Digital Signal Processing

A Practical Approach

Second Edition

Emmanuel C. Ifeachor
Barrie W. Jervis

Prentice Hall

Digital Signal Processing

A Practical Approach

Second edition

Emmanuel C. Ifeachor
University of Plymouth

Barrie W. Jervis
Sheffield Hallam University

Prentice Hall

An imprint of Pearson Education

Harlow, England · London · New York · Reading, Massachusetts · San Francisco · Toronto · Don Mills, Ontario · Sydney
Tokyo · Singapore · Hong Kong · Seoul · Taipei · Cape Town · Madrid · Mexico City · Amsterdam · Munich · Paris · Milan

Pearson Education Limited
Edinburgh Gate
Harlow
Essex CM20 2JE
England

and Associated Companies around the World.

Visit us on the World Wide Web at:
www.pearsoneduc.com

First published under the Addison Wesley imprint 1993
Second edition 2002

ISBN 0201-59619-9

British Library Cataloguing-in-Publication Data
A catalogue record for this book can be obtained from the British Library

Library of Congress Cataloging-in-Publication Data
Ifeachor, Emmanuel C.
Digital signal processing : a practical approach / Emmanuel C. Ifeachor, Barrie W. Jervis.
p. cm.
Includes bibliographical references and index.
ISBN 0–201–59619–9
1. Signal processing–Digital techniques. I. Jervis, Barrie W. II. Title.
TK5102.9.I34 2001
621.382'2–dc21 2001021116

10 9 8 7 6 5 4 3 2 1
05 04 03 02 01

Typeset in 10/12pt Times by 35
Printed and bound in the United States of America

Contents

Introduction

1

The aims of this chapter are to explain the meaning and benefits of digital signal processing (DSP), to introduce basic DSP operations on which much of DSP is founded, and to make the reader aware of the wide range of application areas for DSP. Specific real-world application examples are presented, drawn from areas with which most readers can relate.

1.1 Digital signal processing and its benefits

By a signal we mean any variable that carries or contains some kind of information that can, for example, be conveyed, displayed or manipulated. Examples of the types of signals of particular interest are

- speech, which we encounter for example in telephony, radio and everyday life,
- biomedical signals, such as the electroencephalogram (brain signals),
- sound and music, such as reproduced by the compact disc player,
- video and image, which most people watch on the television, and
- radar signals, which are used to determine the range and bearing of distant targets.

Digital signal processing is concerned with the digital representation of signals and the use of digital processors to analyze, modify, or extract information from signals. Most signals in nature are analog in form, often meaning that they vary continuously with time, and represent the variations of physical quantities such as sound waves. The signals used in most popular forms of DSP are derived from analog signals which have been sampled at regular intervals and converted into a digital form.

The specific reason for processing a digital signal may be, for example, to remove interference or noise from the signal, to obtain the spectrum of the data, or to transform the signal into a more suitable form. DSP is now used in many areas where analog methods were previously used and in entirely new applications which were difficult or impossible with analog methods. The attraction of DSP comes from key advantages such as the following.

- *Guaranteed accuracy*. Accuracy is only determined by the number of bits used.
- *Perfect reproducibility*. Identical performance from unit to unit is obtained since there are no variations due to component tolerances. For example, using DSP techniques, a digital recording can be copied or reproduced several times over without any degradation in the signal quality.
- No drift in performance with temperature or age.
- Advantage is always taken of the tremendous advances in semiconductor technology to achieve greater reliability, smaller size, lower cost, low power consumption, and higher speed.
- *Greater flexibility*. DSP systems can be programmed and reprogrammed to perform a variety of functions, without modifying the hardware. This is perhaps one of the most important features of DSP.
- *Superior performance*. DSP can be used to perform functions not possible with analog signal processing. For example, linear phase response can be achieved, and complex adaptive filtering algorithms can be implemented using DSP techniques.
- In some cases information may already be in a digital form and DSP offers the only viable option.

DSP is not without disadvantages. However, the significance of these disadvantages is being continually diminished by new technology.

- *Speed and cost.* DSP designs can be expensive, especially when large bandwidth signals are involved. At the present, fast ADCs/DACs (analog-to-digital converters/digital-to-analog converters) either are too expensive or do not have sufficient resolution for wide bandwidth DSP applications. Currently, only specialized ICs can be used to process signals in the megahertz range and these are quite expensive. Furthermore, most DSP devices are still not fast enough and can only process signals of moderate bandwidths. Bandwidths in the 100 MHz range are still processed only by analog methods. Nevertheless, DSP devices are becoming faster and faster.
- *Design time.* Unless you are knowledgeable in DSP techniques and have the necessary resources (software packages and so on), DSP designs can be time consuming and in some cases almost impossible. The acute shortage of suitable engineers in this area is widely recognized. However, the situation is changing as many new graduates now possess some knowledge of digital techniques and commercial companies are beginning to exploit the advantages of DSP in their products.
- *Finite wordlength problems.* In real-time situations, economic considerations often mean that DSP algorithms are implemented using only a limited number of bits. In some DSP systems, if an insufficient number of bits is used to represent variables serious degradation in system performance may result.

1.2 Application areas

DSP is one of the fastest growing fields in modern electronics, being used in any area where information is handled in a digital form or controlled by a digital processor. Application areas include the following:

- Image processing
  - pattern recognition
  - robotic vision
  - image enhancement
  - facsimile
  - satellite weather map
  - animation
- Instrumentation/control
  - spectrum analysis
  - position and rate control
  - noise reduction
  - data compression

- Speech/audio
  - speech recognition
  - speech synthesis
  - text to speech
  - digital audio
  - equalization
- Military
  - secure communication
  - radar processing
  - sonar processing
  - missile guidance
- Telecommunications
  - echo cancellation
  - adaptive equalization
  - ADPCM transcoders
  - spread spectrum
  - video conferencing
  - data communication
- Biomedical
  - patient monitoring
  - scanners
  - EEG brain mappers
  - ECG analysis
  - X-ray storage/enhancement
- Consumer applications
  - digital, cellular mobile phones
  - universal mobile telecommunication system
  - digital television
  - digital cameras
  - Internet phones, music and video
  - digital answer machines, fax and modems
  - voice mail systems
  - interactive entertainment systems
  - active suspension in cars

A look at the list, which is by no means complete, will confirm the importance of DSP. A testimony to the recognition of the importance of DSP is the continual introduction of powerful DSP devices by semiconductor manufacturers. However, there are insufficient engineers with adequate knowledge in this area. An objective of this book is to provide an understanding of DSP techniques and their implementation, to enable the reader to gain a working knowledge of this important subject.

1.3 Key DSP operations

Several DSP algorithms exist and many more are being invented or discovered. However, all these algorithms, including the most complex, require similar basic operations. It is instructive to examine some of these operations at the outset so as to appreciate the implementational simplicity of DSP. The basic DSP operations are convolution, correlation, filtering, transformations, and modulation. Table 1.1 summarizes these operations and a brief description of each is given below. An important point to note in the table is that all the basic DSP operations require only simple arithmetic operations of multiply, add/subtract, and shifts to carry out. Notice also the similarity between most of the operations.

1.3.1 Convolution

Convolution is one of the most frequently used operations in DSP. For example, it is the basic operation in digital filtering. Given two finite and causal sequences, $x(n)$ and $h(n)$, of lengths N_1 and N_2, respectively, their convolution is defined as

$$y(n) = h(n) \circledast x(n) = \sum_{k=-\infty}^{\infty} h(k)x(n-k) = \sum_{k=0}^{\infty} h(k)x(n-k), \qquad n = 0, 1, \ldots, (M-1)$$

where the symbol $\circledast$ is used to denote convolution and $M = N_1 + N_2 - 1$. As we shall see in later chapters, DSP device manufacturers have developed signal processors that perform efficiently the multiply–accumulate operations involved in convolution. An example of the linear convolution of the two sequences depicted in Figures 1.1(a) and 1.1(b) is given in Figure 1.1(c). In this example, $h(n)$, $n = 0, 1, 2, \ldots$, can be viewed as the impulse response of a digital system, and $y(n)$ the system's response to the input sequence, $x(n)$. The numerical values for the convolution, that is $y(n)$, were obtained by direct evaluation of Equation 1.1. For example, $y(1)$ is obtained as follows:

$$\begin{aligned} y(1) &= h(0)x(1) + h(1)x(0) + h(2)x(-1) + \ldots + h(12)x(-11) \\ &= 0 \times 1 + (-0.02) \times 1 + 0 \times 0 + \ldots + 0 \times 0 \\ &= -0.02 \end{aligned}$$

The significance of convolution is more apparent when it is observed in the frequency domain, and use is made of the fact that convolution in the time domain is equivalent to multiplication in the frequency domain. A more detailed discussion of convolution including its properties and graphical interpretation is given in Chapter 5.

Table 1.1 Summary of key DSP operations.

