receiver by being placed directly into the lowpass medium without any frequency translation. Such a communication system is referred to as a *baseband communication system*.

The only practical lowpass media are wire pair and coaxial cable. Thus, baseband systems are *always* wireline connected systems. Although the radio spectrum stretches all the way down to frequencies near 0 Hz and it is theoretically possible to convey a baseband signal (e.g. speech signal) by radio, this is not feasible in practice. Because it contains low-frequency components, the baseband signal cannot be efficiently placed into the radio medium at the transmitter or extracted from it at the receiver. We will have more to say on this later.

1.5.3.1 Analogue Baseband Communication System

The simplest example of baseband transmission is the voice intercom system. The baseband voice output signal of a microphone is conducted along a wire pair to a loudspeaker at some distance. Other examples of analogue baseband communication systems include the fixed (i.e. nonmobile) telephone connection between two local subscriber terminals in the PSTN and a closed-circuit television (CCTV) system.

It could be argued that the communication system block diagram of Figure 1.5 does not apply in this case, since there is neither transmitter nor receiver and the transmission medium links the message signal directly from information source to information sink. However, in most cases a transmitter is incorporated that performs very basic signal processing such as frequency-dependent *amplification* to compensate for attenuation and distortion in the transmission medium. The identifying feature of a baseband system is that any type of frequency translation is specifically excluded. The system usually includes a receiver that performs *lowpass filtering* to minimise the effect of noise, and *amplification* to provide for output volume control.

A block diagram of a CCTV system is shown in Figure 1.12. A lens in the video camera forms an image of the scene on a light-sensitive camera sensor, which converts the light intensity into an electrical signal called the *video signal*. This has a peak-to-peak value of 1 V and contains frequencies in the range 0–10 MHz, depending on the horizontal resolution of the camera. It is obvious that the video signal is a baseband signal. For a large distance between camera and console sites, a video amplifier is inserted to boost the video signal level. If there is only one camera then its output is connected directly via coaxial cable to a monitor. Otherwise, the outputs of several cameras are connected to a video switcher. This is an electronic device that selects different cameras automatically or manually for display on one or more monitors. The monitor is like a television set but has no tuning/demodulation circuits and therefore cannot receive a television broadcast off air. The monitor is simply a CRT or other type of visual display device that converts the baseband video signal into a visible display on its screen. A VCR may also be used for a permanent record of the video signal, and a printer to obtain a hard copy of a selected scene.

There is no frequency translation in this CCTV system, and therefore it is a baseband system. There are some implementations of a CCTV system that use radio, infrared, or optical fibre connection between camera and monitor. Such systems must necessarily implement some form of *carrier modulation* and therefore are not baseband systems.

One of the main difficulties with analogue baseband transmission is that a separate transmission medium (wire pair or coaxial cable) is required for each signal. One medium cannot be shared simultaneously by multiple users without interference, since each user's signal is continuous in time and occupies (or at least partly overlaps) the same frequency band. The cost would be prohibitive to provide a separate wire pair for all anticipated simultaneous telephone calls between two exchanges in the PSTN. If the analogue signal is transformed into a discrete-time function then an analogue baseband transmission system can be implemented that has the capability to accommodate multiple users on one link.

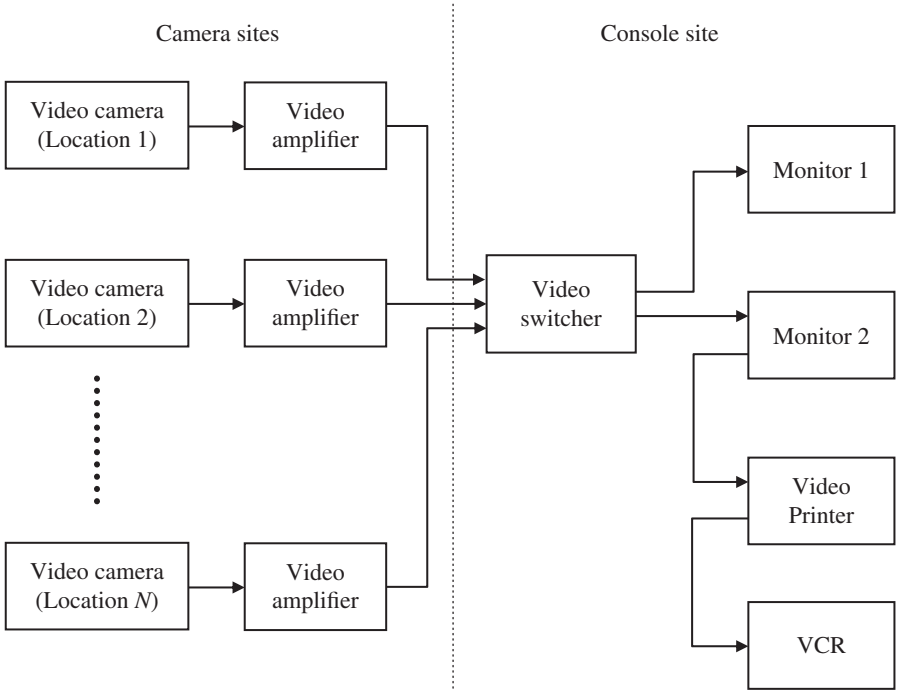


Figure 1.12 Closed-circuit television system (CCTV). Example of analogue baseband transmission system.

1.5.3.2 Discrete Baseband Communication System

If we take regular samples of an analogue signal $v(t)$ at intervals of T_s to obtain a sequence of continuous-value samples $v(nT_s)$ then we can perfectly reconstruct the original signal $v(t)$ from the samples $v(nT_s)$ by passing these samples through a suitable lowpass filter (LPF), provided that T_s is no more than half the period of the highest frequency component of $v(t)$. This is a statement of the *sampling theorem*. A discrete baseband communication system transmits $v(t)$ by sending one voltage pulse at the instance of each sample $v(nT_s)$. The value of each sample may be conveyed in the amplitude, duration, or position of this pulse.

PAM, PWM, and PPM: if the pulses are sent with equal width at regular intervals T_s but the *height* of the nth pulse (occurring at time $t = nT_s$) is varied in proportion to the sample value $v(nT_s)$ then we have what is referred to as *pulse amplitude modulation* (PAM). If the pulses are sent with equal height at regular intervals T_s but the *duration* or *width* of the nth pulse is varied in proportion to $v(nT_s)$, we refer to this as *pulse duration modulation* (PDM), also called *pulse width modulation* (PWM). Finally, sending the pulses with equal height and width but at irregular intervals such that the time of occurrence or *position* of the nth pulse is delayed relative to the nth sampling instant by an amount that is proportional to $v(nT_s)$ gives rise to *pulse position modulation* (PPM). Figure 1.13 illustrates these waveforms. In Figure 1.13d, the sampling instants are indicated by thin vertical lines. Note from this figure that the longest delay τ_{max} occurs at the third pulse where the sample value is maximum. The first pulse corresponds to the smallest sample value and has the shortest delay τ_{min}. The remaining pulses have delays ranging between τ_{min} and τ_{max}. In this way, information regarding the sample values is correctly conveyed by the positions of transmitted pulses.

The block diagram of a PAM generator and receiver is shown in Figure 1.14 and a suitable arrangement for generating PDM and PPM waveforms is given in Figure 1.15. Signal waveforms have been sketched at various points of these block diagrams in order to clarify the function of each element. The waveform of $v(t)$ is as earlier shown in Figure 1.13a and has been omitted. The effect of the transmission medium is also not shown. Note in

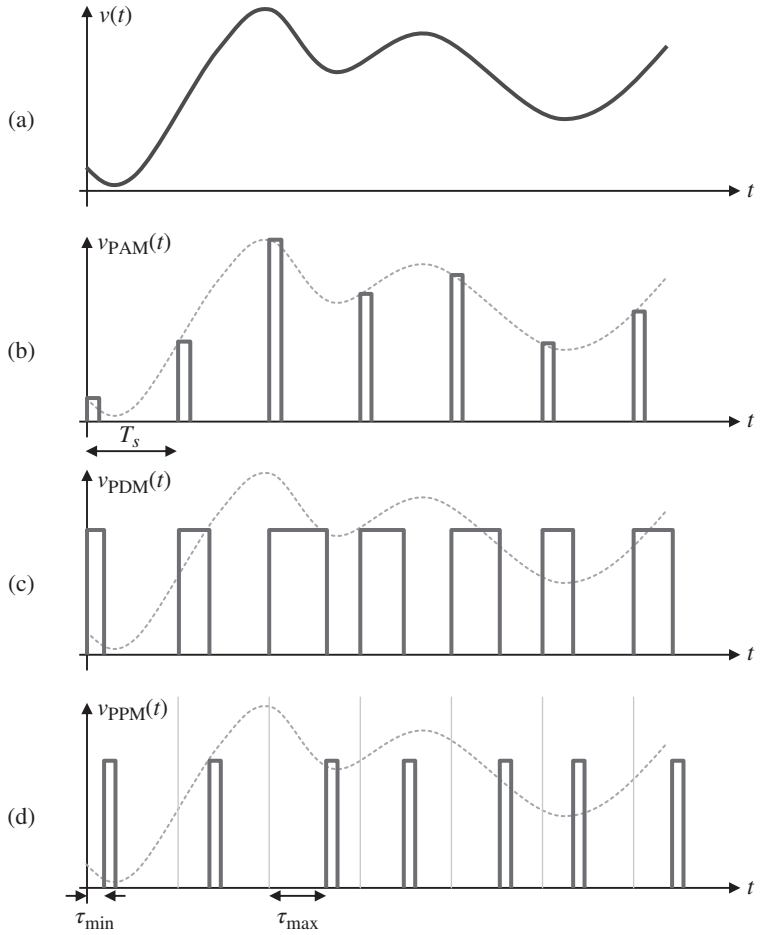


Figure 1.13 Analogue signal and its discrete representation as PAM, PDM, & PPM.

