Part I

Foundations

1

Introduction

1.1 What is a signal?

A *signal* is a varying quantity whose value can be measured and which conveys information.

For example, we can consider temperature to be a signal. It can vary over time, we can measure it using a thermometer, and it conveys information: knowing the temperature outside will inform our decision as to which clothes to wear.

In a digital signal processing system we represent a signal as a sequence of numbers either on a computer or in digital hardware. For example, we could store the temperature at various times of the day as a sequence of numbers in an array on a computer: each number might be a temperature reading in Celsius.

Digital signal processing involves transforming one signal into another signal, represented digitally throughout. The transformation is achieved using simple operations on the numbers representing the signal. For example, we might want to know the average temperature over a day: we could calculate this by adding up the elements in the array of temperature data and dividing the total by the size of the array.

1.2 Domain and range of a signal

Temperature is a function of a single real-valued variable, time: see Figure 1.1. We say that the *domain* of the signal is one-dimensional. Some signals are functions of more than one variable. For example, a black-and-white photograph can be regarded as a signal: the brightness u of a point on the photograph is a function of two variables, the x and y coordinates of the point on the photograph: see Figure 1.2. In this case the domain of the signal is two-dimensional.

In a black-and-white photograph, the brightness of a point on the photograph can be represented as a single real number, and so we call it a real-valued signal, and say that the *range* of the signal is one-dimensional. In a colour photograph,

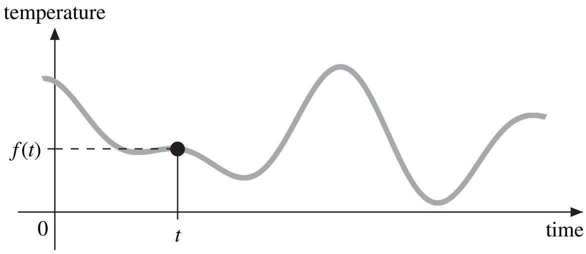

Figure 1.1 A signal with one-dimensional range and one-dimensional domain.

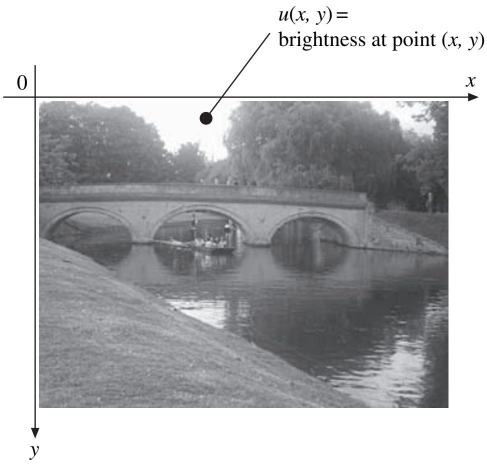

Figure 1.2 A signal with one-dimensional range and two-dimensional domain.

however, one real number will not do. The colour of a point can be expressed as three real numbers, separately giving the amount of red, green and blue that go to make up the colour. In this case, we would say that the range of the signal is three-dimensional.

Now let us consider a colour movie. The colour of a point on the screen with given x and y coordinates can be expressed as three real numbers for the amounts of red, green and blue as described above. However, the picture also changes with time, which adds an extra dimension to the domain of the signal. In total we therefore have three dimensions in the domain (x, y and time) and three in the range (red, green and blue).

1.3 Converting signals from one form to another

A device that converts a signal from one form to another is called a *transducer*. Often the signal on one side of the conversion will be an electrical one.

A loudspeaker is an example of a transducer, in this case converting an electrical signal into a varying air pressure to create a sound. The variation in air pressure is a real-valued function of time. A microphone is a transducer that converts in the opposite direction, from variations in air pressure to variations in an electrical signal.

1.4 Processing signals

Suppose we want to

– modify the amount of bass and treble in an audio signal
– analyse an image to determine what objects are present in it
– compute seasonally adjusted temperature values
– make a photograph sharper or increase its contrast
– measure the pitch of a musical instrument

All of these are examples of signal processing tasks. Later in this book we shall look at how we might go about these kinds of task.

Another common signal processing task is data compression, where we take advantage of special characteristics of a signal to reduce the resources required to store or transmit it. For example, recorded speech often contains long pauses. We can identify these pauses and delete them. To take another example, one frame of a movie is often very similar to the previous one. We can process the signal to find which parts of the picture are changing, and only record those.

Why digital?

