1 Introduction

Signal processing is a discipline concerned with the acquisition, representation, manipulation, and transformation of signals required in a wide range of practical applications. In this chapter, we introduce the concepts of signals, systems, and signal processing. We first discuss different classes of signals, based on their mathematical and physical representations. Then, we focus on continuous-time and discrete-time signals and the systems required for their processing: continuous-time systems, discrete-time systems, and interface systems between these classes of signal. We continue with a discussion of analog signal processing, digital signal processing, and a brief outline of the book.

Study objectives

After studying this chapter you should be able to:

- Understand the concept of signal and explain the differences between continuous-time, discrete-time, and digital signals.
- Explain how the physical representation of signals influences their mathematical representation and vice versa.
- Explain the concepts of continuous-time and discrete-time systems and justify the need for interface systems between the analog and digital worlds.
- Recognize the differences between analog and digital signal processing and explain the key advantages of digital over analog processing.

1.1 Signals

For our purposes a *signal* is defined as any physical quantity that varies as a function of time, space, or any other variable or variables. Signals convey information in their patterns of variation. The manipulation of this information involves the acquisition, storage, transmission, and transformation of signals.

There are many signals that could be used as examples in this section. However, we shall restrict our attention to a few signals that can be used to illustrate several important concepts and they will be useful in later chapters. The speech signal, shown as a *time waveform* in Figure 1.1, represents the variations of acoustic pressure converted into an electric signal by a microphone. We note that different sounds correspond to different patterns of temporal pressure variation.

To better understand the nature of and differences between analog and digital signal processing, we shall use an analog system which is near extinction and probably unknown to many readers. This is the magnetic tape system, used for recording and playback of sounds such as speech or music, shown in Figure 1.2(a). The recording process and playback process, which is the inverse of the recording process, involve the following steps:

- Sound waves are picked up by a microphone and converted to a small analog voltage called the audio signal.
- The audio signal, which varies continuously to "mimic" the volume and frequency of the sound waves, is amplified and then converted to a magnetic field by the recording head.
- As the magnetic tape moves under the head, the intensity of the magnetic field is recorded ("stored") on the tape.
- As the magnetic tape moves under the read head, the magnetic field on the tape is converted to an electrical signal, which is applied to a linear amplifier.
- The output of the amplifier goes to the speaker, which changes the amplified audio signal back to sound waves. The volume of the reproduced sound waves is controlled by the amplifier.

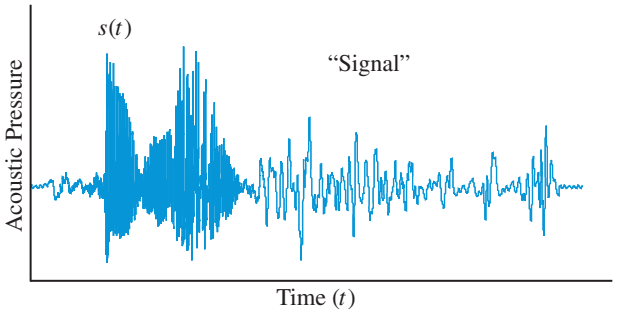

Figure 1.1 Example of a recording of speech. The time waveform shows the variation of acoustic pressure as a function $s(t)$ of time for the word "signal."