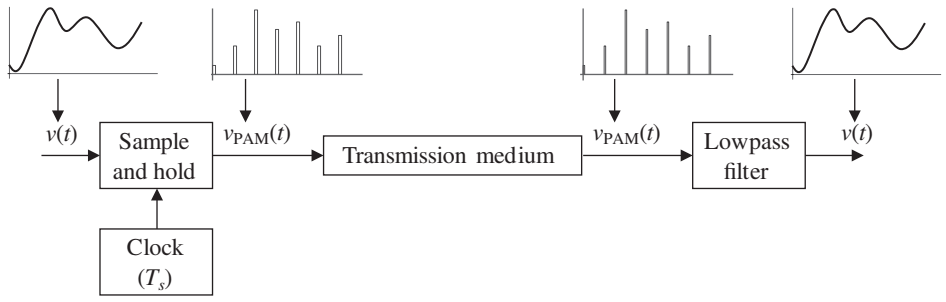


Figure 1.14 Block diagram of a PAM generator and receiver.

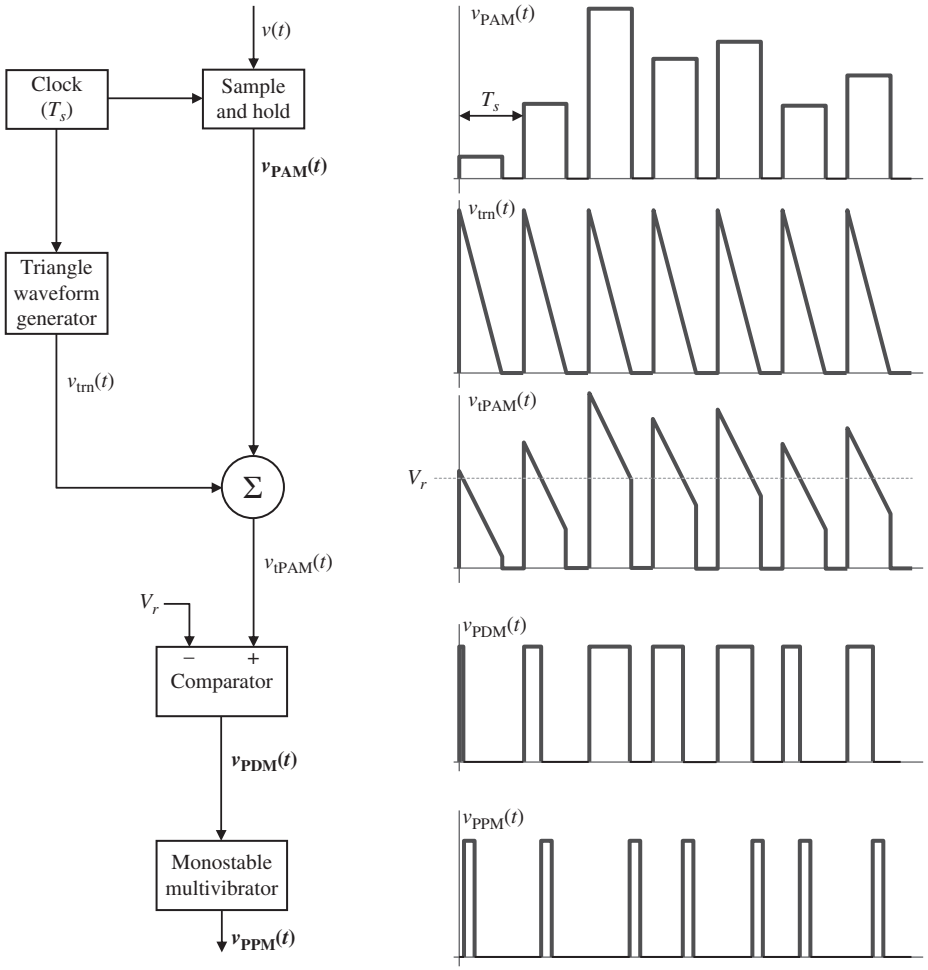


Figure 1.15 Block diagram of PDM and PPM generator.

Figure 1.14 the simplicity of the discrete baseband system, especially the PAM receiver, which is just an LPF. This filter, often referred to as a *reconstruction filter*, may have a frequency response that is shaped in such a way as to correct for a small distortion due to the action of the sample-and-hold circuit. PWM and PPM signals can be received (i.e. converted back to the original analogue signal) by using an integrator to convert pulse width or pulse position to voltage level. This process essentially converts PWM and PPM to a PAM waveform, which is then processed in an LPF to recover the original signal.

The following salient features of a discrete baseband system should be noted.

- Although it is sometimes erroneously viewed as a form of modulation employing a pulse train as the carrier, a discrete baseband communication system is actually a baseband transmission because the spectra of PAM, PDM, and PPM signals contain frequencies down to 0 Hz. In fact, the spectrum of an instantaneously sampled PAM signal is the spectrum of the original (baseband) signal plus exact duplications at regular intervals along the frequency axis.
- A discrete baseband system is an analogue communication system since the parameter of the pulse train that is varied may take on a continuum of values in a specified range. For example, the precise value of the pulse height is significant in a PAM system and any variation due to noise will distort the received signal.

- The bandwidth required to transmit a discrete baseband signal (PAM, PDM, or PPM) far exceeds the bandwidth of the original analogue signal. An instantaneously sampled PAM signal has infinite bandwidth. A discrete baseband signal fills up all the bandwidth available on the transmission medium. To share the medium among multiple signals, each must be sent in a separate time slot.
- PAM is very susceptible to noise. PDM and PPM have a better noise performance, like that of FM systems but are inferior in this respect to digital baseband transmission systems. If the pulses were perfectly rectangular then PDM and PPM would be completely immune to additive noise, as this would only alter the unused height parameter, without affecting the zero crossings of the pulse which determine pulse width or pulse location. Unfortunately, perfectly rectangular pulses, with zero rise time, are not only impossible to generate, they are impossible to maintain in a lowpass transmission medium.
- PDM is wasteful of power compared to PPM. Long pulses in PDM expend more power but carry no additional information. PPM on the other hand transmits pulses of equal energy.
- The main *advantage* of a discrete baseband system is that transmission medium sharing by multiple users, known as *multiplexing*, is possible. The intervals between the samples of one signal are used to transmit samples of other signals. This type of multiplexing is known as TDM and is further discussed below.

TDM: Figure 1.16 shows the block diagram of an N-channel TDM system that allows simultaneous transmission of several independent PAM signals over a single transmission medium. The commutator is shown as a rotating arm simply for the purpose of an easier illustration of the sampling and interleaving process. It is an electronic

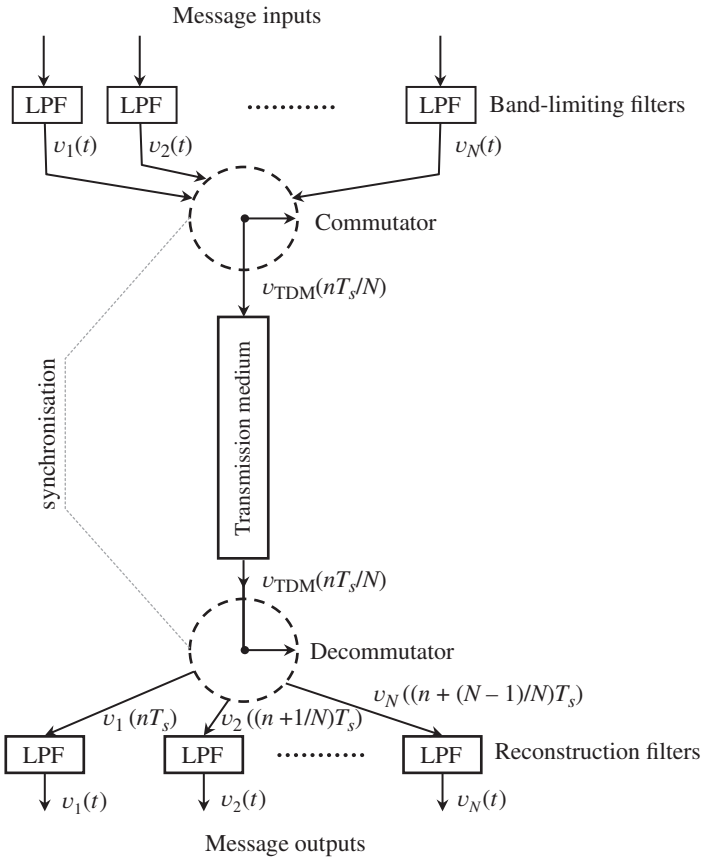


Figure 1.16 Block diagram of an analogue TDM system.