(1) *Convolution*. Given two finite length sequences, $x(k)$ and $h(k)$, of lengths N_1 and N_2, respectively, their linear convolution is

$$y(n) = h(n) \circledast x(n) = \sum_{k=-\infty}^{\infty} h(k)x(n-k) = \sum_{k=0}^{M-1} h(k)x(n-k),\ n = 0, 1, \ldots, M-1 \quad (1.1)$$

where $M = N_1 + N_2 - 1$.

(2) *Correlation*.

(a) Given two N-length sequences, $x(k)$ and $y(k)$, with zero means, an estimate of their cross-correlation is given by

$$\rho_{xy}(n) = \frac{r_{xy}(n)}{[r_{xx}(0)r_{yy}(0)]^{1/2}} \qquad n = 0, \pm 1, \pm 2, \ldots \quad (1.2)$$

where $r_{xy}(n)$ is an estimate of the cross-covariance and defined as

$$r_{xy}(n) = \begin{cases} \dfrac{1}{N}\displaystyle\sum_{k=0}^{N-n-1} x(k)y(k+n) & n = 0, 1, 2, \ldots \\ \dfrac{1}{N}\displaystyle\sum_{k=0}^{N+n-1} x(k-n)y(k) & n = 0, -1, -2, \ldots \end{cases}$$

$$r_{xx}(0) = \frac{1}{N}\sum_{k=0}^{N-1}[x(k)]^2,\ r_{yy}(0) = \frac{1}{N}\sum_{k=0}^{N-1}[y(k)]^2$$

(b) An estimate of the autocorrelation, $\rho_{xx}(n)$, of an N-length sequence, $x(k)$, with zero mean is given by

$$\rho_{xx}(n) = \frac{r_{xx}(n)}{r_{xx}(0)} \qquad n = 0, \pm 1, \pm 2, \ldots \quad (1.3)$$

where $r_{xx}(n)$ is an estimate of the autocovariance and defined as

$$r_{xx}(n) = \frac{1}{N}\sum_{k=0}^{N-n-1} x(k)x(k+n) \qquad n = 0, 1, 2, \ldots$$

(3) *Filtering*. The equation for finite impulse response (FIR) filtering is

$$y(n) = \sum_{k=0}^{N-1} h(k)x(n-k) \quad (1.4)$$

where $x(k)$ and $y(k)$ are the input and output of the filter, respectively, and $h(k)$, $k = 0, 1, \ldots, N-1$, are the filter coefficients.

(4) *Discrete transform*.

$$X(n) = \sum_{k=0}^{N-1} x(k)W^{kn},\ \text{where } W = \exp(-j2\pi/N) \quad (1.5)$$

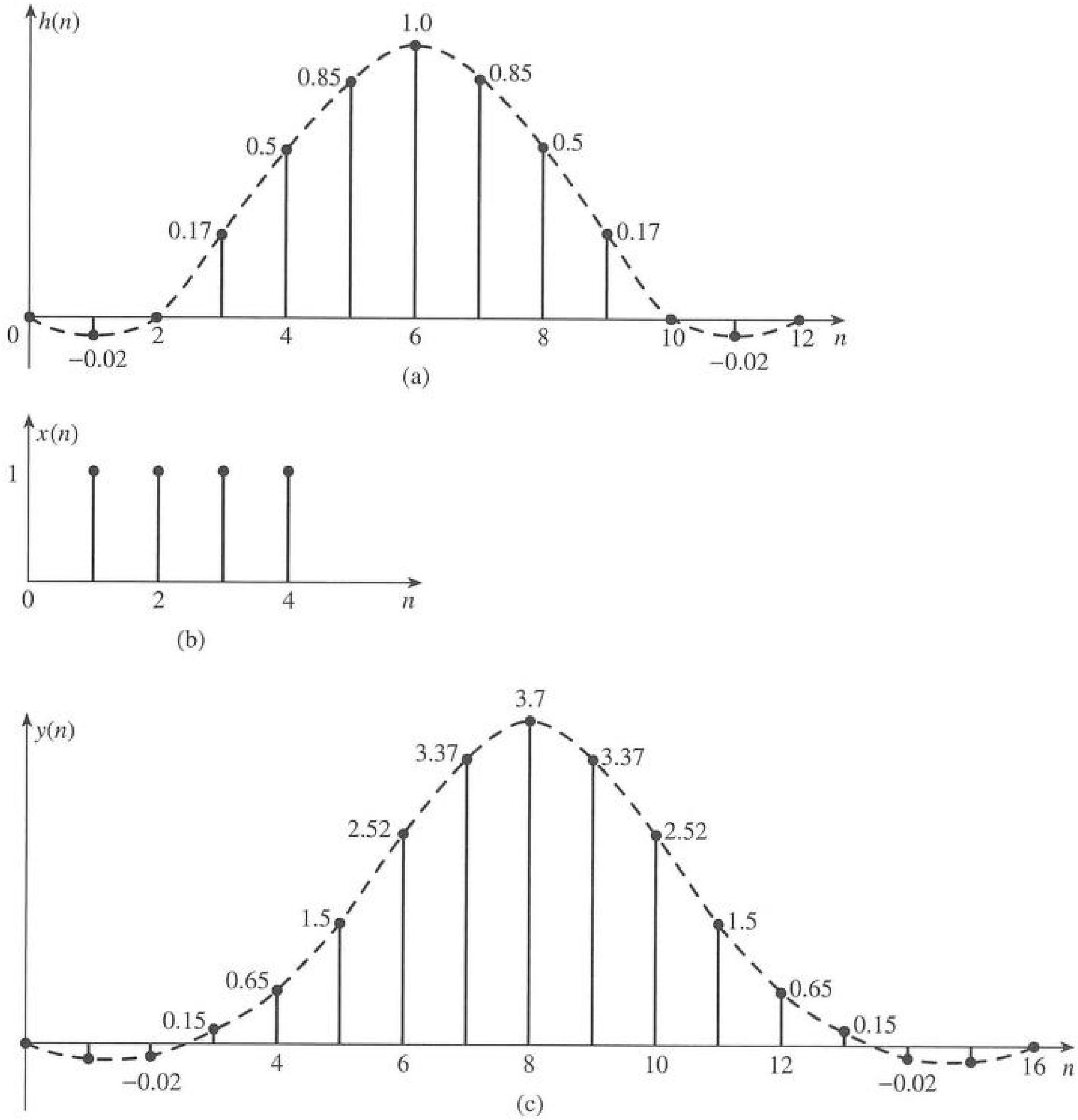

Figure 1.1 An example of the convolution of two sequences. $y(n)$ is the convolution of $h(n)$ and $x(n)$. If $h(n)$ is considered the impulse response of a system, then $y(n)$ is the system's output in response to the input $x(n)$. The values of $y(n)$ above were obtained directly from Equation 1.1.

1.3.2 Correlation

There are two forms of correlations: auto- and cross-correlations.

(1) The cross-correlation function (CCF) is a measure of the similarities or shared properties between two signals. Applications of CCFs include cross-spectral analysis, detection/recovery of signals buried in noise, for example the detection of radar return signals, pattern matching, and delay measurements. CCF is defined in Equation 1.2 in Table 1.1.