1.1 Signals

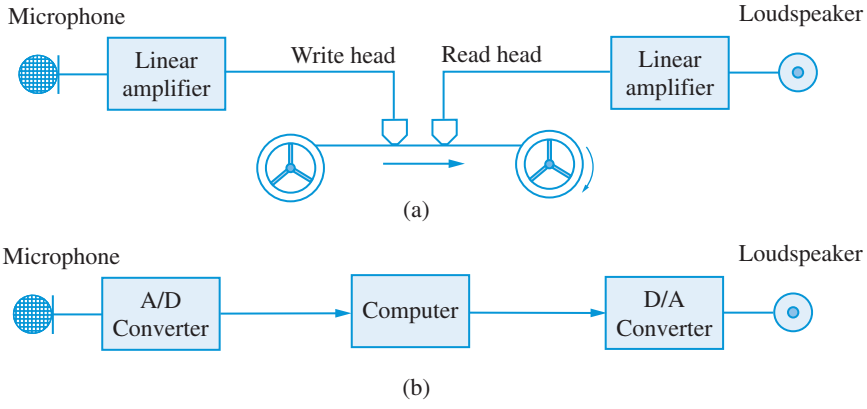

Figure 1.2 Block diagrams of (a) an analog audio recording system using magnetic tape and (b) a digital recording system using a personal computer.

Consider next the system in Figure 1.2(b), which is part of any personal computer. Sound recording and playback with this system involve the following steps:

- The sound waves are converted to an electrical audio signal by the microphone. The audio signal is amplified to a usable level and is applied to an analog-to-digital converter.
- The amplified audio signal is converted into a series of numbers by the analog-to-digital converter.
- The numbers representing the audio signal can be stored or manipulated by software to enhance quality, reduce storage space, or add special effects.
- The digital data are converted into an analog electrical signal; this signal is then amplified and sent to the speaker to produce sound waves.

The major limitation in the quality of the analog tape recorder is imposed by the recording medium, that is, the magnetic tape. As the magnetic tape stretches and shrinks or the speed of the motor driving the tape changes, we have distortions caused by variations in the time scale of the audio signal. Also, random changes in the strength of the magnetic field lead to amplitude distortions of the audio signal. The quality of the recording deteriorates with each additional playback or generation of a copy. In contrast, the quality of the digital audio is determined by the accuracy of numbers produced by the analog-to-digital conversion process. Once the audio signal is converted into digital form, it is possible to achieve error-free storage, transmission, and reproduction. An interesting discussion about preserving information using analog or digital media is given by Bollacker (2010). Every personal computer has a sound card, which can be used to implement the system in Figure 1.2(b); we shall make frequent use of this system to illustrate various signal processing techniques.

1.1.1 Mathematical representation of signals

To simplify the analysis and design of signal processing systems it is almost always necessary to represent signals by mathematical functions of one or more independent variables. For example, the speech signal in Figure 1.1 can be represented mathematically by a function $s(t)$ that shows the variation of acoustic pressure as a function of time. In contrast,

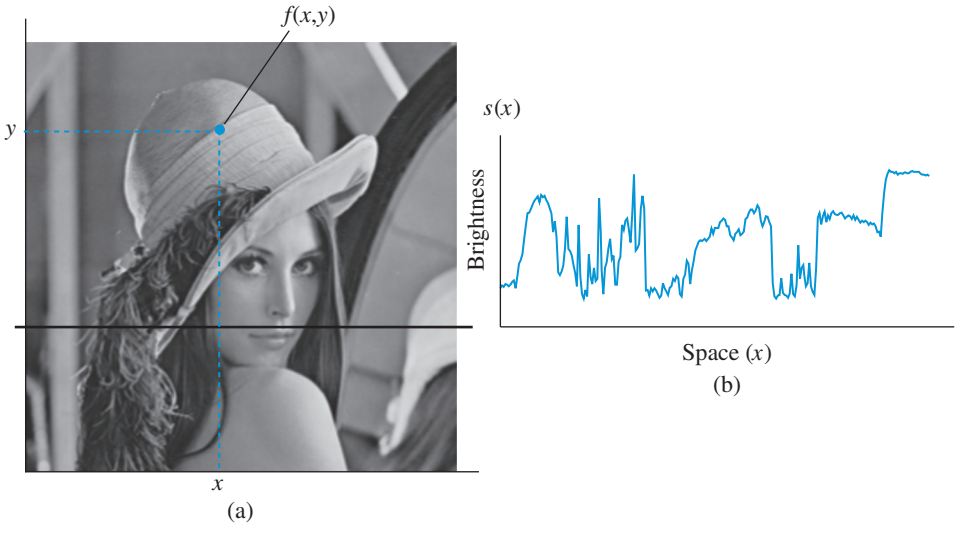

Figure 1.3 Example of a monochrome picture. (a) The brightness at each point in space is a scalar function $f(x, y)$ of the rectangular coordinates x and y. (b) The brightness at a horizontal line at $y = y_0$ is a function $s(x) = f(x, y = y_0)$ of the horizontal space variable x, only.