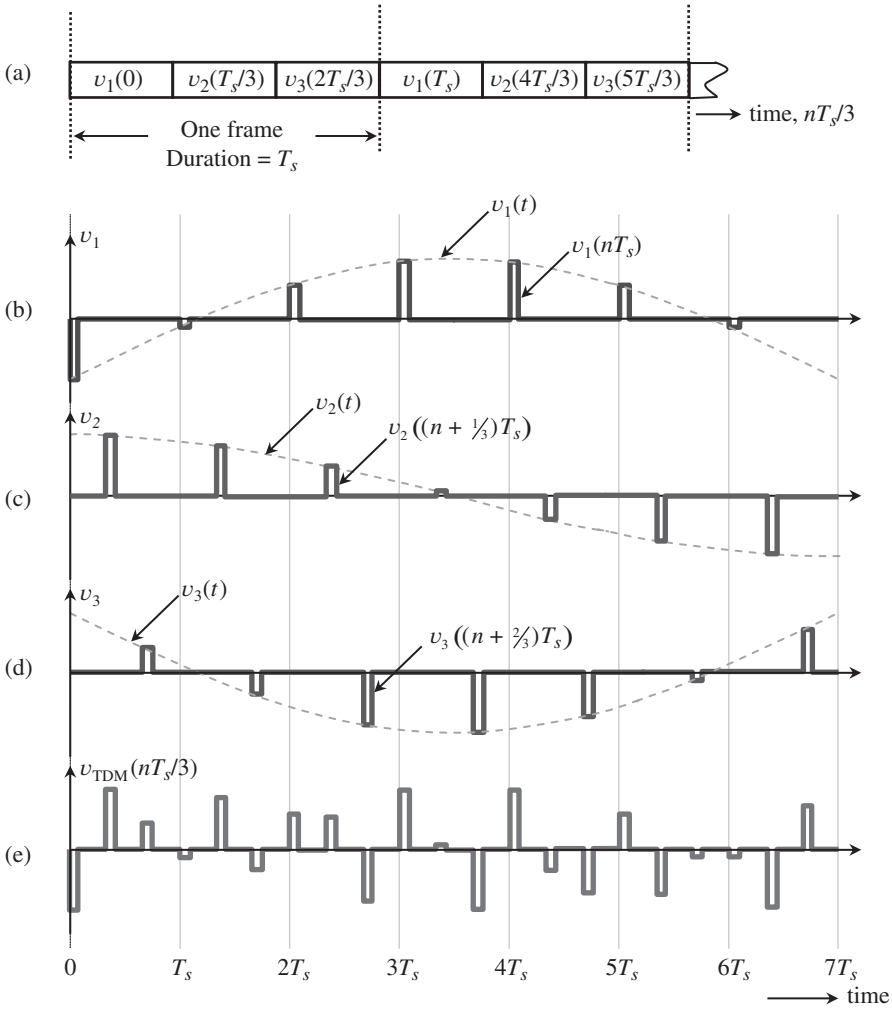


Figure 1.17 3-Channel TDM.

switching circuit that samples each input at a rate $f_s = 1/T_s$ and interleaves the N samples inside the sampling interval T_s. LPFs remove insignificant high-frequency components of the input signals and limit their bandwidth to at most $f_s/2$. These bandlimited analogue waveforms can then be correctly reconstructed at the receiver by passing their respective sequence of samples through an LPF, as shown in Figure 1.16. The waveforms referred to in Figure 1.16 are sketched in Figure 1.17 for $N = 3$. It must be emphasised that this is an analogue system. Its main advantage of simplicity is outweighed by its susceptibility to noise and distortion and therefore the analogue TDM system shown in Figures 1.16 and 1.17 is rarely used in practice. However, this system forms the basis for the TDM of digital signals, which has now become a ubiquitous technique in telecommunication systems. Some of the features of TDM, whether of PAM or digital signals, include:

- The bandwidth requirement of an N-channel TDM signal expands by a factor N, the number of multiplexed signals or channels. This happens because N samples are squeezed into one sampling interval, reducing the sampling pulse period by a factor N and hence increasing its frequency by the same factor. In practice, the bandwidth increase will be slightly larger than the factor N since some time slots must be reserved for system management and synchronisation.

- The transmitter and receiver must be synchronised in order that the interleaved samples in the TDM signal are correctly distributed by the decommutator to their respective channels.
- TDM is sensitive to transmission medium dispersion, which arises because the transmission medium differently attenuates or delays various frequency components of the transmitted pulse. As a result, the pulse may broaden out sufficiently to overlap adjacent time slots, an undesirable situation known as *intersymbol interference* (ISI).
- TDM is, however, immune to system nonlinearity as a source of crosstalk between independent signals, since at any given time instant only one signal is present. This feature is an important advantage that allows amplifiers to be operated near their maximum rating, a typically nonlinear region.

1.5.3.3 Digital Baseband Communication System

Digital signals, originating from coded textual information or from digitised analogue signals, may also be transmitted at baseband. This is a more common type of baseband transmission as it has all the advantages of the digital communication systems discussed earlier. The baseband transmitter performs line coding on the input bit stream, and the baseband receiver has the task of decoding the received (and channel-distorted) coded waveforms back into the original bit stream with minimum error. Before looking at these two operations, let us first consider how an analogue signal is converted into a digital bit stream.

Analogue-to-digital conversion (ADC): various techniques have been devised to convert analogue signals to digital. The earliest and still widespread technique is *pulse code modulation* (PCM). Figure 1.18 shows four signal processing steps involved in converting an analogue signal to a PCM signal. The LPF removes nonessential frequency components in the message signal and limits the highest frequency component to $f_{\max}$. This serves to reduce transmission bandwidth and to ensure that the sampling theorem is satisfied when the message signal is sampled (in the sample-and-hold circuit block) at a manageable rate $f_s \geq 2f_{\max}$. Next, the continuous-value sample sequence (PAM signal) is converted to a discrete-value sequence. This process is known as *quantisation*. It introduces irrecoverable errors as each sample is approximated to the nearest of a set of quantisation levels. In a practical system with many quantisation levels, the errors are referred to as *quantisation noise* since their effect is like that of white noise. If quantisation levels are uniformly spaced, we refer to this as *uniform* or *linear quantisation* and to the signal conversion process simply as uniform ADC. The term PCM is usually applied to the case of nonuniform quantisation, which spaces the quantiser levels nonuniformly in such a way as to make the *signal-to-quantisation-noise ratio* (SQNR) approximately constant over all analogue signal values. Finally, the encoder converts each quantised signal level to a unique binary word of length k bits. Note that k bits can be used to uniquely represent $N = 2^k$ quantisation levels. Taking f_s samples of the signal per second and representing each sample with k bits, the number of bits generated each second, or the *bit rate*, is given by

$$\text{Bit rate} = kf_s \text{ bits per second (b/s)} \tag{1.3}$$

In PCM telephony, $f_s = 8000$ samples per second (or Hz) and $k = 8$ bits/sample, yielding a bit rate of 64 kb/s. In certain bandwidth-limited communication services such as *mobile communications*, this PCM bit rate requirement is excessive.

Reducing the redundancy inherent in a PCM signal allows the bit rate and hence bandwidth and storage requirements to be significantly reduced. In a technique called *differential pulse code modulation* (DPCM), it is the difference $e(nT_s)$ between the actual sample and a predicted value that is quantised and encoded, rather than

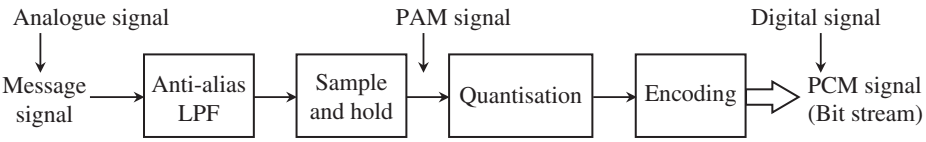


Figure 1.18 Digitisation of an analogue signal: PCM signal generation.

the sample itself. If the predictor is properly designed and an adequate sampling rate ($f_s = 1/T_s$) is used then the range of $e(nT_s)$ will be very small, allowing fewer quantisation levels and hence a smaller k (bits/sample) to be used to achieve a SQNR comparable to that of a PCM system. Assuming the same sampling rate as in PCM, we see from Eq. (1.3) that the bit rate of a DPCM system will be lower than that of a PCM system of comparable SQNR. The ITU-T has adopted for voice telephony a 32 kbit/s DPCM system, obtained by using only $k = 4$ bits to code each sample taken at the rate $f_s = 8$ kHz. A 64 kbit/s DPCM system has also been adopted for wideband audio (of 7 kHz bandwidth). This uses $k = 4$ and $f_s = 16$ kHz.

Further bit rate reduction can be achieved by using sophisticated *data compression* algorithms and, for the digitisation of speech, a range of methods known as *low bit rate speech coding*. These techniques are inherently lossy. They exploit the features of the message signal, and the characteristics of human hearing and vision, to eliminate redundant as well as insignificant information and produce a modified signal of greatly reduced bit rate. After this signal has been decompressed or otherwise processed at the receiver, a human observer finds it acceptably close to the original in quality and information content.

