SIGNAL PROCESSING AND INTEGRATED CIRCUITS

'The scribes full of wisdom,
Their names will last forever.
They leave for an inheritance,
Their teachings and their books.
Their teachings are their Pyramids,
Their magical power touches all those who read their writings.'

Egyptian Hieratic Papyrus in the British Museum

'The instinct of constructiveness, which is one of the chief incentives to artistic creation, can find in scientific systems a satisfaction more massive than any epic poem. Disinterested curiosity, which is the essence of almost all intellectual effort, finds with astonished delight that science can unveil secrets which might well have seemed for ever undiscoverable . . . A life devoted to science is therefore a happy life, and its happiness is derived from the very best sources that are open to dwellers on this troubled and passionate planet.'

Bertrand Russell
'*The Place of Science in a Liberal Education*'

SIGNAL PROCESSING AND INTEGRATED CIRCUITS

Hussein Baher
Professor Emeritus of Electronic Engineering
Alexandria Institute of Engineering and Technology, Egypt

A John Wiley & Sons, Ltd., Publication

This edition first published 2012
© 2012, John Wiley & Sons, Ltd

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book. This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

MATLAB® is a trademark of The MathWorks, Inc. and is used with permission. The MathWorks does not warrant the accuracy of the text or exercises in this book. This book's use or discussion of MATLAB® software or related products does not constitute endorsement or sponsorship by The MathWorks of a particular pedagogical approach or particular use of the MATLAB® software.

Library of Congress Cataloging-in-Publication Data

Baher, H.
Signal processing and integrated circuits / Hussein Baher.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-71026-5 (cloth)
1. Signal processing – Equipment and supplies. 2. Signal processing – Mathematics. 3. Integrated circuits – Design and construction. 4. Electric filters. I. Title.
TK5102.9.B353 2012
621.382′2 – dc23

2011053421

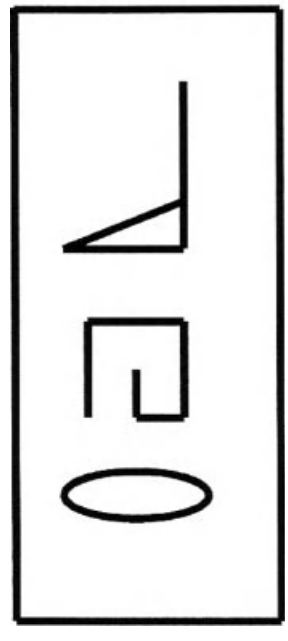
A catalogue record for this book is available from the British Library.

ISBN: 9780470710265

Set in 10/12pt Times by Laserwords Private Limited, Chennai, India

For

Darja



Contents

About the Author

Professor Hussein Baher was born in Alexandria and received his early education at the University of Alexandria and the American University in Cairo. He obtained his Ph.D. in Electronic Engineering from University College Dublin and has held academic positions at universities worldwide, including University College Dublin, Virginia Polytechnic Institute and State University, the Professorship of Electronic Engineering at Dublin City University, and the prestigious *Analog Devices* Professorship of Microelectronics in Massachusetts, United States. He has also been Visiting Professor at the *Technische Universitaet Wien* (TUW), Vienna, Austria.

He has published extensively on circuit design and signal processing including five books: *Synthesis of Electrical Networks* (John Wiley & Sons, Ltd, 1984), *Selective Linear-phase Switched-capacitor and Digital Filters* (Kluwer, 1993), *Microelectronic Switched-capacitor Filters: with ISICAP: a Computer-aided Design Package* (John Wiley & Sons, Ltd, 1996) and *Analog and Digital Signal Processing* (John Wiley & Sons, Ltd, 1990; 2nd Edition, 2001).

Prof. Baher spends his time in Dublin, Vienna and Alexandria as Professor Emeritus of Electronic Engineering. He is also a Member of the *Archaeological Society of Alexandria* and lectures in Dublin and Vienna on Ancient Egyptian civilization.

Preface

'Exact Calculation: The Gateway to Everything.'

Ahmes
'An Egyptian Mathematical Papyrus[1], 1850 BC'

In 2006, Austria was celebrating the 250th anniversary of Mozart's birth. While enjoying the festivities, I gave two graduate-level courses at the Technical University of Vienna (TUW). One course was on *Digital Signal Processing* and the other dealt with *Analog Integrated Circuits for Signal Processing* with application to the design of *Switched-capacitor Filters and Sigma-Delta Data Converters*. The two courses complemented each other to such an extent that the idea of writing a book combining the material of both courses was quite attractive. As the idea became more compelling, the material was updated and the result is this book.

The objective of this book is to provide a coherent and harmonious account of both *analog and digital signal processing*. In the case of digital systems, the design is at the relatively high level of adders, multipliers and delays. In the case of analog systems, the emphasis is laid on *integrated circuit* implementations of both *continuous-time* and *sampled-data (discrete)* circuits and systems, reaching all the way to the transistor level. This provides a comprehensive treatment of analog MOS integrated circuits for signal processing, with application to the design of microelectronic switched-capacitor circuits and extension to the design of mixed-mode processors in the form of integrated sigma-delta data converters. In this context, integrated circuit realizations which have been used successfully in *submicron* and *deep submicron* implementations for *ultra high frequency* applications are also discussed. Finally, MATLAB®[2] is used throughout as a useful aid to the analysis and design problems.

The level of treatment is at the senior to first-year graduate and professional levels while providing enough coverage of fundamental junior-level material to make the book self-contained. The book is divided into four parts.

Part I contains one chapter, which is a general introduction. *Chapter 1* gives a general overview and perspective of the general area of signal processing and the related disciplines, mentioning several applications. The growing areas of *Systems on a Chip* (SoC) and *mobile communications* are used for illustration of the wealth of knowledge required to design a complex signal processing system and to demonstrate the complementary relationship between analog and digital systems.

[1] With remarkable honesty, Ahmes states that he was copying this papyrus in 1650 BC from an older version written 200 years earlier in 1850 BC. This papyrus contains the first ever account of binary arithmetic and calculates the area of a circle assuming $\pi = 4(8/9)^2 = 3.16$. Ahmes states that the area of a circle of diameter 9 equals the area of a square of side 8. In a different context, the dimensions of the Khufu Pyramid lead to $\pi = 22/7$.

[2] MATLAB® is a registered trademark of the Mathworks Inc.

Part II contains eight chapters dealing with the techniques of *signal processing in the analog and digital domains at the system and circuit levels while not reaching the transistor level. Chapter 2* is a review of the fundamental concepts and mathematical tools of analog signal and system analyses. This review can be regarded as a comprehensive summary of the fundamentals of analog signals and systems. It is the distillation of courses on these topics which are usually covered at the junior undergraduate level. Therefore, the discussion is quite *compact*, and the material can be used as an *easy reference* for later chapters and as a *short revision course. Chapter 3* discusses the general theory and techniques of analog continuous-time filter design. These are important in themselves and are also of direct relevance to the design of all types of filter, including those which are of the sampled-data type such as digital and switched-capacitor filters. This is because the filtering operation is based on the same principles and, very often, analog continuous-time models are used as starting points for the design of other types. The chapter concludes with a guide to the use of MATLAB® in the design of analog filters. Extensive use of this material will also be made in later chapters. *Chapter 4* gives a brief and concise review of the process of analog to digital conversion and the representation of discrete signals and systems. This should serve as a revision of the fundamentals of discrete signal and system analyses. In *Chapter 5*, the design techniques of digital filters are discussed in detail. Emphasis is, at first, laid on the conceptual organization and analytical methods of design. Then the chapter concludes with the details of how to use MATLAB® as a computer-aided deign tool. Numerous examples are given throughout the chapter of both analytical and computer-aided design methods. *Chapter 6* provides a discussion of the computational algorithms that have come to be known collectively as the *fast Fourier transform* (FFT). The discrete Fourier transform is introduced and its properties are examined. The applications of the FFT are discussed in relation to spectral analysis, convolution, correlation and filtering of signals. *Chapter 7* introduces the concepts and techniques suitable for the description of stochastic (random) signals. The discussion encompasses both analog and digital signals. However the systems which perform the processing of these signals are themselves *deterministic*. *Chapter 8* deals with the effects of using binary words with finite lengths in representing the various quantities in digital signal processors. The degradation caused by these effects is examined and the resulting errors are assessed quantitatively. In *Chapter 9*, a central problem in signal processing is addressed, namely: the estimation of some signal of interest from a set of received noisy data signals. This leads to the area of adaptive filtering. A closely related area is that of the *modelling* or *simulation* of the behaviour of an unknown system (or process) by a linear system. Initially, the principles of linear estimation and modelling are discussed, then it is shown how these can be implemented using adaptive algorithms.

Part III is devoted to the design of *analog MOS integrated circuits for signal processing*. In *Chapter 10* a brief review of MOS transistor fundamentals and integrated circuit fabrication techniques are given. *Chapter 11* provides a discussion of the basic integrated circuit building blocks such as amplifiers, current mirrors, and load devices. In *Chapter 12* the two-stage CMOS operational amplifier is introduced and complete design examples are given. *Chapter 13* deals with high performance operational amplifiers and operational transconductance amplifiers which are used in G_m-C circuits. Integrated circuit realizations which have been used successfully in *submicron* and *deep submicron* implementations for *ultra high frequency* applications are also discussed in this chapter. *Chapter 14* deals with integrated resistors, capacitors and switches which are building blocks in analog signal processing systems.

Part IV is devoted to the design of signal processing systems using *switched-capacitor and mixed-mode* (i.e. both analog and digital) circuits. *Chapter 15* is a detailed account of the design techniques of *microelectronic switched-capacitor filters*. These are *analog sampled-data circuits* which have established themselves as viable alternatives to digital circuits in many applications. Furthermore, they are particularly amenable to implementation using the same CMOS integrated circuit technology which is used in digital processing, and consequently they can be easily integrated on the same chip with the digital circuits. Both the theoretical foundation and practical considerations are discussed in detail. Of particular importance in analog systems is the non-ideal behaviour of the building blocks, since this can lead to deteriorated performance if not understood and taken into account early in the design. These are treated in detail in *Chapter 16* together with many practical considerations in the design of analog integrated circuits. *Chapter 17* gives a detailed discussion of a highly instructive class of signal processor: the Σ-Δ converter. Its analysis and design require knowledge of both analog and digital signal processing, as well as most of the analytical and computational techniques discussed in this book. Therefore, it is ideal for inclusion in this book, which attempts to unify the two fields in one volume and should serve as a good illustration of the validity of the adopted approach.

Numerous applications in the electronic communications field are given throughout the book at the appropriate points in the chapters. These include: pulse shaping networks for data transmission, switched-capacitor filters for speech CODECs, full duplex data MODEMs, adaptive echo cancellation in the satellite transmission of speech signals, linear estimation, system modelling and adaptive filtering. In addition, the final chapter on sigma-delta data converters constitutes a *comprehensive application*, bringing together all the signal processing techniques in the book (switched-capacitor techniques, digital filters, decimators, FFT, and analog COMS integrated circuits) to design a mixed-mode processor with a wide range of applications as an analog to digital converter.

Finally, the enthusiasm and professionalism of Alexandra King and Liz Wingett of John Wiley and Sons, Ltd (Chichester, UK) have been of great help in the completion of this book.

H. Baher
Vienna and Dublin, 2012

Part I

Perspective

'Science as it exists at present is partly agreeable, partly disagreeable. It is agreeable through the power which it gives us of manipulating our environment, and to a small but important minority, it is agreeable because it affords intellectual satisfaction. It is disagreeable because, however we may seek to disguise the fact, it assumes a determinism which involves, theoretically, the power of predicting human actions; in this respect, it seems to lessen human power.'

Bertrand Russell
'Is Science Superstitious?' (in *'Sceptical Essays'*)

1

Analog, Digital and Mixed-mode Signal Processing

1.1 Digital Signal Processing

The widespread use of digital signal processing systems is due to many factors including reliability, reproducibility, high precision, freedom from aging and temperature effects, low cost and efficient computational algorithms. Furthermore, the revolution in the microelectronics field [1–3] has been characterized by a continuous increase in the level of integration leading to complete systems being integrated on a single chip, that is, systems on a chip (SoC) [3–5].

1.2 Moore's Law and the "Cleverness" Factor

The integrated circuit dates back to around 1960. Since then, the number of devices on a chip has increased dramatically in line with an observation [1, 2] predicting a doubling every year. Now, millions of transistors can be manufactured on a single chip allowing phenomenal processing capability. If we define a pixel as the smallest spot on a chip that can be controlled in the fabrication process, then this will determine the contribution of device miniaturization and chip area to the content of the chip. This contribution can be measured by the quantity A/S where A is the chip area and S is the pixel area. As progress continued, it was found that the number of devices on a chip was actually increasing *faster* than A/S. This additional growth was a result of "clever" techniques of exploiting the space on the chip. These include forming thin-film capacitors on the side holes etched into a chip instead of on the surface, and self-aligned structures where part of the device is used as the mask in the fabrication process. Next came the effect of the wiring on limiting the size of the chip. This, again, has been tackled [1] by the "cleverness" of increasing the number of wire layers.

1.3 System on a Chip

Such a system comprises *application specific integrated circuits* (ASICs). Examples are the single-chip TV or the single chip camera, and the ever-emerging new generations

Signal Processing and Integrated Circuits, First Edition. Hussein Baher.
© 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.

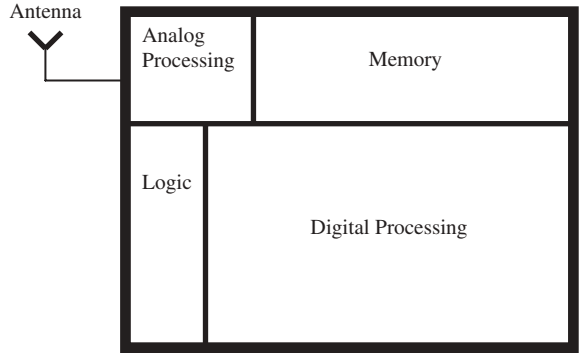


Figure 1.1 System on a chip (SoC)

of integrated telecommunication systems particularly in the mobile communication area. Such systems include *analog* and *digital* sections on the same chip where the technology of choice has been CMOS and possibly BiCMOS. Most functions on these chips are implemented using *digital signal processing* circuits. However, analog circuits are needed as an interface between the system and the real world which is, of course, analog in nature. Figure 1.1 shows a typical SoC containing embedded digital signal processors, embedded memory, reconfigurable logic, and analog circuits to interface with the analog continuous-time world.

The design of signal processing systems with low-power requirements is one of the most important areas of research [6, 7] which together with the need for high speed and density of integration have led to great advances in technology and clever circuit design methods [8].

1.4 Analog and Mixed-mode Signal Processing

The trend to replace, for example, analog filters by digital filters is understandable in view of the advantages of digital filters. However, there are some functions on the processor which will always remain analog [4]. These are the following:

(a) At the input of the system, signals from a sensor, microphone, antenna or cable must be received, amplified and filtered, that is *processed* and brought to a level that permits digitization with acceptable signal to noise ratio and low distortion. Here, we need low-noise amplifiers (LNAs), variable gain amplifiers (VGAs), filters, oscillators and mixers. Applications are:
   - Data and biomedical instrumentation.
   - Sensor interfaces such as airbags and accelerometers.
   - Telecommunications receivers such as telephone or cable modems and wireless telephones.

(b) At the output of the system the signal is reconverted from digital to analog form and strengthened so that it can drive an external load such as an antenna or a loudspeaker with low distortion. Here we also need buffers, filters, oscillators and mixers. Applications are the following
   - Telecommunications transmitters

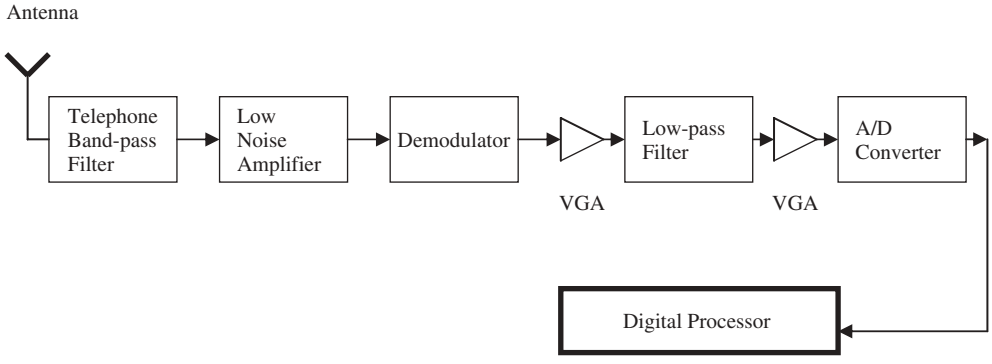


Figure 1.2 The analog and digital parts of a mobile telephone/Bluetooth receiver section

- Audio and video, such CD, SACD, DVD and Blueray
- Loudspeakers
- TV
- PC monitors
- Hearing aids

(c) Mixed-mode circuits are also needed for the interface between the analog and digital parts. These include sample and hold circuits for the sampling of signals, analog to digital (A/D) converters as well as digital to analog converters for signal reconstruction. These are mixed-mode circuits.

(d) The integrated circuits discussed above need stable references for their operation which are analog voltage and current sources and crystal oscillators.

Figure 1.2 illustrates the above points with the block diagram of a mobile telephone/Bluetooth receiver section [9]. This highlights the fact that both analog and digital circuits coexist on the same chip employing CMOS technology, and also the interrelationship between analog and digital signal processing.

1.5 Scope

Now, what do we need to know in order to be able to design a system on a chip? Our knowledge must include the following:

1. Methods of description of both analog and digital signals in the time and frequency domains.
2. Methods of description of the systems which process the signals. We need to do this for both analog and digital systems.
3. Design techniques for analog circuits such as amplifiers, integrators, differentiators, and most importantly: filters taking into account the non-ideal effects.
4. Integrated circuit implementations of analog circuits using CMOS technology.
5. Design of digital filters taking into account the finite word-length effects inherent in all digital processors.
6. Random signals require special methods for their description and processing, leading to the subject of adaptive filters. These, together with the related topics of linear prediction, estimation, and system modelling are essential.

7. Modern design techniques of discrete-time filtering using switched-capacitor techniques, since these are particularly amenable to implementation using VLSI techniques.
8. Design of A/D and D/A converters since these act as the interfaces between the digital and analog parts of the system.

Detailed treatment of the above topics is the aim of this book. To facilitate the numerical calculations, and to be able to study the responses of systems and evaluate their performances, we need a powerful software package. MATLAB® is a good choice, and it is used throughout the book.

Part II

Analog (Continuous-time) and Digital Signal Processing

'It is very desirable in instruction, not merely to persuade the student of the accuracy of important theorems, but to persuade him in the way which itself has, of all possible ways, the most beauty.'

Bertrand Russell
'The Study of Mathematics'

2

Analog Continuous-time Signals and Systems

2.1 Introduction

In this chapter the fundamental concepts and mathematical tools of analog signal and system analyses are reviewed. This review can be regarded as a comprehensive summary of the fundamentals of analog signals and systems. It is the distillation of courses on these topics which are usually covered at the junior to senior undergraduate levels. Therefore, the discussion is quite *compact*, and the material can be used as an *easy reference* for later chapters and as a *short revision course*.

2.2 The Fourier Series in Signal Analysis and Function Approximation

2.2.1 Definitions

A signal $f(x)$ defined over an interval $[-l, l]$ and satisfying conditions of considerable generality, can be represented as an infinite series of pure sine and cosine signals as

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \left[a_k \cos\left(\frac{k\pi x}{l}\right) + b_k \sin\left(\frac{k\pi x}{l}\right) \right] \tag{2.1}$$

where the coefficients are given by

$$a_k = \frac{1}{l} \int_{-l}^{l} f(x) \cos\left(\frac{k\pi x}{l}\right) \mathrm{d}x \quad k = 0, 1, 2, \ldots \tag{2.2a}$$

$$b_k = \frac{1}{l} \int_{-l}^{l} f(x) \sin\left(\frac{k\pi x}{l}\right) \mathrm{d}x \quad k = 1, 2, 3, \ldots \tag{2.2b}$$

A simplified notation results by letting

$$x \to \frac{l\theta}{\pi} \tag{2.3}$$

Signal Processing and Integrated Circuits, First Edition. Hussein Baher.
© 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.

so that the signal is defined over the range $[-\pi, \pi]$ and the series becomes

$$f(\theta) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos k\theta + b_k \sin k\theta) \tag{2.4}$$

with

$$a_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(\theta) \cos k\theta \, \mathrm{d}\theta \tag{2.5a}$$

$$b_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(\theta) \sin k\theta \, \mathrm{d}\theta. \tag{2.5b}$$

or the alternative versions in

$$f(\theta) = \frac{d_0}{2} + \sum_{k=1}^{\infty} d_k \cos(k\theta + \phi_k) \tag{2.6}$$

or

$$f(\theta) = \frac{d_0}{2} + \sum_{k=1}^{\infty} d_k \sin(k\theta + \psi_k) \tag{2.7}$$

with

$$\psi_k = \phi_k + \tfrac{1}{2}\pi. \tag{2.8}$$

$$d_k = (a_k^2 + b_k^2)^{1/2} \tag{2.9}$$

and

$$\phi_k = -\tan^{-1}(b_k/a_k). \tag{2.10}$$

If the signal is periodic, the representation is valid for *all values* of the independent variable. If the signal is non-periodic, then the representation is valid only over the *fundamental range* $[-l, l]$ or $[-\pi, \pi]$.

2.2.2 The Time and Discrete Frequency Domains

If the signal is a function of time t, examples of which are shown in Figure 2.1, then the representation is given by

$$f(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos k\omega_0 t + b_k \sin k\omega_0 t) \tag{2.11}$$

$$a_k = \frac{2}{T} \int_{-T/2}^{T/2} f(t) \cos k\omega_0 t \, \mathrm{d}t \tag{2.12a}$$

$$b_k = \frac{2}{T} \int_{-T/2}^{T/2} f(t) \sin k\omega_0 t \, \mathrm{d}t \tag{2.12b}$$

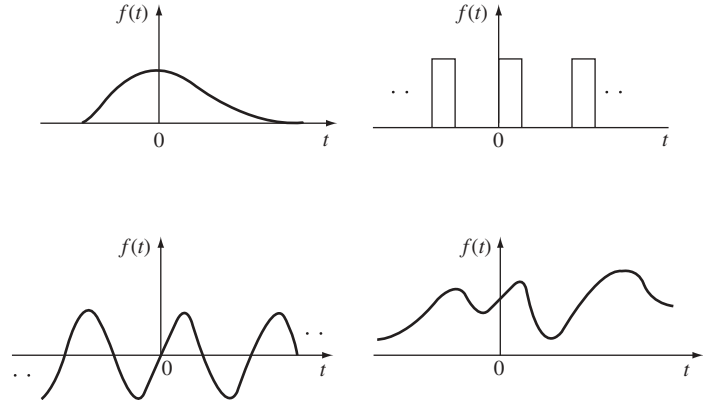


Figure 2.1 Examples of signals as functions of time

where

$$\omega_0 = 2\pi/T \tag{2.13}$$

with T being the *period* and ω_0 is the fundamental radian frequency.

A more compact (complex) Fourier series uses the exponential signal and is given by

$$f(\theta) = \sum_{k=-\infty}^{\infty} c_k \exp(\mathrm{j}k\theta) \tag{2.14}$$

with

$$c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\theta)\exp(-\mathrm{j}k\theta)\,\mathrm{d}\theta \tag{2.15}$$

For a function of time, the series is given by

$$f(t) = \sum_{k=-\infty}^{\infty} c_k \exp(\mathrm{j}k\omega_0 t) \tag{2.16}$$

with

$$c_k = \frac{1}{T}\int_{-T/2}^{T/2} f(t)\exp[-j(2\pi k/T)t]\,\mathrm{d}t \tag{2.17}$$

or

$$c_k = \frac{\omega_0}{2\pi}\int_{-\pi/\omega_0}^{\pi/\omega_0} f(t)\exp(-\mathrm{j}k\omega_0 t)\,\mathrm{d}t \tag{2.18}$$

The coefficients of the series give a representation in the *frequency domain*. The amplitudes of the coefficients give the *amplitude spectrum* whereas their phases give the *phase spectrum*. For a periodic signal these are *line spectra*, and give the frequency domain representation of the signal.

2.2.3 Convolution

The *convolution* of two signals $f_1(\theta)$ and $f_2(\theta)$ is given by

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} f_1(\theta-\psi)f_2(\psi)\,\mathrm{d}\psi = \sum_{k=-\infty}^{\infty} c_k d_k \exp(\mathrm{j}k\theta) \tag{2.19}$$

If we put $\theta = \omega_0 t$ for the two functions $f_1(t)$ and $f_2(t)$ which are periodic with period $T = 2\pi/\omega_0$, the convolution relation becomes

$$\frac{1}{T}\int_{-T/2}^{T/2} f_1(t-\tau)f_2(\tau)\,\mathrm{d}\tau = \sum_{k=-\infty}^{\infty} c_k d_k \exp(\mathrm{j}k\theta_0 t) \tag{2.20}$$

and the complex Fourier series of the convolution has as its coefficients the products of the corresponding ones in the series of the individual signals.

2.2.4 Parseval's Theorem and Power Spectrum

Parseval's theorem relates the *average power* of a signal to the *sum of the squares* of the amplitudes of the complex Fourier coefficients as expressed by

$$\sum_{k=-\infty}^{\infty} |c_k|^2 = \frac{1}{T}\int_{-T/2}^{T/2} [f(t)]^2\,\mathrm{d}t \tag{2.21}$$

The squared amplitudes of the complex Fourier coefficients are called the *power spectral amplitudes* and a plot of these versus frequency is called the *power spectrum of the signal*.