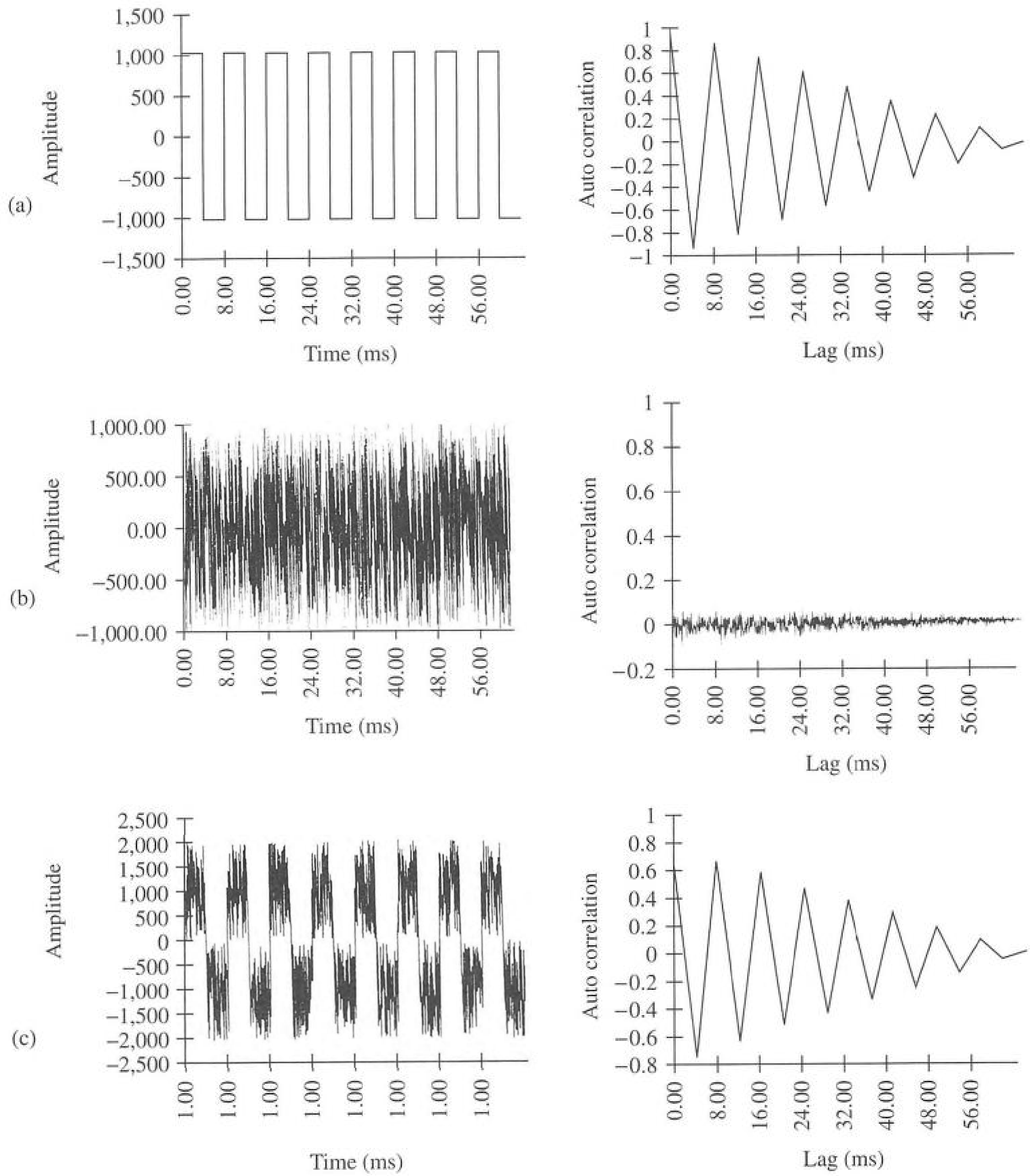

Figure 1.2 Autocorrelations of (a) a periodic signal, (b) noise and (c) periodic signal plus noise. Note that in (c) the periodic nature of the signal buried in noise is still evident, illustrating why autocorrelation is used in detecting hidden periodicity.

(2) The autocorrelation function (ACF) involves only one signal and provides information about the structure of the signal or its behaviour in the time domain. It is a special form of CCF and is used in similar applications. It is particularly useful in identifying hidden periodicities. The ACF is defined in Equation 1.3 in Table 1.1.

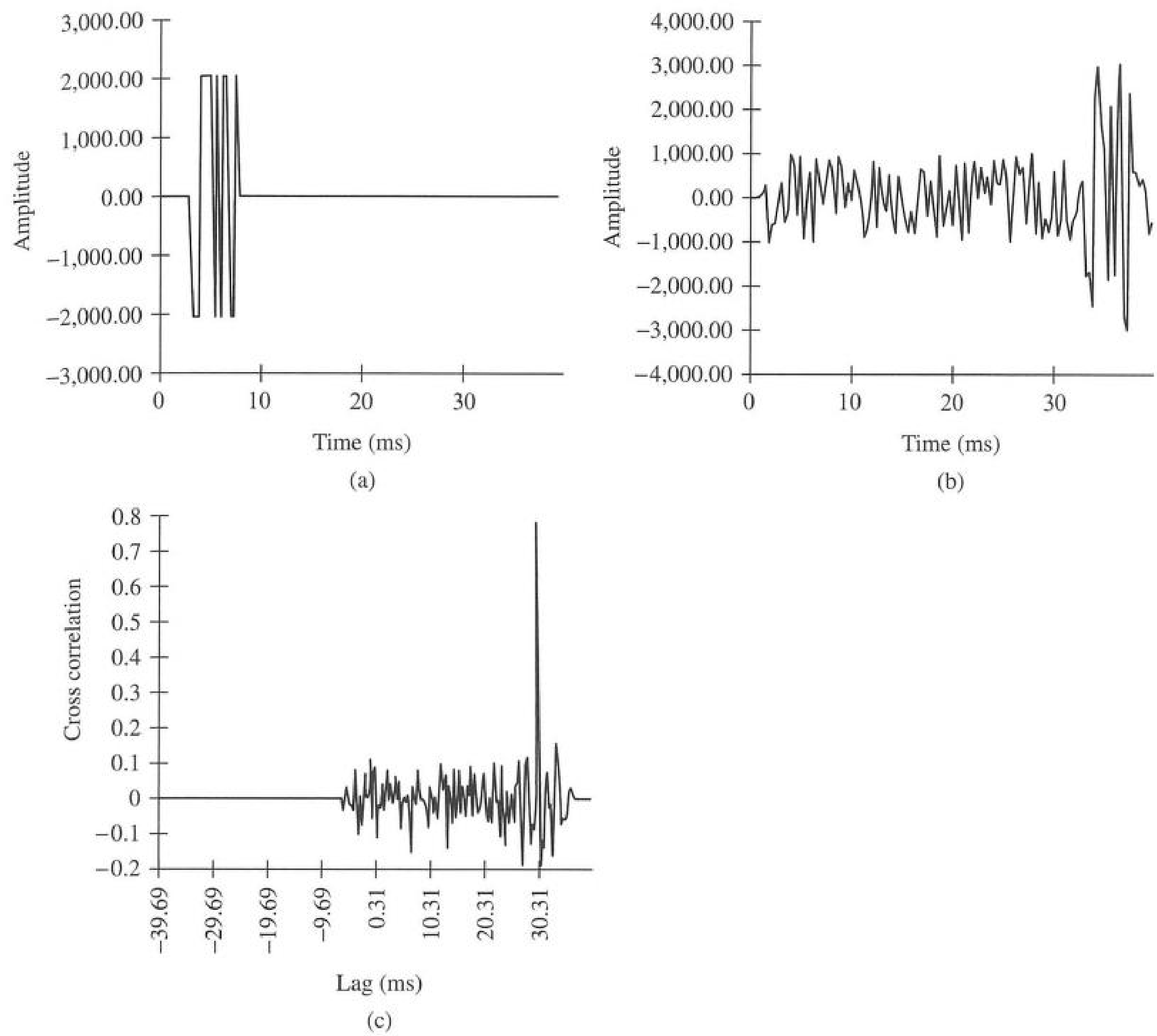

Figure 1.3 Cross-correlation of a random signal, $x(t)$, and a delayed noisy version of the same signal, $y(t)$. The delay between the two signals is the time from the origin to the time where the peak occurred in their cross-correlation in (c).

Examples of CCF and ACF for certain signals are given in Figures 1.2 and 1.3. Notice, for example, that the ACF of the noise-corrupted signal shows clearly that there is a periodic signal buried in noise (Figure 1.2). Figure 1.3 illustrates how to measure delays. The amount of delay introduced by the system is clearly evident from the CCF and can be measured from the time origin to the large peak.

1.3.3 Digital filtering

Digital filtering is one of the most important operations in DSP as will become clear in subsequent chapters. The digital filtering operation for an important class of filters is defined as