Many signal processing tasks can be done using conventional analogue electronics. Our first example above, a tone control which modifies the amount of bass and treble in an audio signal, is particularly simple using analogue technology. If our signal is in the form of a varying voltage, the necessary circuit consists of just a couple of components costing a few pence in total!

Analogue processing systems suffer from several disadvantages, however. Our tone control circuit will be fine as long as we are not too demanding about its performance: if we manufacture several copies of the circuit, the chances are that its characteristics will vary, probably by as much as several per cent from unit to unit. If we needed less variation from unit to unit we could use better-quality components, but that would increase the cost. Alternatively, we could add an adjustment to the circuit so that we can accurately trim each unit to compensate for manufacturing variations, but that would make the assembly process more complicated and so more expensive. The characteristics of analogue systems also tend to drift slowly over time and with temperature.

Any processing which involves storing a lot of information – comparing successive frames of a movie, for example – will almost certainly be complicated and expensive when implemented using analogue circuitry.

Further disadvantages become apparent when we turn to integrated circuits (ICs). It is perfectly possible to make ICs with analogue circuits on them, but processing involving low frequencies, such as the bass part of an audio signal, usually requires physically large components which are difficult to make on an IC. There are also technical difficulties, as well as extra cost, associated with mixing analogue and digital circuitry on a single IC.

Many electronic devices already necessarily incorporate digital circuitry. For example, consider adding a tone control to a CD (compact disc) player which already uses a digital integrated circuit to decode the information stored on the disc. In this case, the cheapest option could easily be to implement the tone control digitally on the same IC: this would add very slightly to the size and hence the cost of the IC, but no extra components would be needed and assembly costs would remain the same.

Once the decision is made to implement the tone control digitally, it becomes very simple to add extra features, such as tone settings optimised for different types of music, all at very little extra cost compared with an analogue implementation.

There are, of course, disadvantages to doing things digitally. As we shall see in later chapters, there are some hazards to avoid when converting a signal between analogue and digital forms. If the processing you want to do is not demanding, and the signal is not already in digital form, the overhead of converting from analogue to digital and back can easily outweigh the advantages of digital processing.

1.5 Notation

Before proceeding further we will make a few remarks on the notation used in this book.

Complex numbers

Sometimes in this book we will be considering signals which take on complex values. They can be thought of as having a two-dimensional range, the two dimensions being the real and imaginary parts of the complex values. They arise more frequently in the intermediate steps in processing a signal than they do naturally as physical quantities (although of course you can think of any two-dimensional quantity as a complex number if you like).

In this book we will write j for the square root of -1. (Engineers often use i to stand for electrical current.) We will write the real part of a complex number z

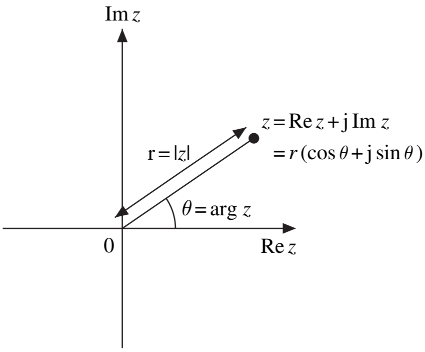

Figure 1.3 A point z in the complex plane can be represented in terms of its real and imaginary parts, or in terms of its magnitude and argument.

as $\mathrm{Re}\, z$ and the imaginary part as $\mathrm{Im}\, z$. The *magnitude* of z, $|z|$, is the distance between the point z and the origin of the complex plane; and the *argument* of z, $\arg z$, is the angle between the real axis and the line from the origin to z, measured anticlockwise.

Figure 1.3 shows that we can write a complex number z in terms of its real part $x = \mathrm{Re}\, z$ and imaginary part $y = \mathrm{Im}\, z$ as

$$z = x + \mathrm{j}y$$

or in terms of its magnitude $r = |z|$ and argument $\theta = \arg z$:

$$\begin{aligned} z &= r\cos\theta + \mathrm{j}r\sin\theta \\ &= r\mathrm{e}^{\mathrm{j}\theta}. \end{aligned}$$

We will frequently switch between these two representations of complex numbers.

One handy use of complex numbers is to represent angles. An angle θ can be represented by the complex number $z = \cos\theta + \mathrm{j}\sin\theta = \mathrm{e}^{\mathrm{j}\theta}$. The magnitude of z is 1, and so this is a point in the complex plane that lies on a circle of radius 1 centred on the origin (the *unit circle*). The argument of z is θ: see Figure 1.4. The advantage of this representation is that the angle 359° (very nearly a complete revolution) is represented by a point very near to the one that represents 0°, which simplifies calculations in some applications: Exercise 1.4 gives an example.