the monochromatic picture in Figure 1.3 is an example of a signal that carries information encoded in the spatial patterns of brightness variation. Therefore, it can be represented by a function $f(x, y)$ describing the brightness as a function of two spatial variables x and y. However, if we take the values of brightness along a horizontal or vertical line, we obtain a signal involving a single independent variable x or y, respectively. In this book, we focus our attention on signals with a single independent variable. For convenience, we refer to the dependent variable as *amplitude* and the independent variable as *time*. However, it is relatively straightforward to adjust the notation and the vocabulary to accommodate signals that are functions of other independent variables.

Signals can be classified into different categories depending on the values taken by the amplitude (dependent) and time (independent) variables. Two natural categories, that are the subject of this book, are continuous-time signals and discrete-time signals.

The speech signal in Figure 1.1 is an example of a *continuous-time signal* because its value $s(t)$ is defined for every value of time t. In mathematical terms, we say that $s(t)$ is a function of a continuous independent variable. The amplitude of a continuous-time signal may take any value from a continuous range of real numbers. Continuous-time signals are also known as *analog signals* because their amplitude is "analogous" (that is, proportional) to the physical quantity they represent.

The mean yearly number of dark spots visible on the solar disk (sunspots), as illustrated in Figure 1.4, is an example of a discrete-time signal. *Discrete-time signals* are defined only at discrete times, that is, at a discrete set of values of the independent variable. Most signals of practical interest arise as continuous-time signals. However, the use of digital signal processing technology requires a discrete-time signal representation. This is usually done by *sampling* a continuous-time signal at isolated, equally spaced points in time

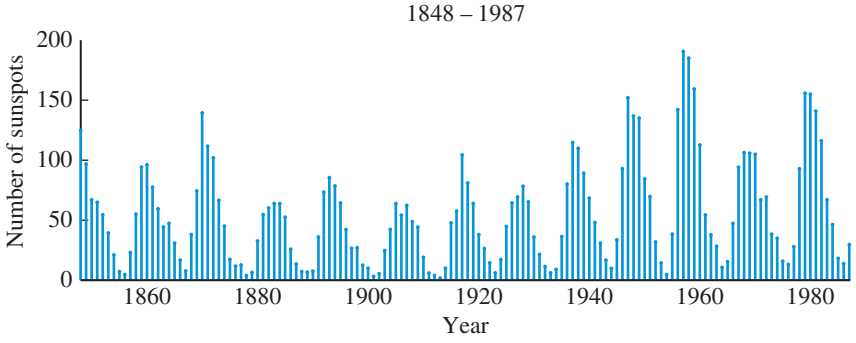

Figure 1.4 Discrete-time signal showing the annual mean sunspot number determined using reliable data collected during the 13 cycles from 1848 to 1987.

(periodic sampling). The result is a sequence of numbers defined by

$$s[n] \triangleq s(t)|_{t=nT} = s(nT), \tag{1.1}$$

where n is an integer $\{\ldots, -1, 0, 1, 2, 3, \ldots\}$ and T is the *sampling period*. The quantity $F_s \triangleq 1/T$, known as *sampling frequency* or *sampling rate*, provides the number of samples per second. The relationship between a continuous-time signal and a discrete-time signal obtained from it by sampling is a subject of great theoretical and practical importance. We emphasize that the value of the discrete-time signal in the interval between two sampling times is not zero; simply, it is *not* defined. Sampling can be extended to two-dimensional signals, like images, by taking samples on a rectangular grid. This is done using the formula $s[m, n] \triangleq s(m\Delta x, n\Delta y)$, where Δx and Δy are the horizontal and vertical sampling periods. The image sample $s[m, n]$ is called a *picture element* or *pixel*, for short.