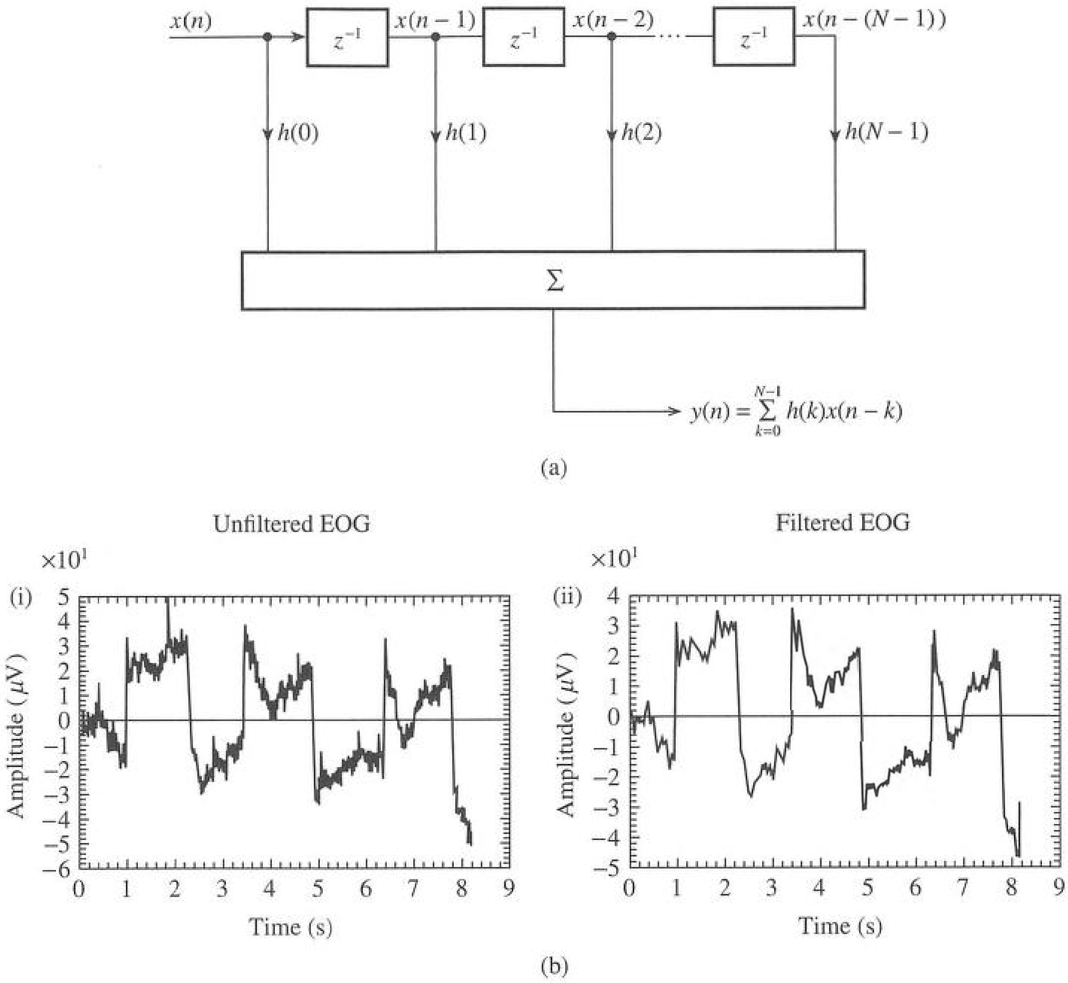

Figure 1.4 (a) Block diagram representation of the transversal filter. $h(k)$, $k = 0, 1, \ldots, N-1$, are the filter coefficients, and each box containing z^{-1} represents a delay of one sampling period. (b) Digital lowpass filtering of a biomedical signal to remove noise.

$$y(n) = \sum_{k=0}^{N-1} h(k)\,x(n-k)$$

where $h(k)$, $k = 0, 1, \ldots, N-1$, are the coefficients of the filter, and $x(n)$ and $y(n)$, respectively, the input and output of the filter. For a given filter, the values of its coefficients are unique to it and determine the filter's characteristics.

We note that filtering is in fact the convolution of the signal and the filter's impulse response in the time domain, that is $h(k)$. Figure 1.4(a) shows a block diagram representation of the filter defined above. In this form, the filter is popularly known as the transversal filter. In the figure, z^{-1} represents a delay of one sample time.

A common filtering objective is to remove or reduce noise from a wanted signal. For example, Figure 1.4(b) shows the effects of digital lowpass filtering of a certain biomedical signal to remove high frequency distortion. The use of a digital filter in this application was especially important to minimize the distortion of the in-band signal components.

1.3.4 Discrete transformation

Discrete transforms allow the representation of discrete-time signals in the frequency domain or the conversion between time and frequency domain representations. The spectrum of a signal is obtained by decomposing it into its constituent frequency components using a discrete transform. A knowledge of such a spectrum is invaluable in, for example, determining the bandwidth required to transmit the signal. Conversion between time and frequency domains is necessary in many DSP applications. For example, it allows for a more efficient implementation of DSP algorithms, such as those for digital filtering, convolution and correlation.

Many discrete transformations exist, but the discrete Fourier transform (DFT) is the most widely used and is defined as

$$X(k) = \sum_{n=0}^{N-1} x(n)W^{nk}, \text{ where } W = \mathrm{e}^{-\mathrm{j}2\pi/N}$$

An example of the use of the DFT is given in Figure 1.5. Here, the impulse response of a filter, $h(n)$, $n = 0, 1, \ldots, N-1$, is transformed to give the frequency response of the filter using the DFT. Details of the DFT and its applications are given in Chapters 3, 4 and 11.

1.3.5 Modulation

Digital signals are rarely transmitted over long distances or stored in large quantities in their raw form. The signals are normally modulated to match their frequency characteristics to those of the transmission and/or storage media to minimize signal distortion, to utilize the available bandwidth efficiently, or to ensure that the signals have some desirable properties. Perhaps the two application areas where modulation is extensively employed are telecommunications and digital audio engineering.

The process of modulation often involves varying a property of a high frequency signal, known as the carrier, in sympathy with the signal we wish to transmit or store, called the modulating signal. The three most commonly used digital modulation schemes for transmitting digital data over a bandpass channel (for example a microwave link) are amplitude shift keying (ASK), phase shift keying (PSK), and frequency shift keying (FSK). When digital data is transmitted over an all-digital network, a scheme known as pulse code modulation (PCM) is commonly used (see, for example, Bellamy, 1982). Several other modulation schemes have been developed for digital audio, details of which can be found in Watkinson (1987).

Correlation and convolution

5

The nature of the correlation process is first described in this chapter followed by an explanation using worked examples of the calculation of cross- and autocorrelations. The attenuating effects of correlation on the noise content of signals is described, as are a number of applications of correlation. The technique of fast correlation utilizing the FFT is then explained. The topic of convolution is covered in a similar manner to correlation. The treatment includes circular and linear convolution, fast linear convolution, and the sectioning methods (overlap–add, overlap–save) needed to handle large amounts of input data. Deconvolution is also included. The relationship between correlation and convolution is established. The chapter finishes with a section on implementation and some worked application examples.

5.1 Introduction

It is frequently necessary to be able to quantify the degree of interdependence of one process upon another, or to establish the similarity between one set of data and another. In other words, the correlation between the processes or data is sought. Correlation can be defined mathematically and can be quantified. The process of correlation occupies a significant place in signal processing. Applications are found in image processing for robotic vision or remote sensing by satellite in which data from different images is compared, in radar and sonar systems for range and position finding in which transmitted and reflected waveforms are compared, in the detection and identification of signals in noise, in control engineering for observing the effect of inputs on outputs, in the identification of binary codewords in pulse code modulation systems using correlation detectors, as an integral part of the ordinary least squares estimation technique, in the computation of the average power in waveforms, and in many other fields, such as, for example, climatology. Correlation is also an integral part of the process of convolution. The convolution process is essentially the correlation of two data sequences in which one of the sequences has been reversed. This means that the same algorithms may be used to compute correlations and convolutions simply by reversing one of the sequences. The process of convolution gives the output from a system which filters the input. The spectrum of a recorded signal consists of the convolution of the spectrum of the signal with the spectrum of its window function.

The determination of an unknown system impulse response is known as system identification. The determination of an unknown input from the system impulse response and the output signal is known as deconvolution. When the impulse response is unknown, the determination of the unknown input signal is known as blind deconvolution. Each of these important topics is described.

5.2 Correlation description

Consider how two data sequences, each consisting of simultaneously sampled values taken from the two corresponding waveforms, might be compared. If the two waveforms varied similarly point for point, then a measure of their correlation might be obtained by taking the sum of the products of the corresponding pairs of points. This proposal becomes more convincing when the case of two independent and random data sequences is considered. In this case the sum of the products will tend towards a vanishingly small random number as the number of pairs of points is increased. This is because all numbers, positive and negative, are equally likely to occur so that the product pairs tend to be self-cancelling on summation. By contrast, the existence of a finite sum will indicate a degree of correlation. A negative sum will indicate negative correlation, that is an increase in one variable is associated with a decrease in the other variable. The cross-correlation $r_{12}(n)$ between two data sequences $x_1(n)$ and $x_2(n)$ each containing N data might therefore be written as

$$r_{12} = \sum_{n=0}^{N-1} x_1(n) x_2(n)$$

This definition of cross-correlation, however, produces a result which depends on the number of sampling points taken. This is corrected for by normalizing the result to the number of points by dividing by N. Alternatively this may be regarded as averaging the sum of products. Thus, an improved definition is

$$r_{12} = \frac{1}{N}\sum_{n=0}^{N-1} x_1(n)x_2(n)$$

Example 5.1

The calculation of r_{12} is illustrated in the following example, in which the point numbers in the data sequences are the n, and the sequences are x_1 and x_2.

n	1	2	3	4	5	6	7	8	9
x_1	4	2	−1	3	−2	−6	−5	4	5
x_2	−4	1	3	7	4	−2	−8	−2	1

$$r_{12} = \frac{1}{9}(4\times -4 + 2\times 1 + -1\times 3 + 3\times 7 + -2\times 4 + -6\times -2 + -5\times -8 + 4\times -2 + 5\times 1)$$
$$= 5$$

However, this definition needs modification to be useful. In some cases it may indicate zero correlation although the two waveforms are 100% correlated. This may occur, for example, when the two waveforms are out of phase, which will often be the case. The situation is illustrated by the waveforms of Figure 5.1. From this figure it is seen that each pair product in the correlation is zero, and hence the correlation is zero, because one of either x_1 or x_2 is always zero. However, the waveforms are clearly highly correlated, although they are out of phase. The phase difference could, for example, occur because x_1 is the reference signal while x_2 is the delayed output from a circuit. To overcome such phase differences it is necessary to shift, or lag, one of the waveforms with respect to the other. Typically x_2 is shifted to the left to align the

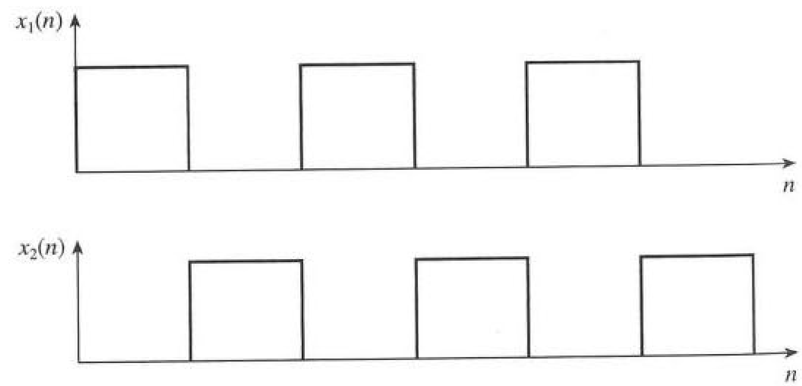

Figure 5.1 Out-of-phase 100% correlated waveforms with zero correlation at lag zero.