2.2.5 The Gibbs' Phenomenon

The approximation of a function by a truncated Fourier series results in *minimization* of the *mean-square error* between the values of the function and the approximating finite series. The smoother the function, the faster the convergence of the series to the function. If the function has a discontinuity, such as the one shown in Figure 2.2, the approximation yields a fixed error at the discontinuity, which does not decrease by increasing the number of terms in the approximating series; this is the *Gibbs' phenomenon*. For example the

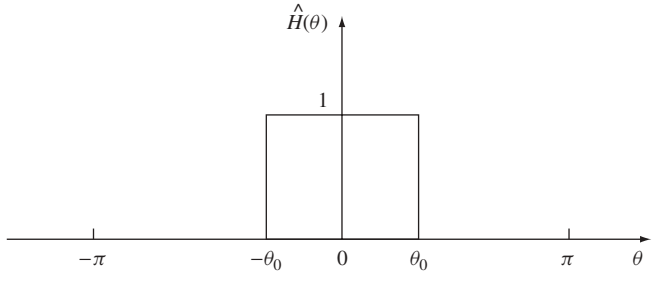


Figure 2.2 A signal with a discontinuity

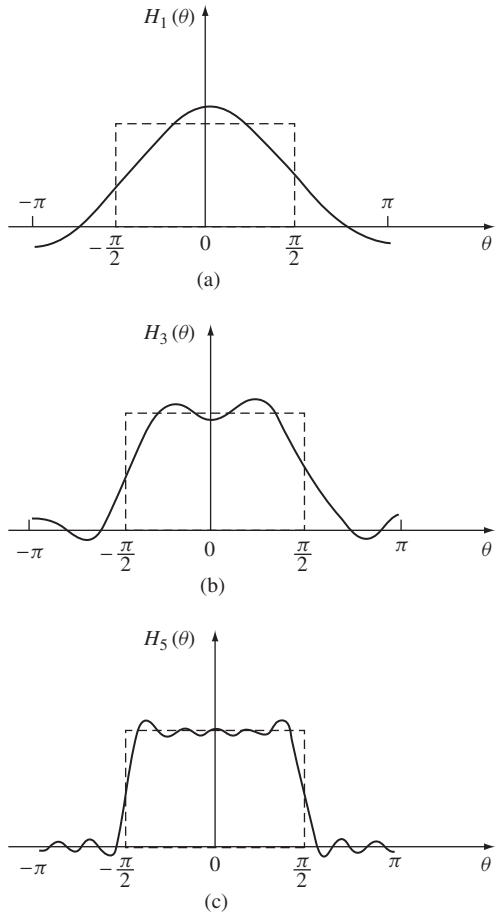


Figure 2.3 Examples of the approximation of the signal of Figure 2.2 by a truncated Fourier series: (a) n = 1, (b) n = 3, (c) n = 5; all for $\theta_0 = \pi/2$

approximation of the signal shown in Figure 2.2 by a truncated Fourier series $H_n(\theta)$, which is expression (2.14) taking only n terms of the series, is shown in Figure 2.3.

2.2.6 Window Functions

A modification of the Fourier coefficients using *window functions* improves the convergence of the series at the discontinuity. This is obtained by forming the new coefficients according to

$$d_k = w_k c_k \tag{2.22}$$

where the following are commonly used *windows*:

(i) The Fejer window

$$w_k = 1 - k/n \tag{2.23}$$

(ii) The Lanczos window

$$w_k = \frac{\sin(k\pi/n)}{(k\pi/n)} \tag{2.24}$$

(iii) The von Hann window

$$w_k = 0.5[1 + \cos(k\pi/n)] \tag{2.25}$$

(iv) The Hamming window

$$w_k = 0.54 + 0.46\cos(k\pi/n) \tag{2.26}$$

(v) The Kaiser window

$$w_k = \frac{I_0\{\beta[1-(k/n)^2]^{1/2}\}}{I_0(\beta)} \tag{2.27}$$

where $I_0(x)$ is the zero$^{\text{th}}$-order Bessel function of the first kind. The truncated series becomes

$$\begin{aligned} S_n(\theta) &= \sum_{k=-n}^{n} w_k c_k \exp(\mathrm{j}k\theta) \\ &= \sum_{k=-n}^{n} d_k \exp(\mathrm{j}k\theta) \end{aligned} \tag{2.28}$$

2.3 The Fourier Transformation and Generalized Signals

2.3.1 Definitions and Properties

The *Fourier transform* of a function of time gives the frequency-domain representation of the function [10–12]. This relation between the time-domain and frequency-domain representations is given by the expressions

$$F(\omega) = \int_{-\infty}^{\infty} f(t)\exp(-\mathrm{j}\omega t)\,\mathrm{d}t \tag{2.29}$$

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)\exp(\mathrm{j}\omega t)\,\mathrm{d}\omega \tag{2.30}$$

$$F(\omega) = \Im[f(t)] \tag{2.31}$$

$$f(t) = \Im^{-1}[F(\omega)] \tag{2.32}$$

and the notation

$$f(t) \leftrightarrow F(\omega) \tag{2.33}$$

is used to signify that $f(t)$ and $F(\omega)$ form a *Fourier transform pair*.

The Fourier transform $F(\omega)$ of $f(t)$ is a *complex* function of ω, so that we may write

$$F(\omega) = |F(\omega)|\exp(\mathrm{j}\phi(\omega)) \tag{2.34}$$

where ω is a continuous frequency variable. This means that a plot of $|F(\omega)|$ against ω now gives the (continuous) *amplitude spectrum* of $f(t)$, while $\phi(\omega)$ plotted against ω gives the (continuous) *phase spectrum* of $f(t)$.

The basic properties of the Fourier transform are given below:

(i) Symmetry

$$F(\pm t) \leftrightarrow 2\pi f(\mp\omega). \tag{2.35}$$

(ii) Conjugate pairs

$$F(\omega) = F^*(-\omega). \tag{2.36}$$

(iii) Linearity

$$af_1(t) + bf_2(t) \leftrightarrow aF_1(\omega) + bF_2(\omega) \tag{2.37}$$

where a and b are arbitrary constants.

(iv) Scaling

$$f(\alpha t) \leftrightarrow \frac{1}{|\alpha|} F\left(\frac{\omega}{\alpha}\right) \tag{2.38}$$

(v) Shifting in time

$$f(t-\alpha) \leftrightarrow \exp(-\mathrm{j}\,\alpha\omega)F(\omega) \tag{2.39}$$

(vi) Shifting in frequency

$$f(t)\exp(\pm\mathrm{j}\,\omega_0 t) \leftrightarrow F(\omega \mp \omega_0) \tag{2.40}$$

(vii) Modulation

$$f(t)\cos\omega_0 t \leftrightarrow \tfrac{1}{2}[F(\omega-\omega_0) + F(\omega+\omega_0)] \tag{2.41}$$

An illustration of this property is shown in Figure 2.4.

(viii) Differentiation in frequency

$$(-\mathrm{j}t)^n f(t) \leftrightarrow \frac{\mathrm{d}^n F(\omega)}{\mathrm{d}\omega^n} \tag{2.42}$$

with

$$\begin{aligned} f_1(t) &\leftrightarrow F_1(\omega) \\ f_2(t) &\leftrightarrow F_2(\omega) \end{aligned} \tag{2.43}$$

the convolution in the time domain of the two signals is defined as

$$f_1(t) * f_2(t) \triangleq \int_{-\infty}^{\infty} f_1(t-\tau) f_2(\tau)\,\mathrm{d}\tau \tag{2.44}$$

or equivalently

$$f_1(t) * f_2(t) \triangleq \int_{-\infty}^{\infty} f_1(\tau) f_2(\tau-\tau)\,\mathrm{d}\tau \tag{2.45a}$$

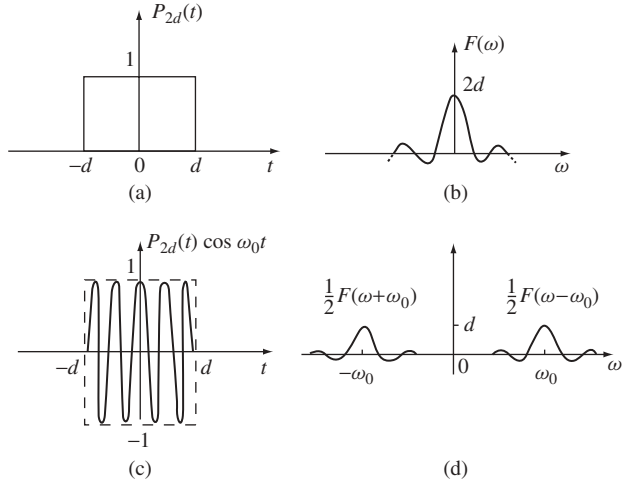


Figure 2.4 (a) A pulse, (b) its spectrum, (c) the modulated pulse, (d) the spectrum of the modulated pulse

then

$$f_1(t) * f_2(t) \leftrightarrow F_1(\omega)F_2(\omega) \tag{2.45b}$$

That is, the Fourier transform of the convolution of two signals is the product of the transforms of the individual signals. The convolution in the frequency domain corresponds to multiplication of the two signals in the time domain as expressed in

$$f_1(t)f_2(t) \leftrightarrow \frac{1}{2\pi}F_1(\omega) * F_2(\omega) \tag{2.46}$$

where

$$\begin{aligned} F_1(\omega) * F_2(\omega) &\triangleq \int_{-\infty}^{\infty} F_1(\mu)F_2(\omega-\mu)\,\mathrm{d}\mu \\ &= \int_{-\infty}^{\infty} F_1(\omega-\mu)F_2(\mu)\,\mathrm{d}\mu. \end{aligned} \tag{2.47}$$

2.3.2 *Parseval's Theorem and Energy Spectra*

The *energy spectral density* (or *energy spectrum*) of a signal is the square of the modulus of its Fourier transform

$$E(\omega) \triangleq |F(\omega)|^2 \tag{2.48}$$

so that

$$W = \frac{1}{2\pi}\int_{-\infty}^{\infty} E(\omega)\,\mathrm{d}\omega = \int_{-\infty}^{\infty} |f(t)|^2\,dt \tag{2.49}$$

This is Parseval's theorem, which highlights the fact that the energy in the spectrum equals the energy in the parent signal of time. Since the integral in (2.49) is assumed to exist, such signals are called *finite-energy* signals.

2.3.3 Correlation Functions

The *autocorrelation* of a signal is defined by

$$\rho_{ff}(\tau) = \int_{-\infty}^{\infty} f(t) f(t+\tau)\,\mathrm{d}\tau. \tag{2.50}$$

For a finite energy signal, the *autocorrelation* and the *energy spectrum* form a Fourier transform pair

$$\rho_{ff}(\tau) \leftrightarrow E(\omega). \tag{2.51}$$

For two finite energy signals, the *cross-correlation* is defined by

$$\rho_{fg}(\tau) = \int_{-\infty}^{\infty} f(t) g(t+\tau)\,\mathrm{d}t \tag{2.52}$$

The *cross-energy spectrum* of the two signals is defined by

$$\Im[\rho_{fg}(\tau)] = F^*(\omega) G(\omega) = E_{fg}(\omega). \tag{2.53}$$

which is the Fourier transform of the cross-correlation, and this is a measure of the similarity between the two signals.

2.3.4 The Unit Impulse and Generalized Signals

The only satisfactory way to include signals with *impulses* and discontinuities in transform theory is by means of the *theory of distributions* or *generalized functions*. A distribution or generalized function $D(t)$ is the process of associating with (or assigning to) an arbitrary function $\phi(t)$ a number V_D depending on the function. This number is normally written as the integral

$$V_D\{\phi(t)\} = \int_{-\infty}^{\infty} D(t)\phi(t)\,\mathrm{d}t \tag{2.54}$$

Indeed, a distribution $D(t)$ is the process of associating with $\phi(t)$ *any* other quantity depending on $\phi(t)$.

The *unit impulse* or *Dirac delta function* is shown in Figure 2.5 and is a distribution (see Figure 2.6) defined by

$$\int_{-\infty}^{\infty} \phi(t)\delta(t)\,\mathrm{d}t = \phi(0). \tag{2.55}$$

The integral of the product of any function with the unit impulse gives the value of the function at $t = 0$. The basic properties of the unit impulse are given by

$$\delta(at) = \frac{1}{|a|}\delta(t) \tag{2.56}$$

$$\delta(t) = \delta(-t). \tag{2.57}$$

$$\delta(t) \leftrightarrow 1. \tag{2.58}$$

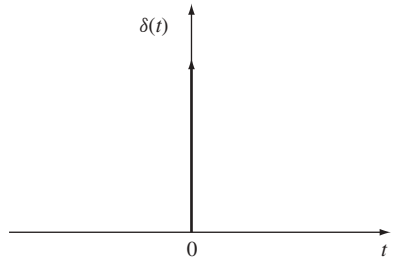


Figure 2.5 The Dirac delta function or unit impulse

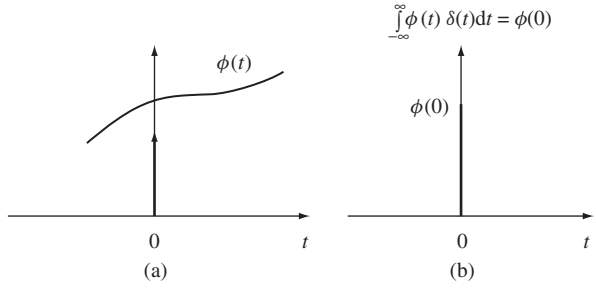


Figure 2.6 Illustrating the definition of the unit impulse

that is, the Fourier transform of the unit impulse is unity.

$$\int_{-\infty}^{\infty} \phi(\tau) \left(\frac{\mathrm{d}^n \delta(\tau)}{\mathrm{d}\tau^n} \right) \mathrm{d}\tau = (-1)^n \left. \frac{\mathrm{d}^n \phi(t)}{\mathrm{d}t^n} \right|_{t=0} \tag{2.59}$$

$$g(t)\delta(t-\alpha) = g(\alpha)\delta(t-\alpha) \tag{2.60}$$

$$\delta(t) = \frac{\mathrm{d}u(t)}{\mathrm{d}t} \tag{2.61}$$

where $u(t)$ is the unit step function.

2.3.5 *The Impulse Response and System Function*

The system function is the ratio of the Fourier transform of the response to the Fourier transform of the excitation (Figure 2.7).

The *system function* $H(j\omega)$ is the Fourier transform of the *impulse response* $h(t)$ of the system, that is, its response to an excitation equal to the unit impulse.

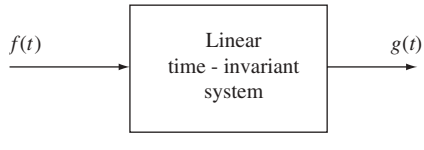


Figure 2.7 A linear system with the excitation $f(t)$ and response $g(t)$

A *causal signal* is one that is zero for negative values of time. A *causal system* is a system whose impulse response is a causal signal.

The Fourier transform of a periodic train of impulses is another train of impulses as given by

$$\sum_{k=-\infty}^{\infty} \delta(t - kT) \leftrightarrow \omega_0 \sum_{k=-\infty}^{\infty} \delta(\omega - k\omega_0) \tag{2.62}$$

2.3.6 Periodic Signals

The Fourier transform of a periodic function, as that shown in Figure 2.8, is an infinite train of equidistant impulses as expressed in

$$F_{\mathrm{p}}(\omega) = \omega_0 \sum_{k=-\infty}^{\infty} F(k\omega_0)\delta(\omega - k\omega_0) \tag{2.63}$$

where $F(k\omega_0)$ is the Fourier transform of $f(t)$ evaluated at the discrete set of frequencies $k\omega_0$ that is

$$F(k\omega_0) = \int_{-T/2}^{T/2} f(t) \exp(-\mathrm{j}k\omega_0 t)\, \mathrm{d}t. \tag{2.64}$$

2.3.7 The Uncertainty Principle

This is written as

$$\Delta t \Delta\omega \geq K \tag{2.65}$$

stating that the *duration* $\Delta(t)$ of a signal and its *bandwidth* $\Delta(\omega)$ cannot be simultaneously small, so that the smaller the duration the wider the bandwidth and vice versa. For example, the *unit impulse*, which exists at a specific point only, has *infinite bandwidth*. Of course K is a constant, which depends on the definitions or interpretations of the terms *duration* and *bandwidth*.

2.4 The Laplace Transform and Analog Systems

2.4.1 The Complex Frequency

The Laplace transform of a function of time is a function of the complex frequency variable. The transform is defined by the integral in

$$F(s) = L[f(t)] = \int_{0^-}^{\infty} f(t) \exp(-st)\, \mathrm{d}t \tag{2.66}$$

where s is the complex frequency variable taking the place of the restricted $\mathrm{j}\omega$ in the Fourier integral, that is

$$s = \sigma + \mathrm{j}\,\omega \tag{2.67}$$

and defines a complex frequency plane as shown in Figure 2.9.

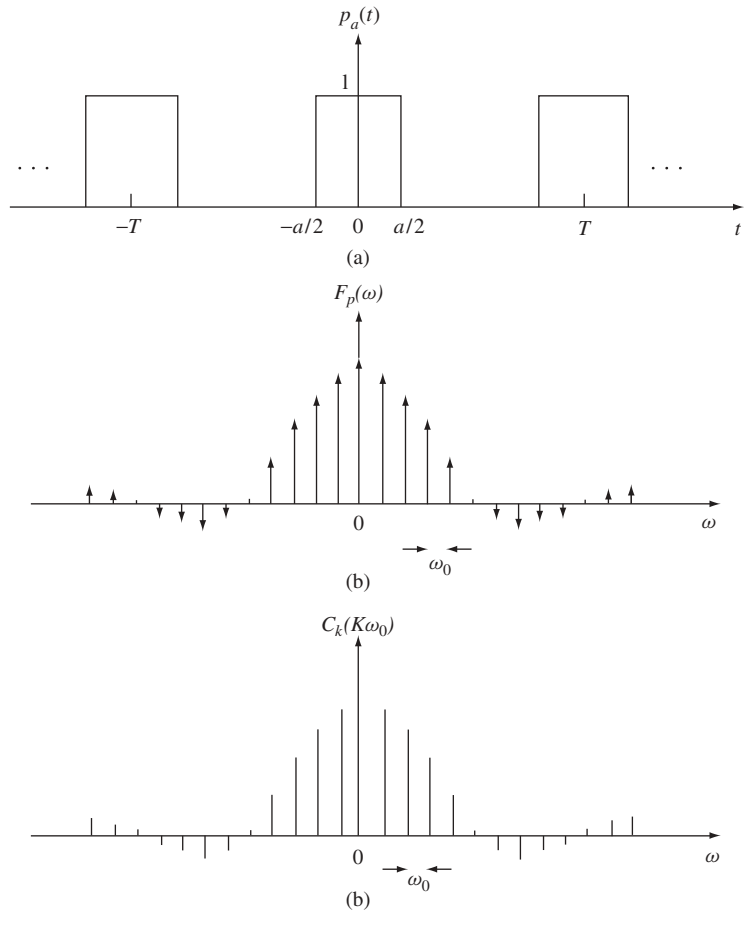


Figure 2.8 (a) A periodic train of rectangular pulses, (b) its spectrum as a train of impulses and (c) its spectrum as a plot of the Fourier series coefficients

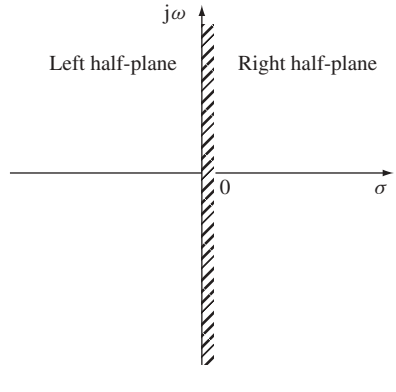


Figure 2.9 The complex frequency s-plane

2.4.2 Properties of the Laplace Transform

The basic properties of the Laplace transform are given below:

(i) *Linearity*. This is a general property of all integral transforms and states that if

$$L[f_i(t)] = F_i(s) \quad i = 1, 2, 3, \ldots n \tag{2.68a}$$

then

$$L\left(\sum_{i=1}^{n} a_i f_i(t)\right) = \sum_{i=1}^{n} a_i F_i(s) \tag{2.68b}$$

where $a_i (i = 1, 2, \ldots n)$ are arbitrary constants.

(ii) *Scaling in time*

$$L[f(\alpha t)] = \frac{1}{\alpha} F(s/\alpha). \tag{2.69}$$

(iii) *Differntiation in time*

$$L[f^{(n)}(t)] = s^n F(s) - s^{n-1} f(0^-) - s^{n-2} f'(0^-) - \cdots f^{(n-1)}(0^-) \tag{2.70}$$

where $f^{(n)}$ denotes the nth derivative.

(iv) *Integration in time*

$$L\left(\int_{0^-}^{t} f(\tau)\, \mathrm{d}\tau\right) = \frac{F(s)}{s} \tag{2.71}$$

(v) *Differentiation in frequency*

$$L[-tf(t)] = \frac{\mathrm{d}F(s)}{\mathrm{d}s} \tag{2.72}$$

(vi) *Integration in frequency*

$$\int_{s}^{\infty} F(s)\, \mathrm{d}s = L\left[\frac{f(t)}{t}\right] \tag{2.73}$$

(vii) *Shifting in time*

$$L[f(t-\alpha)u(t-\alpha)] = \exp(-\alpha s) F(s) \tag{2.74}$$

where $u(x)$ is the unit step function.

(viii) *Shifting in frequency*

$$L[\exp(\alpha t) f(t)] = F(s - \alpha) \tag{2.75}$$

The *convolution* of two signals is defined by

$$\begin{aligned} f_1(t) * f_2(t) &\triangleq \int_{0^-}^{\infty} f_1(t-\tau) f_2(\tau)\, \mathrm{d}\tau \\ &\triangleq \int_{0^-}^{t} f_1(t-\tau) f_2(\tau)\, \mathrm{d}\tau \end{aligned} \tag{2.76}$$

The Laplace transform of the resulting signal is the product of the transforms of the individual signals as expressed by

$$L[f_1(t) * f_2(t)] = F_1(s)F_2(s) \tag{2.77}$$

The inverse Laplace transform of a rational function can be obtained by breaking up the function as the sum of its partial fractions and taking the inverse of the individual simpler functions. Table 2.1 gives some Laplace transform pairs.
The correspondence between the time-domain representations and frequency-domain representations of a function is of paramount importance in signal and system analysis and design.

The Laplace transform converts a linear differential equation with constant coefficients into a linear algebraic equation that can be solved very easily. To go back to the time domain, the inverse Laplace transform of the desired variable is taken, and this makes the Laplace transform the ideal technique for the analysis of linear networks and systems because these are described by equations which, in the time domain, are linear differential equations in the time variable.

2.4.3 The System Function

A linear time invariant system, as depicted in Figure 2.10, is described by a linear differential equation with constant coefficient

$$\begin{aligned} &a_m \frac{\mathrm{d}^m f(t)}{\mathrm{d}t^m} + a_{m-1}\frac{\mathrm{d}^{m-1} f(t)}{\mathrm{d}t^{m-1}} + \cdots a_1 \frac{\mathrm{d}f(t)}{\mathrm{d}t} + a_0 f(t) \\ &\quad = b_n \frac{\mathrm{d}^n g(t)}{\mathrm{d}t^n} + b_{n-1}\frac{\mathrm{d}^{n-1} g(t)}{\mathrm{d}t^{n-1}} + \cdots + b_1 \frac{\mathrm{d}g(t)}{\mathrm{d}t} + b_0 g(t). \end{aligned} \tag{2.78}$$

with

$$\begin{aligned} L[f(t)] &= F(s) \\ L[g(t)] &= G(s) \end{aligned} \tag{2.79}$$

the transformed Equation (2.78) becomes the *algebraic* relation between $F(s)$ and $G(s)$

$$\begin{aligned} &a_m s^m F(s) + a_{m-1}s^{m-1}F(s) + \cdots + a_1 sF(s) + a_0 F(s) \\ &\quad = b_n s^n G(s) + b_{n-1}s^{n-1}G(s) + \cdots b_1 sG(s) + b_0 G(s) \end{aligned} \tag{2.80}$$

or

$$(a_m s^m + a_{m-1}s^{m-1} + \cdots as_1 + a_0)F(s) = (b_n s^n + b_{n-1}s^{n-1} + b_1 s + b_0)G(s) \tag{2.81}$$

from which it follows that

$$G(s) = \left(\frac{a_m s^m + a_{m-1}s^{m-1} + \cdots + a_1 s + a_0}{b_n s^{n-1} + b_{n-1}s^{n-1} + \cdots + b_1 s + b_0}\right) F(s). \tag{2.82}$$

Table 2.1 Some Laplace transform pairs

$f(t)$	$F(s)$
$f'(t)$	$sF(s) - f(0^-)$
$f^{(n)}(t) = \dfrac{\mathrm{d}^n f(t)}{dt^n}$	$s^n F(s) - \sum_{k=1}^{n} s^{n-k} f^{k-1}(0^-)$
$\int_{0^-}^{t} f(\tau)\,\mathrm{d}\tau$	$\dfrac{F(s)}{s}$
$(-t)^n f(t)$	$\dfrac{\mathrm{d}^n}{\mathrm{d}s^n} F(s)$
$f(t-\alpha)u(t-\alpha)$	$\mathrm{e}^{-\alpha s} F(s)$
$\mathrm{e}^{\alpha t} f(t)$	$F(s-\alpha)$
$u(t)$	$\dfrac{1}{s}$
$\delta(t)$	1
$\delta^{(n)}(t) = \dfrac{\mathrm{d}^n \delta(t)}{\mathrm{d}t^n}$	s^n
T	$\dfrac{1}{s^2}$
t^n (n an integer)	$\dfrac{n!}{s^{n+1}}$
$\mathrm{e}^{-\alpha t}$	$\dfrac{1}{s+\alpha}$
$\sin\omega_0 t$	$\dfrac{\omega_0}{s^2+\omega_0^2}$
$\cos\omega_0 t$	$\dfrac{s}{s^2+\omega_0^2}$
$\sinh \beta t$	$\dfrac{\beta}{s^2-\beta^2}$
$\cosh \beta t$	$\dfrac{s}{s^2-\beta^2}$
$t^{-1/2}$	$(\pi/s)^{1/2}$
t^k (k may not be an integer)	$\dfrac{\Gamma(k+1)}{s^{k+1}}$

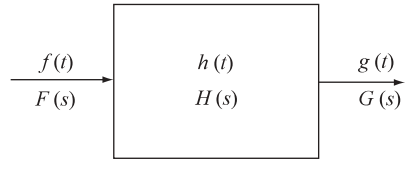


Figure 2.10 A linear system

Defining the *system function* $H(s)$ as

$$H(s) = \frac{G(s)}{F(s)} = \frac{a_m s^m + a_{m-1}s^{m-1} + \cdots + a_1 s + a_0}{b_n s^n + b_{n-1}s^{n-1} + \cdots + b_1 s + b_0} \tag{2.83}$$

we obtain a *real rational function* of the complex frequency s, that is $H(s)$ is the ratio of two polynomials in s with *real constant coefficients*.

The *impulse response* $h(t)$ and the *system function* $H(s)$ form a Laplace transform pair.

A system is *wide-sense stable* if all the poles of the system function lie in the closed left-half plane with those on the imaginary axis being simple (i.e. not multiple). It is *strictly stable* if all the poles occur only in the open left-half plane, that is, excluding the imaginary axis.