Line coding: whatever its origin, whether from Unicode-coded textual information or from a digitised analogue signal, we have a bit stream to be transmitted. A digital baseband transmitter chooses suitable voltage symbols (e.g. rectangular or shaped pulses) to represent the string of 1's and 0's, a process known as *line coding*, and places these symbols directly into a transmission line system. Figure 1.19 shows the line codes used in Europe for connections between equipment (often within one exchange building) in the interfaces of the digital transmission hierarchy. Line codes are designed to have certain desirable characteristics and to fulfil several important functions.

- The spectral characteristics of coded data must be matched to the characteristics or frequency response of the transmission medium. A mismatch may result in significant distortion of the transmitted voltage pulses. In particular, the code should have no DC offset. Line transmission systems are easier to design when different parts of the system are capacitor or transformer coupled to separate their DC bias voltage levels. These coupling elements pass higher-frequency (AC) voltages but block zero-frequency (DC) voltages. The coded data must therefore be void of DC content to prevent *droop* and *baseline wander*, whereby the received waveform drifts significantly relative to the decision threshold, which is 0 V for the case of a bipolar binary code. See Figure 1.20.
- The line code must combine data and timing information in one signal. It would be very expensive if a separate wire pair or coaxial cable had to be employed to carry the timing information needed at the receiver for setting decision or sampling instants. Furthermore, the line code must have a reasonable amount of clock content: the

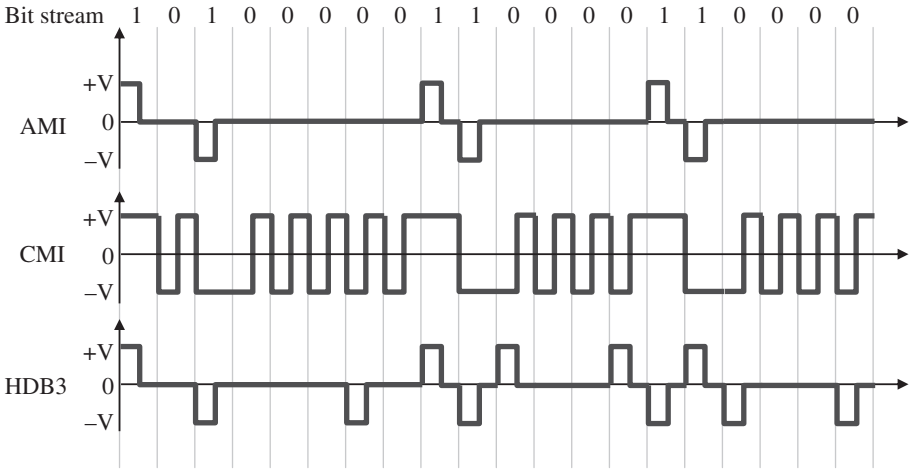


Figure 1.19 Examples of Line Codes: Alternate mark inversion (AMI), coded mark inversion (CMI), and high density bipolar with 3 zero maximum (HDB3).

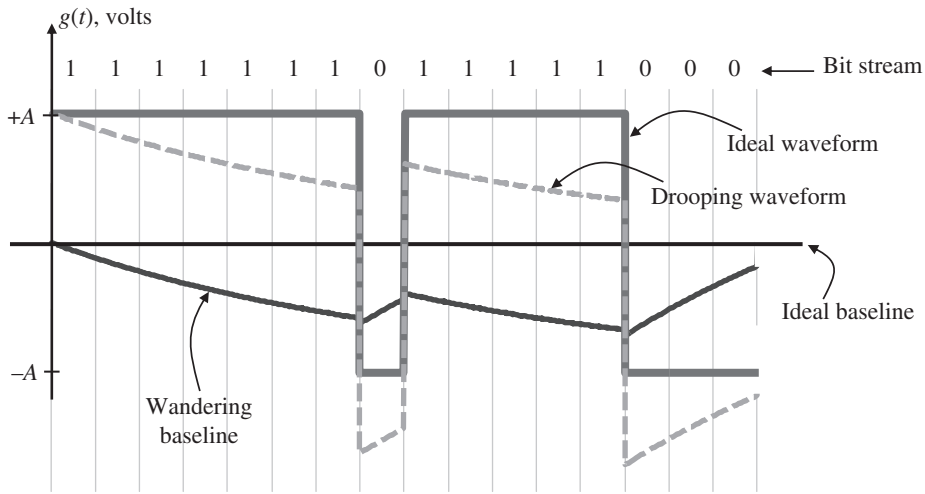


Figure 1.20 Droop and baseline wander in a binary line code.

timing content of a code is the maximum number of symbols that can occur together without a level transition – a small number indicating a high timing content. Ideally, there should be at least one transition in every symbol, but the major penalty is an increased bandwidth requirement.

- Vulnerability of the data to noise and ISI must be minimised. Sudden changes in a signal imply high frequencies in its spectrum. A rectangular pulse (with sharp transitions) transmitted through a lowpass transmission medium will spread out, with a potential for ISI. Thus, pulse shaping is frequently employed to reduce high-frequency components, which also reduces crosstalk since higher frequencies are more readily radiated. Pulse shaping also reduces the bandwidth necessary to correctly transmit the coded waveforms. The larger this bandwidth, the larger will be the amount of noise power that the receiver inevitably 'admits' in the process of receiving the waveforms.
- The line code should allow some amount of error detection. This usually involves the use of redundancy in which some codewords or symbol patterns are forbidden. A received codeword that violates the coding rule in force would then indicate some error.
- The line code should maximise code efficiency to allow a lower symbol rate to be used for a given bit rate. In long-distance cable systems, a lower symbol rate allows increased repeater spacing and reduces overall system cost. It turns out, however, that codes of high efficiency may lack certain other desirable characteristics. The code selected in practice will involve some compromise and will depend on the priorities of the system. Code efficiency is the ratio of actual information content (or bits) per code symbol to potential information content per code symbol. Potential information content per code symbol being given by $\log_2$(Code radix), where *code radix* is the number of signalling levels or voltage levels used by the code symbols. For example, the potential information content per code symbol of a binary code (radix = 2) is $\log_2(2) = 1$ bit, that of a ternary code (radix = 3) is $\log_2(3) = 1.585$ bits, and that of a quaternary code (radix = 4) is $\log_2(4) = 2$ bits, etc. Codes with higher radix can therefore convey more information per symbol, but there is increased codec complexity and a higher probability of error. Although multilevel codes (of radix $\geq$4) are very common in modulated communication systems to cope with restricted bandwidth, only codes with radix $\leq$4 are employed in baseband systems.
  One example of a quaternary code is the 2B1Q line code, which was adopted by ANSI in 1986 for use on basic ISDN lines. It is also the line code used on DSL local loops. As the name suggests, the 2B1Q code represents two binary digits using one quaternary symbol, i.e. one of four voltage levels. More specifically, the *dibits* 00, 01, 11, and 10 are represented by the voltage levels −3 V, −V, +V and +3 V, respectively.

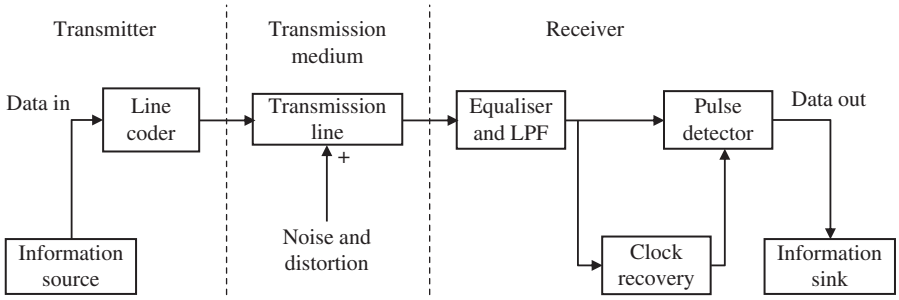


Figure 1.21 Digital baseband transmission system.

- Finally, the complexity of the encoder and decoder circuits (codec) should be kept to a minimum in order to reduce costs. In general, line codes that can be implemented by simple codecs are used for short links, whereas more efficient but complex and costly codecs are used for long-distance links because they can work with fewer repeaters and hence reduce overall system cost.

Line decoding: the transmission medium distorts the transmitted pulses by adding noise and by differently attenuating and delaying various frequency components of the pulses. A baseband receiver or repeater takes the distorted pulses as input and produces the original clean pulses at its output. It does this through three 'R' operations. First, a *reshaping* circuit comprising an equaliser and an LPF is used to reshape the pulse and ensure that its spectrum has a raised cosine shape. This operation is important to minimise ISI. Next, a *retiming* circuit recovers the clock signal from the stream of reshaped pulses. Level transitions within the pulse stream carry the clock information. Finally, a *regenerating* circuit detects the pulses at the sampling instants provided by the recovered clock signal. Occasionally, an error will occur when a noise voltage causes the pulse amplitude to cross the detection threshold. The frequency of occurrence of such errors, or BER, can be maintained at an acceptable level by ensuring that the noisy pulses are detected before the ratio of pulse energy to noise power density falls below a specified threshold. Figure 1.21 is a block diagram of a digital baseband system showing the basic operations discussed here.