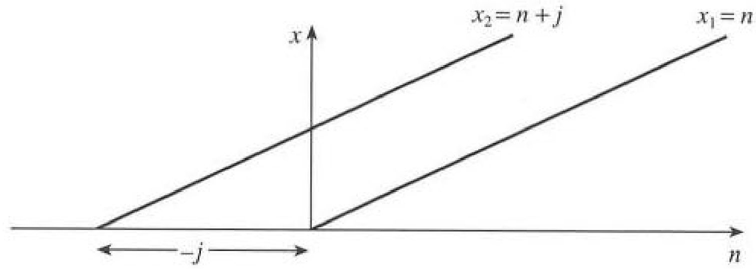

Figure 5.2 Waveform $x_2 = x_1 + j$ shifted j lags to the left of waveform x_1.

waveforms prior to correlation. As illustrated in Figure 5.2 this is equivalent to changing $x_2(n)$ to $x_2(n+j)$, where j represents the amount of lag which is the number of sampling points by which x_2 has been shifted to the left. An alternative, but equivalent, procedure is to shift x_1 to the right. The formula for the cross-correlation thus becomes

$$r_{12}(j) = \frac{1}{N}\sum_{n=0}^{N-1} x_1(n)x_2(n+j)$$

$$= r_{21}(-j) = \frac{1}{N}\sum_{n=0}^{N-1} x_2(n)x_1(n-j) \qquad (5.1)$$

In practice when two waveforms are correlated their phase relationship will probably not be known and so the correlation will be computed for a number of different lags in order to establish the largest value of the correlation which is then taken to be the correct value.

Example 5.2

Consider the cross-correlation of the above two sequences $x_1(n)$ and $x_2(n)$ at a lag of $j = 3$, that is consider $r_{12}(3)$. The two sequences become

n	1	2	3	4	5	6	7	8	9
x_1	4	2	−1	3	−2	−6	−5	4	5
x_2	7	4	−2	−8	−2	−1			

so

$$r_{12}(3) = \frac{1}{9}(4 \times 7 + 2 \times 4 + -1 \times -2 + 3 \times -8 + -2 \times -2 + -6 \times -1)$$

$$= 2.667$$

Of course, it is also possible to consider correlation in the continuous time domain, and some analog signal correlation is implemented this way. In the continuous domain $n \rightarrow t$ and $j \rightarrow \tau$ and

$$r_{12}(\tau) = \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} x_1(t)x_2(t+\tau)\,\mathrm{d}t \tag{5.2}$$

However, if $x_1(t)$ and $x_2(t)$ are periodic with period T_0 Equation 5.2 simplifies to

$$r_{12}(\tau) = \frac{1}{T_0}\int_{-T_0/2}^{T_0/2} x_1(t)x_2(t+\tau)\,\mathrm{d}t \tag{5.3}$$

If the waveforms are finite energy waveforms, for example nonperiodic pulse-type waveforms, then the average evaluated over time T as $T \to \infty$ is not taken because then $1/T \to 0$ and $r_{12}(\tau)$ is always vanishingly small. For this case Equation 5.4 is used in principle:

$$r_{12}(\tau) = \int_{-\infty}^{\infty} x_1(t)x_2(t+\tau)\,\mathrm{d}t \tag{5.4}$$

In practice, a finite record length will be processed and so Equation 5.5 or 5.1 will be applied:

$$r_{12}(\tau) = \frac{1}{T}\int_{0}^{T} x_1(t)x_2(t+\tau)\,\mathrm{d}t \tag{5.5}$$

There is another difficulty associated with cross-correlating finite lengths of data. This can be seen in the above example in which $r_{12}(3) = 2.667$ was determined. As x_2 is shifted to the left the waveforms no longer overlap and data at the ends of the sequences no longer form pair products. This is known as the end effect. In the example the number of pairs has dropped from nine to six for a lag of three. The result is a linear decrease in $r_{12}(j)$ as j increases, leading to debatable values of $r_{12}(j)$. One possible solution is to make one of the sequences twice as long as the required length for correlation. This could be achieved by recording more data, or, if one of the sequences were periodic, by repeating the sequence (taking care to match the two ends). Another possibility is to add a correction to all computed values. Figure 5.3 shows how $r_{12}(j)$ decreases with j purely as a result of the end effect, that is actual variations in $r_{12}(j)$ are not included. At $j = 0$, $r_{12}(j) = r_{12}(0)$, which can be computed. At $j = N$, $r_{12}(N) = 0$, because the waveforms no longer overlap. In between, at some lag j, the true value of $r_{12}(j)$ is $r_{12}(j)_{\text{true}}$ while the actual value caused by the end effect is $r_{12}(j)$. Then, from the figure

$$\frac{r_{12}(j)_{\text{true}} - r_{12}(j)}{j} = \frac{r_{12}(0)}{N}$$

whence

$$r_{12}(j)_{\text{true}} = r_{12}(j) + \frac{j}{N}r_{12}(0) \tag{5.6}$$

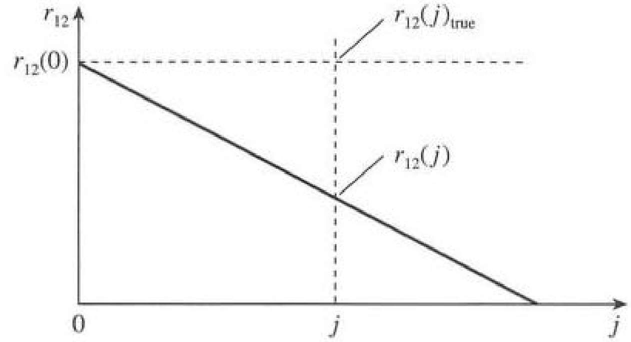

Figure 5.3 The effect of the end-effect on the cross-correlation $r_{12}(j)$.

Computed values of the cross-correlation are therefore easily corrected for end effects by adding $jr_{12}(0)/N$ to the values of $r_{12}(j)$.

The cross-correlation values computed according to the above formulae depend on the absolute values of the data. It is often necessary to measure cross-correlations according to the fixed scale between −1 and +1. This can be achieved by normalizing the values by an amount depending on the energy content of the data. For example, consider the two pairs of waveforms $x_1(n)$, $x_2(n)$, and $x_3(n)$, $x_4(n)$. The data values are given in the table below:

n	0	1	2	3	4	5	6	7	8
$x_1(n)$	0	3	5	5	5	2	0.5	0.25	0
$x_2(n)$	1	1	1	1	1	0	0	0	0
$x_3(n)$	0	9	15	15	15	6	1.5	0.75	0
$x_4(n)$	2	2	2	2	2	0	0	0	0

As may be seen from Figure 5.4, waveforms $x_1(n)$ and $x_3(n)$ are alike, differing only in magnitude. The same is true of the pair $x_2(n)$ and $x_4(n)$. The correlation between $x_1(n)$ and $x_2(n)$ is therefore the same as that between $x_3(n)$ and $x_4(n)$. However, the cross-correlations $r_{12}(1)$ and $r_{34}(1)$ are 1.47 and 8.83 respectively. They are different because they depend on the absolute values of the data. This situation can be rectified by normalizing the cross-correlation $r_{12}(j)$ by the factor

$$\left[\frac{1}{N}\sum_{n=0}^{N-1} x_1^2(n) \times \frac{1}{N}\sum_{n=0}^{N-1} x_2^2(n)\right]^{1/2} = \frac{1}{N}\left[\sum_{n=0}^{N-1} x_1^2(n) \sum_{n=0}^{N-1} x_2^2(n)\right]^{1/2} \tag{5.7}$$

and similarly for $r_{34}(j)$. The normalized expression for $r_{12}(j)$ then becomes

$$\rho_{12}(j) = \frac{r_{12}(j)}{\frac{1}{N}\left[\sum_{n=0}^{N-1} x_1^2(n) \sum_{n=0}^{N-1} x_2^2(n)\right]^{1/2}} \tag{5.8}$$