A real polynomial is *Hurwitz* if all its zeros are in the closed left half-plane with those on the imaginary axis being simple; it is *strictly Hurwitz* if imaginary-axis zeros are also excluded. The denominator of the transfer function of a wide-sense stable system is a Hurwitz polynomial. Strict stability (bounded-input–bounded-output) requires the denominator of the system function to be a strict Hurwitz polynomial.

2.5 Elementary Signal Processing Building Blocks

Examination of the differential equation (2.78) relating the response and excitation of a system, reveals that there are three basic operations involved. These are: addition, multiplication by a constant and differentiation. If (2.78) is integrated n times (assuming $n \geq m$), then the operations expressed by the resulting equation become: addition, multiplication by a constant and *integration*. Each one of these elementary operations can be represented *symbolically* as shown in Figure 2.11, which can be viewed as elementary subsystems or basic *building blocks*. These are now considered separately, then it is shown how the behaviour of the entire system described by (2.78), or equivalently the function (2.83), can be described using these basic building blocks.

2.5.1 Realization of the Elementary Building Blocks using Operational Amplifier Circuits

For the realization of the building blocks of Figure 2.11, we shall use an *active* circuit employing the basic operational amplifier (Op Amp) configuration shown in Figure 2.12 where $Z_1(s)$ and $Z_2(s)$ are arbitrary complex frequency domain impedances. Assuming a near ideal Op Amp of sufficient band-width to accommodate the operating frequencies, we have for this circuit

$$V_{\text{out}}(s) \simeq -(Z_2(s)/Z_1(s))V_{\text{in}}(s) \tag{2.84}$$

which is obtained by a generalization of the familiar analysis of the Op Amp circuit with sinusoidal frequency impedances $Z_1(\mathrm{j}\omega)$ and $Z_2(\mathrm{j}\omega)$ then extending the results to complex frequencies. A particularly useful special case of Figure 2.12 is the *inverter* shown in Figure 2.13. It produces an output which is the negative of the input. When the circuit of Figure 2.12 is followed by an inverter (i.e. connected in *cascade*) the input–output relation of the overall circuit is identical to that given by (2.84) *without* the negative sign. The cascade is shown in Figure 2.14.

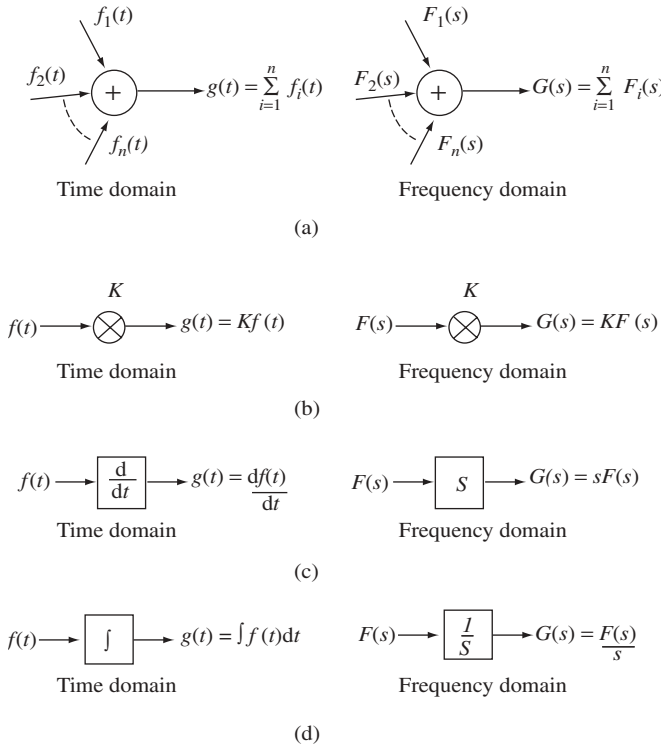


Figure 2.11 Basic analog signal processing building blocks: (a) adder, (b) multiplier, (c) differentiator, (d) integrator

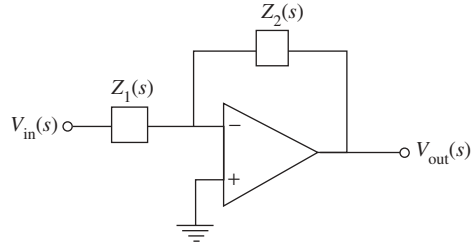


Figure 2.12 Basic inverting operational amplifier configuration

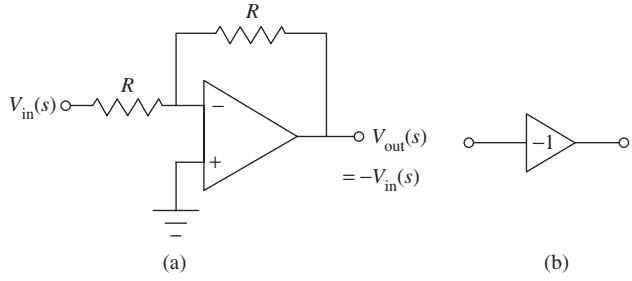


Figure 2.13 Inverter: (a) Op Amp circuit, (b) symbolic representation

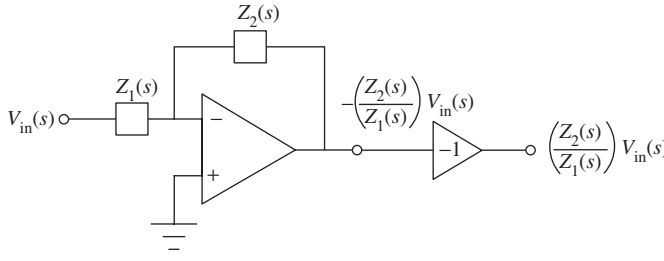


Figure 2.14 Realization of the transfer function $H(s) = Z_2(s)/Z_1(s)$

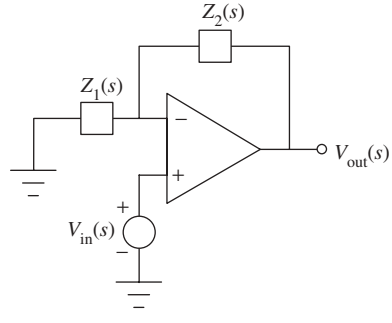


Figure 2.15 Non-inverting operational amplifier configuration

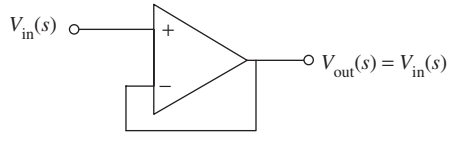


Figure 2.16 Unity-gain buffer, or voltage follower

A direct *non-inverting* Op Amp circuit is shown in Figure 2.15 whose transfer function is given by

$$\frac{V_{out}(s)}{V_{in}(s)} = \left(\frac{Z_2(s)}{Z_1(s)} + 1\right). \tag{2.85}$$

A special case of Figure 2.15 is obtained by taking $Z_2 = 0$ and removing Z_1 replacing it by an open circuit. This results in the *voltage follower* or *unity-gain buffer* shown in Figure 2.16. It is used to effect isolation between stages, that is, performing the connection while preventing interaction in the form of the second stage drawing current from the first.

We now examine the realization of the multiplier, adder, differentiator and integrator by means of the basic Op Amp configuration discussed above:

(i) Multiplier

In the time domain we require an input–output relation of the form

$$g(t) = Kf(t) \tag{2.86a}$$

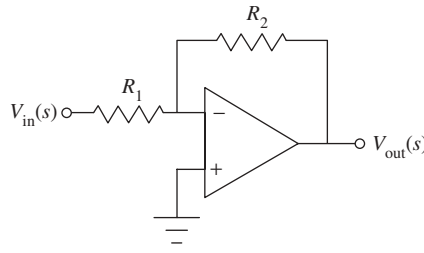


Figure 2.17 Multiplier (inverting)

where K is a constant. In the Laplace domain we have

$$G(s) = KF(s). \tag{2.86b}$$

This operation can be implemented using the Op Amp circuit shown in Figure 2.12 in which the quantities $F(s)$ and $G(s)$ are simulated by the voltage $V_{in}(s)$ and $V_{out}(s)$, respectively, and we take Z_1 and Z_2 to be pure resistances R_1 and R_2. This gives the circuit of Figure 2.17, for which

$$V_{out}(s) \simeq -\left(\frac{R_2}{R_1}\right) V_{in}(s) \tag{2.87}$$

and, if K is positive, the minus sign can be cancelled by employing an inverter in cascade.

Alternatively, for positive K (and $K > 1$) the circuit of Figure 2.15 can be used with $Z_1 = R_1$, $Z_2 = R_2$, that is pure resistances and their values are chosen according to (2.86) and (2.87).

(ii) Adder

In the time domain, this building block adds the signals $f_1(t), f_2(t), \ldots f_n(t)$ to produce an output $g(t)$ given by

$$g(t) = f_1(t) + f_2(t) + \cdots + f_n(t) \tag{2.88a}$$

or in the Laplace domain

$$G(s) = F_1(s) + F_2(s) + \cdots + F_n(s). \tag{2.88b}$$

This operation can be implemented by the Op Amp circuit shown in Figure 2.18 in which all the quantities entering in the above equations are simulated by voltages and

$$V_{out}(s) = -R(R_1^{-1}V_1(s) + R_2^{-1}V_2(s) + \cdots + R_n^{-1}V_n(s)). \tag{2.89}$$

Clearly, this circuit realizes the combined operations of addition and multiplication. For cancellation of the negative sign in (2.89) the circuit is followed by an inverter. For direct summation we set $R_1 = R_2 = \cdots = R_n = R$.

(iii) Differentiator (with a scaling factor)

For this operation we require, as shown in Figure 2.11(c),

$$g(t) = K\frac{\mathrm{d}f(t)}{\mathrm{d}t} \tag{2.90a}$$

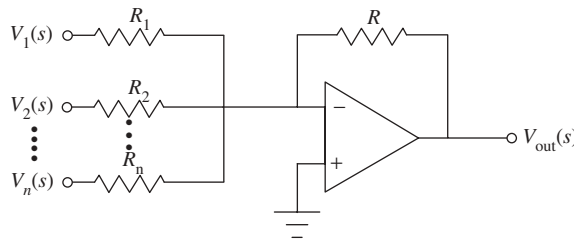


Figure 2.18 Adder-multiplier (inverting)

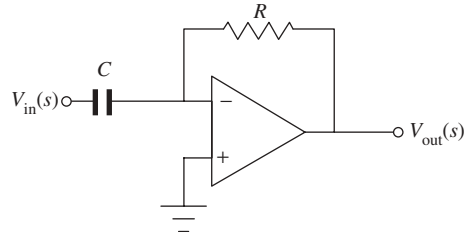


Figure 2.19 Differentiator (with a scaling factor)

or in the Laplace domain

$$G(s) = KsF(s). \tag{2.90b}$$

A realization of this operation can be obtained using the basic Op Amp circuit of Figure 2.19 with Z_1 as a capacitor C and Z_2 as a resistor R. Thus Figure 2.12 gives the differentiator circuit of Figure 2.19 in which the quantities $F(s)$ and $G(s)$ are simulated by the voltages $V_{\text{in}}(s)$ and $V_{\text{out}}(s)$ respectively, and

$$V_{\text{out}}(s) = -RCsV_{\text{in}}(s). \tag{2.91}$$

Again, to eliminate the negative sign the circuit is followed by an inverter. The RC product can be chosen such that $RC = K$, which can be adjusted to unity if required.

(iv) Integrator (with a scaling factor)

This is represented symbolically in Figure 2.11(d), and in the time domain produces an output $g(t)$ related to its input $f(t)$ by

$$g(t) = K \int f(t)\,\mathrm{d}t \tag{2.92a}$$

or, in the Laplace domain

$$G(s) = K\frac{F(s)}{s} \tag{2.92b}$$

The circuit of Figure 2.12 implements the above operation with Z_1 as a resistor R and Z_2 as a capacitor C. Thus for this circuit with $F(s)$ and $G(s)$ simulated by $V_{\text{in}}(s)$ and $V_{\text{out}}(s)$, respectively we have the integrator of Figure 2.20 for which

$$V_{\text{out}}(s) = -\frac{1}{RCs}V_{\text{in}}(s) \tag{2.93}$$

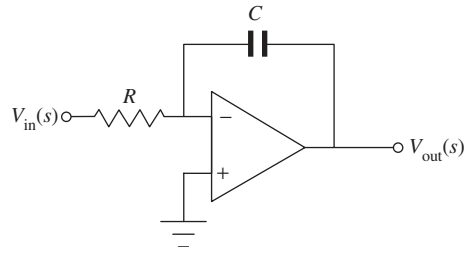


Figure 2.20 Integrator (with a scaling factor)

and as before, the negative sign may be cancelled by cascading the circuit with an inverter at the output.

2.6 Realization of Analog System Functions

2.6.1 General Principles and the Use of Op Amp Circuits

Now, given a system function $H(s)$ we wish to find the functional block-diagram representation as the interconnection of the building blocks described so far. This process is called the *simulation of the system or the realization of the system function*. Before considering this procedure in general, it is first illustrated by the second-order transfer function

$$H(s) = \frac{b_0 + b_1 s + b_2 s^2}{1 + a_1 s + a_2 s^2} \tag{2.94}$$

Write the function as

$$H(s) = H_1(s) H_2(s) \tag{2.95}$$

where

$$H_1(s) = b_0 + b_1 s + b_2 s^2 \tag{2.96}$$

and

$$H_2(s) = \frac{1}{1 + a_1 s + a_2 s^2} \tag{2.97}$$

Now, if $H_1(s)$ is considered separately as the transfer function of a subsystem with input $F_1(s)$ and output $G_1(s)$, then

$$\begin{aligned} G_1(s) &= (b_0 + b_1 s + b_2 s^2) F_1(s) \\ &= b_0 F_1(s) + b_1 s F_1(s) + b_2 s^2 F_1(s) \end{aligned} \tag{2.98}$$

which leads directly to the functional block-diagram realization shown in Figure 2.21(a).

Similarly, if $H_2(s)$ is considered separately as the transfer function of a subsystem with input $F_2(s)$ and output $G_2(s)$, then

$$G_2(s) = H_2(s) F_2(s) \tag{2.99}$$

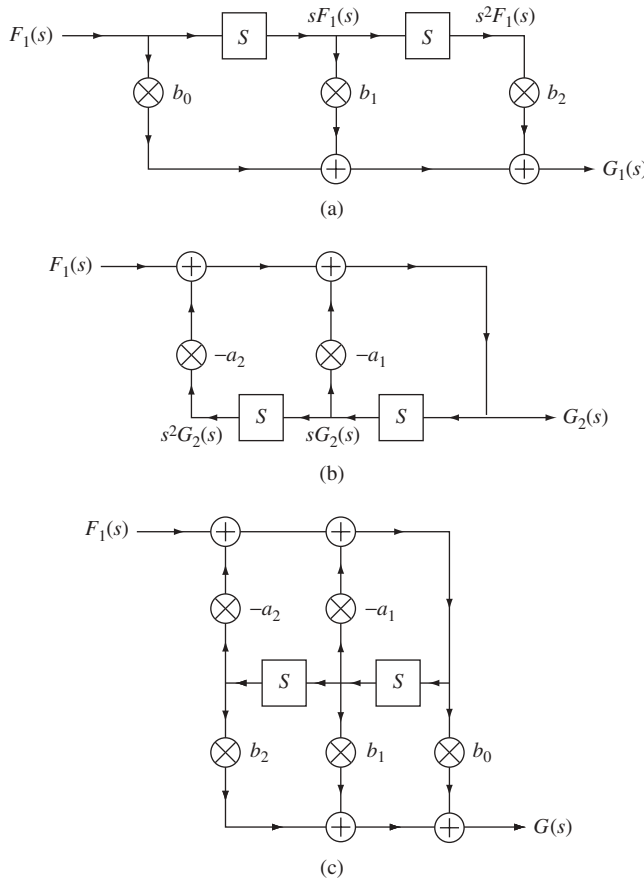


Figure 2.21 Realization of a second-order system function: (a) subsystem for realizing (2.96), (b) subsystem for realizing (2.97), (c) connection of (a) and (b) to realize (2.94)

or

$$G_2(s) = F_2(s) - a_1 s G_2(s) - a_2 s^2 G_2(s) \tag{2.100}$$

which leads directly to the realization shown in Figure 2.21(b).

Now, the decomposition (2.95) means that

$$H(s) = \frac{G_1(s)}{F_1(s)} \frac{G_2(s)}{F_2(s)} = \frac{G(s)}{F(s)} \tag{2.101}$$

so that if the output $G_2(s)$ of the second system is fed as the input $F_1(s)$ of the first, we put

$$G_2(s) = F_1(s), F_2(s) = F(s), G_1(s) = G(s). \tag{2.102}$$

so that

$$G(s) = (b_0 + b_1 s + b_2 s^2) G_2(s) \tag{2.103}$$

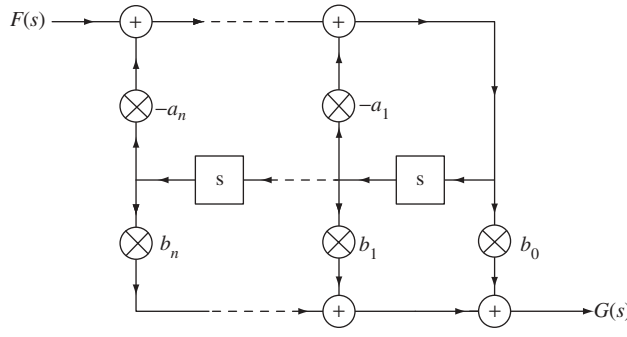


Figure 2.22 Realization of the system function in (2.105)

where

$$G_2(s) = F(s) - a_1 s G_2(s) - a_2 s^2 G_2(s) \tag{2.104}$$

which are combined into the realization of the entire second-order function shown in Figure 2.21(c), in which the differentiators are shared symbolically between the two subsystems.

The general case described by (2.82) can be treated in exactly the same way. Thus, a general transfer function of the form

$$H(s) = \frac{b_0 + b_1 s + b_2 s^2 + \cdots + b_n s^n}{1 + a_1 s + a_2 s^2 + \cdots a_n s^n} \tag{2.105}$$

has the realization shown in Figure 2.22. If the degree of the numerator of $H(s)$ is less than that of the denominator, then the corresponding multipliers and paths in Figure 2.22 are deleted.

It is also possible to start from the transfer function (2.105), and divide its numerator and denominator by s^n, which converts the function into a ratio of two polynomials in s^{-1} of the form

$$H(s) = \frac{b_n + b_{n-1} s^{-1} + \cdots + b_0 s^{-n}}{a_n + a_{n-1} s^{-1} + \cdots + s^{-n}} \tag{2.106}$$

Then, a realization can be found employing only *integrators* instead of differentiators. This is shown in Figure 2.23,which is more desirable from the practical viewpoint since integrators are preferable to differentiators and therefore almost exclusively used instead.

Either of the realizations in Figures 2.22 or 2.23 is referred to as an *analog computer simulation* of the system. The multipliers, adders, differentiators or integrators can be implemented using the Op Amp circuits discussed earlier.

An alternative method of realization of the transfer function in (2.105) which has practical advantages, begins by factorizing the function $H(s)$ as the product of second order factors (with a possible first-order term for an odd-degree function) as

$$H(s) = \prod_k H_k(s). \tag{2.107}$$

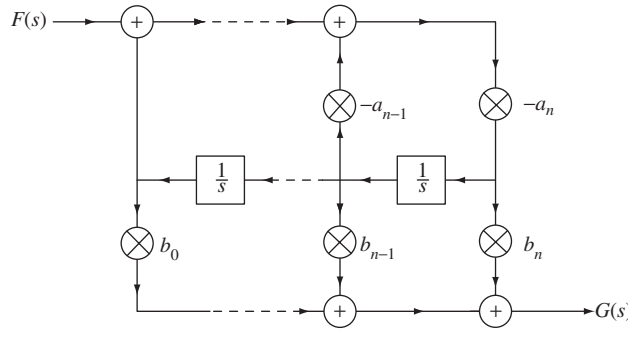


Figure 2.23 Realization of the system function in (2.106)

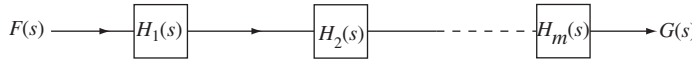


Figure 2.24 Cascade form of realization

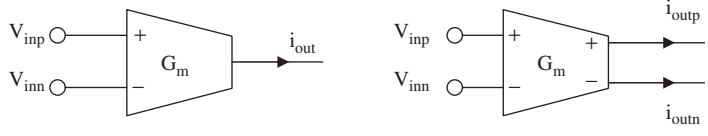


Figure 2.25 An ideal transconductance amplifier

Then, each factor is realized separately and the resulting subsystems are connected in cascade as shown in Figure 2.24. Each second-order subsystem is of the general form shown in Figure 2.22 with $n = 2$, and the possible first-order factor is obtained from the same circuit by deleting the appropriate parts corresponding to the s^2 terms and the corresponding paths. Naturally, this method of realization assumes that the cascade connections do not disturb the individual subsystems. The fulfilment of this requirement is facilitated by the use of buffers which effect isolation between the stages. However, the result is not always satisfactory and many modifications of this conceptually simple method are needed to guarantee a practically viable realization. In this regard, a more suitable realization of the integrators is now considered in terms of *operational transconductance amplifiers* (OTAs).

2.6.2 Realization Using OTAs and $G_m - C$ Circuits

An ideal transconductance amplifier is a voltage controlled current source with infinite input and output impedances [25, 26]. Thus with reference to Figure 2.25

$$i_{out} = G_m(v_{\text{inp}} - v_{\text{inn}}) \tag{2.108}$$

in which G_m is the transconductance.

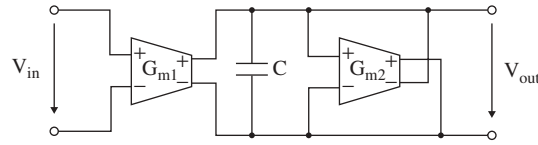


Figure 2.26 First-order section using two transconductors and a capacitor: a $G_m - C$ circuit

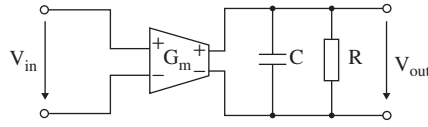


Figure 2.27 An alternative first-order section using a transconductor, a capacitor and a resistor

Using this building block, two possible realizations of a first-order section are shown in Figures 2.26 and 2.27. The former has the transfer function

$$H(s) = V_{out}/V_{in} = \frac{G_{m1}}{G_{m2}} \cdot \frac{1}{1 + sC/G_{m2}} \tag{2.109}$$

while the latter has the transfer function

$$H(s) = \frac{G_m R}{1 + RCs} \tag{2.110}$$

Either can be used for the realization of arbitrary first-order transfer functions. Furthermore, a simple integrator can be implemented using the same circuits with the removal of G_{m2} in Figure 2.26 or the removal of R in Figure 2.27, leading to the transfer function of an integrator as

$$H(s) = G_m/Cs \tag{2.111}$$

Therefore, in principle, an arbitrary transfer function can be realized as an all-integrator circuit using either of these circuits.

Another practical realization is the cascade of first-order section if the type of Figure 2.26 and second-order transfer sections employing a number of OTAs as shown in Figure 2.28. The transfer function of this second-order section is

$$H(s) = \frac{A\omega_0^2}{s^2 + (\omega_0/Q)s + \omega_0^2} \tag{2.112}$$

with

$$A = \frac{G_{m0}}{G_{m2}}, \omega_0 = \sqrt{\frac{G_{m1}G_{m2}}{C_1C_2}}, Q = \sqrt{\frac{C_2}{C_1}G_{m1}G_{m2}R_2^2} \tag{2.113}$$

In the context of filter design these circuits are referred to as $G_m - C$ circuits. The integrated circuit realizations of these circuits, together with the practical non-ideal effects, will be considered in a later chapter dealing with analog MOS integrated circuits for signal processing. In filter design, the cascade connection of lower-order sections to implement the entire filter is quite popular.

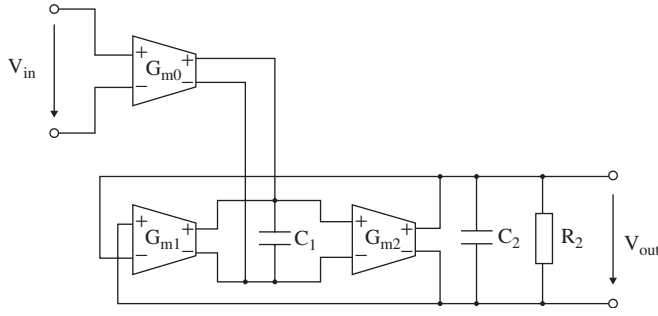


Figure 2.28 A second-order $G_m - C$ circuit

2.7 Conclusion

This chapter may be regarded as a compact review of the fundamental tools of analog signal and system analyses. The material is given for easy reference in later chapters and also to make the treatment in the book self-contained requiring no supplementary reading. The section on operational transconductance amplifiers is also very useful in filter design applications particularly using MOS integrated circuit realizations.