1.5.3.4 Modulated Communication Systems

In very many situations it is necessary to translate the baseband signal (without distorting its information content) to a frequency band centred at a frequency f_c that is well above 0 Hz. A communication system in which the information signal undergoes this *modulation* process before being placed into the transmission medium is referred to as a *modulated communication system*. There are numerous examples of modulated systems. All satellite communication systems, mobile communication systems, radio and TV broadcast systems, radio relay systems, and optical communication systems are modulated systems.

Formally, we define *modulation* as the process of imposing the variations (or information) in a lower-frequency electrical signal (called the *modulating* or *baseband* or *message* signal) onto a higher frequency signal (called the *carrier*). The carrier signal is usually a sinusoidal signal of frequency f_c. It effectively gives the message signal a 'ride' through the transmission medium because, for several reasons, it is impossible or undesirable for the message signal to make the 'journey' on its own.

Role of modulation: there are several reasons why modulation is extensively used in modern communication:

- Modulation is used to obtain a more efficient exploitation of the transmission medium by accommodating more than one user in the same medium at the same time. In most cases, the bandwidth that is available in the medium is much larger than what is required by one user or message signal. For example, the bandwidth available on a coaxial cable is more than 10 000 times the bandwidth of one telephone speech channel. The bandwidth of an optical fibre medium exceeds that of an analogue TV signal by a factor of up to one million; and the radio

spectrum is much wider than the bandwidth required by one radio station for its broadcasts. Modulation allows the implementation of FDM in which each user's signal is placed in a separate frequency band by modulating an appropriate carrier. If the carrier frequencies are sufficiently far apart, the different signals do not interfere with each other. A signal can be recovered at the receiver by filtering (to exclude the unwanted channels) and demodulation (to detect the message signal in the carrier signal). Providers of radio services can transmit and receive within the bands allocated to them by using a suitable modulation technique.

- Modulation allows us to select a frequency that is high enough to be efficiently radiated by an antenna in radio systems. The power radiated by an antenna may be expressed as $P = I_{rms}^2 R_r$, where I_{rms} is the root mean square (rms) value of the current signal fed into the antenna and R_r is the antenna's radiation resistance. It turns out that R_r depends on the size of the antenna measured in wavelength units. In general, the size of the antenna must be at least one-tenth of the signal wavelength if the antenna is to radiate an appreciable amount of power. Consider the minimum size of antennas required to radiate signals at three different frequencies, 3 kHz, 3 MHz, and 3 GHz. The wavelengths of these signals (given by the ratio between the speed of light 3×10^8 m/s and the signal's frequency) are 100 km, 100 m, and 10 cm, respectively. Thus, if we attempted to radiate a 3 kHz speech signal, we would need an antenna that is at least 10 km long. Not only is such an antenna prohibitively expensive, it is hardly suited to portable applications such as in handheld mobile telephone units. If, on the other hand, we use our 3 kHz speech signal to modulate a 3 GHz carrier signal then it can be efficiently radiated using very small and hence affordable antennas of minimum size 1 cm.
- The use of modulation to transmit at higher frequencies also provides a further important advantage. It allows us to exploit the higher bandwidths available at the top end of the radio spectrum in order to accommodate more users or to transmit signals of large bandwidth. For example, an AM radio signal has a bandwidth of 10 kHz. Thus, the maximum number of AM radio stations that can be operated at low frequency (LF ≡30–300 kHz) in one locality is given by

$$\text{Maximum number of AM stations at LF} = \frac{300 - 30}{10} = 27$$

  Observe that there is a tenfold increase in the number of AM radio stations when we move up by just one band to medium frequency (MF ≡ 300−3000 kHz)

$$\text{Maximum number of AM stations at MF} = \frac{3000 - 300}{10} = 270$$

  The NTSC television signal requires a radio frequency bandwidth of 6 MHz. Frequency bands at MF and below can therefore not be used for TV transmission because they do not have enough bandwidth to accommodate the signal. The high frequency (HF ≡ 3−30 MHz) band can accommodate a maximum of only four such TV channels. However, as we move to higher bands, we can accommodate 45 TV channels at very high frequency (VHF ≡ 30−300 MHz), 450 TV channels at ultra high frequency (UHF ≡ 300−3000 MHz), 4500 TV channels at super high frequency (SHF ≡ 3−30 GHz), 45000 TV channels at extra high frequency (EHF ≡ 30−300 GHz); and the optical frequency band (300 GHz − 860 THz) can accommodate literally millions of TV channels.
- Another important function of modulation is that it allows us to transmit at a frequency that is best suited to the transmission medium. The behaviour of all practical transmission media is frequency dependent. Some frequency bands are passed with minimum distortion, some are heavily distorted, and some are blocked altogether. Modulation provides us with the means of placing the signal within a band of frequencies where noise, signal distortion, and attenuation are at an acceptable level within the transmission medium. Satellite communication was pioneered in C-band inside the 1–10 GHz window where both noise (celestial and atmospheric) and propagation impairments are minimum. Ionospheric reflection and absorption become increasingly significant the lower you go below this band, until at about 12 MHz the signal is completely blocked by the ionosphere. Furthermore, attenuation by tropospheric constituents such as rain, atmospheric gases, fog, cloud water droplets, etc., becomes significant and eventually very severe at higher-frequency bands. For this reason, modulation must be used in satellite communication to translate the baseband signal to a congenial higher-frequency band.

- Another example of a bandpass transmission medium is the optical fibre medium. This blocks signals at radio wave frequencies, but passes signals in the near-infrared band, particularly the frequencies around 194 and 231 THz. Thus, the only way to use this valuable medium for information transmission is to modulate an optical carrier signal with the baseband information signal.

Types of modulation: there are three basic methods of modulation depending on which parameter of the carrier signal is varied (or modulated) by the message signal. Consider the general expression for a sinusoidal carrier signal

$$v_c(t) = A_c \cos(2\pi f_c t + \phi) \tag{1.4}$$

There are three parameters of the carrier, which may be varied in step with the message signal. The (unmodulated) carrier signal $v_c(t)$ has a constant amplitude A_c, a constant frequency f_c, and a constant initial phase ϕ. Varying the amplitude according to the variations of the message signal, while maintaining the other parameters constant, gives what is known as *amplitude modulation* (AM). *Frequency modulation* is obtained by varying the frequency f_c of a constant-amplitude carrier, and *phase modulation* is the result of varying only the phase ϕ of the carrier signal.

An analogue modulating signal will cause a continuous variation of the carrier parameter, the precise value of the varied parameter being significant always. It is obvious that this is then an analogue modulation and the resulting system is an *analogue modulated communication system*.

A digital modulating signal (consisting of a string of binary 1's and 0's) will, on the other hand, cause the carrier parameter to change (or shift) in discrete steps. Information is then conveyed, not in the continuous precise value of the parameter but rather in the interval of the parameter value at discrete decision (or sampling) instants. This is therefore digital modulation and the resulting system is a *digital modulated communication system* having all the advantages of digital communication. In this case, the three modulation methods are given the special names *amplitude shift keying* (ASK), *frequency shift keying* (FSK), and *phase shift keying* (PSK) to emphasise that the parameters are varied in discrete steps. The number of steps of the parameter generally determines the complexity of the digital modulation scheme. The simplest situation, and the one most robust to noise, is *binary modulation* (or binary shift keying) where the carrier parameter can take on one of two values (or steps). This is shown in Figure 1.22.

To transmit a bit stream using binary ASK, we take one bit at each clock instant and transmit a sinusoidal carrier of frequency f_c for the duration of the bit. The carrier frequency f_c is chosen to suit the transmission medium. The carrier amplitude is set to A_1 for bit '1' and to A_0 for bit '0'. If either A_1 or A_0 is zero, we have a special type of binary ASK known as *on–off keying* (OOK), which is the digital modulation scheme used in optical fibre communication. In binary FSK, same-amplitude and same-phase carriers of frequencies f_1 and f_0 are transmitted for bit 1 and bit 0, respectively. In binary PSK, the carrier amplitude and frequency are fixed, but the carrier is transmitted with a phase ϕ_1 for bit 1 and ϕ_0 for bit 0. Usually, the difference between the two phases is 180° in order to use the maximum possible spacing between them.

A useful way to look at digital modulation is to treat each transmitted carrier state as a code symbol. Thus, OOK represents bits 1 and 0 with the symbols S_{OOK1} and S_{OOK0} shown in Figure 1.22d; FSK uses the symbols S_{FSK1} and S_{FSK0} shown in Figure 1.22e; and PSK uses the symbols S_{PSK1} and S_{PSK0} shown in Figure 1.22f. The symbol rate (or signalling rate) is limited by the transmission bandwidth available. The theoretical maximum symbol rate is twice the transmission bandwidth. In a telephone channel with a bandwidth of about 3.1 kHz, the maximum possible symbol rate is therefore 6200 Bd.