Block diagrams

We will often use block diagrams to help explain how digital signal processing systems are built up from simple modules. Each module is thought of as carrying out an operation on a sequence of incoming numbers one at a time, producing

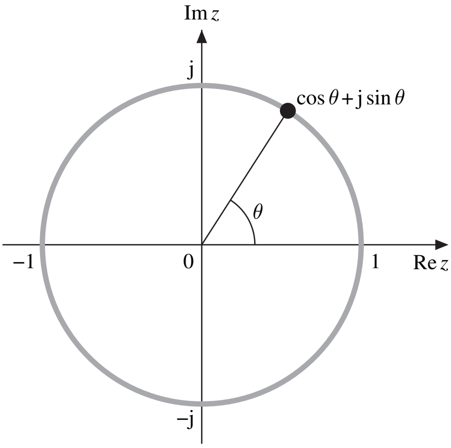

Figure 1.4 Representing an angle as a point on the unit circle in the complex plane.

a processed sequence of numbers at its output. Lines with arrows show the interconnections between modules.

The most basic modules are those that carry out the arithmetic operations of addition, subtraction, multiplication and division. The symbols we use for these modules are shown in Figure 1.5(a) to (d). The label on the input to the subtractor indicates that it subtracts its bottom input from its top one; likewise, the divider divides its bottom input into its top one.

Modules that carry out other operations are shown as boxes labelled with the relevant function. A box labelled with a ‘D’ delays its input by one time unit: see Figure 1.5(e).

These basic modules can be put together to construct higher-level building blocks such as filters, as we will describe in Chapter 5. Filters are often used as modules in more complicated signal-processing systems and so have their own

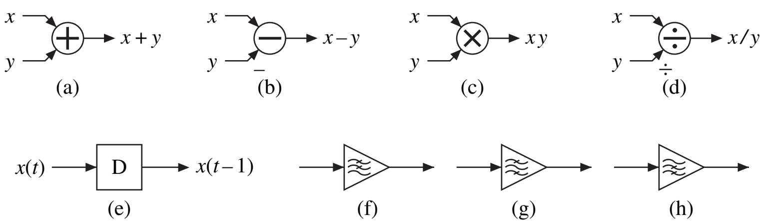

Figure 1.5 Block diagram symbols: (a) adder; (b) subtractor; (c) multiplier; (d) divider; (e) delay element; (f) low-pass filter; (g) band-pass filter; (h) high-pass filter.

symbols. The symbols for the three main types of filter, known as ‘low-pass’, ‘band-pass’, and ‘high-pass’, are shown in Figure 1.5(f) to (h).

Exercises

1.1 At the beginning of the chapter we used outside temperature of an example of a signal, considering it only as a function of time. A weather forecast, however, will talk about temperature not only as a function of time, but as a function of location in the country: how many dimensions does the domain of this signal have?

An aircraft pilot whose hobby is signal processing thinks of the outside temperature as a signal. How many dimensions would you imagine he thinks its domain has?

1.2 Consider a black-and-white movie as a signal. How many dimensions do its domain and range have?

1.3 Use the following procedure to determine roughly how accurately you can represent real numbers on your computer. Set a floating-point variable `a` to 1. Set a variable `b` to $1+10^{-k}$ for various integer values of k, and compare `a` for equality with `b`. Find the largest value of k for which your computer reports that `a` is not equal to `b`. The machine precision is then roughly one part in 10^k.

A capacitor is manufactured using metal plates of length 1 millimetre, and the capacitance is directly proportional to this length. If the capacitor is to be made to the same precision as the numbers you can represent on your computer, what variation can we allow in the length of the metal plates? If an atom is 10^{-10} metres across, how many atoms does this correspond to?

1.4 Five hikers with varying levels of expertise at reading a compass attempt to measure the bearing of a landmark (i.e., the angle between a line pointing due north from where they are standing and a line to the landmark, measured clockwise when viewed from above). Each produces an answer in degrees from 0° to 359° inclusive. Using complex numbers, devise a procedure to take the five readings and average them to produce a more reliable result. Try it on the following sets of bearings: (a) 12°; 15°; 13°; 9°; 16°; (b) 358°; 1°; 359°; 355°; 2°; (c) 210°; 290°; 10°; 90°; 170°.