$\rho_{12}(j)$ is known as the cross-correlation coefficient. Its value always lies between −1 and +1. +1 means 100% correlation in the same sense, −1 means 100% correlation in

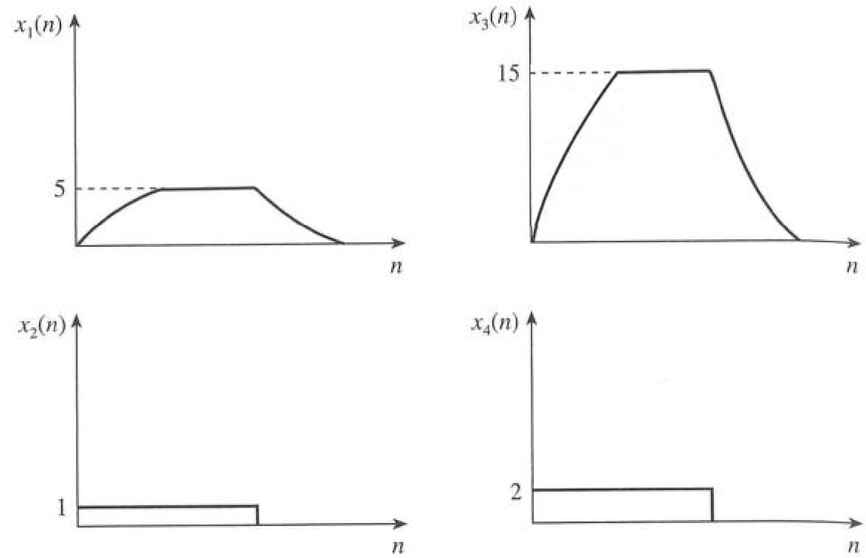

Figure 5.4 Pairs of waveforms $\{x_1(n), x_2(n)\}$, $\{x_3(n), x_4(n)\}$ of different magnitudes but equal cross-correlations.

the opposing sense, for example signals in antiphase. A value of 0 signifies zero correlation. This means the signals are completely independent. This would be the case, for example, if one of the waveforms were completely random. Small values of $\rho_{12}(j)$ indicate very low correlation. The normalizing factor for $r_{12}(j)$ in the above illustration is

$$\frac{1}{N}\left[\sum_{n=0}^{N-1} x_1^2(n) \sum_{n=0}^{N-1} x_2^2(n)\right]^{1/2} = \frac{1}{9}(88.31 \times 6)^{1/2} = 2.56$$

and for $r_{34}(j)$ it is

$$\frac{1}{N}\left[\sum_{n=0}^{N-1} x_3^2(n) \sum_{n=0}^{N-1} x_4^2(n)\right]^{1/2} = \frac{1}{9}(794.8 \times 24)^{1/2} = 15.35$$

Therefore

$$\rho_{12}(1) = \frac{r_{12}(1)}{2.56} = \frac{1.47}{2.56} = 0.57$$

and

$$\rho_{34}(1) = \frac{r_{34}(1)}{15.34} = \frac{8.83}{15.35} = 0.58$$

Now $\rho_{12}(1) = \rho_{34}(1)$ which demonstrates that this normalization process indeed allows a comparison of cross-correlations independently of the absolute data values.

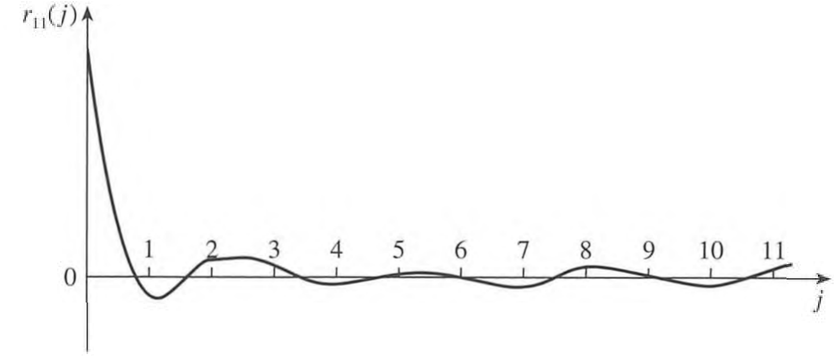

Figure 5.5 Autocorrelation function of a random waveform.

A special case occurs when $x_1(n) = x_2(n)$. The waveform is then cross-correlated with itself. This process is known as autocorrelation. The autocorrelation of a waveform is given by

$$r_{11}(j) = \frac{1}{N}\sum_{n=0}^{N-1} x_1(n)x_1(n+j)$$

The autocorrelation function has one very useful property in that

$$r_{11}(0) = \frac{1}{N}\sum_{n=0}^{N-1} x_1^2(n) = S$$

where S is the normalized energy of the waveform. This provides a method for calculating the energy of a signal. If the waveform is completely random, for example corresponding to that of white, gaussian noise in an electrical system, then the autocorrelation will have its peak value at zero lag and will reduce to a random fluctuation of small magnitude about zero for lags greater than about unity (see Figure 5.5). This constitutes a test for random waveforms. This topic will be more fully covered in Section 5.2.1. It is also true that

$$r_{11}(0) \geqslant r_{11}(j)$$

5.2.1 Cross- and autocorrelation

Care has to be exercised when cross-correlating two unequal length sequences when they are periodic. This is because the result of the correlation will be cyclic with the period of the shorter sequence. This result does not represent the full periodicity of the longer sequence and is, therefore, incorrect. This may be demonstrated by cross-correlating the sequences $a = \{4, 3, 1, 6\}$ and $b = \{5, 2, 3\}$ to obtain $r_{ab}(j)$. The sequence b is placed below sequence a, and b is shifted left by one lag on each of the subsequent rows, with the value of the cross-correlation appearing in the final column on the right.

Sequence				*Lag*	$r_{ab}(j)$	
4	3	1	6			
3	5	2	3	0	47	
5	2	3	5	1	59	
2	3	5	2	2	34	
3	5	2	3	3	47	$r_{ab}(j)$ repeats
5	2	3	5	4	59	
etc.						

The result shows that $r_{ab}(j)$ is cyclic, repeating every third lag, that is $r_{ab}(j)$ has the same period as that of the shorter sequence, b. This procedure is known as cyclic correlation. To obtain the correct value in which each value in a is multiplied by each value in b, all the elements in b have to be shifted in turn below each value in a as shown below:

```
          4 3 1 6
                5 2 3
              5 2 3
            5 2 3
          5 2 3
        5 2 3
      5 2 3
    5 2 3
```

This is seen to require 6 lags before the b sequence repeats. The sequence lengths are 4 and 3 and the number of lags necessary is $4 + 3 - 1 = 6$. This reveals the general rule for obtaining the linear cross-correlation of two periodic sequences of lengths N_1 and N_2: add augmenting zeros to each sequence to make the lengths of each sequence $N_1 + N_2 - 1$. This may be expressed as adding $N_2 - 1$ zeros to the sequence of length N_1 and adding $N_1 - 1$ zeros to the sequence of length N_2. This is now demonstrated for the given sequences a and b:

Sequence						*Lag*	$r_{ab}(j)$	
4	3	1	6	0	0			
5	2	3	0	0	0	0	29	
2	3	0	0	0	5	1	17	
3	0	0	0	5	2	2	12	
0	0	0	5	2	3	3	30	
0	0	5	2	3	0	4	17	
0	5	2	3	0	0	5	35	
5	2	3	0	0	0	6	29	$r_{ab}(j)$ repeats
etc.								

Thus, the required linear cross-correlation of a and b is

$$r_{ab}(j) = \{29, 17, 12, 30, 17, 35\}$$

So far, the instances of cross-correlation taken have all assumed digitized data, but cross-correlation may also be performed analytically when analytical expressions can be written for the waveforms, including when this requires sectioning of the waveforms. In practice the analytical procedure has its equivalent in the use of analog circuits to effect the cross-correlation. An example of analytical cross-correlation follows.

Example 5.3

Obtain the cross-correlation $r_{12}(-\tau)$ between the waveforms $v_1(t)$ and $v_2(t)$ of Figure 5.6.

It is easy to express the waveforms analytically by dividing them into straight-line sections. It is only necessary to do this over one period, T, of the waveforms because $r_{12}(-\tau)$ will be periodic in τ with period T. Therefore, for $0 \leqslant t \leqslant T$, $v_1(t) = t/T$, and for $0 \leqslant t \leqslant T/2$, $v_2(t) = 1.0$, while for $T/2 \leqslant t \leqslant T$, $v_2(t) = -1.0$. The requirement is to obtain an expression for $r_{12}(-\tau)$, that is $v_2(t)$, the rectangular waveform, is to be shifted right with respect to $v_1(t)$. For $0 \leqslant \tau \leqslant T/2$, the situation is described by Figure 5.7

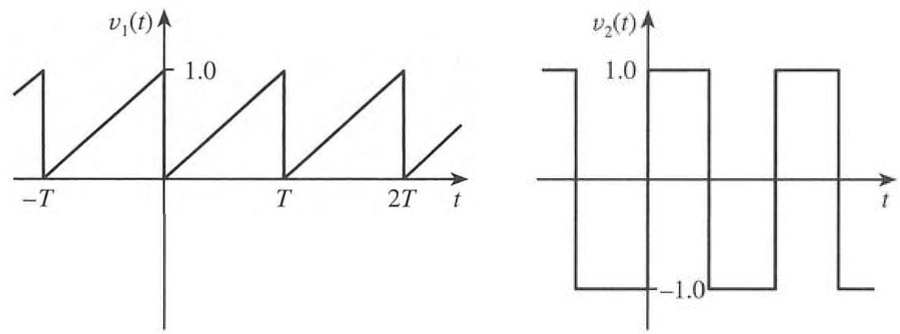

Figure 5.6 The waveform $v_1(t)$ and $v_2(t)$ for cross-correlation example.