The main drawback of binary shift keying is that bit rate and symbol rate are equal since each symbol represents one bit. In many applications, a much higher bit rate is desired than can be obtained by increasing the symbol rate, which is limited by the available bandwidth. A higher bit rate can be achieved by using multilevel (also called *M-ary*, pronounced em-aary) digital modulation, in which M distinct carrier states are transmitted. Binary

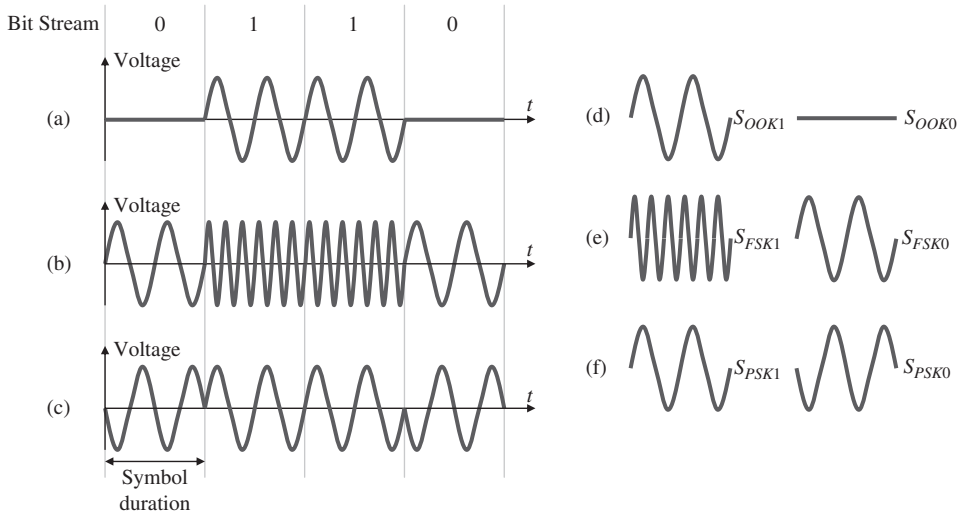


Figure 1.22 Binary digital modulation schemes: (a) On-Off Keying (OOK), a special type of ASK; (b) Frequency Shift Keying (FSK); (c) Phase Shift Keying (PSK); (d) OOK symbols; (e) FSK symbols; (f) PSK symbols.

modulation is the special case $M = 2$. To do this, we take a group of k bits at a time and represent them with a unique carrier state or code symbol, where $k = \log_2 M$, or $M = 2^k$. For example, taking $k = 3$ bits at a time, we have $M = 2^3 = 8$ possible states, namely 000, 001, 010, 011, 100, 101, 110, and 111, which must each be represented by a unique carrier state. Each symbol now carries three bits of information and the bit rate is therefore three times the symbol rate. In general, M-ary modulation increases the bit rate according to the relation

$$\text{Bit rate} = \log_2 M \times (\text{Symbol rate}) \tag{1.5}$$

However, the symbol states are closer together than in binary modulation, making it easier for noise effects to shift one state sufficiently close to an adjacent state to cause a symbol error. For example, the phase difference between adjacent states in an M-ary PSK is $(360/M)°$, which for $M = 16$ is only 22.5°. Bearing in mind that these PSK states all have the same energy, carrier amplitude being constant, we see that there is great potential for error if the transmission medium is prone to phase distortion. A combination of *amplitude and phase shift keying* (APSK), also called *quadrature amplitude modulation* (QAM), is often used, which increases the phase difference between symbol states. For example, it is possible to design for the minimum phase difference between states of equal energy in 16-APSK to be 90°, compared to only 22.5° in 16-PSK.

1.5.4 Circuit Versus Packet Switching

A transmission path is required for information to flow between two users (i.e. between source and destination) in a communication system. It is only in a tiny minority of simple communication systems that this path is a permanent setup. For the rest, the communication system is a network of transceivers or nodes, and a path is a connected sequence of links which must be found as needed to convey information from the sending node through one or more intervening nodes to the destination node. We may classify communication systems either as *circuit switched* or *packet switched* according to how such a path is established.

In both cases, switching is required at various points in the network (e.g. at a local exchange in a PSTN) to establish the concatenation of links that constitute the transmission path. This switching may be implemented in space or time or as a combination of the two. In *space switching*, a message arriving on one input port of the switch is routed or connected in its entirety to a designated output port or line of the switch without buffering. In *time*

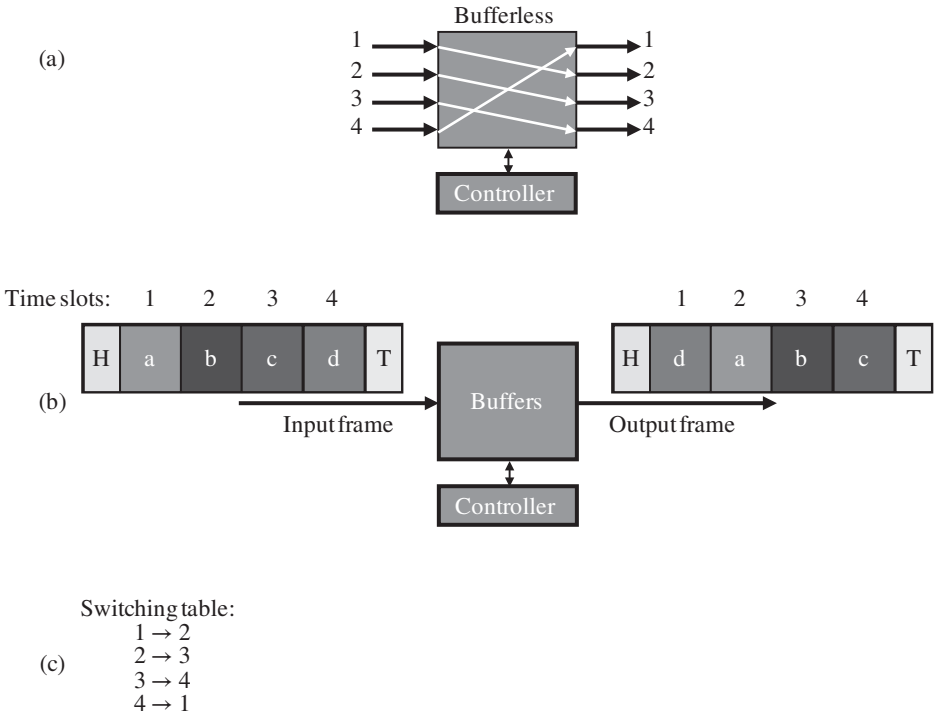


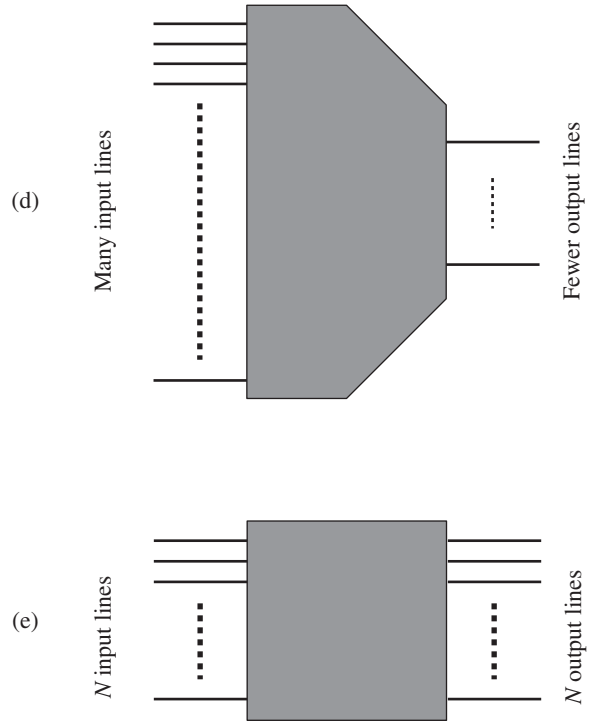
Figure 1.23 (a) Space switching; (b) Time switching; (c) Switching table; (d) Concentrator switching; (e) Route switching.

switching, the input and output of the switch are multi-slot data frames. The input data frame is buffered, and the contents of its time slots are switched to designated respective time slots in the output frame. The specification of input and output pairings is usually given in a routing table. Furthermore, space switching may be either *concentrator switching*, in which the switch has a larger number of input than output lines, or it may be *route switching*, where the switch has the same number of input and output lines. A concentrator switch is typically used to switch subscriber lines at a local exchange. It improves line occupancy but introduces the possibility of *lost* or *blocked calls* if an input message arrives at the switch while all output lines are engaged. Route switching on the other hand is typically used when switching trunk lines between exchanges. Figure 1.23 illustrates space and time switching in (a) and (b), respectively, under the control of the same switching table given in (c). The slots labelled H and T in the data frames of (b) are the header and trailer bits of each frame. Note, for example, that the contents of time slot 1 in the input frame are identified with the letter 'a' and these contents are switched to time slot 2 in the output frame according to the specification of the routing table. Concentrator and route switching are also illustrated in Figure 1.23d and e, respectively.

1.5.4.1 Circuit Switching

In a circuit switched communication system, before communication begins between two users, a dedicated channel or trunk or transmission path is set up end-to-end for the exclusive use of the two users for the entirety of their connection. The dedicated channel may be a separate frequency band or time slot or code sequence or combinations of these, depending on the channel access plan in operation. If a free path or circuit cannot be set up for a new connection between a pair of users then the requesting user is either denied access (i.e. blocked) or is placed on a queue until a free circuit becomes available.

Figure 1.23 (*Continued*)



The circuit switching approach described above has several advantages.