In this book continuous independent variables are enclosed in parentheses ( ), and discrete-independent variables in square brackets [ ]. The purpose of these notations is to emphasize that parentheses enclose real numbers while square brackets enclose integers; thus, the notation in (1.1) makes sense. Since a discrete-time signal $s[n]$ is a sequence of real numbers, the terms "discrete-time signal" and "sequence" will be used interchangeably. We emphasize that a discrete-time signal $s[n]$ is defined *only* for integer values of the independent variable.

A discrete-time signal $s[n]$ whose amplitude takes values from a finite set of K real numbers $\{a_1, a_2, \ldots, a_K\}$, is known as a *digital signal*. All signals stored on a computer or displayed on a computer screen are digital signals.

To illustrate the difference between the different signal categories, consider the continuous-time signal defined by

$$s(t) = \begin{cases} e^{-2t}\cos(3\pi t), & t \geq 0 \\ 0, & t < 0. \end{cases} \tag{1.2}$$

The continuous-time character of $s(t)$ is depicted graphically using a solid line, as shown in Figure 1.5(a).

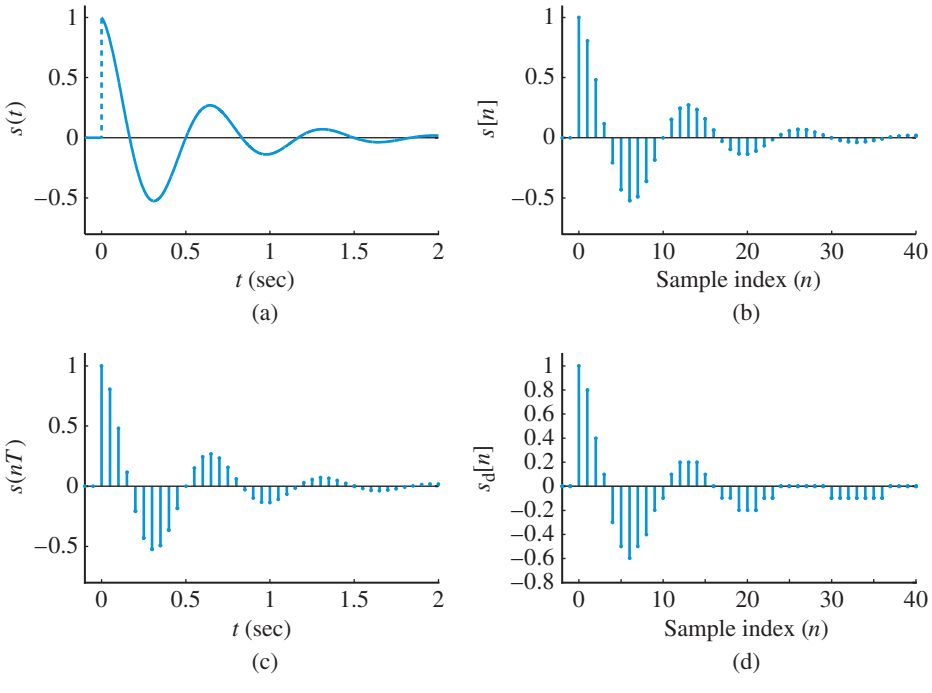

Figure 1.5 Plots illustrating the graphical representation of continuous-time signals (a), discrete-time signals (b) and (c), and digital signals (d).