Problems

2.1 Find both the complex and sine-cosine Fourier series for each of the signals shown in Figure 2.29.

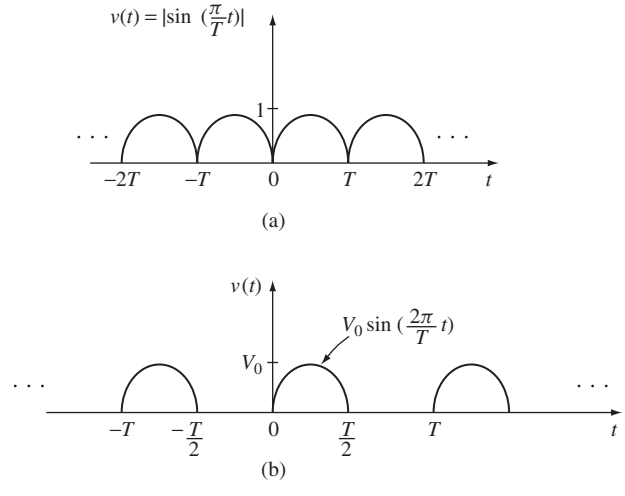


Figure 2.29

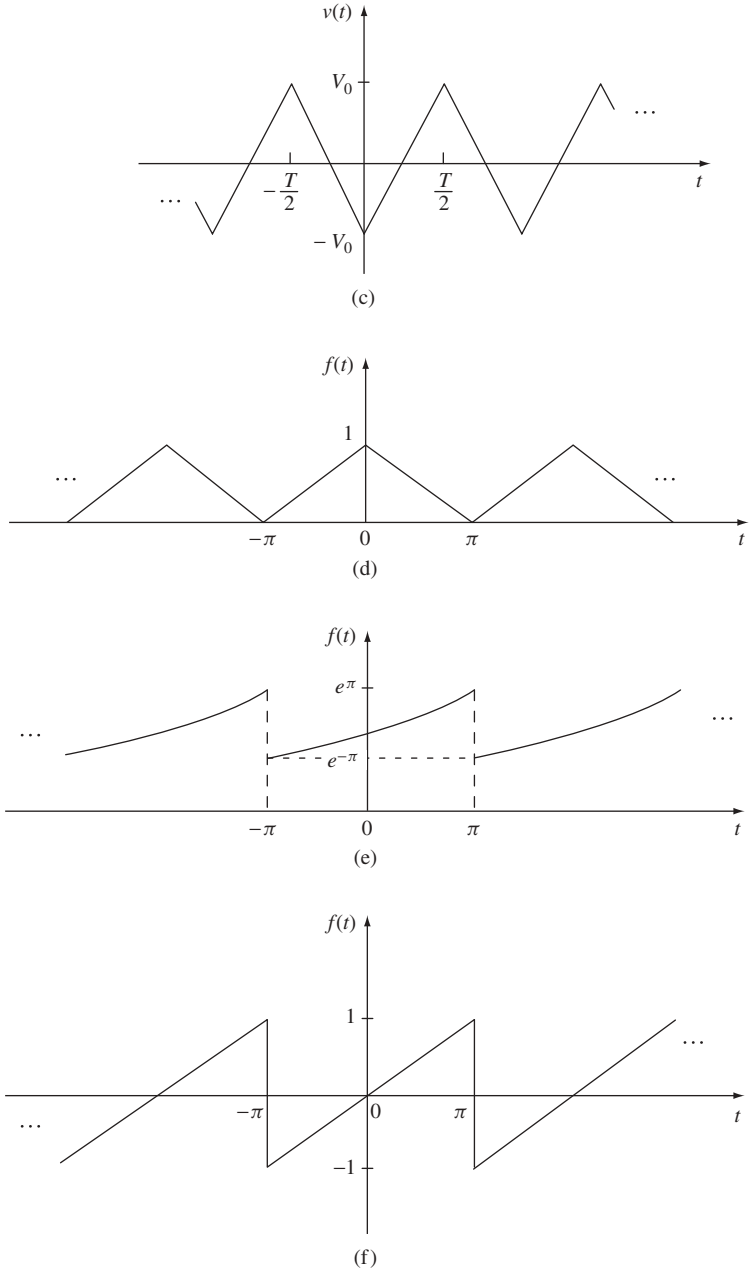


Figure 2.29 (*continued*)

2.2 Find the Fourier transform of the signal shown in Figure 2.30.

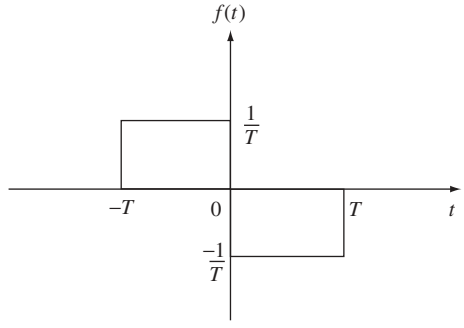


Figure 2.30

2.3 A linear system has an impulse response given by

$$h(t) = te^{-t}u(t).$$

If the excitation is the signal

$$f(t) = e^{-t}u(t)$$

find the Fourier transform of the response.

2.4 A system has a transfer function given by

$$H(s) = \frac{1}{(s+1)(s^2+s+1)}.$$

Find the response $g(t)$ of the system when the excitation is

$$f(t) = (1 - e^{-t} + e^{-3t})u(t).$$

2.5 Test the following functions for wide-sense and strict-sense (BIBO) stability

(a) $H(s) = \dfrac{(s+1)(s-2)(s+4)}{s^3+s^2+2s+2}$

(b) $H(s) = \dfrac{s+1}{s^4+s^2+s+1}$

(c) $H(s) = \dfrac{1}{s^5+2s^3+s}$

(d) $H(s) = \dfrac{s^2+2s+3}{s^7+s^5+s^3+s}$

(e) $H(s) = \dfrac{s^3}{s^3+4s^2+5s+2}$

(f) $H(s) = \dfrac{s^3}{s^6+7s^4+5s+4}.$

2.6 Realize each of the following transfer functions in direct and cascade forms using all-integrator circuits and also $G_m - C$ circuits.

(a) $H(s) = \dfrac{1}{s^4 + 2.613s^3 + 3.414s^2 + 2.613s + 1}$

(b) $H(s) = \dfrac{1}{(s^7 + 4.494s^6 + 10.103s^5 + 14.605s^4) + 14.606s^3 + 10.103s^2 + 4.494s + 1}$

(c) $H(s) = \dfrac{1}{s^5 + 15s^4 + 105s^3 + 420s^2 945s + 945}$

(d) $H(s) = \dfrac{s^3}{s^3 + 6s^2 + 15s + 15}$

(e) $H(s) = \dfrac{s^2 + 0.5}{0.5s^4 + 3s^3 + 4s^2 + 2s + 1}$

3

Design of Analog Filters

3.1 Introduction

The general theory and techniques of analog continuous-time filter design [13, 14] are discussed. These design techniques are by now *classical* in the sense of being *timeless*. They are important in themselves and are also of direct relevance to the design of all types of filter, including those which are of the sampled-data type such as digital and switched capacitor filters. This is because the filtering operation is based on the same principles and, very often, analog continuous-time models are used as starting points for the design of other types. The heavy numerical calculations which used to be a distraction from the conceptual organization of the subject can now be easily performed by powerful software tools. Therefore the chapter contains a detailed guide and numerous examples of the use MATLAB® in the design of analog filters and analysing their responses. The chapter concludes with a very important application in telecommunications namely: that of the design of pulse shaping filters for data transmission.

3.2 Ideal Filters

Consider a system or network, as shown in Figure 3.1 whose input is $f(t)$ and output is $g(t)$. With the original explicit notation of the Fourier transform we write

$$f(t) \longleftrightarrow F(\mathrm{j}\omega) \tag{3.1}$$

and

$$g(t) \longleftrightarrow G(\mathrm{j}\omega).$$

The transfer function of the system is

$$\begin{aligned} H(\mathrm{j}\omega) &= \frac{G(\mathrm{j}\omega)}{\mathrm{F}(\mathrm{j}\omega)} \\ &= |H(\mathrm{j}\omega)| \exp(\mathrm{j}\psi(\omega)) \end{aligned} \tag{3.2}$$

where $|H(\mathrm{j}\omega)|$ is the *amplitude response* and $\psi(\omega)$ is the *phase response*. The system is called an *ideal filter* if its amplitude response is constant (unity for simplicity) within

Signal Processing and Integrated Circuits, First Edition. Hussein Baher.
© 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.

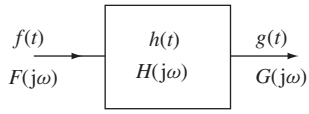


Figure 3.1 A linear system with input $f(t)$ and output $g(t)$

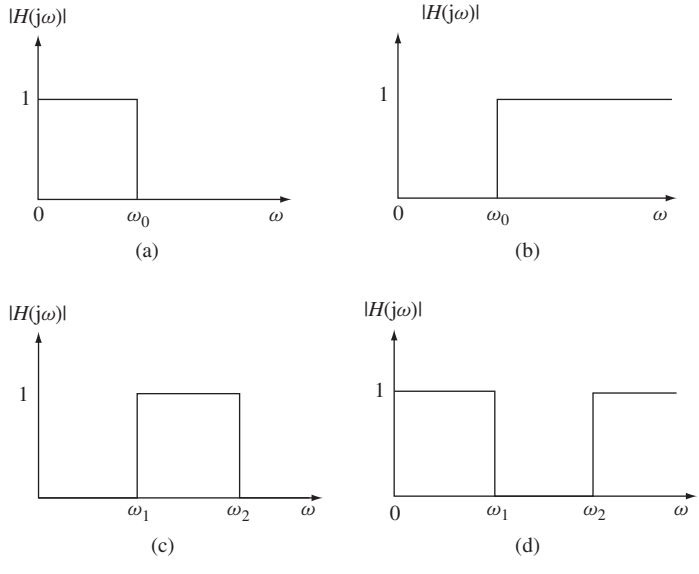


Figure 3.2 The ideal filter amplitude characteristics: (a) low-pass, (b) high-pass, (c) band-pass, (d) band-stop

certain frequency bands, and exactly zero outside these bands. In addition, in the bands where the amplitude is constant, the phase is a linear function of ω. The amplitude response of the ideal filter is shown in Figure 3.2 for the four main types of low-pass, high-pass, band-pass and band-stop filters. The ideal phase response for the low-pass case is shown in Figure 3.3 and is given by

$$\psi(\omega) = -k\omega \quad |\omega| \leq \omega_0 \tag{3.3}$$

where k is a constant.

To appreciate why these are the desirable ideal characteristics, consider the low-pass case described by the transfer function

$$\begin{aligned} H(\mathrm{j}\omega) &= \exp(-\mathrm{j}k\omega) \quad 0 \leq |\omega| \leq \omega_0 \\ &= 0 \qquad\qquad\quad |\omega| > \omega_0. \end{aligned} \tag{3.4}$$

Then

$$G(\mathrm{j}\omega) = H(\mathrm{j}\omega)F(\mathrm{j}\omega) \tag{3.5}$$

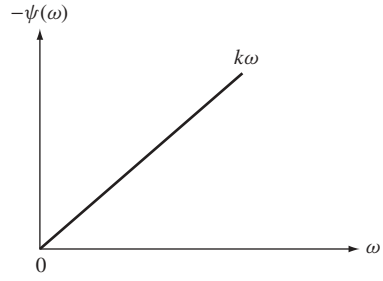


Figure 3.3 The ideal low-pass filter phase characteristic

so that for $|\omega| > \omega_0$, $G(j\omega) = 0$. In the passband $(0 \leq |\omega| \leq \omega_0)$

$$G(\mathrm{j}\omega) = \exp(-\mathrm{j}k\omega)F(\mathrm{j}\omega). \tag{3.6}$$

Taking the inverse Fourier transform of the above expression we obtain

$$g(t) = f(t-k) \tag{3.7}$$

which means that the output is an exact replica of the input, but delayed by a constant time value k. It follows that any input signal with spectrum lying within the pass-band of the ideal filter will be transmitted without attenuation and without distortion in its phase spectrum; the signal is merely delayed by a constant time value.

Now the ideal filter characteristics cannot be obtained using realizable (causal) transfer functions and must, therefore, be approximated. This statement can be appreciated if we consider again the ideal low-pass characteristic described by (3.4).

Taking the inverse Fourier transform of $H(\mathrm{j}\omega)$ we obtain the impulse response of the ideal filter as

$$\begin{aligned} h(t) &= \mathfrak{I}^{-1}[H(\mathrm{j}\omega)] \\ &= \frac{\sin\omega_0(t-k)}{\pi(t-k)} \end{aligned} \tag{3.8}$$

which is shown in Figure 3.4, and clearly exists for negative values of time, so that the required system is *non-causal*, which means that it is unrealizable by physical components.

Similarly for an ideal band-pass filter we require

$$\begin{aligned} H(\mathrm{j}\omega) &= \exp(-\mathrm{j}k\omega) \quad \omega_1 \leq |\omega| \leq \omega_2 \\ &= 0 \qquad\qquad\qquad \text{elsewhere} \end{aligned} \tag{3.9}$$

which, upon taking its inverse Fourier transform, gives the impulse response of the required filter as

$$h(t) = \frac{2}{\pi(t-k)} \cos\left(\frac{\omega_2+\omega_1}{2}\right)(t-k) \sin\left(\frac{\omega_2-\omega_1}{2}\right)(t-k) \tag{3.10}$$

which is, again, non-causal.

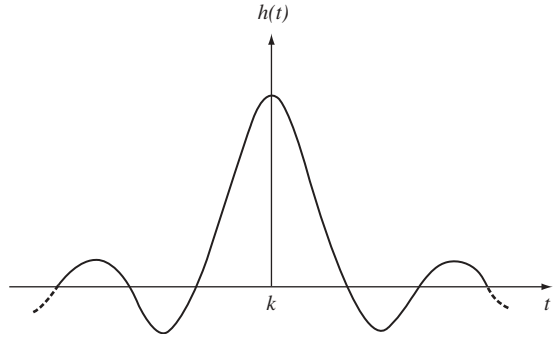


Figure 3.4 The impulse response of the ideal filter

Any deviation from the ideal amplitude characteristic is called *amplitude distortion*, while any deviation from the ideal (linear) phase characteristic is called *phase distortion*. In some applications, such as voice communication, filters are designed on amplitude basis only since it is claimed that the human ear is relatively insensitive to phase distortion. Other applications tolerate some amplitude distortion while requiring a close approximation to the ideal linear phase response. However, in modern high-capacity communication systems, filters are required to possess good amplitude (high selectivity) as well as good phase characteristics.

Since the ideal filter characteristics cannot be achieved exactly by realizable (causal) transfer functions, these must be approximated under the restriction of realizability. This means that a transfer function must be derived which, on the one hand, meets the specifications on the filter response within given tolerance and, on the other hand, it must meet the realizability conditions as a physical system. This is called the *approximation problem*. Then, the design can be completed by realizing the transfer function using the desired building blocks; this is the *synthesis problem*. The steps involved in the filter design problem are illustrated in Figure 3.5.

In the remainder of this chapter we shall give an introductory outline of the approximation problem for amplitude-oriented design and phase-oriented design. In doing so, the resulting transfer functions are realizable in a wide variety of forms, including passive as well as active networks. However, a study of these realization techniques as well as the associated realizability theory belong to the area of *network synthesis* which is one

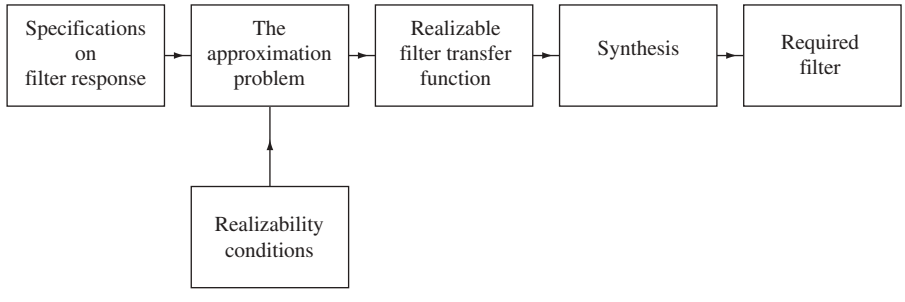


Figure 3.5 The steps to be followed for the design of a filter

of the most exciting and rigorous areas of circuit theory and signal processing, but lies outside the scope of this book. Therefore, we shall derive the filter transfer functions for the most commonly used types of filter, and rely on the realization techniques discussed in Chapter 2 as conceptually simple and sometimes viable synthesis techniques in so far as they illustrate some of the basic principles involved. Furthermore, the filter transfer functions discussed in this chapter will serve as *prototypes* from which digital and switched-capacitor filters can be designed as we shall see in later chapters.

Before embarking on the discussion of the filter design problem, we note that *causality* of the filter transfer function $H(s)$ is implied throughout the exposition. Consequently, the sinusoidal steady state response of the filter is obtained by letting $s \rightarrow j\omega$ in $H(s)$. Conversely, we can obtain $H(s)H(-s)$ from the magnitude squared function $|H(j\omega)|^2 = H(j\omega)H(-j\omega)$ by letting $\omega \rightarrow s/j$. A stable transfer function is then obtained by assigning the open left half-plane poles to $H(s)$.

3.3 Amplitude-oriented Design

We now discuss the amplitude approximation problem. This consists in finding a realizable magnitude function $|H(j\omega)|$, or equivalently a magnitude squared function $|H(j\omega)|^2$ which is capable of meeting arbitrary specifications on the amplitude response of the filter. That is, the function is capable of approximating, arbitrarily closely, the ideal amplitude filter characteristic.

Now, the attenuation (or loss) function of a filter described by $H(j\omega)$ is defined as

$$\alpha(\omega) = 10 \log \frac{1}{|H(j\omega)|^2} \text{dB} \tag{3.11}$$

It is convenient to begin our discussion by considering the design of low-pass filters, then proceed with the methods of obtaining other types. Thus, the low-pass approximation problem is to determine $|H(j\omega)|^2$ such that the typical specifications shown in Figure 3.6 are met. This is a *tolerance scheme* such that in the passband and stopband the response lies within the shaded areas in Figure 3.6(a) and the corresponding areas in Figure 3.6(b). In the transition band, lying between ω_0 and ω_s the amplitude is assumed to decrease monotonically, hence the attenuation increases monotonically. The frequency ω_0 is called the *passband edge or cut-off frequency* while ω_s is referred to as the *stopband edge*.

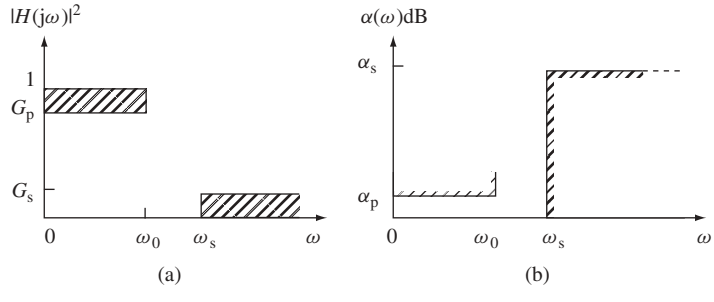


Figure 3.6 Tolerance schemes for amplitude-oriented filter design: (a) magnitude-squared, (b) attenuation

Thus for the low-pass response, the passband extends from 0 to ω_0 while the stopband extends from ω_s to infinity.

Now, the amplitude-squared function of the filter may be written as

$$|H(\mathrm{j}\omega)|^2 = \frac{\sum_{r=0}^{m} a_r \omega^{2r}}{1 + \sum_{r=1}^{n} b_r \omega^{2r}} \tag{3.12}$$

and the problem may be posed as one of determining the coefficients a_r and b_r such that the above function is capable of meeting an arbitrary set of specifications.

3.3.1 Maximally Flat Response in both Pass-band and Stop-band

This type of approximation leads to the so-called Butterworth response, the general appearance of which is shown in Figure 3.7. This is obtained by forcing the maximum possible number of derivatives of $|H(\mathrm{j}\omega)|^2$, with respect to ω, to vanish at $\omega = 0$ and $\omega = \infty$.

For $(2n - 1)$ zero derivatives around $\omega = 0$ we have

$$a_r = b_r \quad r = 1, 2, \ldots (n - 1). \tag{3.13}$$

For $(2n - 1)$ zero derivatives at $\omega = \infty$ we obtain

$$a_r = 0 \quad r = 1, 2, \ldots (n - 1). \tag{3.14}$$

Combining (3.13) with (3.14), expression (3.12) becomes

$$|H(\mathrm{j}\omega)|^2 = \frac{a_0}{1 + b_n \omega^{2n}}$$

in which a_0 and b_n are constants which can be taken as unity for convenience without affecting the properties of the response. Thus

$$|H(\mathrm{j}\omega)|^2 = \frac{1}{1 + \omega^{2n}} \tag{3.15}$$

where n is the degree of the filter and the 3 dB point occurs at $\omega = 1$ for all n. This point can be arbitrarily scaled, at a later stage, to any arbitrary value. Typical responses of filters

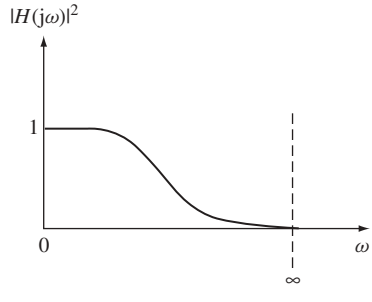


Figure 3.7 General appearance of the maximally flat amplitude response in both bands (Butterworth)

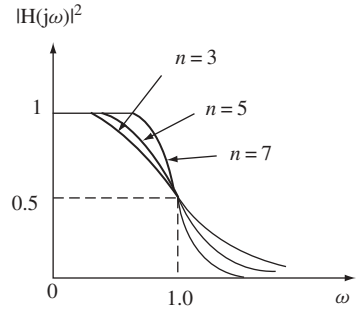


Figure 3.8 Typical responses of filters with maximally flat response in both bands

are shown in Figure 3.8 for varying degree n. These show that all aspects of the response improve with increasing the degree of the filter (hence requiring more elements for its realization), and as $n \to \infty$ the response approaches the ideal amplitude characteristic.

In order to determine the expression for $H(s)$ we note that the poles of expression (3.15) occur at

$$\omega^{2n} = -1 = \exp[\,\mathrm{j}(2r-1)\pi\,] \tag{3.16}$$

that is, at

$$\omega = \exp\left(\frac{\mathrm{j}(2r-1)\pi}{2n}\right) \tag{3.17}$$

or, by analytic continuation, letting $\omega^2 = -s^2$, the poles of $H(s)H(-s)$ occur at

$$\begin{aligned} s &= \mathrm{j}\,\exp(\mathrm{j}\theta_r) \\ &= -\sin\theta_r + \mathrm{j}\,\cos\,\theta_r \end{aligned} \tag{3.18}$$

where

$$\theta_r = \frac{(2r-1)}{2n}\pi \quad r = 1, 2, \ldots \tag{3.19}$$

For a stable (realizable) $H(s)$ we select the open left half-plane poles to form a strictly Hurwitz denominator. Hence the resulting transfer function is obtained as

$$H(s) = \frac{1}{\prod_{r=1}^{n} [s - \mathrm{j}\,\exp(\mathrm{j}\theta_r)]} \tag{3.20}$$

It now remains to determine the degree n of the required filter from a set of specifications. These may be expressed in either of two forms:

(a) 3 dB-point at $\omega = 1$,
Stopband edge at $\omega = \omega_s$ with $\alpha(\omega) \geq \alpha_s$ dB for $\omega \geq \omega_s$.
To obtain the required degree in this case, (3.15) is used to write at the stopband edge

$$10\log(1 + \omega_s^{2n}) \geq \alpha_s \tag{3.21}$$

which gives

$$n \geq \frac{\log(10^{0.1\alpha_s} - 1)}{2\log\omega_s} \tag{3.22}$$

in which ω_s is the actual stopband-edge frequency *normalized* with respect to the 3 dB point.

(b) An alternative format for the specifications may be as follows:
Maximum passband attenuation $= \alpha_p$, $\omega \leq \omega_p$
Minimum stopband attenuation $= \alpha_s$, $\omega \geq \omega_s$
Ratio of stopband to passband edges $\omega_s/\omega_p = \gamma$.

In the above format, the passband edge is defined at a frequency which is not necessarily the 3 dB point. In this case we require at the passband edge

$$10\log(1 + \omega_p^{2n}) \leq \alpha_p$$

so that

$$2n\log\omega_p \leq (10^{0.\alpha_p} - 1). \tag{3.23}$$

At the stopband edge, (3.21) is still valid, which when combined with (3.23) the result is

$$n \geq \frac{\log[(10^{0.1\alpha_s} - 1)/(10^{0.1\alpha_p} - 1)]}{2\log(\omega_s/\omega_p)} \tag{3.24}$$

3.3.2 Chebyshev Response

For the same filter degree n, a considerable improvement in the rate of cut-off, over the Butterworth response, results if we require $|H(j\omega)|^2$ to be equiripple in the passband while retaining the maximally flat response in the stopband. Typical responses of this type of approximation are shown in Figure 3.9.

For a low-pass prototype with passband edge normalized to $\omega = 1$, we still use (3.14) to force the function to have $(2n - 1)$ zero derivatives at $\omega = \infty$. The resulting function takes the form

$$|H(j\omega)|^2 = \frac{1}{1 + \varepsilon^2 T_n^2(\omega)} \tag{3.25}$$

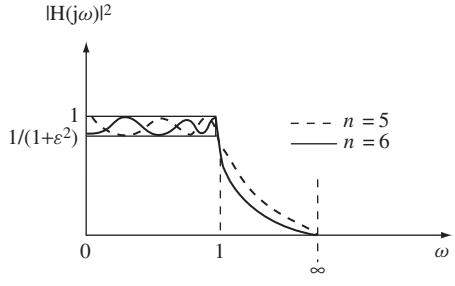


Figure 3.9 Typical responses of low-pass prototype Chebyshev filters

where $T_n(\omega)$ is chosen to be an odd or even polynomial which oscillates between -1 and $+1$ the maximum number of times in the passband $|\omega| \leq 1$ and is monotonically increasing outside this interval. This leads to a filter response which oscillates between the values 1 and $1/(1+\varepsilon^2)$ in the passband $|\omega| \leq 1$. The size of the oscillations or *ripple* can be controlled by a suitable choice of the parameter ε. The polynomial $T_n(\omega)$ which leads to these desired properties is the *Chebyshev polynomial of the first kind* defined by

$$\begin{aligned} T_n(\omega) &= \cos(n \cos^{-1} \omega) \qquad 0 \leq |\omega| \leq 1 \\ &= \cosh(n \cosh^{-1} \omega) \quad |\omega| > 1. \end{aligned} \tag{3.26}$$

This polynomial can be generated using the recurrence formula

$$T_{n+1}(\omega) = 2\,\omega T_n(\omega) - T_{n-1}(\omega) \tag{3.27}$$

with

$$T_0(\omega) = 1 \quad T_1(\omega) = \omega.$$

The above formula gives

$$\begin{aligned} T_2(\omega) &= 2\,\omega^2 - 1 \\ T_3(\omega) &= 4\,\omega^3 - 3\,\omega \\ T_4(\omega) &= 8\,\omega^4 - 8\,\omega^2 + 1 \\ T_5(\omega) &= 16\,\omega^5 - 20\,\omega^3 + 5\,\omega. \end{aligned} \tag{3.28}$$

It is observed that

$$\begin{aligned} T_n(0) &= 0 \quad \text{for } n \text{ odd} \\ T_n(0) &= 1 \quad \text{for } n \text{ even} \end{aligned} \tag{3.29}$$

which means that for n odd $|H(0)|^2 = 1$ while for n even $|H(0)|^2 = 1/(1+\varepsilon^2)$.