- Once a path has been established, i.e. once the call setup phase has been completed, a message will transit quickly through the system. There are no extra delays at intermediate nodes arising from queuing and the time taken to absorb portions of the message or its packets into a buffer at each node.
- The transmission system or network can manage congestion arising from excessive demand for simultaneous connections. New connection requests are simply denied or placed on a queue to protect the quality of active connections.
- The message has a smooth transit through the network from source to destination. There is no jitter arising from variations in the transit time of portions of the message, which would be the case if sections of the path were shared with other active connections. Such jitter is not desirable in real-time voice and video communications.
- There is no need for extra bits (called *overhead bits*) to be added to each portion of the message to convey source and destination addresses and sequencing information (for use in reassembling the message in the right order at the destination). As a result, the number of bits conveyed through the network is significantly reduced.
- The overall computational and processing demand on the network is low since the message will transit seamlessly through the network without the need for any further processing at intermediate nodes once a path has been established.

However, circuit switching has several significant shortcomings.

- It can lead to a highly inefficient utilisation of network resources. The dedicated channel cannot be used for another call during idle intervals of the active call. An interactive phone conversation between two people will be punctuated by periods of silence and an Internet browsing session may contain long intervals of inactivity, for example while examining downloaded content. The circuit switching approach locks away and wastes channel capacity during such idle intervals.

- The call setup phase, when a free channel is being arranged prior to commencement of the call, may be responsible for a significant portion of the delay from the time a request is made for a channel to the time the message is fully delivered at the destination. In fact, for short messages the call setup phase will account for the bulk of the delay, which may lead to an unacceptably low *data throughput* – this being the number of bits successfully delivered divided by the total time taken.
- The dedicated connection provides for a constant data rate between the two users, and this makes it difficult to interconnect users with a variety of data rates.
- The impact of a technical failure or breakdown in any portion of the path or route from source to destination is catastrophic and results in the termination of the active call. Once a call has been set up, circuit switching has no mechanism to modify the route mid-call. The probability of call termination is therefore increased on (e.g. long-distance) connections involving the concatenation of many links since the failure of any one link will lead to call termination.

The features itemised above make circuit switching the ideal approach to ensure the best service quality for real-time voice and video communication through a network with reliable constituents. The approach is also suitable for transmitting sufficiently long messages of whatever content or type through such networks. However, a circuit switched network could be likened to a community that generously divides its cake among the few. Packet switching offers the possibility of a different approach in which everyone can have a smaller piece of the community cake.

1.5.4.2 Packet Switching

In a packet-switched communication system or network, each message (which must be in digital form) is first broken into one or more packets of bits prior to transmission. Overhead bits in the form of a packet header are added to each packet to fulfil a variety of purposes, such as identification (e.g. source and/or destination address and/or route ID), error control, and transmission and network management functions. A data *packet* therefore consists of two sections, namely a *header* made up of overhead bits and a *payload* made up of source bits. Note that the source bits will include message bits, any redundant bits inserted by the channel encoder for error control, and overhead bits from any previous round of data packetisation or framing.

If the packets are of fixed length (usually specified in bytes ≡ 8 bits) they are referred to as *cells*. An example of this fixed-length packet approach is ATM switching. An ATM cell is of length 53 bytes, which consists of a 5-byte header and a 48-byte payload. In other cases, such as IP (*Internet protocol*) switching, the packets are of variable length. IPv4 (version 4) has a header length that is typically 24 bytes but may vary from 20 to 60 bytes depending on header options. Since a path may be a concatenation of different link types, each with its own maximum transmission unit (MTU) which is the maximum number of bytes in a packet that the link can support, and the lowest MTU is typically 576 bytes for all link types, the recommended maximum IPv4 packet length is therefore 576 bytes to avoid packet fragmentation in transit. The latest IP version is IPv6, which has a fixed 40-byte header and a variable payload length constrained by a lowest MTU spec of 1280 bytes. IPv4 makes a 32-bit provision for source address (and similarly for destination address), which is enough to uniquely identify $2^{32} \approx 4.3$ billion items such as nodes. In the early days of computer networking this provision was thought to be more than enough, but with the explosion of the Internet, including IoT (the Internet of Things), that early thinking proved extremely short-sighted. IPv6 therefore makes a 128-bit address provision, which is enough to uniquely identify 340 undecillion items—one undecillion being one trillion trillion trillion, i.e. one followed by 36 zeroes. IPv6 is thus considered the ultimate IP version with enough capacity for all future needs, but its uptake has been slow due to the need to invest in new Internet routers.

There are two modes of packet switching, namely *connection-oriented* (CO) packet switching, also called *virtual circuit*, and *connectionless* (CL) packet switching, sometimes called *datagram*.

Connection-oriented (CO) packet switching: in CO packet switching, a free but nonexclusive path is first set up through the network from source to destination and all user packets follow this path and arrive at the destination strictly in the same order as they left the source. It is important to observe that, unlike circuit

switching, the path set up prior to the commencement of transmission in this virtual-circuit approach is not a dedicated path. Other packets from other users may simultaneously use portions of the same path to reach separate destinations. Therefore, packets may need to be queued at each node until their designated outgoing link is free. Each packet header carries a logical connection identifier or *virtual circuit identifier* (VCI), which identifies a predefined route to be followed by the packets from source to destination. Intervening nodes do not therefore make any routing decisions.

Examples of connection-oriented packet switching include X.25 – the first public data network launched in the 1970s, Frame Relay – a wide area network (WAN) protocol adopted in the 1980s that largely replaced X.25, and ATM – launched in the 1990s with a lot of hype but now, due to the runaway global success of IP, mainly confined to use by telephone carriers for high-speed internal data transport. CO packet switching addresses some of the shortcomings of the circuit-switching approach and delivers the following improvements (when compared to circuit switching):

- CO packet switching allows node-to-node links to be dynamically shared by multiple users over time so that network resources are used much more efficiently.
- It supports the exchange of packets between two users operating at different data rates.
- It allows new calls to be accepted during heavy traffic, although with an increased packet transfer delay, and hence reduced data throughput, due to a higher incidence of queuing at intermediate nodes. Circuit switching would simply block new calls during heavy traffic if a free path cannot be established.
- A prioritisation scheme may be efficiently implemented whereby important packets (e.g. those conveying network status information) are catapulted to the front of the queue at each node so that they have a much faster transit through the network. The only way to implement such prioritisation in a circuit-switched approach would be to dedicate some paths for exclusive use by priority packets, but this would be inefficient and wasteful because such packets are few and far between. Think of it this way: you wouldn't build exclusive roads for ambulance vehicles and fire engines because there are too few of them for this to be cost-effective. However, the very fact that there are only a few of them means they can be prioritised and even allowed to interrupt other road users' green lights at road junctions. In this way, their journey time is massively reduced, yet the added delay to other road users is negligible.

However, in addition to retaining the drawbacks of circuit switching associated with requiring an initial call setup phase and the impact of failure of any portion of the path after call commencement, CO packet switching introduces the following disadvantages (when compared to circuit switching):

- CO packet switching increases the transit time of messages due to queuing delays at network nodes. As a minimum, a node delays each packet by the time taken to fully receive the packet into the node's buffer before routing to the designated output port can commence.
- It introduces a variation in overall packet delay (called *jitter*), which is not desirable for real-time voice and video. Although all packets of one message connection follow the same route, the transit time of each of these packets will be a function of time as the queue state and workload of shared nodes change due to varying network conditions.
- It increases the number of bits transported in the network because a message is broken into packets, and to each of these packets a header is added. For example, in ATM, around 10% of all transported bits are overhead bits from the header.
- There is increased processing at network nodes. Although nodes do not perform independent routing, each node must at least read the VCI of each packet in order to correctly forward the packet according to the route established when the circuit was set up.

Nevertheless, because CO packet switching solved what was the biggest shortcoming of circuit switching, namely the inefficient utilisation of network resources, it was an important step towards building a global high-capacity integrated digital network.

Connectionless (CL) packet switching: in CL packet switching, there is no initial call or connection setup phase. Packets are sent as individual datagrams, each containing source and destination IDs in its header. The packets are independently routed through the network at each router, which makes its own decisions on a packet-by-packet basis according to a dynamic routing table that is regularly updated as network conditions change.

IP is a CL packet switching implementation which has become the single most dominant networking technology and has left its closest rivals, such as ATM, in relative obscurity. As the name implies, IP is the routing protocol that powers the entire Internet and since 4G in 2011 it was also adopted for mobile broadband connectivity. In the 1990s, phrases like *voice over Internet protocol* (VoIP) and *television over Internet protocol* (TVoIP) were coined, but such phrases have become superfluous because it is now everything over IP. CL packet switching addresses some of the shortcomings of the virtual circuit packet switching approach, including the following improvements (when compared to virtual circuit switching):

- CL packet switching has no call setup phase, so message transmission through the network is an instant-start process. Therefore, the delay associated with circuit setup is eliminated and transmission is much quicker for users sending only a few packets, which is often the case in social media interactions and SMS (short messaging service).
- It is much more flexible and agile in reacting to network conditions and, unlike virtual circuits, can do so mid-call or mid-connection. If congestion develops at a given node, subsequent packets from the source can be routed away from that node, whereas in virtual circuits all packets of a call must follow a predefined route, which may include the congested node.
- It is also inherently more reliable. If a node fails then only packets that are lost in the crash are affected. A route may be easily found through the network for subsequent packets to bypass the failed node. In virtual circuits, all connections set up prior to the node failure and that include the failed node in their paths would continue to attempt to pass through the node after its failure and would therefore be terminated.