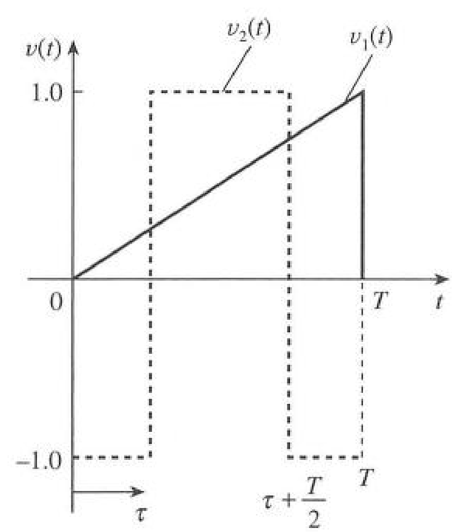

Figure 5.7 Sections of $v_2(t)$ for $\theta \leqslant \tau \leqslant T$.

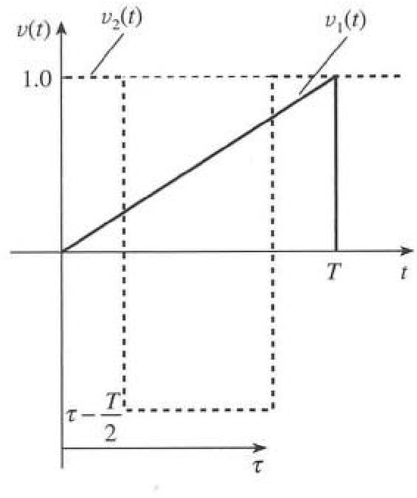

Figure 5.8 Sections of $v_2(t)$ for $T/2 \leqslant \tau \leqslant T$.

which shows that v_1(t) has to be multiplied by three consecutive sections of $v_2(t)$ in which $v_2(t)$ has the consecutive values −1, 1, −1. For $T/2 \leqslant \tau \leqslant T$, Figure 5.8 applies in which the consecutive values of the set of $v_2(t)$ have changed to 1, −1, +1. This means there are two parts to the solution which must match at $\tau = T/2$.

Referring to Figure 5.7, the cross-correlation is split into the three sections with boundaries at $t = \tau$, $t = \tau + T/2$, and $t = T$. Hence

$$r_{12}(-\tau) = \frac{1}{T}\int_0^T v_1(t)v_2(t-\tau)\,\mathrm{d}t$$

$$= \frac{1}{T}\int_0^{\tau} \frac{t}{T}(-1)\,\mathrm{d}t + \frac{1}{T}\int_{\tau}^{\tau+T/2} \frac{t}{T}(1)\,\mathrm{d}t + \frac{1}{T}\int_{\tau+T/2}^{T} \frac{t}{T}(-1)\,\mathrm{d}t$$

$$= \frac{-1}{T^2}\left[\frac{t^2}{2}\right]_0^{\tau} + \frac{1}{T^2}\left[\frac{t^2}{2}\right]_{\tau}^{\tau+T/2} - \frac{1}{T^2}\left[\frac{t^2}{2}\right]_{\tau+T/2}^{T}$$

$$r_{12}(-\tau) = -\frac{1}{4} + \frac{\tau}{T} \quad \text{for } 0 \leqslant \tau \leqslant \frac{T}{2} \tag{5.9}$$

For $T/2 \leqslant \tau \leqslant T$, and referring to Figure 5.8, it is seen that

$$r_{12}(-\tau) = \frac{1}{T}\int_0^{\tau-T/2} \frac{t}{T}(1)\,\mathrm{d}t + \frac{1}{T}\int_{\tau-T/2}^{\tau} \frac{t}{T}(-1)\,\mathrm{d}t + \frac{1}{T}\int_{\tau}^{T} \frac{t}{T}(1)\,\mathrm{d}t$$

$$r_{12}(-\tau) = \frac{3}{4} - \frac{\tau}{T} \quad \text{for } \frac{T}{2} \leqslant \tau \leqslant T \tag{5.10}$$

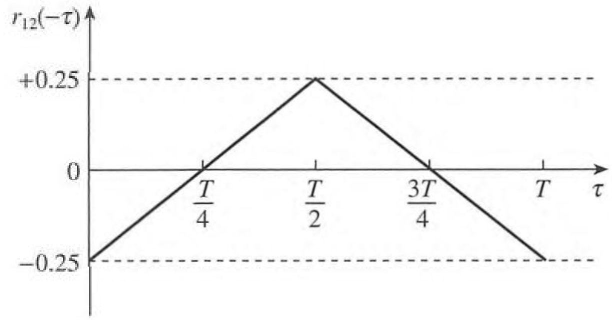

Figure 5.9 $r_{12}(-\tau)$ as a function of τ.

Substituting $\tau = T/2$ into Equations 5.9 and 5.10 gives $r_{12}(-\tau) = 1/4$ in both cases, confirming that the two functions match correctly. Figure 5.9 shows a plot of $r_{12}(-\tau)$ versus τ for $0 \leqslant \tau \leqslant T$.

It is of interest to give some consideration to the consequences of using finite lengths of data in the calculation of the correlation. In other words, what is the effect of using Equation 5.5, in which T is finite, instead of Equation 5.2?

This question can be answered by considering just one sinusoidal Fourier harmonic component of the signal. Equation 5.2 will give the correct autocorrelation, in which $T \gg T_p$, where T_p is the period of the sinusoid. Thus

$$r_{11}(\tau) = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} A\sin(\omega t)A\sin(\omega t + \tau)\,\mathrm{d}t$$

$$= \lim_{T\to\infty} \frac{A^2}{2}\left[\cos(\omega\tau) - \frac{\cos(\omega T)}{2\omega T}\sin(\omega\tau)\right] \tag{5.11}$$

Inspection of this equation shows that the second term in the bracket $\to 0$ when $T \to \infty$, so when $T \neq \infty$ it represents an error. The $\cos(\omega T)$ term represents periodic error effects, while the term $1/2\omega T$ gives the trend in the error. Thus, as far as the correlation length, T, is concerned, the errors are greater the shorter the sequence, and are also largest for the lower frequency components of the waveform. The errors are also periodic in τ.

The $\cos(\omega T)$ term gives least errors when $\omega T = [(2n + 1)/2]\pi$. Since $\omega = 2\pi/T_p$ and large values of T are sought, this corresponds to

$$T \geqslant (2n + 1)\frac{T_p}{4} \tag{5.12}$$

The $\sin(\omega\tau)$ term is least when $\omega\tau = m\pi$, where m is integer. Hence,

$$\tau = \frac{m}{2} T_p \qquad (5.13)$$

It is now necessary to make some reasonable assumptions. Assume the condition for large T is satisfied by $n \geqslant 10$. Then $T \geqslant nT_p/2$, or

$$T \geqslant 5T_p \qquad (5.14)$$

From Equation 5.13, the largest allowable value of τ for the lowest frequency component ($m = 1$) satisfies

$$\tau < T_p \qquad (5.15)$$

Combining Equations 5.14 and 5.15,

$$\tau \leqslant T/5$$

This means that when correlating waveforms the errors due to finite data lengths may be minimized by

(1) ensuring that $T \geqslant 5T_p$, where T_p corresponds to the lowest frequency component of interest, and
(2) overlapping the data by no more than 20% of their length.

Thus, for example, if telephone speech signals with a bandwidth of 300 Hz to 3.4 kHz and sampled at 40 kHz are to be correlated, $T_p = 1/300 = 3.3 \times 10^{-3}$ s. The least acceptable data length would be $5 \times 3.3 \times 10^{-3}$ s = 16.7 ms and the largest correlation shift would be 3.33 ms, or 133 data points.