The Chebyshev approximation is known to be the optimum solution to the problem of determining an $|H(\mathrm{j}\omega)|^2$ which is constrained to lie in a band for $0 \leq |\omega| \leq 1$ and attain the maximum value for all ω in the range $1 < \omega \leq \infty$ for a given degree n. We shall see, shortly, that a combination of ε and n can always be found to meet arbitrary specifications. First let us find the expression for $H(s)$. The poles of (3.25) occur at

$$\varepsilon^2 T_n^2 = -1. \tag{3.30}$$

Let an *auxiliary parameter* η be defined as

$$\eta = \sinh\left(\frac{1}{n} \sinh^{-1} \frac{1}{\varepsilon}\right). \tag{3.31}$$

Then, from (3.26) the pole locations satisfy

$$\cos^2(n \cos^{-1} \omega) = -\sinh^2(n \sinh^{-1} \eta) \tag{3.32}$$

or

$$n \cos^{-1} \omega = n \sin^{-1} \mathrm{j}\eta + (2r-1)\pi/2 \tag{3.33}$$

that is the poles occur at

$$s = -\mathrm{j}\cos(\sin^{-1}\mathrm{j}\eta + \theta_r) \tag{3.34}$$

where

$$\theta_r = \frac{(2r-1)\pi}{2n}. \tag{3.35}$$

Thus, a stable (realizable) transfer function is obtained by selecting the open left half-plane poles from those given by (3.34) to give

$$H(s) = \frac{\prod_{r=1}^{n}[\eta^2 + \sin^2(r\pi/n)]^{1/2}}{\prod_{r=1}^{n}\{\mathrm{s} + [\eta\sin\theta_r + \mathrm{j}(1+\eta^2)^{1/2}\cos\theta_r]\}} \tag{3.36}$$

in which the numerator is just a constant chosen such that $H(0) = 1$ for n odd and $H(0) = 1/(1+\varepsilon^2)^{1/2}$ for n even.

Now, it remains to determine the expression for the minimum filter degree required to meet a typical set of specifications. Let these be expressed as

Passband attenuation $\alpha(\omega) \leq \alpha_\mathrm{p},\ 0 \leq \omega \leq 1$
Stopband attenuation $\alpha(\omega) \geq \alpha_\mathrm{s},\ \omega \geq \omega_\mathrm{s}$.

Therefore, from (3.25), we require in the passband

$$10\log(1+\varepsilon^2) \leq \alpha_\mathrm{p}$$

or

$$\varepsilon^2 \leq 10^{0.1\alpha_\mathrm{p}} - 1 \tag{3.37}$$

while at the stopband edge we must have

$$10\log\{1 + [\varepsilon\cosh\{n\cosh^{-1}\omega_\mathrm{s})]^2\} \geq \alpha_\mathrm{s}\}$$

which, upon solving for n and making use of (3.37) gives for the required degree

$$n \geq \frac{\cosh^{-1}[(10^{0.1\alpha_\mathrm{s}} - 1)/(10^{0.1\alpha_\mathrm{p}} - 1)]^{1/2}}{\cosh^{-1}\omega_\mathrm{s}} \tag{3.38}$$

in which ω_s is the actual stopband edge frequency normalized to the passband edge since the latter is assumed to be at $\omega = 1$.

3.3.3 Elliptic Function Response

This is the optimum amplitude response in the sense of minimizing the maximum deviation from specified values in each band of a rational function $|H(\mathrm{j}\omega)|^2$, for the two-band approximation of Figure 3.6. It gives rise to equiripple responses in both the passband

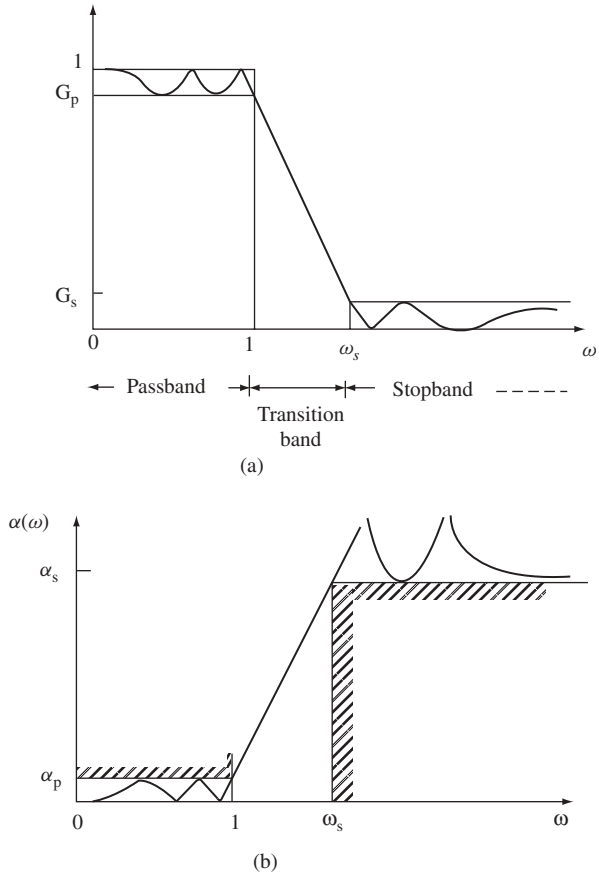


Figure 3.10 Typical response of the optimum equiripple low-pass prototype filter: (a) magnitude-squared, (b) attenuation

and stopband as shown by the typical example in Figure 3.10. The expressions for the transfer functions involve Jacobian elliptic integrals and elliptic functions [15], hence the name *elliptic filters*. These will not be discussed here, but due to the fact that these filters provide the optimum amplitude response for a given degree, the expressions for the transfer functions as well tables for the element values in many types of realizations are given in handbooks [16] and can also be obtained using a software such as MATLAB®. This point will be illustrated by numerous design examples.

3.4 Frequency Transformations

Our discussion so far, has concentrated on low-pass prototype filters in which the pass-band edge (cut-off frequency) is normalized to $\omega = 1$. We now consider the process of denormalization to arbitrary cutoff as well as the design of high-pass, band-pass and band-stop filters relying on the results of the low-pass prototype.

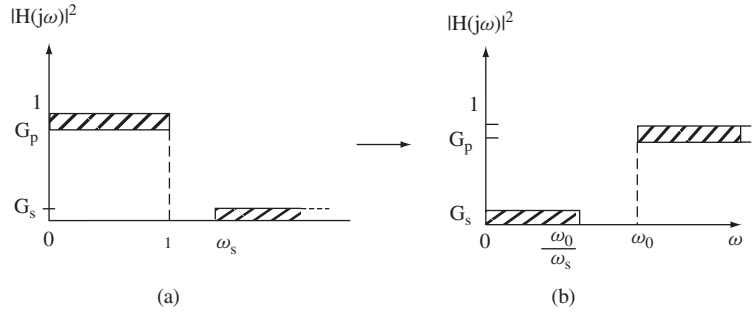


Figure 3.11 Low-pass to high-pass transformation: (a) low-pass prototype specifications, (b) high-pass specifications

3.4.1 Low-pass to Low-pass Transformation

In the prototype transfer function, denormalization to an arbitrary cutoff ω_0 can be achieved by means of the transformation

$$\omega \rightarrow \omega/\omega_0 \tag{3.39a}$$

or

$$s \rightarrow s/\omega_0. \tag{3.39b}$$

3.4.2 Low-pass to High-pass Transformation

A high-pass response with passband edge at ω_0 can be obtained from the low-pass prototype transfer functions by letting

$$\omega \rightarrow \omega_0/\omega \tag{3.40a}$$

or

$$s \rightarrow \omega_0/s \tag{3.40b}$$

which is illustrated in Figure 3.11.

3.4.3 Low-pass to Band-pass Transformation

Starting from the low-pass prototype specifications shown in Figure 3.12(a) in which the negative frequency side is included, we seek a transformation to a band-pass response with passband extending from ω_1 to ω_2 as shown in Figure 3.12(b).

In the prototype low-pass transfer function, let

$$\omega \rightarrow \beta\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right) \tag{3.41}$$

or

$$s \rightarrow \beta\left(\frac{s}{\omega_0} + \frac{\omega_0}{s}\right)$$

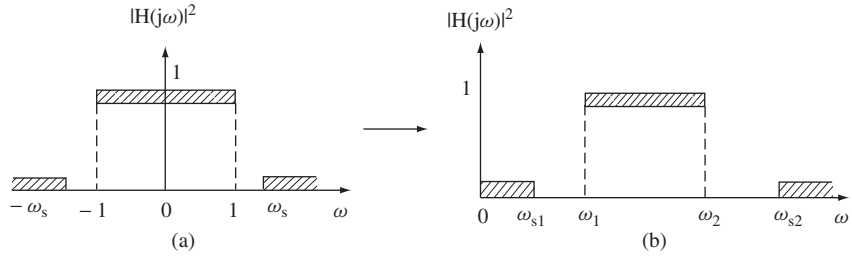


Figure 3.12 Low-pass to band-pass transformation: (a) low-pass prototype specifications, (b) band-pass specifications

where β and ω_0 are to be determined from ω_1 and ω_2. By reference to Figure 3.12, we use (3.41) to impose the conditions

$$\begin{aligned} -1 &= \beta\left(\frac{\omega_1}{\omega_0} - \frac{\omega_0}{\omega_1}\right) \\ 1 &= \beta\left(\frac{\omega_2}{\omega_0} - \frac{\omega_0}{\omega_2}\right) \end{aligned} \tag{3.42}$$

which, when solved simultaneously, yield

$$\begin{aligned} \omega_0 &= (\omega_1\,\omega_2)^{1/2} \\ \beta &= \frac{\omega_0}{\omega_2 - \omega_1} = \frac{\omega_0}{BW}. \end{aligned} \tag{3.43}$$

For example, the low-pass Chebyshev response function is transformed to the band-pass one given by

$$|H(\,\mathrm{j}\omega)|^2 = \left\{1 + \varepsilon^3 T_n^2\left[\beta\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right)\right]\right\}^{-1} \tag{3.44}$$

Examination of the transformation in (3.41) reveals that the resulting band-pass response has *geometric symmetry* around ω_0, which is referred to as the *band-centre*. This means that the amplitude has the same value at every pair of frequencies $\hat{\omega}$ and $\hat{\hat{\omega}}$ related by $\hat{\omega}\hat{\hat{\omega}} = \omega_0^2$, that is

$$\alpha(\hat{\omega}) = \alpha(\hat{\hat{\omega}}) \tag{3.45}$$

for

$$\hat{\omega}\hat{\hat{\omega}} = \omega_0^2.$$

3.4.4 Low-pass to Band-stop Transformation

This is illustrated in Figure 3.13 and can be achieved by starting from the low-pass prototype functions and letting

$$\omega \longrightarrow 1/\beta\left(\frac{\omega}{\omega_0} - \frac{\omega_0}{\omega}\right) \tag{3.46a}$$

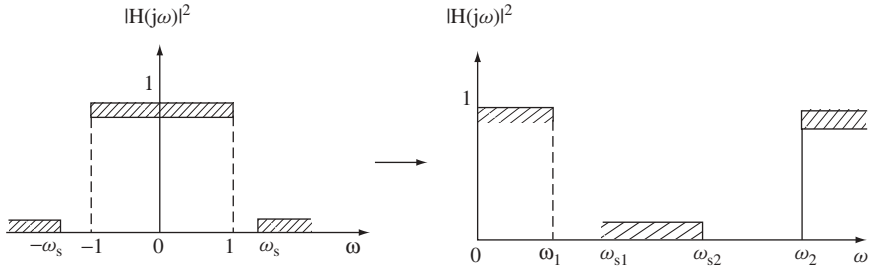


Figure 3.13 Low-pass to band-stop transformation: (a) low-pass prototype specifications, (b) band-stop specifications

or

$$s \longrightarrow 1/\beta \left(\frac{s}{\omega_0} + \frac{\omega_0}{s} \right) \tag{3.46b}$$

with β and ω_0 given by the same expressions (3.43), where ω_1 is the lower passband edge and ω_2 is the upper passband edge.

3.5 Examples

Example 3.1 Find the transfer function of a maximally flat low-pass filter with the following specifications

Passband: 0 to 1 kHz, attenuation $\leq$3 dB.
Stopband edge: 1.5 kHz, attenuation $\geq$40 dB.

Solution. The normalized stopband edge is $\omega_s = 1.5$ relative to the 3 dB point. Thus, the degree of the required filter is obtained from (3.22) as

$$n \geq \frac{\log(10^4 - 1)}{2 \log 1.5} \geq 11.358$$

so that we take $n = 12$. The transfer function of the filter is obtained from (320) with $n = 12$, then the transformation (3.39) is used with $\omega_0 = 2\pi \times 10^3$ for denormalization to the actual cut-off.

Example 3.2 Find the transfer function of a Chebyshev low-pass filter which meets the following specifications

Passband: 0 to 0.5 MHz, with 0.2 dB ripple.
Stopband edge: 1.0 MHz, attenuation $\geq$50 dB.

Solution. Normalizing the frequencies to the given cut-off of 0.5 MHz, we have $\omega_s = 1/0.5 = 2$. Substituting in (3.38) for $\alpha_p = 0.2$, $\alpha_s = 50$ and $\omega_s = 2$, we obtain

$n \geq 6.06$ and therefore, we take $n = 7$. Also from (3.37)

$$\varepsilon = (10^{0.02} - 1)^{1/2} = 0.2171$$

and the auxiliary parameter in is

$$\eta = \sinh\left(\frac{1}{7}\sinh^{-1}\frac{1}{0.2171}\right).$$

The above values are used in (3.36) to obtain the normalized prototype transfer function, which is then denormalized to the actual cutoff of $2\pi \times 0.5 \times 10^6$ using the transformation in (3.39).

Example 3.3 Find the transfer function of the band-pass Chebyshev filter with the specifications shown in the tolerance scheme of Figure 3.14.

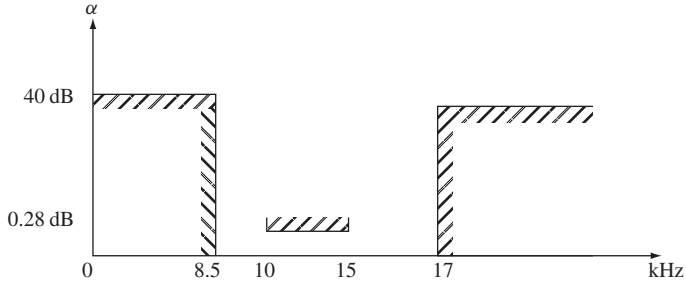


Figure 3.14 Specifications on the filter of Example 3.3

Solution. The specifications do not display the geometric symmetry inherent in the transformation (3.41). Therefore, the filter has to be designed according to the more severe of the two requirements on the lower and upper transition bands. First note that $\omega_0 = 2\pi(10 \times 15)^{1/2} = 2\pi \times 12.247 \times 10^3$ rad/s. Also, using (3.42) we have

$$\beta = \frac{\omega_0}{\omega_2 - \omega_1} = \frac{12.247}{15 - 10} = 2.4219.$$

Then, with $f_0^2 = 150 \times 10^6$, the filter has to be designed according to the more severe of the two requirements: (a) 40 dB at 8.5 kHz or (b) 40 dB at 17 kHz. If we require $\alpha \geq 40$ dB for $f \leq 8.5$ kHz we also obtain [according to (2.150)] $\alpha \geq 40$ dB for $f \geq (150/8.5) \geq 17.65$ kHz, and we have failed to satisfy the upper stopband requirement. But, requiring $\alpha \geq 40$ dB for $f \geq 17$ kHz gives $\alpha \geq 40$ dB for $f \leq (150/17) \leq 8.82$ kHz. Thus, the lower stopband requirement at 8.5 kHz is oversatisfied. Therefore we use the 40 dB requirement at 17 kHz to determine the prototype. This frequency corresponds to ω_s of the low-pass prototype

$$\begin{aligned}\omega_s &= \beta\left(\frac{\omega_{s2}}{\omega_0} - \frac{\omega_0}{\omega_{s2}}\right)\\ &= 2.4219\left(\frac{17}{12.247} - \frac{12.247}{17}\right)\\ &= 1.635.\end{aligned}$$

Also, from (3.37) with $\alpha_p = 0.28$ we obtain $\varepsilon = 0.258$, and (2.143) gives $n \geq 6.19$, so that we take $n = 7$. The auxiliary parameter η is obtained from (3.31) as $\eta = 0.2992$. Thus the required prototype transfer function can now be obtained using (3.36). Then, the transformation of (3.41) is used to find the final band-pass filter transfer function.

3.6 Phase-oriented Design

3.6.1 Phase and Delay Functions

The ideal (no distortion) phase characteristic is a linear function of ω as shown in Figure 3.3 for the low-pass case. In the treatment of phase approximation to the ideal characteristic, the problem may be stated directly in terms of the required phase. Thus, if we write

$$H(\mathrm{j}\omega) = |H(\mathrm{j}\omega)| \exp(\mathrm{j}\psi(\omega)) \tag{3.47}$$

then we require

$$\psi(\omega) \approx -k\omega. \tag{3.48}$$

Alternatively, the problem may be formulated in terms of either the *group delay* $T_g(\omega)$ or the *phase delay* $T_{ph}(\omega)$ defined by

$$T_g(\omega) = -\frac{\mathrm{d}\psi(\omega)}{\mathrm{d}\omega} \tag{3.49}$$

$$T_{ph}(\omega) = -\frac{\psi(\omega)}{\omega}. \tag{3.50}$$

Evidently, for an approximation to the ideal phase characteristic in the passband $|\omega| < \omega_0$, the group delay in (3.49) is required to approximate a constant within the passband.

Taking logarithms of both sides of (3.47) we obtain

$$\begin{aligned} \ln H(\mathrm{j}\omega) &= \ln |H(\mathrm{j}\omega)| + \mathrm{j}\psi(\omega) \\ &= \frac{1}{2} \ln(H(\mathrm{j}\omega)H(-\mathrm{j}\omega)) + \mathrm{j}\psi(\omega) \end{aligned} \tag{3.51}$$

$$\psi(\omega) = -\frac{1}{2}\,\mathrm{j} \ln \frac{H(\mathrm{j}\omega)}{H(-\mathrm{j}\omega)} \tag{3.52}$$

so that

$$\begin{aligned} -\frac{\mathrm{d}\psi(\omega)}{\mathrm{d}\omega} &= -\frac{1}{2}\left(\frac{\mathrm{d}}{\mathrm{d}(\mathrm{j}\omega)} \ln H(j\omega) + \frac{\mathrm{d}}{\mathrm{d}(-\mathrm{j}\omega)} \ln H(-\mathrm{j}\omega)\right) \\ &= -\mathrm{Re}\left(\frac{\mathrm{d}}{\mathrm{d}(\mathrm{j}\omega)} \ln H(\mathrm{j}\omega)\right) \end{aligned} \tag{3.53}$$

and we can write

$$T_g(\omega) = -\mathrm{Ev}\left(\frac{\mathrm{d}}{\mathrm{d}s} \ln H(s)\right)\Bigg|_{s=\mathrm{j}\omega}. \tag{3.54}$$

Furthermore, if $H(\mathrm{j}\omega)$ is written in the form

$$H(\mathrm{j}\omega) = \frac{E_1(\omega) + \mathrm{j}O_1(\omega)}{E_2(\omega) + \mathrm{j}O_2(\omega)} \tag{3.55}$$

where $E_{1,2}(\omega)$ are even and $O_{1,2}(\omega)$ are odd polynomials, then the phase $\psi(\omega)$ is an odd function given by

$$\psi(\omega) = \tan^{-1}\frac{O_1(\omega)}{E_1(\omega)} - \tan^{-1}\frac{O_2\,\omega}{E_2(\omega)}. \tag{3.56}$$

Defining the *generalized phase function* as

$$\psi(s) = -\frac{1}{2}\ln\left(\frac{H(s)}{H(-s)}\right) \tag{3.57}$$

so that

$$\psi(\omega) = -\mathrm{j}\psi(s)|_{s=\mathrm{j}\omega} \tag{3.58}$$

and the *generalized group delay* can be defined as

$$T_\mathrm{g}(s) = -\frac{\mathrm{d}\psi(s)}{\mathrm{d}s} \tag{3.59}$$

so that

$$\begin{aligned} T_\mathrm{g}(s) &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}s}\left(\ln\frac{H(s)}{H(-s)}\right) \\ &= \frac{1}{2}\left(\frac{\mathrm{d}}{\mathrm{d}s}\ln H(s) - \frac{\mathrm{d}}{\mathrm{d}s}\ln H(-s)\right) \end{aligned} \tag{3.60}$$

that is

$$T_\mathrm{g}(s) = \mathrm{Ev}\left(\frac{\mathrm{d}}{\mathrm{d}s}\ln H(s)\right) \tag{3.61}$$

and

$$T_\mathrm{g}(\omega) = T_\mathrm{g}(s)|_{s-\mathrm{j}\omega}. \tag{3.62}$$

If $H(s)$ is written as

$$H(s) = \frac{P(s)}{Q(s)} \tag{3.63}$$

then, use of (3.61) gives

$$T_\mathrm{g}(s) = \frac{1}{2}\left(\frac{P'(s)}{P(s)} + \frac{P'(-s)}{P(-s)} - \frac{Q'(s)}{Q(s)} - \frac{Q'(-s)}{Q(-s)}\right) \tag{3.64}$$

where the prime denotes differentiation. The expressions obtained so far allow us to evaluate the delay and phase responses of any system.

A function $H(s)$ with all its zeros in the closed left half-plane is termed a *minimum-phase function*. This is because under this constraint, the phase is a minimum for a given amplitude.

3.6.2 Maximally Flat Delay Response

We now consider one method for approximating the ideal constant group delay characteristic. Consider a function of the form

$$H(s) = \frac{1}{Q(s)} \tag{3.65}$$

and write

$$Q(s) = E(s) + O(s) \tag{3.66}$$

where $E(s)$ is even and $O(s)$ is odd. The generalized phase function is then given by

$$\psi(s) = \tanh^{-1} \frac{O(s)}{E(s)} \tag{3.67}$$

so that

$$\psi(\omega) = -\tan^{-1} \frac{O(\mathrm{j}\omega)}{\mathrm{j}E(\mathrm{j}\omega)}. \tag{3.68}$$

In the maximally flat group delay approximation, it is required to derive the nth degree polynomial $Q_n(s)$ which results in a group delay function possessing the maximum possible number of zero derivatives at $\omega = 0$. To this end we write

$$\tanh \psi(s) = \frac{O(s)}{E(s)} \tag{3.69}$$

and note that, if the right side of the above expressions were to be *identical to the function*

$$\tanh s = \sinh s / \cosh s \tag{3.70}$$

then the group delay would be exactly constant at all frequencies. However, since $O(s)$ and $E(s)$ are *real* polynomials whereas $\tanh s$ is transcendental, we must find a method of approximating (3.70) by a real rational function. A possible method is to write

$$\begin{aligned} \sinh s &= s + \frac{s^3}{3!} + \frac{s^5}{5!} + \cdots \\ \cosh s &= 1 + \frac{s^2}{2!} + \frac{s^4}{4!} + \cdots \end{aligned} \tag{3.71}$$

then perform a *continued fraction expansion* of ($\sinh s$/$\cosh s$) to obtain

$$\tanh s = \cfrac{1}{\cfrac{1}{s} + \cfrac{1}{\cfrac{3}{s} + \cfrac{1}{\cfrac{5}{s} + \ddots \cfrac{1}{\cfrac{(2n-1)}{s} + \ddots}}}} \tag{3.72}$$

Finally, $O(s)/E(s)$ is identified with the nth approximant in the above continued fraction expansion. This means that the expansion is truncated at the nth step, then the resulting terms are remultiplied to form a rational function approximation to $\tanh s$. This is equated to $O(s)/E(s)$ in (3.69) and the required polynomial $Q_n(s)$ is defined. In fact, it is possible to use the theory of continued fractions to show that the polynomial $Q_n(s)$ can be generated using the recurrence formula

$$Q_{n+1}(s) = Q_n(s) + \frac{s^2}{(4n^2 - 1)} Q_{n-1}(s) \tag{3.73}$$

with

$$Q_0(s) = 1 \quad Q_1(s) = 1 + s.$$

The polynomial defined by (3.73) is related to the well known *Bessel polynomial* $B(s)$ by

$$Q_n(s) = s^n B_n(1/s). \tag{3.74}$$

Therefore the resulting filter is often called a *Bessel filter*. The maximally flat character of the group delay of the transfer function can be verified using (3.64) and the properties of $Q_n(s)$. Furthermore, since $Q_n(s)$ is obtained according to (3.72), it is strictly Hurwitz and the filter is guaranteed to be stable. Figure 3.15 shows examples of the delay response of Bessel filters. However, since all the available degrees of freedom have been placed on the delay response, we expect the amplitude response of these filters to be rather poor.

Finally, it is to be noted that the low-pass to low-pass transformation (3.20) can be used to adjust the cut-off to an arbitrary value. However, none of the other transformations to high-pass, band-pass or band-stop are valid since they distort the delay characteristics. Therefore these cases must be designed independently.

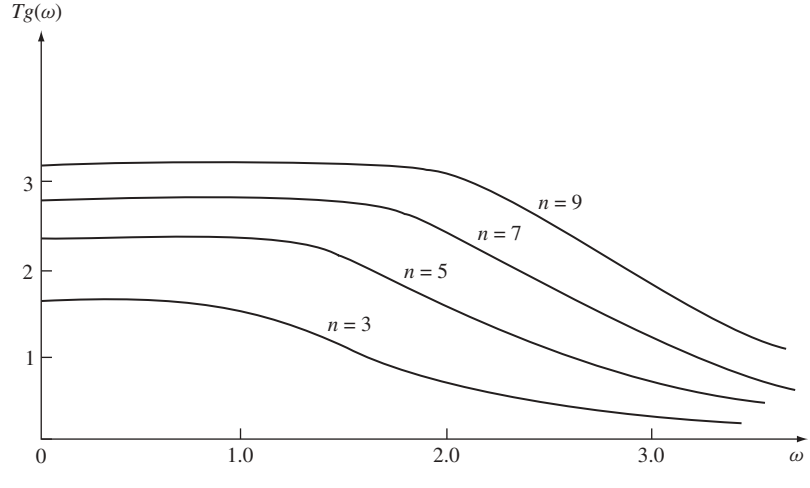


Figure 3.15 Typical delay responses of Bessel filters

3.7 Passive Filters

All the transfer functions generated in the previous section can be realized in passive form as ladder structures of the general form shown in Figure 3.16.