To plot $s(t)$ on a computer screen, we can only compute its values at a finite set of discrete points. If we sample $s(t)$ with a sampling period $T = 0.05$ s, we obtain the discrete-time signal

$$s[n] = s(nT) = \begin{cases} e^{-0.2n} \cos(0.3\pi n), & n \geq 0 \\ 0, & n < 0 \end{cases} \tag{1.3}$$

which is shown graphically as a stem plot in Figure 1.5(b). Each value of the sequence is represented by a vertical line with a dot at the end (stem). The location of each sample is labeled by the value of the discrete-time index n. If we wish to know the exact time instant $t = nT$ of each sample, we plot $s(nT)$ as a function of t, as illustrated in Figure 1.5(c).

Suppose now that we wish to represent the amplitude of $s[n]$ using only one decimal point. For example, the value $s[2] = 0.4812$ is approximated by $s_d[2] = 0.4$ after truncating the remaining digits. The resulting digital signal $s_d[n]$, see Figure 1.5(d), can only take values from the finite set $\{-0.6, -0.5, \ldots, 1\}$, which includes $K = 17$ distinct signal amplitude levels. All signals processed by computers are digital signals because their amplitudes are represented with finite precision fixed-point or floating-point numbers.

1.1.2 Physical representation of signals

The storage, transmission, and processing of signals require their representation using physical media. There are two basic ways of representing the numerical value of physical quantities: analog and digital:

1.1 Signals

1. In *analog representation* a quantity is represented by a voltage or current that is *proportional* to the value of that quantity. The key characteristic of analog quantities is that they can vary over a continuous range of values.
2. In *digital representation* a quantity is represented *not* by a proportional voltage or current but by a combination of ON/OFF pulses corresponding to the digits of a binary number. For example, a bit arrangement like $b_1 b_2 \cdots b_{B-1} b_B$ where the B *bi*nary digi*ts* (*bits*) take the values $b_i = 0$ or $b_i = 1$ can be used to represent the value of a binary integer as

$$D = b_1 2^{B-1} + b_2 2^{B-2} + \cdots + b_{B-1} 2^1 + b_B 2^0, \tag{1.4}$$

or the value of a B-bit fraction as

$$D = b_1 2^{-1} + b_2 2^{-2} + \cdots + b_{B-1} 2^{-(B-1)} + b_B 2^{-B}. \tag{1.5}$$

The physical representation of analog signals requires using the physical characteristics of the storage medium to create two "continuous analogies:" one for the signal amplitude, and the other for time. For example, in analog tape recording, time is represented by increasing linear distance along magnetic tape; the amplitude of the original signal is represented by the magnetic field of the tape. In practice, all analog physical representation techniques suffer from two classes of problem: those which affect the "analog of time" (for example, variations in the speed of motor driving the tape), and those which affect the "analog of amplitude" (for example, variations in the magnetic field of the tape). The meaning of analog in this connotation is "continuous" because its amplitude can be varied continuously or in infinitesimally small steps. Theoretically, an analog signal has infinite resolution or, in other words, can represent an uncountably infinite number of values. However, in practice, the accuracy or resolution is limited by the presence of noise.

Binary numbers can be represented by any physical device that has only two operating states or physical conditions. There are numerous devices that satisfy this condition: switch (on or off), diode (conducting or nonconducting), transistor (cut off or saturated), spot on a magnetic disk (magnetized or demagnetized). For example, on a compact disc binary data are encoded in the form of pits in the plastic substrate which are then coated with an aluminum film to make them reflective. The data are detected by a laser beam which tracks the concentric circular lines of pits.

In electronic digital systems, binary information is represented by two nominal voltages (or currents) as illustrated in Figure 1.6. The exact value of the voltage representing the binary 1 and binary 0 is not important as long as it remains within a prescribed range. In a digital signal, the voltage or current level represents no longer the magnitude of a variable, because there are only two levels. Instead, the magnitude of a variable is represented by a combination of several ON/OFF levels, either simultaneously on different lines (parallel transmission) or sequentially in time on one line (serial transmission). As a result, a digital signal has only a finite number of values, and can change only in discrete steps. A digital signal can always provide any desired precision if a sufficient number of bits is provided for each value.