However, CL packet switching does introduce the following disadvantages (when compared to virtual circuit switching):

- CL packet switching causes packets to transit the network more slowly since routing decisions must be made at every node. It is therefore a slower mechanism for users sending many packets. It could be said that CL is an instant-start jogger, whereas CO is a slow-start sprinter.
- Packets may follow different routes to the same destination and therefore may arrive out of order. This necessitates extra provision in the CL packet header and processing at the destination to allow received packets to be sorted into the right order to form the correct message.
- It is inherently difficult in CL packet switching to implement measures that guarantee a desired quality of service (QoS) and achieve congestion control. CL being an instant-start algorithm, it means that any source can instantly start sending packets through the network at any time. Each such new entrant can contribute towards both network congestion and degrading the QoS of existing users. In contrast, QoS and congestion control can be more readily managed in virtual circuit switching by negotiating the allocation of enough resources in advance for each connection during the call setup phase.

Despite the above shortcomings, CL's unmatched routing flexibility and elimination of the significant delay overhead in call set-up have made it the undisputed technology of choice for the ubiquitous broadband communication networks of the twenty-first century, including the Internet and 4G mobile networks and beyond. The significant weakness of CL packet switching in relation to congestion control and QoS have been cleverly mitigated by overlaying it with smart oversight and control algorithms such as transmission control protocol (TCP), real-time transport protocol (RTP), and user datagram protocol (UDP). For example, TCP provides the congestion and flow control mechanisms with which IP is supervised to ensure a smooth and reliable data transfer between sender and receiver.

1.6 Epilogue

Our detailed overview of communication systems is now complete. We had a twofold purpose in this chapter. First, we wanted to carefully and nonmathematically lay a foundation of communication engineering and erect important knowledge pegs on which you can hang the more detailed study presented in subsequent chapters. Second, we wanted to make a brief and highly selective historical sketch of the journey of telecom from telegraphy in 1837 to ubiquitous broadband Internet and 5G mobile communication in 2020. We believe that your study of communication engineering should begin with allowing yourself to be informed and inspired in equal measure by this history. We hope that seeing such potential for seemingly endless progress will stir in you a genuine hunger for competence in the subject and for a complete mastery of its principles and concepts. We hope that you will go on to become one of the architects of a yet-unseen broadband communication future.

Back in the 1940s during the heyday of copper line communication, who saw the rise of optical fibre communication using glass as a transmission medium? In the 1950s when the launch of *Sputnik 1* set off the space race, who saw today's satnav ubiquity? When computer networking began with ARPANET in 1971, who imagined broadband Internet and its social media offspring? And today, who sees the future of telecom? Is there anything beyond today's smartphone? Will it be an artificial intelligence device (AID)? Is there anything beyond today's *multimedia* broadband Internet with its predominantly audio-visual capability? Will it be a *multisensory* Internet that adds tactile and olfactory functionality so that after 150 years of telecoms it can finally directly cater to our senses of touch and smell as much as to our sight and hearing?

We hope that this chapter has inspired you to freely imagine our telecom future and that subsequent chapters will equip you with the skills to help shape it. We will now embark on the latter task and begin a more in-depth study of the subject in the next chapter with a detailed introduction to signals and systems in the context of telecommunications.

References

1 Live-Counter.com (2020). Worldwide Automobile Productions. https://www.live-counter.com/number-of-cars (accessed on 11th June 2020).

2 Statista (2020). Number of smartphone users worldwide from 2016 to 2021 (in billions). https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide (accessed on 11th June 2020).

3 Gartner (2020). Gartner says worldwide smartphone sales will grow 3% in 2020. https://www.gartner.com/en/newsroom/press-releases/2020-01-28-gartner-says-worldwide-smartphone-sales-will-grow-3 (accessed 11th June 2020).

Review Questions

1.1 (a) Discuss the drawbacks of verbal and visual nonelectrical telecommunications.
(b) In spite of its many drawbacks, nonelectrical telecommunication remains indispensable in society. Discuss various situations in which nonelectrical telecommunication has some important advantage over (modern) electrical telecommunications. In your discussion, identify the type of nonelectrical telecommunication and the advantage it provides in each situation.

1.2 Sketch the flags that represent the following messages in semaphore code. Remember to include an end of signal code in each case.
(a) NO
(b) TAKE 5.

1.3 Sketch the voltage pulse sequence for the Morse code representation of each of the following complete messages.
(a) WE WON 2-0
(b) I LOVE YOU.

1.4 Baudot code (Appendix, Table A.2) requires seven transmitted bits per character, which includes a start bit (binary 0) and a stop bit (binary 1). Write out the complete bit sequence for the transmission (least significant bit (LSB) first) of the following messages using Baudot code.
(a) DON'T GIVE UP
(b) 7E;Q8.

1.5 In asynchronous transmission of ASCII-coded data, characters are transmitted one frame at a time starting with the LSB of the character. A frame is formed as follows: (i) take the 7-bit ASCII code for the character; (ii) insert one odd parity check bit in the MSB position to form a byte; (iii) write the byte from LSB to MSB, so transmission will start from LSB then insert bit 0 as a frame header (i.e. start bit) and bits 11 as a frame trailer (i.e. stop bits) to complete the frame.
Repeat Question 1.4 for asynchronous transmission of the same messages using ASCII code. Compare the number of bits required by Baudot and ASCII codes in each case and comment on your results. Calculate the transmission efficiency of both coding schemes in (a) and (b).

1.6 The last century witnessed revolutionary developments in telecommunications far beyond what could have been contemplated by the pioneers of telegraphy. Discuss four key areas of these developments.

1.7 Draw a clearly labelled block diagram that is representative of all modern communication systems. List 12 different devices that could serve as the information source of a communication system. Identify a suitable information sink that may be used at the receiver in conjunction with each of these sources.

1.8 With the aid of a suitable diagram where possible, discuss the operation of the following information sources or sinks.
(a) Dynamic microphone
(b) Loudspeaker
(c) CRT
(d) Plasma-panel display
(e) Liquid crystal display.

1.9 Figure 1.8 shows the ISO standard 226 contours of equal loudness as a function of frequency and SPL.
(a) What is the loudness of sound of frequency 200 Hz and SPL 40 dB?
(b) If a tone of SPL 55 dB is perceived at a loudness 50 phon, what are the possible frequencies of the tone?
(c) Determine the amount by which the vibration (i.e. SPL in dB) of a 50 Hz tone would have to be increased above that of a 1 kHz tone in order that both tones have equal loudness 10 phon.

1.10 Discuss the signal processing tasks performed by the transmitter in a communication system. Indicate why each process is required and how it is reversed at the receiver to recover the original message signal.

1.11 Audio and television broadcasting have both gone digital and there has been a virtually complete digitalisation of the telecommunication networks of most countries. Examine the reasons for this trend of

digitalisation by presenting a detailed discussion, with examples where possible of the advantages and disadvantages of digital communication. Indicate in your discussion how the impact of each of the disadvantages can be minimised in practical systems.

1.12 Give **two** examples of each of the following types of communication systems:
(a) Simplex system
(b) Half-duplex system
(c) Duplex system
(d) Analogue baseband
(e) Analogue modulated
(f) Digital baseband
(g) Digital modulated.

1.13 (a) With the aid of suitable block diagrams, discuss the operation of any **two** examples of an analogue baseband communication system.
(b) What are the most significant disadvantages of an analogue baseband communication system?

1.14 With the aid of suitable block diagrams, discuss the generation of the following discrete baseband signals starting from an analogue message signal.
(a) PAM
(b) PDM
(c) PPM.

1.15 Compare the discrete baseband systems of Question 1.14 in terms of noise performance, bandwidth requirement, power consumption, and circuit complexity.

1.16 Explain how three independent user signals can be simultaneously carried on one transmission link using discrete baseband techniques. What are the major drawbacks of this system?

1.17 Using a suitable block diagram, discuss the steps involved in the ADC process. Identify the major parameters that must be specified by the ADC system designer and explain the considerations involved.

1.18 Give a detailed discussion of the desirable characteristics of line codes, which are used to electrically represent bit streams in digital baseband communication systems.

1.19 Discuss the three 'R' operations of a digital baseband receiver.

1.20 Discuss the roles of modulation in communication systems. Hence, identify the transmission media that can be used for voice communication in the absence of modulation.

1.21 Sketch the resulting waveform when the bit stream 10111001 modulates a suitable carrier using:
(a) OOK
(b) Binary FSK
(c) Binary PSK.
Assume that the carrier frequency is always an integer multiple of $1/T_b$, where T_b is the bit interval.

1.22 If the following digital communication systems operate at the same symbol rate of 4 kBd, determine the bit rate of each system.
(a) OOK
(b) 2B1Q
(c) Binary FSK
(d) 16-APSK.

1.23 Discuss the advantages of circuit switching over packet switching and identify the scenarios where circuit switching would be the preferred technique.

1.24 Discuss the mechanisms used by TCP to ensure a reliable service even though the IP network underneath may be unreliable.