The details of the techniques may be found in [13, 14]. Here we give a brief outline of the results. With reference to Figure 3.17, the Butterworth filter prototype has the element values

$$g_r = \sin\theta_r, \quad r = 1, 2, \ldots n \tag{3.75}$$

where

$$g_r = L_r \text{ or } C_r \tag{3.76}$$

and

$$R_L = R_g = 1\,\Omega \tag{3.77}$$

For the Chebyshev response

$$g_1 = \frac{2}{\eta}\sin\left(\frac{\pi}{2n}\right) \tag{3.78}$$

$$g_r g_{r+1} = \frac{4\sin\theta_r \sin\left(\dfrac{2r+1}{2n}\right)\pi}{\eta^2 + \sin^2\left(\dfrac{r\pi}{n}\right)} \tag{3.79}$$

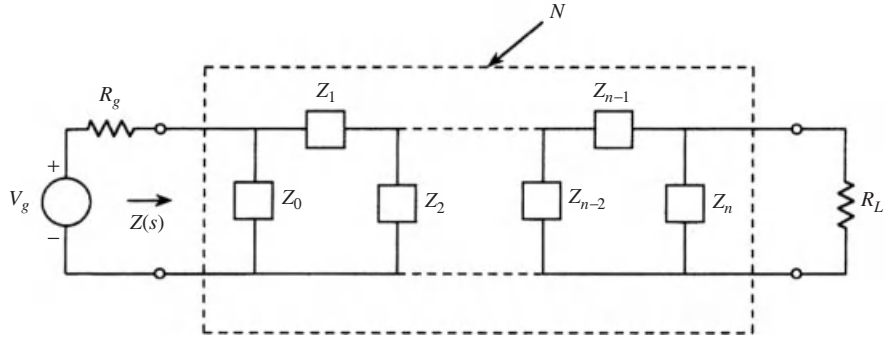


Figure 3.16 General passive ladder terminated in resistors

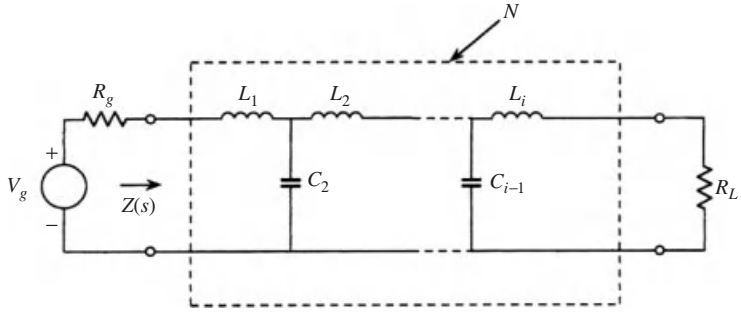


Figure 3.17 Low-pass ladder for the realization of Butterworth, Chebyshev and Bessel filters

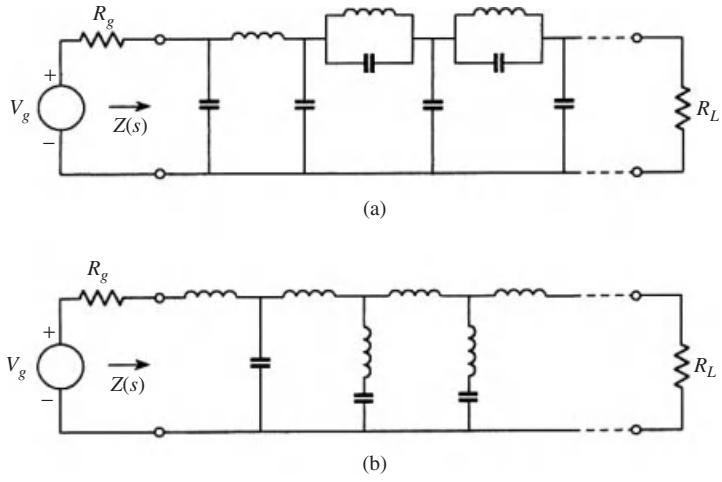


Figure 3.18 Alternative ladder forms for the realization of elliptic filters (a) mid-shunt, (b) mid-series

with the same notation as in (3.76). Here, for n odd $R_g = R_L = 1\Omega$ while for n even $R_g = 1$ (Figure 3.17): and R_L is determined from

$$|H(0)|^2 = \frac{4R_L}{(R_L + 1)^2} \tag{3.80}$$

For the elliptic filter prototype shown in Figure 3.18, there are no explicit formulae, so that the element values are determined from tables [16].

Finally, Figure 3.19 shows the transformations to obtain the element values of the filter types from the normalized prototype.

3.8 Active Filters

The design techniques of analog continuous-time active filters can follow either of two possible approaches. The first begins by designing a passive filter meeting the required specifications, then, an active equivalent is derived from the designed passive network. This approach has the advantage of possessing low sensitivity properties with respect to element value variations, which is a desirable attribute from the practical viewpoint. The second approach starts by finding the transfer function of the required filter from the specifications, as indicated earlier in this chapter, this is then factored into second-order sections (and a possible first-order section), each is realized separately by a simple active network and the resulting sections are connected in cascade. This approach was discussed in Section 2.6 and is much simpler, but in the case of high-order filters can lead to responses which are quite sensitive to element value variations due to tolerance and other possible errors inherent in the fabrication process.

Consider the general passive ladder shown in Figure 3.20 where the branches are arbitrary impedances. Write the *state equations* of the ladder, relating the series currents

Normalized low-pass prototype with cut-off at $\omega = 1$	Low-pass with cut-off at $\omega = \omega_0$	High-pass with cut-off at $\omega = \omega_0$	Band-pass with passband edges at ω_1 and ω_2	Band-stop with stopband edges at ω_1 and ω_2 (ω_s is the stopband edge in the low-pass prototype)
L_r	L_r/ω_0	$I/L_r\omega_0$	$\frac{\beta L_r}{\omega_0}$ $\frac{I}{\beta L_r \omega_0}$	$\frac{L_r\omega_s}{\beta\omega_0}$ $\frac{\beta}{L_r\omega_0\omega_s}$
C_r	C_r/ω_0	$I/C_r\omega_0$	$\frac{I}{\beta C_r\omega_0}$ $\frac{\beta C_r}{\omega_0}$	$\frac{\beta}{C_r\omega_0\omega_s}$ $\frac{C_r\omega_s}{\beta\omega_0}$
For denormalization to arbitrary source resistor R_g, $L \to R_g L$, $C \to CIR_g$, $R_L \to R_g R_L$				

Figure 3.19 Frequency transformations and impedance scaling

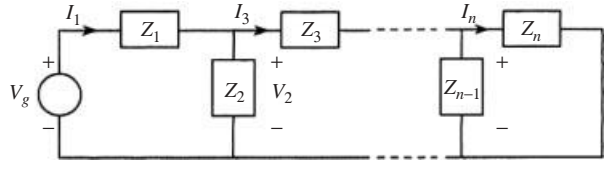


Figure 3.20 A general passive ladder

to the shunt voltages. For the sake of specificity, we assume n to be odd.

$$
\begin{aligned}
I_1 &= Z_1^{-1}(V_2 - V_g) \\
V_2 &= Z_2(I_1 - I_3) \\
I_3 &= Z_3^{-1}(V_2 - V_4) \\
&\cdots\cdots\cdots\cdots \\
&\cdots\cdots\cdots\cdots \\
I_n &= Z_n^{-1} V_{n-1}
\end{aligned}
\tag{3.81}
$$

Consider the simulation of this ladder by means of the structure shown in Figure 3.21. Let the currents $I_1, I_3, I_5, \ldots . I_n$ in (3.81) be simulated by the voltages $\hat{V}_1, \hat{V}_3, \hat{V}_5, \ldots . \hat{V}_n$.

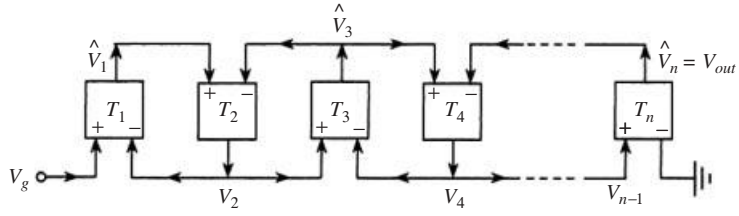


Figure 3.21 State-variable (leap-frog) ladder

The state variable (leap-frog) ladder simulation is described by

$$\begin{aligned}
\hat{V}_1 &= T_1(V_o - V_2) \\
V_2 &= T_2(\hat{V}_1 - \hat{V}_3) \\
V_3 &= T_3(\hat{V}_2 - V_4) \\
&\cdots\cdots\cdots\cdots\cdots \\
&\cdots\cdots\cdots\cdots\cdots \\
\hat{V}_n &= T_n V_{n-1}
\end{aligned} \tag{3.82}$$

The internal structure of each box in Figure 3.21 is chosen such that

$$\begin{aligned}
T_1 &= \alpha Z_1^{-1} \\
T_2 &= \alpha^{-1} Z_2 \\
T_3 &= \alpha Z_3^{-1} \\
&\cdot \\
&\cdot \\
&\cdot \\
T_n &= \alpha Z_n^{-1}
\end{aligned} \tag{3.83}$$

where α is a constant. Then, the transfer function of the leap-frog ladder of Figure 3.21

$$\hat{H}_{21} = \frac{V_{out}}{V_g} \tag{3.84}$$

differs from H_{21} of the passive ladder only by a constant. Consequently, given a ladder of the general form in Figure 3.20, with the specific type and values of elements, it is possible to determine the state variable simulation provided we can find the necessary building blocks with transfer functions satisfying (3.83). For the ladders which realize the Butterworth, Chebyshev and Bessel low pass filters, we require a state variable simulation with integrators (elements with frequency dependence $1/Ls$ or $1/Cs$). So, an all-integrator network is possible. Integrators and integrator-summers can, in principle, be implemented using the Op Amp resistor–capacitor combinations discussed earlier. However from the practical viewpoint, passive resistors are to be avoided in integrated circuit realizations. Therefore, operational transconductance amplifiers (OTAs) provide an attractive alternative for the implementation of integrators and $G_m - C$ circuits result which employ integrators as explained in Section 2.6. Again the integrated circuit realizations of Op Amps, OTAs and other building blocks will be discussed in great detail in the part on analog MOS integrated circuits for signal processing.

3.9 Use of MATLAB® for the Design of Analog Filters

The filter transfer functions derived and studied in the previous section, can be easily generated using the *Signal Processing Toolbox* in MATLAB®. The relevant commands and functions are given in this section and may be used to reinforce the readers knowledge and as a labour-saving device. These functions generate the transfer function from the filter specifications and give the result in either of two forms:

1. The poles and zeros of the filter transfer function in the form [z,p,k]. In this case, MATLAB® returns a column matrix [z] with the zero locations, a column matrix [p] with the pole locations, and a value k of the gain of the filter. In the case of an all-pole transfer function such as that of the Butterworth or Chebyshev filter, [z] is returned as an empty matrix.
2. The coefficients of the numerator and denominator are given in the form [a,b] where [a] is an array of the numerator coefficients in descending powers of s, while [b] is an array of the denominator coefficients in the same order with the first coefficient as 1. In the case of an all-pole transfer function such as that of the Butterworth or Chebyshev filter, [a] is returned as an empty matrix.

3.9.1 Butterworth Filters

The required degree of the filter is calculated from

```
n=buttord(wp,ws,ap,as,'s')
```

where for a low-pass and high-pass filters, wp and ws are scalars. For band-pass and band-stop filters each is a vector of two numbers [wp1 wp2] and [ws1 ws2] denoting the two passband edges and the two stopband edges, respectively.

Low-pass Prototype

```
[z,p,k]=buttap(n)
```

Low-pass with Cutoff at wo

```
[z,p,k]=butter(n,wo,'s')
[a,b]=butter(n,wo,'s')
```

High-pass with Cutoff at wo

```
[z,p,k]=butter(n,wo,'high','s')
[a,b]=butter(n,wo,'high','s')
```

Band-pass with wn = [w1 w2] a vector defined by the passband edge frequencies w1 and w2

```
[z,p,k]=butter(n,wn,'s')
[a,b]=butter(n,wn,'s')
```

Band-stop with wn = [w1 w2] a vector defined by the stopband edge frequencies w1 and w2

```
[z,p,k]=butter(n,wn,'stop','s')
[z,p,k]=butter(n,wn,'stop','s')
```

3.9.2 Chebyshev Filters

The required degree of the filter is calculated from

```
n=cheb1ord(wp,ws,ap,as,'s')
```

where for low-pass and high-pass filters, wp and ws are scalars. For band-pass and band-stop filters each is a vector of two numbers [wp1 wp2] and [ws1 ws2] denoting the two passband edges and the two stopband edges, respectively.

Low-pass Prototype

[z,p,k]=cheby1ap(n,α_p)

Low-pass with Cutoff at wo

[z,p,k]=cheby1(n,α_p, wo,'s')
[a,b]=cheby1(n,α_p, wo,'s')

High-pass with Cutoff at wo

[z,p,k]=cheby1(n,α_p, wo,'high','s')
[a,b]=cheby1(n,α_p, wo,'high','s')

Band-pass with wn = [w1 w2], a Vector Defined by the Passband Edge Frequencies w1 and w2

[z,p,k]=cheby1(n,α_p, wn,'s')
[a,b]=cheby1(n,α_p, wn,'s')

Band-stop with wn = [w1 w2], a Vector Defined by the Stopband Edge Frequencies w1 and w2

[z,p,k]=cheby1(n,α_p, wn,'stop','s')
[a,b]=cheby1(n,α_p, wn,'stop','s')

3.9.3 Elliptic Filters

The required degree of the filter is calculated from

```
n=ellipord(wo,ws,ap,as,'s')
```

where for low-pass and high-pass filters, wo and ws are scalars. For band-pass and band-stop filters each is a vector of two numbers [wp1 wp2] and [ws1 ws2] denoting the two passband edges and the two stopband edges, respectively.

Low-pass Prototype

[z,p,k]=ellipap(n,α_p, α_s)

Low-pass with Cutoff at wo

[z,p,k]=ellip(n,α_p, α_s, wo,'s')
[a,b]=ellip(n,α_p, α_s, wo,s)

High-pass with Cutoff at wo

[z,p,k]=ellip(n,α_p, α_s, wo,'high','s')
[a,b]=ellip(n,α_p, α_s, wo,'high','s')

Band-pass with wn = [w1 w2], a Vector Defined by the Passband Edge Frequencies w1 and w2

[z,p,k]=ellip(n,α_p, α_s, wn,'s')
[a,b]=ellip(n,α_p, α_s, wn,'s')

Band-stop with wn = [w1 w2], a Vector Defined by the Stopband Edge Frequencies w1 and w2

[z,p,k]=ellip(n,α_p, α_s, wn,',stop','s')
[a,b]=ellip(n,α_p, α_s, wn,'stop','s')

3.9.4 Bessel Filters

Low-pass Prototype

[z,p,k]=besselap(n)

Low-pass with Cutoff at wo

[z,p,k]=besself(n,ω_o)
[b,a]=besself(n,ω_o)

Band-pass with wn = [w1 w2], a Vector Defined by the Passband Edge Frequencies w1 and w2

[z,p,k] = besself(n,wn,'bandpass')
[b,a]= besself(n,wn,'bandpass')

High-pass with Cutoff at wo

[z,p,k] = besself(n,wo,'high')
[b,a]= besself(n,wn,'high')

Band-stop with wn = [w1 w2], a Vector Defined by the Stopband Edge Frequencies w1 and w2

[z,p,k] = besself(n,wn,'stop')
[b,a]= besself(n,wn,'stop')

Having obtained the filter transfer function, the amplitude and phase responses can be plotted and analysed using the MATLAB function

freqs(a,b)

For the realization, the numerator and denominator can be factored using the statement

R=roots(a)
Q=roots(b)

Which give the zeros and poles of the transfer function.

3.10 Examples of the use of MATLAB®

Example 3.4 Find the transfer function of a maximally flat low-pass filter with the following specifications

Passband: 0 to 2 kHz, attenuation $\leq$3 dB.
Stopband edge: 4.0 kHz, attenuation $\geq$40 dB.

```
n=buttord(wp,ws,ap,as,'s')
n=buttord(2*pi*2000,2*pi*4000,3,40,'s')
```

which gives n = 7

```
[z,p,k]=butter(n,wo,'s')
```

```
[p] =  1.0e+004 *[  −0.2796 + 1.2251i   −0.2796 − 1.2251i   −0.7835 + 0.9825i   −0.7835
      − 0.9825i   −1.1322 + 0.5452i   −1.1322 − 0.5452i   −1.2566]

k =   4.9484e+028
```

Example 3.5 Find the transfer function of an Elliptic low-pass filter which meets the following specifications

Passband: 0 to 0.5 MHz, with 0.2 dB ripple.
Stopband edge: 1 MHz, attenuation $\geq$50 dB.

```
n=ellipord(500000*2*pi,1000000*2*pi,0.2,50,'s')
```

which gives n=5

```
[z,p,k]=ellip(n,0.2,50, 500000*2*pi,'s')
[a,b]=ellip(n,0.2,50, 500000*2*pi,'s')
```

which give for the zero and pole locations

```
[z]= (1.0e+006)[0.0000 + 7.8891i   0.0000 − 7.8891i   0.0000 + 5.2110i   0.0000 − 5.2110i]

[p] = (1.0e+006)[−0.3341 + 3.2732i   − 0.3341 − 3.2732i   − 1.1401   + 2.2836i   − 1.1401
        − 2.2836i   − 1.6823]
```

and the gain

```
k = 7.0201e+004
```

Figures 3.22 and 3.23 show the responses of the filter.

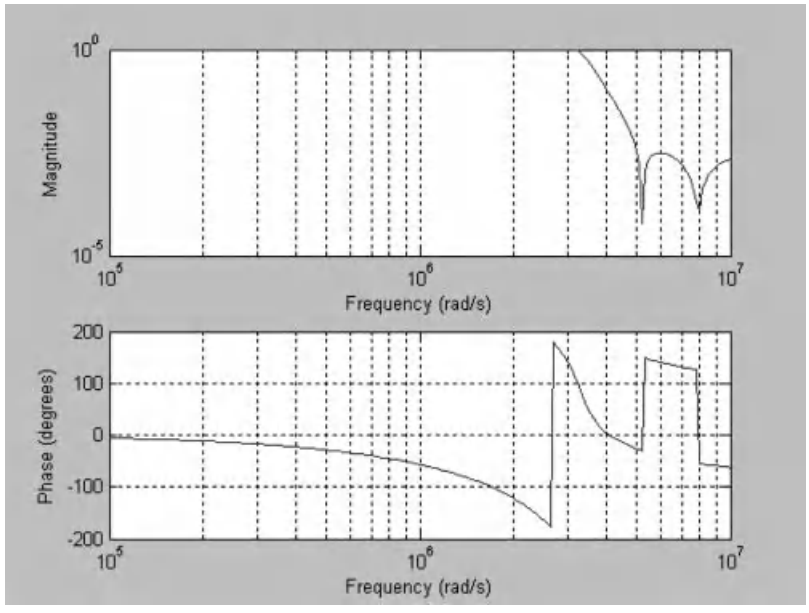


Figure 3.22 Amplitude and phase responses of the elliptic filter of Example 3.5

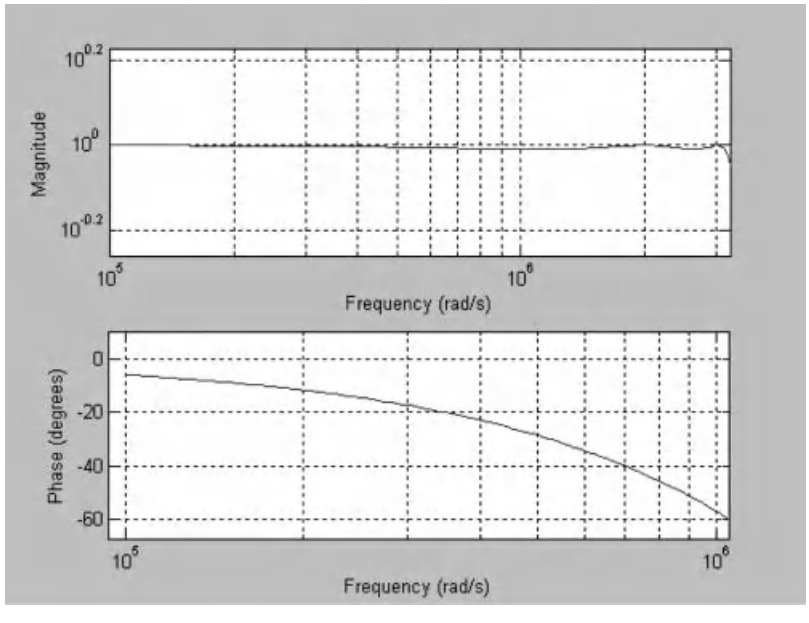


Figure 3.23 Pass-band detail of the responses of the filter of Example 3.5

Example 3.6 Find the transfer function of the band-pass Chebyshev filter with the specifications shown in the tolerance scheme of Figure 3.14 but with the passband attenuation changed to 3 dB

```
n=cheb1ord(wp,ws,ap,as,'s')
[n,wn]=cheb1ord([10000*2*pi 15000*2*pi],[8500*2*pi 17000*2*pi],3,40,'s')
n=5
wn=(1.0e+00) [6.2832   9.4248]
[z,p,k]=cheby1(n,αp, wn,'s')
z = 0 0 0 0 0
p = (1.0e+004) [−0.1028 + 9.3603i  −0.1028−9.3603i  −0.2529 + 8.6867i
−0.2529−8.6867i  −0.2789 + 7.6902i  −0.2789−7.6902i  −0.1983 + 6.8113i
−0.1983−6.8113i  −0.0695 + 6.3257i  −0.0695−6.3257i]
k = 1.9172e+021
[a,b]=cheby1(n,αp, wn,'s')
```

Figure 3.24 shows the responses of the filter.

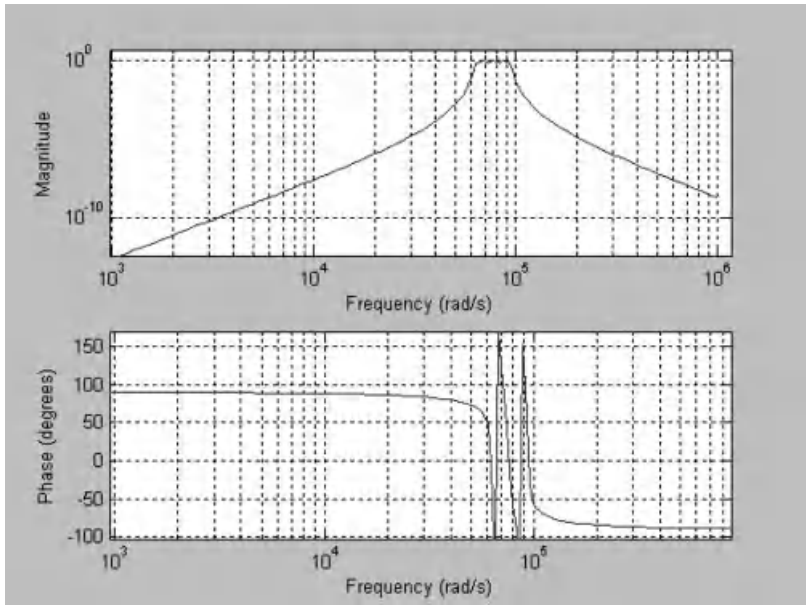


Figure 3.24 Responses of the filter of Example 3.6

3.11 A Comprehensive Application: Pulse Shaping for Data Transmission

Data of all kinds are transmitted as *pulses* which carry the relevant information about the data. The generation of the necessary pulses is one of the most important aspects of the design of data transmission systems. The questions we answer here are: what is the optimum pulse shape and how do we generate such a pulse? First we note that the use of

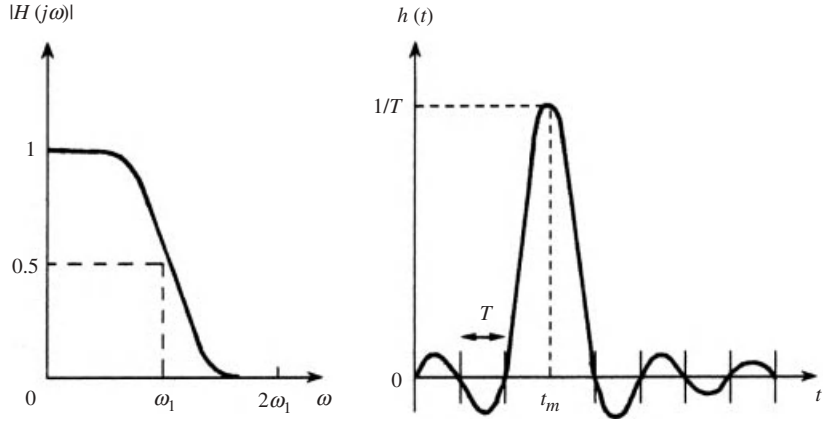


Figure 3.25 Salient features of the frequency and time responses of a pulse shaping filter

pulses which are exactly rectangular in shape is neither possible nor desirable. Second, to pack a large number of pulses into a very narrow time slot, we need a prohibitively wide bandwidth. This is clear from the uncertainty principle. So: what is the optimum pulse shape which conserves bandwidth and allows the transmission of a train of pulses without significant interference between the pulses? It is then clear that, a data transmission filter is a pulse (or impulse) shaping network. It has a time domain pulse (or impulse) response which allows the transmission of a train of shifted versions of this pulse without significant intersymbol interference (ISI). In addition, the filter must have a frequency response with sufficient selectivity and attenuation for band-limiting and suppression of noise and cross-talk. Figure 3.25 illustrates the salient features of the time response and frequency bands of interest of a continuous-time data transmission filter, in which T is the data rate interval and $2\omega_1$ is the highest frequency in the band of interest such that

$$\omega_1 = \frac{\pi}{T} \tag{3.85}$$

Furthermore, the phase response of the filter is important since a deviation from the ideal linear characteristic causes unpredictable distortion of the pulse shape and also such a deviation is incompatible with the minimum ISI requirement.

Now, it has been suggested [17] that the optimum pulse shape is that of a *raised cosine* pulse which has the frequency response

$$\begin{aligned} H(\omega) &= 1, \quad |\omega| \leq \frac{1-\alpha}{2} \\ &= 0.5\left[1 + \cos\left(\frac{\pi}{\alpha}\left\{|\omega| - \frac{1-\alpha}{2}\right\}\right)\right], \\ &= 0 \text{ elsewhere} \end{aligned} \tag{3.86}$$

with

$$0 \leq \alpha \leq 1 \tag{3.87}$$

and the time response of the raised cosine pulse is

$$h(t) = \sin c(t/T)\frac{\cos(\pi\alpha t/T)}{1-(4\alpha^2 t^2/T^2)} \tag{3.88}$$

Obviously, the above response is physically unrealizable (non-causal) and must be approximated.

We shall now develop the design of pulse shaping filters a little further in order to examine the relation between their time and frequency responses and show how they can be designed. The application to digital design is discussed in a later chapter.

The raised cosine function is not the only function that gives rise to the desired characteristics of a data transmission filter. In fact it is a particular property of the raised cosine function which when present in any other function leads to the target response [17]. This is a very simple property and may be stated as symmetry of the real part of the transfer function around the point ω_1 together with phase linearity over the range $2\omega_1$. These, of course can only be approximated.