Suppose you have more confidence in the compass-reading skills of some of the hikers than others. Describe a simple extension to your procedure to give a different weight to each reading.

2
Sampling

In Chapter 1 we looked at various types of signal. We now start to examine how signals can be processed digitally.

2.1 Regular sampling

The first step is to reduce a continuous signal to a finite number of values, in a process called *sampling*. For example, if we have a signal that varies with time, such as an audio signal, then we normally measure its value at equal temporal intervals. The audio signal is then represented by these regularly spaced measurements or *samples* as shown in Figure 2.1. The sample values (+4.7, +3.3, +4.2, etc.) would typically be stored in consecutive elements of a one-dimensional array. The sampling process is similar to the way a movie camera takes a series of pictures of a scene, regularly spaced in time; or to the way a dieter weighs himself every morning.

We can sample a two-dimensional signal in a similar way. For example, a still image can be sampled spatially by overlaying it with a rectangular grid of points. A greyscale image is represented by a two-dimensional array of values giving the intensity at each sample point. Usually the value 0.0 is used to represent black, and 1.0 is used to represent white, although other scales are also used.

Note that sampling the intensity of an image at a single point is not quite the same thing as averaging it over a small area around that point, which is what typical scanners and cameras do. We shall return to this distinction in Section 3.5.

In a colour image, three numbers may be used to represent each sample point, the numbers being proportional to the amount of each of the three primary colours (red, green and blue) present in the colour at that point.

When sampling occurs at regular intervals of time, as in the case of an audio signal, we usually speak of its *sample rate*, measured in units such as samples per second, or Hertz (Hz). If sampling is spatial, as in the case of the still image, we normally talk of its *resolution*, measured in units such as pixels per millimetre or dots per inch (dpi).

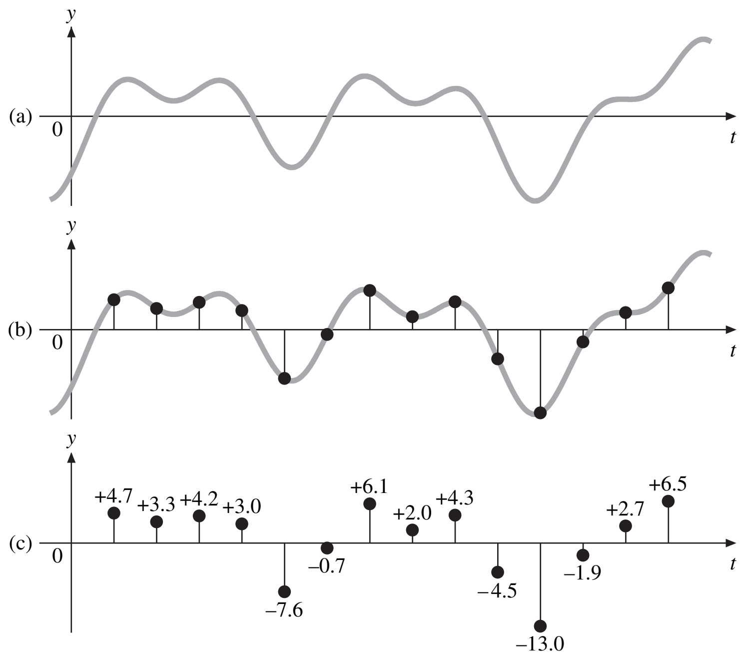

Figure 2.1 A typical audio waveform: (a) the original signal; (b) sample points on the waveform; (c) sampled representation.

The process of sampling discards some information about the signal. If an event occurs in the moment between the frames taken by a movie camera, it will not appear on the film. This can be seen in coverage of (especially outdoor) sporting events, when, because the lighting is very bright, the camera's shutter only opens very briefly for each frame. A fast-moving ball will appear to jump from position to position across the screen. Of course, the ball does not really jump; rather, the fact that the ball was between those positions was just never captured on the film.

In the same way, if an audio signal changes very quickly over time compared with the rate at which we take samples, or if an image includes features which are small compared with the sampling resolution, then we will not catch all the variations in the original. We will now make this idea more precise.

2.2 What is lost in sampling?

Figure 2.2 shows two audio waveforms superimposed. The first is a sine wave of frequency 200 Hz, and the second is a sine wave of frequency 800 Hz. The waveforms have been sampled at 1 kHz. At the points where the samples have been taken, shown by the dark dots, the two waveforms always have the same value. This means that if we are using a sample rate of 1 kHz then, given just the