In analog systems, the exact value of the voltage is important because it represents the value of the quantity. Therefore, analog signals are more susceptible to noise (random fluctuations). In contrast, once the value of the data in a digital representation is determined,

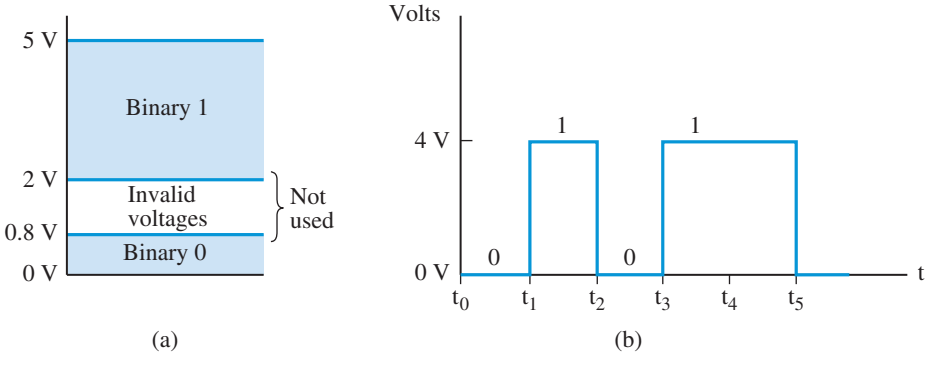

Figure 1.6 Digital signals and timing diagrams. (a) Typical voltage assignments in digital system; (b) typical digital signal timing diagram.

it can be copied, stored, reproduced, or modified without degradation. This is evident if we consider the difference in quality between making a copy of a compact disc and making a copy of an audio cassette.

The digital signals we process and the programs we use to manipulate them are stored as a sequence of bits in the memory of a computer. A typical segment of computer memory might look as follows:

$$\ldots 01101001111010000100101111010101011 10\ldots$$

This collection of bits at this level is without structure. The first step in making sense of this bit stream is to consider the bits in aggregates referred to as *bytes* and *words*. Typically, a byte is composed of 8 bits and a word of 16 or 32 bits. Memory organization allows us to access its contents as bytes or words at a particular address. However, we still cannot speak meaningfully of the contents of a byte or word. To give numerical meaning to a given byte, we must know the type of the value being represented. For example, the byte "00110101" has the value 53 if treated as integer or the value 0.2070 if treated as a fraction. Each computer language has different types of integer and floating representations of numbers. Different types of number representation and their properties are discussed in Chapter 15. We shall use the term *binary code* to refer to the contents of a byte or word or its physical representation by electronic circuits or other physical media.

1.1.3 Deterministic and random signals

The distinction between continuous-time signals and discrete-time signals has important implications in the mathematical tools used for their representation and analysis. However, a more profound implication stems from the distinction between deterministic signals and random signals. The behavior of deterministic signals is completely predictable, whereas the behavior of random signals has some degree of uncertainty associated with them. To make this distinction more precise, suppose that we know all past values of a signal up to the present time. If, by using the past values, we can predict the future values of the signal exactly, we say that the signal is *deterministic*. On the other hand, if we cannot predict the future values of the signal exactly, we say that the signal is *random*. In practice, the distinction between these two types of signal is not sharp because every signal

is corrupted by some amount of unwanted random noise. Nevertheless, the separation into deterministic and random signals has been widely adopted when we study the mathematical representation of signals.