Let the continuous-time filter transfer function [17] be written as

$$H(s) = \frac{P_{2m}(s)}{Q_n(s)}, 2m < n \tag{3.89}$$

with

$$P_{2m}(s) = \sum_{r=0}^{m} b_r s^{2r} \tag{3.90}$$

We now impose the symmetry constraints in such a way as to facilitate the derivation of the transfer function. Define the functions

$$\begin{aligned} \gamma_1(\omega) &= \cos\beta(\omega_1+\omega) \\ \gamma_2(\omega) &= \cos\beta(\omega_1-\omega) \\ F_1(\omega) &= \mathrm{Re}\,H[j(\omega_1+\omega)] \\ F_2(\omega) &= \mathrm{Re}\,H[j(\omega_1-\omega)] \end{aligned} \tag{3.91}$$

where $2\omega_1$ is the highest frequency of the band of interest and β is a parameter. Now, let $H(j\omega)$ be obtained such that it satisfies the following conditions

(a) $Arg[Q(j\omega)]$ interpolates the linear phase function $\beta\omega$ at n equidistant points over the band of interest, that is

$$\beta\omega_i - Arg[Q(j\omega_i)] = 0 \text{ for } \omega_i = i\frac{2\omega_i}{n} \quad i = 0, 1, 2, \ldots n \tag{3.92}$$

so that

$$0 < |\omega| < 2\omega_i \tag{3.93}$$

with an approximation error that can be made arbitrarily small by increasing n.

(b) $F_1/\gamma_1 + F_2/\gamma_2$ approximates unity in the least mean square sense over the band $0 < |\omega| < \omega_1$.

We can show that if conditions (a) and (b) are enforced, then the desired property of minimum ISI is obtained. Thus, the impulse response of the filter will be such that if we define

$$h_\beta(t) = h(t + \beta) \tag{3.94}$$

then

$$h_\beta(kT) \approx 0 \text{ for } k = 1, \pm 2, \pm 3, \ldots\ldots\ldots \tag{3.95}$$

and it deviates from zero by an error that can be made negligibly small by taking n, m, $(n - m)$ sufficiently large. Also

$$h_\beta(0) \approx \frac{1}{T} \text{ with } T = \frac{\pi}{\omega_1} \tag{3.96}$$

Pulse Transmission

Rather than treat the case of arbitrary (or rectangular) pulse transmission as a separate problem, we can reformulate the problem in a manner such that the conditions stated above can be directly applied without modification. To this end, let the transfer function $H_p(s)$ of the filter be formed as the product of two functions as

$$H_p(s) = \prod(s) H(s) \tag{3.97}$$

where $\prod(s)$ is the Laplace transform of a pulse $p(t)$. Then the impulse response of the filter with transfer function $H_p(s)$ is the same as the pulse response of the filter with transfer function $H(s)$. Consequently, if we design the data transmission filter with a transfer function $H(s)$ such that $H_p(s)$ satisfies conditions (a) and (b), then the pulse response of the required filter will have the required properties, those being exactly the same as those of the impulse response corresponding to $H_p(s)$.

Figures 3.26–3.28 show the responses of an 11th order pulse shaping filter designed using the present technique.

3.12 Conclusion

This chapter dealt with the design of analog continuous-time filters. Both passive and active realizations were covered. The use of MATLAB® as a design aid was also highlighted and numerous examples were given. The material in this chapter is important in itself, but is also essential for the design of digital and switched-capacitor filters to be discussed in later chapters. Passive filters act as models and reference designs for all other types of filter. They are also used extensively in radio frequency (RF) applications. A degree of familiarity with these filters is therefore of great benefit to filter designers and researchers. The chapter concluded by a comprehensive application of filter design methods to the design of one of the most important areas of communications namely: data transmission.

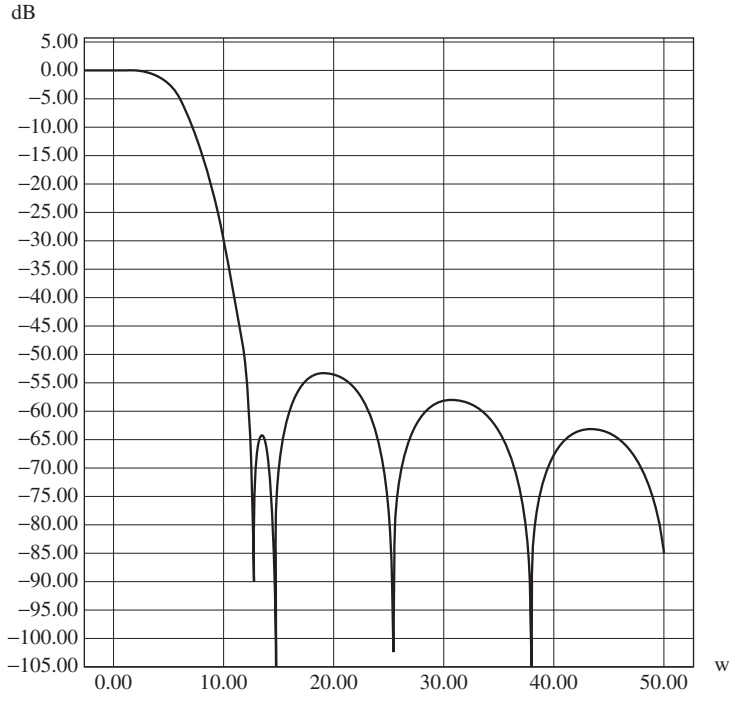


Figure 3.26 Amplitude response (dB) of an 11th order pulse shaping filter

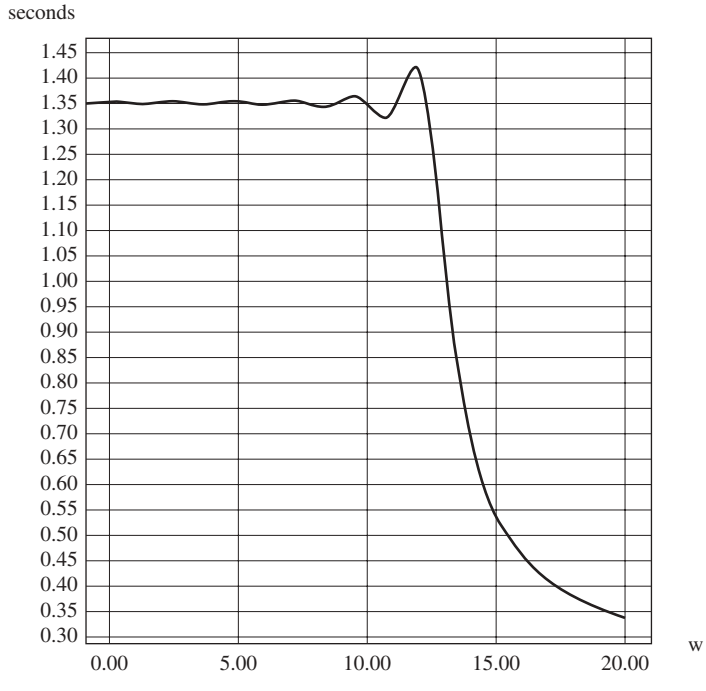


Figure 3.27 Delay response of the pulse shaping filter

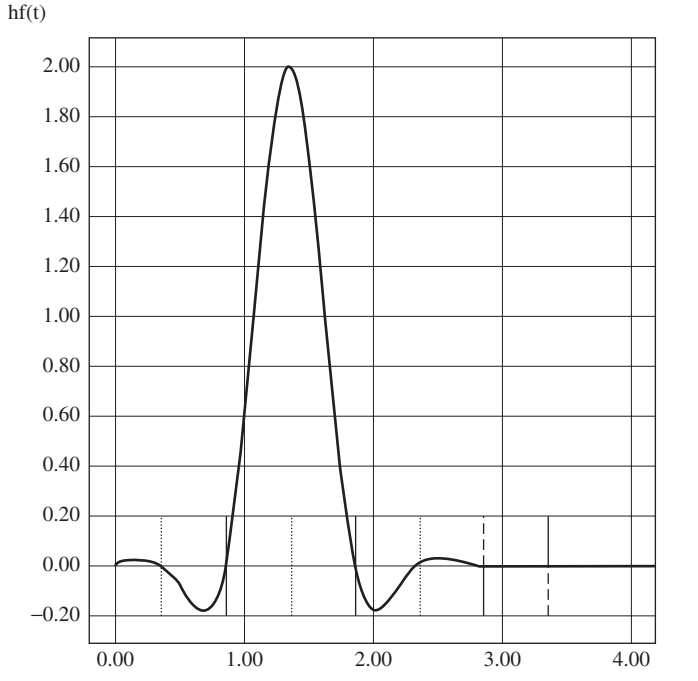


Figure 3.28 Pulse response of the 11th order pulse-shaping filter

Problems

In the following problems, the filter is to be designed using (a) passive components employing the state variable ladder structure and (b) active circuits.

3.1 Design a low-pass maximally flat filter with the following specifications
Passband: 0 to 0.8 MHz, attenuation $\leq$1 dB.
Stopband edge at 1.2 MHz, attenuation $\geq$32 dB.

3.2 Design a low-pass Chebyshev filter with the same specifications given in Problem 3.1. Compare the required degree with that of the maximally flat filter.

3.3 Design a low-pass Chebyshev filter with the following specifications
Passband: 0 to 3.2 kHz, attenuation $\leq$0.25 dB.
Stopband edge at 4.6 kHz, attenuation $\geq$32 dB.

3.4 Design a band-pass maximally flat filter with the following specifications
Passband: 12 to 15 kHz, attenuation $\leq$3 dB.
Stopband edges at 8 kHz and 20 kHz with minimum attenuation of 20 dB in both stopbands.

3.5 Design a band-pass Chebyshev filter with the following specifications
Passband: 0.8 to 1.6 MHz, attenuation $\leq$0.1 dB.
Stopband edges at 0.5 and 1.8 MHz with minimum attenuation of 50 dB in both stopbands.

3.6 Design a band-stop maximally flat filter with the following specifications
Stopband: 0.7 to 0.8 MHz, attenuation $\geq$40 dB.
Passband edges at 0.4 and 1.0 MHz with maximum attenuation of 3 dB.

3.7 Design a band-stop Chebyshev filter with the following specifications
Stopband: 10 to 12 kHz, attenuation $\geq$30 dB.
Passband edges at 8 and 15 kHz with maximum attenuation of 0.5 dB.

4

Discrete Signals and Systems

4.1 Introduction

This chapter gives a brief and compact review of the process of analog to digital conversion and the representation of discrete signals and systems. Then the classification of digital systems is given with emphasis on linear shift-invariant types. The forms of realization of finite duration and infinite duration impulse response filters are discussed [11, 12]. This chapter, as in the case of Chapter 2, can be used as a compact review of the material which is normally given early at the junior level.

4.2 Digitization of Analog Signals

A signal $f(t)$ is called a *continuous-time* or an *analog* signal if it is defined, somehow, for all values of the *continuous* variable t. If $f(t)$ is defined only at discrete values of t, it is called a *discrete-time* signal or an *analog sampled-data* signal. Suppose that in addition to being discrete-time the signal quantities $f(t)$ can assume only discrete values, and that each value is represented by a code such as the binary code. The resulting signal is said to be a *digital signal*.

The first step in the digitization process is to take samples of the signal $f(t)$ at regular time intervals: nT $(n = 0, \pm 1, \pm 2, \ldots)$. This amounts to converting the continuous-time variable t into a discrete one. This way, we obtain a signal $f(nT)$ which is defined only at discrete instants which are integral multiples of the same quantity T, which is called the *sampling period*. Such a signal may be thought of as a sequence of numbers

$$\{f(nT)\} \triangleq \{f(0), f(\pm T), f(\pm 2T), \ldots\} \tag{4.1}$$

representing the values of the function at the sampling instants. If the signal $f(t)$ is causal, that is

$$f(t) = 0 \quad t < 0 \tag{4.2}$$

then the sampled version is denoted by the sequence

$$\{f(nT)\} \triangleq \{f(0), f(T), f(2T), \ldots\}. \tag{4.3}$$

Signal Processing and Integrated Circuits, First Edition. Hussein Baher.
© 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.

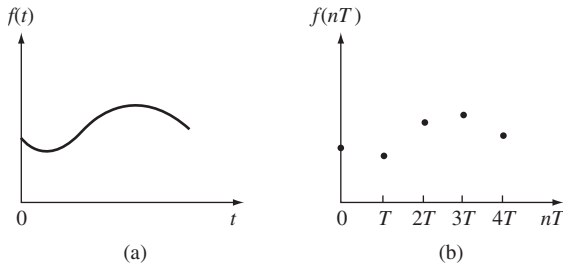


Figure 4.1 (a) A causal analog signal and (b) its sampled version

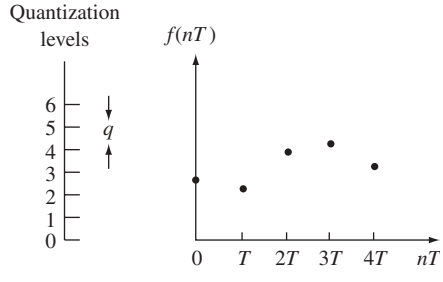


Figure 4.2 The sampled signal of Figure 4.1 and the quantization levels

In the above notation, the curly brackets denote the entire sequence, while $f(nT)$ denotes the nth sample. However, it is often convenient to drop the brackets and let $f(nT)$ denote either the sequence or a single sample, and let the context define which one is meant. This is done wherever no confusion may result. Figure 4.1 shows a causal signal and its sampled version.

Next, the discrete-time signal is *quantized*. That is, the amplitude (vertical) axis is converted into a discrete one as shown in Figure 4.2, and we regard the range of values between successive levels as inadmissible. Then, from the sequence $\{f(nT)\}$, we form a new *quantized* sequence $\{f_q(nT)\}$ by assigning to each $f(nT)$ the value of a quantization level. There are two basic methods for doing this, which will be discussed later. Finally, the discrete-time quantized sequence $\{f_q(nT)\}$ is *encoded* as shown in Figure 4.3. This means that each member of the sequence $\{f_q(nT)\}$ is represented by a code; the most commonly used one is a *binary* code.

The entire process of sampling, quantization and encoding is usually called analog to digital (A/D) conversion (Figure 4.3). In this section, each step in this conversion process is examined in some detail.

4.2.1 Sampling

4.2.1.1 Ideal Impulse Sampling

Although impulses cannot be produced physically, their use in explaining the sampling process is very instructive. Let $f(t)$ be a continuous-time (analog) signal and consider its

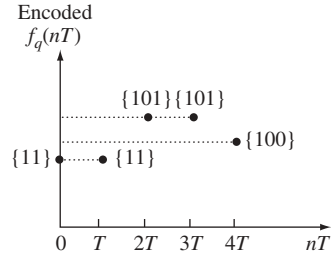


Figure 4.3 The digitized analog signal of Figure 4.1 after quantization and encoding

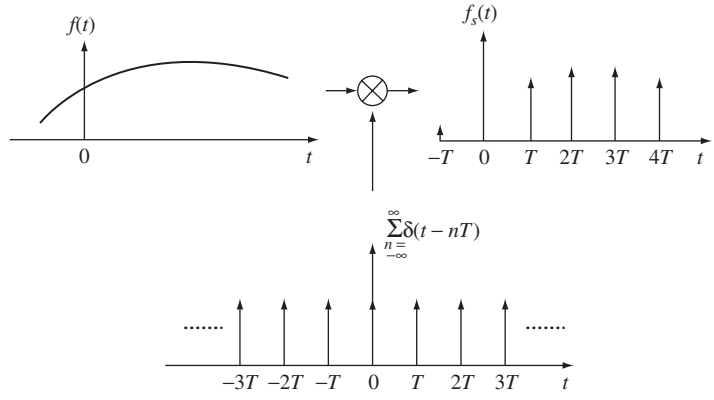


Figure 4.4 Sampling as impulse modulation

multiplication by the periodic train of impulses (see Figure 4.4)

$$\delta_\infty(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT) \tag{4.4}$$

to obtain the signal

$$f_s(t) = f(t) \sum_{n=-\infty}^{\infty} \delta(t - nT). \tag{4.5}$$

or

$$f_s(t) = \sum_{n=-\infty}^{\infty} f(nT)\delta(t - nT) \tag{4.6}$$

which is another periodic train of impulses each of strength $f(nT)$ which is the value of the function $f(t)$ at the nth instant. Figure 4.4 shows a model of this *impulse modulation* process together with the resulting impulse train.

This impulse train is regarded as the sampled signal. This is because the strength of the nth impulse is the sample value of $f(nT)$ and if each impulse is replaced by its strength

(area) then we obtain a set of numbers $\{f(nT)\}$ defining the discrete-time signal and the sampling process has been achieved.

Now, consider the affect of the sampling process on the Fourier spectrum of the original continuous-time signal. Let

$$f(t) \leftrightarrow F(\omega) \tag{4.7}$$

and suppose $F(\omega)$ is band-limited to ω_{m}, that is

$$|F(\omega)| = 0 \quad |\omega| \geq \omega_{\mathrm{m}}.$$

Using the frequency convolution relation applied to $f(t)$ and $\delta_{\infty}(t)$ we have the Fourier transform of the sampled signal

$$\begin{aligned} \mathfrak{I}[f_s(t)] &= \mathfrak{I}[f(t)\delta_{\infty}(t)] \\ &= \frac{1}{2\pi}\mathfrak{I}[f(t)] * \mathfrak{I}\left(\sum_{r=-\infty}^{\infty} \delta(t - nT)\right). \end{aligned} \tag{4.8}$$

However

$$\begin{aligned} \mathfrak{I}\left(\sum_{n=-\infty}^{\infty} \delta(t - nT)\right) &= \omega_0 \sum_{n=-\infty}^{\infty} \delta(\omega - n\omega_0) \\ &= \frac{2\pi}{T} \sum_{n=-\infty}^{\infty} \delta\left(\omega - \frac{2n\pi}{T}\right) \end{aligned} \tag{4.9}$$

so that

$$\mathfrak{I}[f_s(t)] = \frac{1}{T}F(\omega) * \sum_{n=-\infty}^{\infty} \delta\left(\omega - \frac{2n\pi}{T}\right). \tag{4.10}$$

But the unit impulse is the identity element in the process of convolution. Hence, we have (with $T = 2\pi/\omega_0$)

$$\begin{aligned} \mathfrak{I}[f_s(t)] &= F_s(\omega) \\ &= \frac{1}{T} \sum_{n=-\infty}^{\infty} F(\omega - n\omega_0). \end{aligned} \tag{4.11}$$

Thus, the spectrum of the sampled signal consists of the periodic extension of the spectrum $F(\omega)$ of the original continuous-time signal. This is illustrated in Figure 4.5, which also illustrates very important consequences of the sampling process.

Figure 4.5(b) shows the periodic spectrum of the sampled signal in the case where $\omega_0 > 2\omega_{\mathrm{m}}$. This means that the sampling frequency

$$\omega_0 = 2\pi/T \tag{4.12}$$

exceeds twice the highest frequency component of the spectrum $F(\omega)$ of the original signal. In this case, it is clear that $F(\omega)$ can be recovered by passing the spectrum $F_s(\omega)$ through a low-pass filter, which eliminates

$$\sum_{r=-\infty}^{\infty} F(\omega - r\omega_0) \quad \text{for } r = 1, 2, \ldots$$

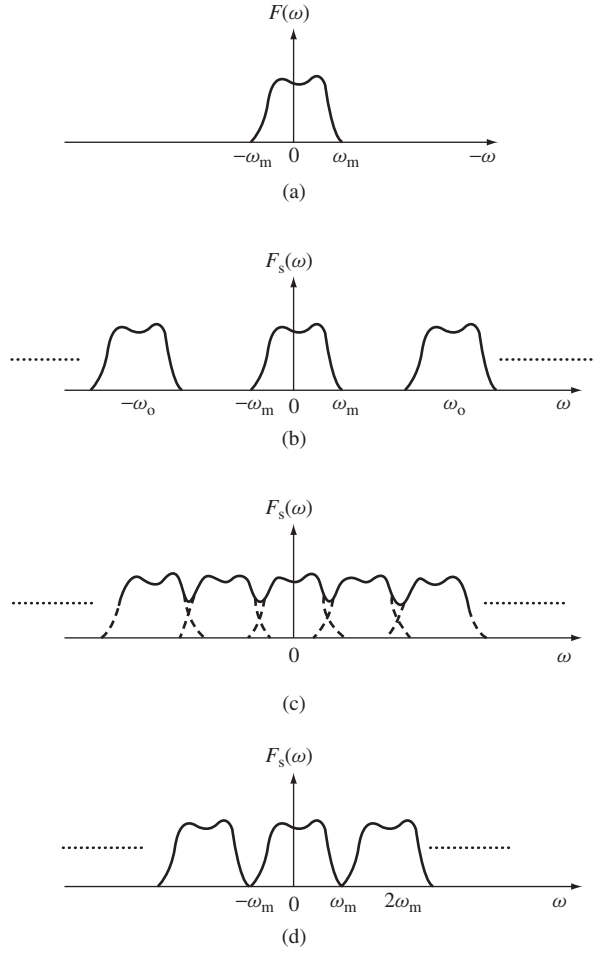


Figure 4.5 Effect of the sampling process: (a) The baseband signal spectrum, (b) the spectrum with over-sampling $\omega_0 > 2\omega_m$, (c) with under-sampling $\omega_0 < 2\omega_m$, (d) with critical sampling: $\omega_0 = 2\omega_m$

Figure 4.5(c) shows the periodic spectrum $F_s(\omega)$ for the case $\omega_0 < 2\omega_m$. This means that the sampling frequency is less than twice the highest frequency component of the band-limited spectrum $F(\omega)$ of the original signal. In this case the periodic parts of the spectrum overlap resulting in the effect called *aliasing*. This makes it impossible to recover the spectrum $F(\omega)$ by filtering.

Figure 4.5(d) shows the case of *critical* sampling with $\omega_0 = 2\omega_m$, that is, the sampling frequency *just* exceeding twice the highest frequency component of $F(\omega)$. In this case, it is possible in principle to recover $F(\omega)$ by passing $F_s(\omega)$ through an *ideal* low-pass filter, with cut-off at ω_m, as shown in Figure 4.6.

The *minimum* sampling rate required to prevent aliasing must *just* exceed *twice* the highest frequency component in the spectrum $F(\omega)$ of $f(t)$ before sampling. This minimum sampling rate is called the (radian) *Nyquist* frequency ω_N. These considerations lead to the following fundamental result.

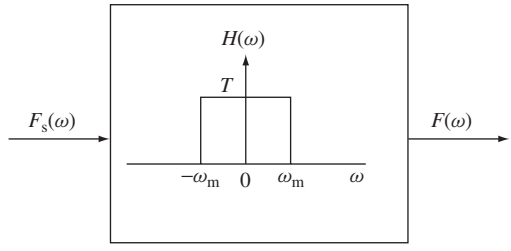


Figure 4.6 Signal recovery of the critically sampled signal using an ideal low-pass filter

4.2.1.2 The Sampling Theorem

A signal $f(t)$ whose spectrum is band-limited to below a frequency ω_m, can be completely recovered from its samples $\{f(nT)\}$ taken at a rate

$$f_N = \frac{\omega_N}{2\pi} (= 1/T) \quad \text{where } \omega_N = 2\ \omega_m. \tag{4.13}$$

The signal $f(t)$ is determined from its sample values $\{f(nT)\}$ by

$$f(t) = \sum_{n=-\infty}^{\infty} f(nT) \frac{\sin \omega_m (t - nT)}{\omega_m (t - nT)} \tag{4.14}$$

where

$$T = \frac{\pi}{\omega_m} = \frac{2\pi}{\omega_N} = \frac{1}{f_N}. \tag{4.15}$$

To prove this result we note that $F(\omega)$ can be recovered from $F_s(\omega)$ by passing it through an ideal low-pass filter of amplitude T and cut-off at ω_m, as shown in Figure 4.6. Thus, assuming critical sampling at twice the highest frequency in $F(\omega)$, the impulse response of the required filter is obtained from

$$h(t) = \mathfrak{I}^{-1}[H(j\omega)] \tag{4.16}$$

where

$$\begin{aligned} H(j\omega) &= T \qquad |\omega| \le \omega_m \\ &= 0 \qquad |\omega| > \omega_m. \end{aligned} \tag{4.17}$$

Thus

$$\begin{aligned} h(t) &= T \frac{\sin \omega_m t}{\pi t} \\ &= \frac{\sin \omega_m t}{\omega_m t} \end{aligned} \tag{4.18}$$

and the output of the filter is

$$f(t) = f_s(t) * h(t)$$

$$= \int_{-\infty}^{\infty} \left(\sum_{n=-\infty}^{\infty} f(nT)\delta(\tau - nT) \right) \frac{\sin \omega_{\mathrm{m}}(t-\tau)}{\omega_{\mathrm{m}}(t-\tau)} \, \mathrm{d}\tau$$

$$= \sum_{n=-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(nT)\delta(\tau - nT) \, \frac{\sin \omega_{\mathrm{m}}(t-\tau)}{\omega_m(t-\tau)} \right) \mathrm{d}\tau \Bigg) \tag{4.19}$$

which upon use of (2.55) gives

$$f(t) = \sum_{n=-\infty}^{\infty} f(nT) \frac{\sin \omega_{\mathrm{m}}(t-nT)}{\omega_{\mathrm{m}}(t-nT)} \tag{4.20}$$

as required. This expression is essentially a formula for the interpolation of the signal values by its values at the sampling points. $F(\omega)$ is, however, recoverable from $F_s(\omega)$ when $F(\omega)$ does not contain any frequency components higher than half the sampling frequency. It follows that $f(t)$ is recoverable from its sample values, at least in principle, using (4.20) if and only if the sampling theorem is satisfied.

From the above discussion of the idealized sampling process we can draw the following important conclusions:

1. The choice of a sampling frequency for a signal, is determined by the highest frequency component of the Fourier spectrum $F(\omega)$ of the signal $f(t)$. In practice, the signal is usually band-limited to a frequency $\omega_{\mathrm{N}}/2$, prior to sampling at a frequency of ω_{N}. This is done by *prefiltering*.
2. Critical sampling with $\omega_{\mathrm{N}} = 2\omega_{\mathrm{m}}$ requires an ideal filter for the reconstruction of a signal having frequency components up to ω_{m}. Such a filter is non-causal and hence, is physically unrealizable. Therefore, in practice the sampling frequency is chosen to be higher than the Nyquist rate in order that the reconstruction filter may have a realizable response. For example speech signals are bandlimited to 3.4 kHz and sampled at a rate of 8.0 kHz instead of the critical rate of 6.8 kHz.
3. So far, we have assumed a signal $f(t)$ whose spectrum extends from $\omega = 0$ to $|\omega| = \omega_{\mathrm{m}}$, that is a *low-pass* signal. In this case the signal is completely determined from its set of values at regularly spaced intervals of period $T = 1/2\ f_m = \pi/\omega_m$. Now consider a *band-pass* signal whose spectrum exists only in the range

$$\omega_1 < |\omega| < \omega_2 \tag{4.21}$$

as shown in Figure 4.7. It is easy to show that the minimum (radian) sampling frequency in this case is given by

$$\omega_{\mathrm{N}} = 2(\omega_2 - \omega_1) \tag{4.22}$$

The reconstruction of the signal, in this case, must be done by a band-pass filter.