Figure 5.10 shows the plot of $\rho_{11}(j)$, the autocorrelation coefficient of a purely random waveform, for example white noise. The expected value of $r_{11}(j)$ can be

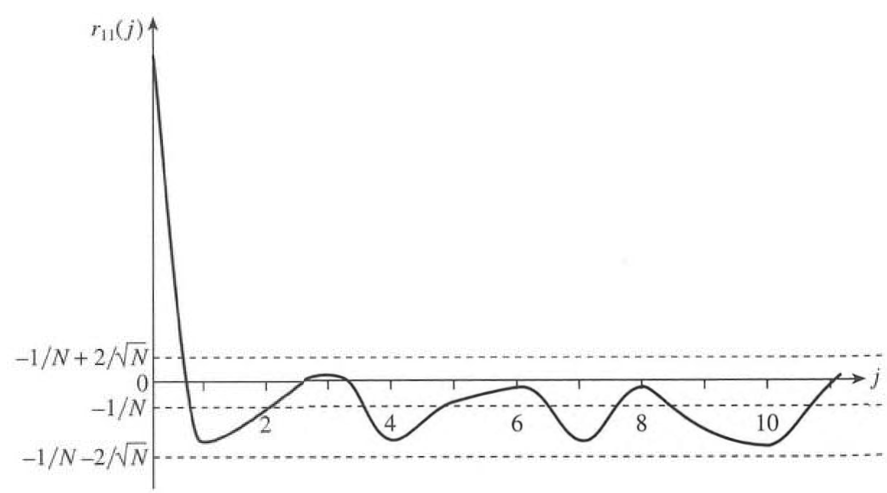

Figure 5.10 The autocorrelation coefficient of a random waveform.

shown to be $E[r_{11}(j)] \approx -1/N$ (Chatfield, 1980), where N is the number of data points, and its variance is $\mathrm{var}[r_{11}(j)] \approx 1/N$. The expected value of $-1/N$ is shown on the figure as are the 95% confidence limits of $-1/N$, which are $\pm 2/N^{1/2}$. Values of $r_{11}(j)$ which fall outside these confidence limits may be significant, that is they may indicate that the waveform was not truly random. However, it should be noted that as many as one point in 20 may lie outside these limits even when the waveform is completely random. For a random waveform $r_{11}(j)$ should fall to within the 95% confidence limits within one or two lags. Experience and more sophistication is required to be sure that a waveform is random. For example, data pre-whitening may be advisable (Jenkins and Watts, 1968).

The autocorrelation function of a periodic waveform is itself a periodic waveform. This is easily proved as follows. The periodic waveform $x(t)$ of period T satisfies

$$x(t) = x(t + nT)$$

so,

$$r_{11}(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)x(t+\tau)\,\mathrm{d}t$$

$$= \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t)x(t+\tau+nT)\,\mathrm{d}t$$

$$r_{11}(\tau) = r_{11}(\tau + nT) \tag{5.16}$$

Thus $r_{11}(\tau)$ is seen to be periodic in τ with period T. This is a useful property because it enables the detection of periodic signals in noise for small signal-to-noise ratios. Autocorrelating the waveform tends to reduce the noise while at the same time developing the periodic autocorrelation function of the signal. Once detected, further processing can be applied to determine its shape if this is required.

Equation 5.11 showed that the autocorrelation function of $A\sin(\omega t)$ is $(A^2/2)\cos(\omega\tau)$. In this case, as in others, the amplitude of the autocorrelation function is related simply to that of the signal, and may be used to estimate the signal amplitude. Another common example is that of the rectangular wave of amplitude A which the reader could show has a triangular autocorrelation function of amplitude A^2. Finally, it should be noted that autocorrelation functions are not unique. This means that a number of different waveforms may share the same autocorrelation function. Hence the shapes of waveforms should not be deduced from the detected autocorrelation functions.

Consider now the case in which the waveform, $v(t)$, is partially random. This represents the case of a noisy signal which may be written as the sum of a signal term, $s(t)$, and a noise term, $q(t)$. Thus

$$v(t) = s(t) + q(t) \tag{5.17}$$

$s(t)$ and $q(t)$ are assumed to be uncorrelated. The sampled autocorrelation function of $v(t)$ is $r_{vv}(j)$ given by

$$r_{vv}(j) = \frac{1}{N}\sum_{n=0}^{N-1}[s(n)+q(n)][s(n+j)+q(n+j)] \tag{5.18}$$

$$= \frac{1}{N}\sum_{n=0}^{N-1}s(n)s(n+j) + \frac{1}{N}\sum_{n=0}^{N-1}s(n)q(n+j) + \frac{1}{N}\sum_{n=0}^{N-1}q(n)s(n+j)$$

$$+\frac{1}{N}\sum_{n=0}^{N-1}q(n)q(n+j) \tag{5.19}$$

$$= r_{ss}(j) + E[s(n)q(n+j)] + E[q(n)s(n+j)] + E[q(n)q(n+j)]$$

$$= r_{ss}(j) + E[s(n)]E[q(n+j)] + E[q(n)]E[s(n+j)] + E[q(n)]E[q(n+j)]$$

$$= r_{ss}(j) + \overline{s(n)}\,\overline{q(n)} + \overline{q(n)}\,\overline{s(n)} + \overline{q(n)}^2$$

$$= r_{ss}(j) + 2\bar{s}\bar{q} + \bar{q}^2 \tag{5.20}$$

Now, $\bar{q} \to 0$ for large N, for which

$$r_{vv}(j) \to r_{ss}(j) \tag{5.21}$$

For smaller N, the cross-correlation terms in Equation 5.19 and the autocorrelation of the noise tend towards zero with increasing lag j.

Thus it is seen that the autocorrelation function of a partially random, or noisy, waveform consists of the autocorrelation function of the signal component superimposed on a noisy decaying function which depends on both the random and signal components and which decays towards the value $2\bar{s}\bar{q} + \bar{q}^2$. Thus the plot of $r_{vv}(j)$ against j will display the periodicity of $s(t)$ provided $|r_{ss}(j)| > |(2\bar{s}\bar{q} + \bar{q}^2)|$: see Figure 5.11. This offers a method for identifying the period of a signal in noise (see Section 5.2.2).

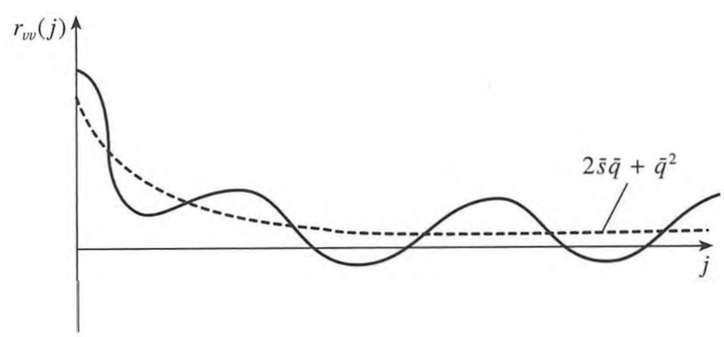

Figure 5.11 The autocorrelation function of a noisy signal.

Example 5.4

Derive the cross-correlation function of two noisy waveforms.

Let the two waveforms be $\{s_1(t) + q_1(t)\}$ and $\{s_2(t) + q_2(t)\}$. Their sampled cross-correlation, $r_{12}(j)$, is given by

$$r_{12}(j) = \frac{1}{N}\sum_{n=0}^{N-1}[\{s_1(n) + q_1(n)\}\{s_2(n+j) + q_2(n+j)\}] \tag{5.22}$$

$$= \frac{1}{N}\sum_{n=0}^{N-1}[s_1(n)s_2(n+j) + s_1(n)q_2(n+j) + q_1(n)s_2(n+j) + q_1(n)q_2(n+j)]$$

$$= \frac{1}{N}\sum_{n=0}^{N-1}s_1(n)s_2(n+j) + \frac{1}{N}\sum_{n=0}^{N-1}s_1(n)q_2(n+j) + \frac{1}{N}\sum_{n=0}^{N-1}q_1(n)s_2(n+j)$$

$$+ \frac{1}{N}\sum_{n=0}^{N-1}q_1(n)q_2(n+j)$$

$$= r_{s_1s_2}(j) + r_{s_1q_2}(j) + r_{q_1s_2}(j) + r_{q_1q_2}(j) \tag{5.23}$$

As in the previous case of autocorrelation the last three terms on the right-hand side of Equation 5.23 decay towards zero with increasing lag j. For large N, Equation 5.23 becomes

$$r_{12}(j) = r_{s_1s_2}(j) + \overline{s_1}\,\overline{q_2} + \overline{q_1}\,\overline{s_2} + \overline{q_1}\,\overline{q_2} \tag{5.24}$$

Thus as N increases $r_{12}(j) \rightarrow r_{s_1s_2}(j)$, the cross-correlation function of the two signals.

The above analyses illustrate that the cross- and autocorrelation processes emphasize signal properties by reducing the noise content.

5.2.2 Applications of correlation

5.2.2.1 Calculations of energy spectral density and energy content of waveforms

It can be shown that

$$F[r_{11}(\tau)] = G_E(f) \tag{5.25}$$

where $G_E(f)$ is the energy spectral density of the waveform, that is the energy spectral density and the autocorrelation function constitute a Fourier transform pair.

It can further be shown that

$$r_{11}(0) = E \tag{5.26}$$

where E is the total energy of the waveform.