Deterministic signals can be described, at least in principle, by mathematical functions. These functions can often take the form of explicit mathematical formulas, as for the signals shown in Figure 1.5. However, there are deterministic signals that cannot be described by simple equations. In principle, we assume that each deterministic signal is described by a function $s(t)$, even if an explicit mathematical formula is unavailable. In contrast, random signals cannot be described by mathematical functions because their future values are unknown. Therefore, the mathematical tools for representation and analysis of random signals are different from those used for deterministic signals. More specifically, random signals are studied using concepts and techniques from the theory of probability and statistics. In this book, we mainly focus on the treatment of deterministic signals; however, a brief introduction to the mathematical description and analysis of random signals is provided in Chapters 13 and 14.

1.2 Systems

In Merriam-Webster's dictionary, a system is broadly defined as a "regularly interacting or interdependent group of items forming a unified whole." In the context of signal processing, a *system* is defined as a process where a signal called *input* is transformed into another signal called *output*. Systems are classified based on the category of input and output signals.

1.2.1 Continuous-time systems

A *continuous-time system* is a system which transforms a continuous-time input signal $x(t)$ into a continuous-time output signal $y(t)$. For example, the continuous-time system described by the formula

$$y(t) = \int_{-\infty}^{t} x(\tau)\mathrm{d}\tau \tag{1.6}$$

produces an output signal which is the integral of the input signal from the start of its operation at $t = -\infty$ to the present time instant t. Symbolically, the input-output relation of a continuous-time system is represented by

$$x(t) \overset{\mathcal{H}}{\longmapsto} y(t) \quad \text{or} \quad y(t) = \mathcal{H}\{x(t)\}, \tag{1.7}$$

where $\mathcal{H}$ denotes the mathematical operator characterizing the system. A pictorial representation of a continuous-time system is shown in Figure 1.7(a).

Continuous-time systems are physically implemented using analog electronic circuits, like resistors, capacitors, inductors, and operational amplifiers. The physical implementation of a continuous-time system is known as an *analog system*. Some common analog systems are audio amplifiers, AM/FM receivers, and magnetic tape recording and playback systems.

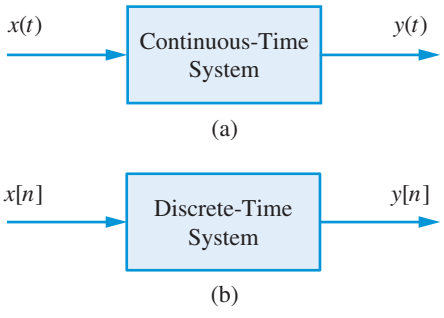

Figure 1.7 Pictorial or block-diagram representation of a continuous-time system (a) and a discrete-time system (b).

1.2.2 Discrete-time systems

A system that transforms a discrete-time input signal $x[n]$ into a discrete-time output signal $y[n]$, is called a *discrete-time system*. A pictorial representation of a discrete-time system, denoted symbolically by

$$x[n] \overset{\mathcal{H}}{\longmapsto} y[n] \quad \text{or} \quad y[n] = \mathcal{H}\{x[n]\}, \tag{1.8}$$

is shown in Figure 1.7(b). The discrete-time equivalent of the continuous-time integrator system (1.6) is the accumulator system

$$y[n] = \sum_{k=-\infty}^{n} x[k]. \tag{1.9}$$

We note that the integral in (1.6), which is an operator applicable to continuous functions, is replaced by summation, which is a discrete operation.

The physical implementation of discrete-time systems can be done either in software or hardware. In both cases, the underlying physical systems consist of digital electronic circuits designed to manipulate logical information or physical quantities represented in digital form by binary electronic signals. Numerical quantities represented in digital form can take on only discrete values, or equivalently are described with finite precision. Therefore, in practice every discrete-time system has to be implemented by a *digital system*. The term digital is derived from the way computers perform operations, by counting digits.

1.2.3 Interface systems

An analog system contains devices that manipulate physical quantities that are represented in analog form. In an analog system, the amplitude of signals can vary over a continuous range of values. In contrast, a digital system is a combination of devices designed to manipulate physical quantities that are represented in digital form using logical operations. Therefore, there is a need for systems that provide the interface between analog and digital signals.