4.2.1.3 Practical Sampling Functions

In the previous section we considered *instantaneous* sampling of signals, by means of impulses. Though very instructive, leading to valid conclusions about the sampling process, impulse sampling is not feasible in practice. Therefore we now give a more practical method of sampling a signal, and show that it leads essentially to the same conclusions.

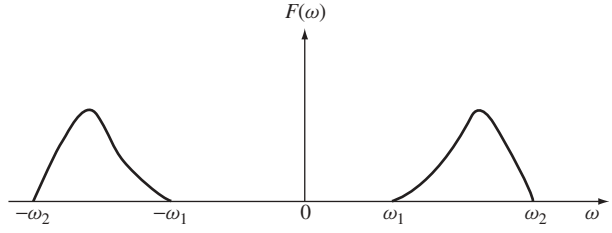


Figure 4.7 The spectrum of a band-pass signal

4.2.1.4 Natural Sampling

Consider the signal $f(t)$, bandlimited to ω_m, and multiply it by the sampling function, $S(t)$, which is a periodic train of rectangular pulses as shown in Figure 4.8.

The rectangular pulse train has a Fourier series given by

$$S(t) = \sum_{k=-\infty}^{\infty} c_k \exp(\mathrm{j}k\omega_0 t) \tag{4.23}$$

where

$$c_k = \frac{\tau}{T} \frac{\sin(k\pi\tau/T)}{(k\pi\tau/T)} \tag{4.24}$$

in which τ is the pulse width and $T = 2\pi/\omega_0$ is the sampling period. The result of multiplying $f(t)$ by $S(t)$ is shown in Figure 4.8 and reveals that the sampled signal consists of a sequence of pulses of width τ and varying amplitudes whose tops also follow the variation of $f(t)$. The sampled signal is given by

$$\begin{aligned} f_s(t) &= f(t)S(t) \\ &= \sum_{k=-\infty}^{\infty} c_k f(t) \exp(\mathrm{j}k\omega_0 t) \end{aligned} \tag{4.25}$$

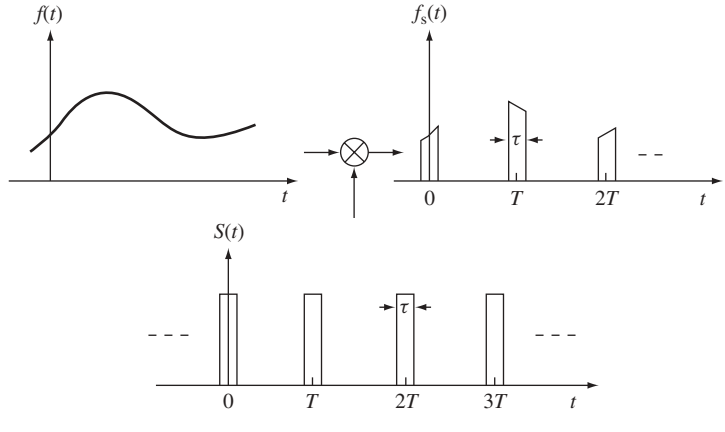


Figure 4.8 Natural sampling

whose Fourier transform is

$$F_s(\omega) = \int_{-\infty}^{\infty} \left(\sum_{k=-\infty}^{\infty} c_k f(t) \exp(\mathrm{j}k\omega_0 t) \exp(-\mathrm{j}\omega t) \right) \mathrm{d}t$$

$$= \sum_{k=-\infty}^{\infty} c_k \left(\int_{-\infty}^{\infty} f(t) \exp[-\mathrm{j}(\omega - k\omega_0)t] \, \mathrm{d}t \right) \tag{4.26}$$

that is

$$F_s(\omega) = \sum_{k=-\infty}^{\infty} c_k F(\omega - k\omega_0). \tag{4.27}$$

The above expression shows that the spectrum of the sampled signal is the same as that obtained with impulse sampling *except* that the displaced side-bands $F(\omega - k\omega_0)$ are variable according to c_k, the Fourier coefficients of the sampling function $S(t)$. The spectra are shown in Figure 4.9. It is clear that for this type of sampling, signal recovery is also possible under precisely the same condition of the sampling theorem obtained in the case of impulse sampling. It is clear that the actual shape of the sampling pulses is immaterial, the only difference is in the Fourier coefficients c_k, and the sampling theorem is always valid.

Nyquist sampling, as discussed above, is the classic technique for band-limited signal representation. More recently [18] it has been shown that a bandlimited signal can also be represented by replacing each variable amplitude Nyquist sample by a variable-width constant-amplitude pulse. Such a use of a two-level waveform to represent a continuous waveform has advantages for power efficient amplification. This is because it eliminates the distortion of amplifiers and obviates the need for precise analog amplitudes; instead all is needed is a precise clock for generating the pulse-widths.

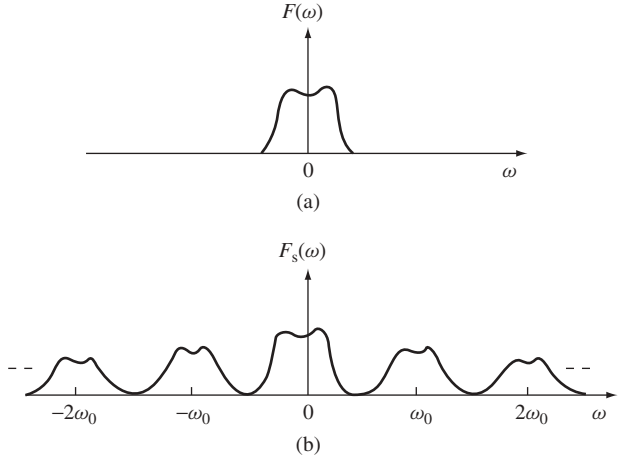


Figure 4.9 (a) The spectrum of a baseband signal. (b) The spectrum of the signal after natural sampling

4.2.2 Quantization and Encoding

After sampling, the next step in analog to digital conversion is *quantization*. This is an approximation of each sample value $f(nT)$ by an integeral multiple of a basic quantity q called the *quantizing step*. This operation is illustrated in Figure 4.10, which shows the input–output relationship of the *quantizer*. When the step size q is constant, the quantization is said to be *uniform*. This is the type considered here.

In Figure 4.10, the quantizer has a *staircase* characteristic, producing the signal $f_q(nT)$. There are two major ways of producing the approximation of $f(nT)$ by $f_q(nT)$. The first is called *rounding*, while the second is *truncation*. The choice of either method defines the centring of the characteristic.

Figure 4.10 represents the case of *rounding*, where each sample value $f(nT)$ lying between $(n - \frac{1}{2})q$ and $(n + \frac{1}{2})q$ is rounded to nq. This method minimizes the power of the error of the approximation as will be shown in a later chapter.

The other method is called *truncation* and consists of approximating by nq any value of $f(nT)$ lying between nq and $(n + 1)q$. To obtain the characteristic of the quantizer in this case, we shift the characteristic of Figure 4.10 to the right by $q/2$.

Naturally, quantization by either of the two methods mentioned above, introduces errors in the representation of the sequence. Since proper sampling introduces no error, the total error in approximating the original signal, after sampling and quantization is due entirely to the quantizer. The study of such errors will be undertaken in a later chapter. It suffices

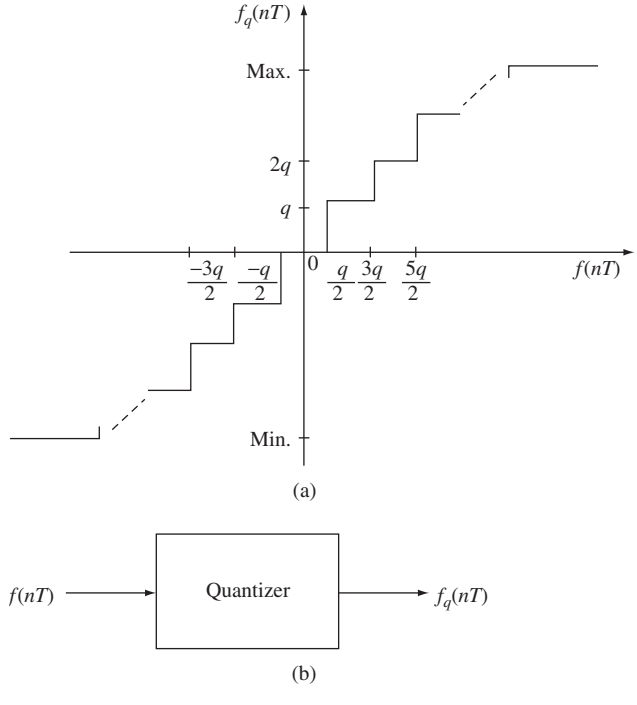


Figure 4.10 Quantization: (a) the transfer characteristic of the quantizer (with rounding), (b) symbolic representation of the process

here to note that the signal $f(nT)$ can be written as the sum of the quantizer output $f_q(nT)$ and an error signal $\varepsilon(nT)$

$$f(nT) = f_q(nT) + \varepsilon(nT) \tag{4.28}$$

Now, the operations of sampling and quantization can be performed in either order, although sampling is usually carried out first. There is, however, no objection to quantization being performed first provided we sample at a frequency ω_s which is somewhat higher than the Nyquist frequency ω_{N}. This is because in this case, we actually sample the signal plus the error, and the error signal can have a spectrum extending beyond the maximum frequency of the error-free signal. Therefore, aliasing will occur if we do not make the appropriate increase in the sampling rate.

The final step in the digitization process is encoding and the representation of the signal by a binary number. This, the reader should be familiar with from an elementary course in binary number systems.

4.3 Discrete Signals and Systems

A discrete-time (or discrete system) is defined as shown in Figure 4.11. Sequences $\{f(t)\}$ or discrete-time signals $\{f(nT)\}$ appear either naturally or as a result of sampling continuous-time signals every T seconds (Figure 4.11) as depicted in Figure 4.12.

The z-transformation plays the same role in the analysis of discrete signals and systems that the Laplace transform does in relation to continuous signals and systems. The one-sided z-transform of a sequence is defined by

$$\begin{aligned} F(z) &= Z\{f(n)\} \\ &\triangleq \sum_{n=0}^{\infty} f(n) z^{-n}. \end{aligned} \tag{4.29}$$

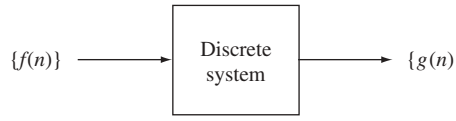


Figure 4.11 A discrete system produces an output sequence due to an input sequence

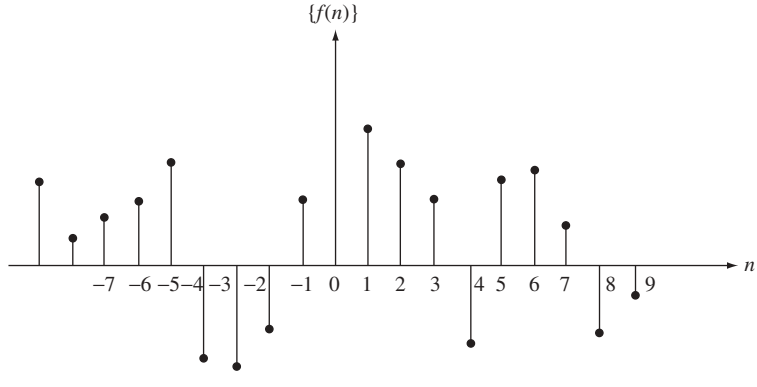


Figure 4.12 Representation of a sequence

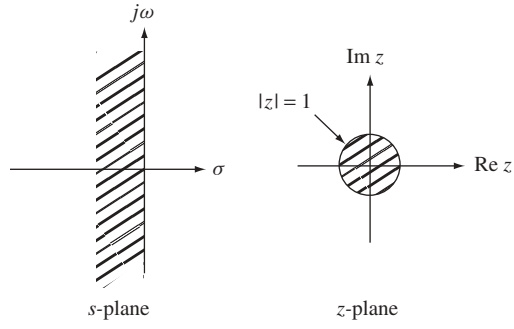


Figure 4.13 Mapping between the s- and z-planes

The Laplace transform and z-transform of a causal sequence are identical, if we make the identification in

$$z^{-1} \equiv \exp(-Ts) \tag{4.30}$$

The correspondence between the complex s-plane and the z-plane is shown in Figure 4.13.

The basic properties of the z-transform include

$$Z[a\{f_1(n)\} + b\{f_2(n)\}] = F_1(z) + F_2(z) \tag{4.31}$$

$$Z\{f(n-m)\} = z^{-m}\sum_{k=0}^{\infty} f(k)z^{-k} + z^{-m}\sum_{k=1}^{m} f(-k)z^{k} \tag{4.32}$$

$$Z\{f(n-m)\} = z^{-m}F(z) + z^{-m}\sum_{k=1}^{m} f(-k)z^{k} \tag{4.33}$$

$$\begin{aligned} Z[\{f_1(n)\} * \{f_2(n)\}] &= F_1(z)F_2(z) \\ &= F_2(z)F_1(z) \end{aligned} \tag{4.34}$$

and it follows that

$$\begin{aligned} Z\left(\sum_{m=0}^{n} f_1(m)f_2(n-m)\right) &= Z\left(\sum_{m=0}^{n} f_1(n-m)f_2(m)\right) \\ &= F_1(z)F_2(z). \end{aligned} \tag{4.35}$$

The inverse z-transform can be obtained using the inversion integral that gives the sequence in terms of its z-transform.

$$f(n) = \frac{1}{2\pi \mathrm{j}}\oint_C F(z)z^{n-1}\,\mathrm{d}z \tag{4.36}$$

The inverse transform of a rational function can also be determined using partial fractions. Consider a rational function $F(z)$ written as

$$F(z) = \frac{P_M(z^{-1})}{D_N(z^{-1})}$$

$$= \frac{P_M(z^{-1})}{\prod_{r=1}^{N}(1 - p_r z^{-1})}. \tag{4.37}$$

Then, obtain the partial fraction expansion of the function as

$$F(z) = \sum_{r=1}^{N} \frac{a_r}{(1 - p_r z^{-1})} \tag{4.38}$$

with the poles at $z = p_r$. Each term in the expansion is, however, the z-transform of a sequence obtained as

$$\begin{aligned} \{f_r(n)\} &= a_r p_r^n \quad \text{for } n \geq 0 \\ &= 0 \qquad \text{for } n < 0. \end{aligned} \tag{4.39}$$

Therefore, the inverse z-transform of the whole function is the sum of such sequences, that is

$$\begin{aligned} \{f(n)\} &= \sum_{r=1}^{N} a_r p_r^n \quad n \geq 0 \\ &= 0 \qquad\qquad n < 0. \end{aligned} \tag{4.40}$$

If two sequences $\{f(n)\}$ and $\{g(n)\}$ have the z-transforms $F(z)$ and $G(z)$, then the complex convolution of $F(z)\,G(z)$ is defined as

$$Q(z) = \frac{1}{2\pi \mathrm{j}} \oint_{c_1} F(v) G\left(\frac{z}{v}\right) \frac{\mathrm{d}v}{v} \tag{4.41a}$$

or

$$Q(z) = \frac{1}{2\pi \mathrm{j}} \oint_{c_2} F\left(\frac{z}{v}\right) G(v) \frac{\mathrm{d}v}{v} \tag{4.41b}$$

which is the z-transform of the product of the two sequences

$$\{q(n)\} = \{f(n)\}\{g(n)\} \tag{4.42}$$

where this is defined as point by point multiplication.

4.4 Digital Filters

An important class of linear shift-invariant systems is the *digital filter* described by a linear difference equation with constant coefficients as

$$g(n) = \sum_{r=0}^{M} a_r f(n - r) - \sum_{r=1}^{N} b_r g(n - r) \text{ with } M \leq N \tag{4.43}$$

Taking the z-transform, the above equation can be converted into an algebraic equation in z, as given by

$$G(z) = F(z)\sum_{r=0}^{M} a_r z^{-1} - G(z)\sum_{r=1}^{N} b_r z^{-r}. \tag{4.44}$$

Thus, a *transfer function* $H(z)$ of the system may be formed as

$$H(z) = \frac{G(z)}{F(z)} \tag{4.45}$$

which upon use of (4.44) takes the general form

$$H(z) = \frac{\sum_{r=0}^{M} a_r z^{-r}}{1 + \sum_{r=1}^{N} b_r z^{-r}}. \tag{4.46}$$

This is a real rational function of z (or z^{-1}) relating the input and output in the z-domain. This equation gives rise to a description of the system in terms of its transfer function that is the ratio of the z-transform of the output to that of the input. It is a real rational function of z that relates the input and output in the z-domain. If the function has only zeros (its denominator is $= 1$), it is of the *finite duration impulse response* (FIR) type. If it has both poles and zeros, it is of the *infinite duration impulse response* (IIR) type.

The frequency response of the system is obtained from the transfer function by letting $z \rightarrow \exp(\mathrm{j}T\omega)$ in the transfer function. The magnitude of the function is the amplitude response whereas its phase is the phase response. Both constitute the frequency response of the system.

A discrete system is stable if all the poles of its transfer function lie inside the unit circle in the z-plane, and this corresponds to the left half-plane in the complex frequency

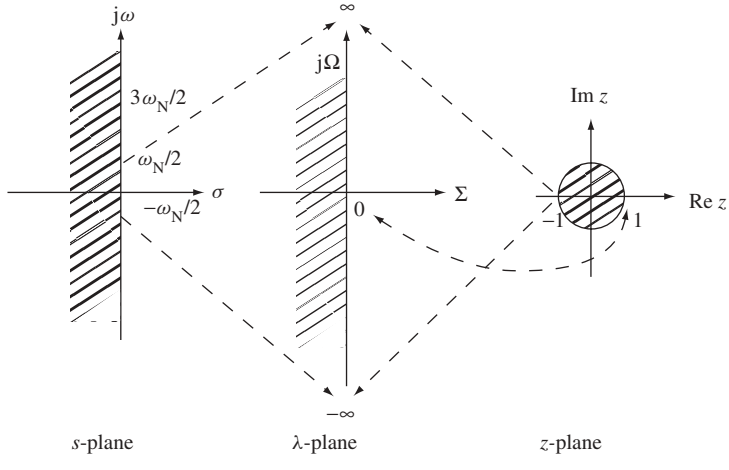


Figure 4.14 Mapping between the three planes of interest

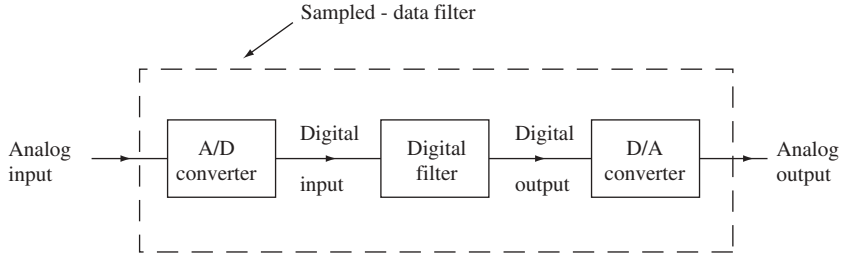


Figure 4.15 The digital filter in an analog environment

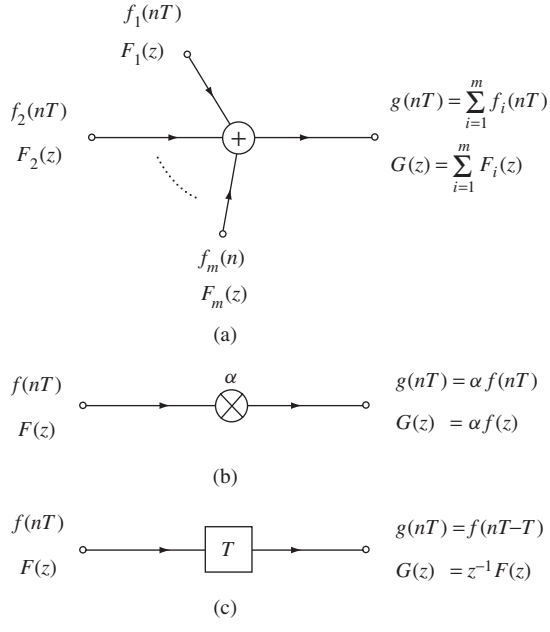


Figure 4.16 The basic building block of digital filters

domain. The mapping between the z-plane and s-plane is shown in Figure 4.14 together with the correspondence with the bilinear variable plane, the latter is particularly useful in filter design and is defined by

$$\lambda = \frac{1 - z^{-1}}{1 + z^{-1}} = \frac{z - 1}{z + 1} \tag{4.47}$$

$$\lambda = \Sigma + \mathrm{j}\Omega \tag{4.48}$$

A digital filter (shown in analog environment in Figure 4.15) uses the basic building blocks shown in Figure 4.16, which are the adder, multiplier and unit delay. These building blocks implement the transfer function of the filter using either software or hardware.

An IIR filter can be realized in direct form as shown in Figure 4.17 or in cascade form or parallel form.

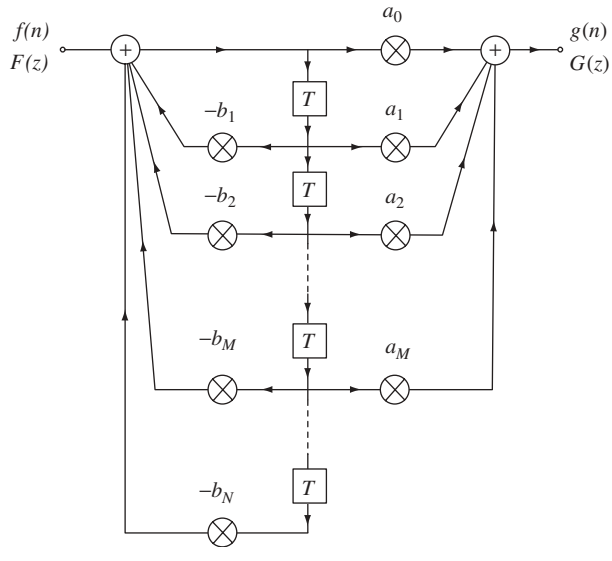


Figure 4.17 Direct realization of an IIR digital filter

For a cascade realization we write

$$H(z) = \prod_k H_k(z) \tag{4.49}$$

where a typical quadratic factor is of the form

$$H_k(z) = \frac{\alpha_{0k} + \alpha_{1k} z^{-1} + \alpha_{2k} z^{-2}}{1 + \beta_{1k} z^{-1} + \beta_{2k} z^{-2}} \tag{4.50}$$

which can be realized as shown in Figure 4.18.

A possible first-order factor is of the form

$$H_k(z) = \frac{\alpha_{0k} + \alpha_{1k} z^{-1}}{1 + \beta_{1k} z^{-1}} \tag{4.51}$$

which can be realized as shown in Figure 4.19. The individual sections are then connected in cascade as shown in Figure 4.20.

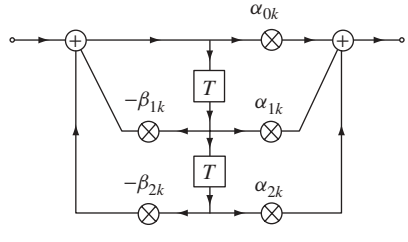


Figure 4.18 Second order section with a transfer function of the form (4.50)

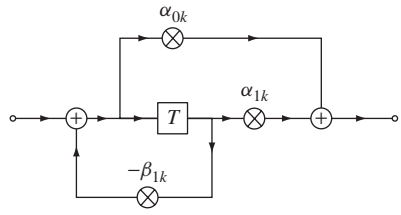


Figure 4.19 First order section with a transfer function of the form (4.51)

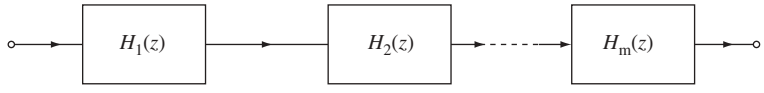


Figure 4.20 Cascade form of realization of a digital transfer function

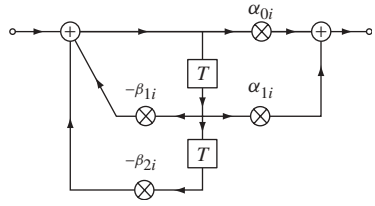


Figure 4.21 Second order section with transfer function in the expansion of (4.52)

Alternatively, the transfer function can be decomposed into the form

$$H(z) = K + \sum_{i=1}^{r-1} \frac{\alpha_{0i} + \alpha_{1i} z^{-1}}{1 + \beta_{1i} z^{-1} + \beta_{2i} z^{-2}} + \frac{\alpha_r}{1 + \beta_r z^{-1}}$$
$$= K + \sum_{i=1}^{r} \hat{H}_i(z) \tag{4.52}$$

Then, each section is realized as shown in Figure 4.21 or Figure 4.22 and the sections are then connected in the parallel form shown in Figure 4.23.

The realization of an FIR transfer function is usually accomplished in the non-recursive form shown in Figure 4.24.

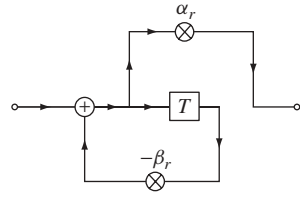


Figure 4.22 First order section with transfer function in the expansion of (4.52)

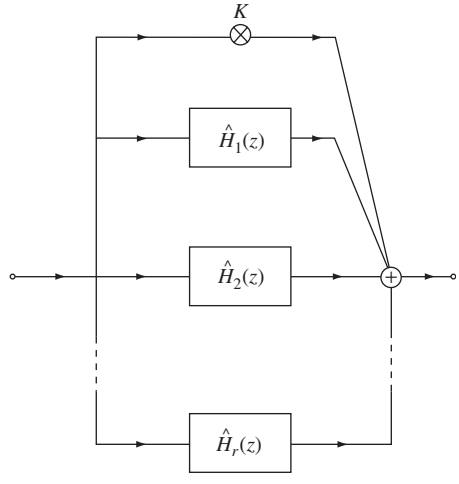


Figure 4.23 Parallel form of realization of a digital transfer function

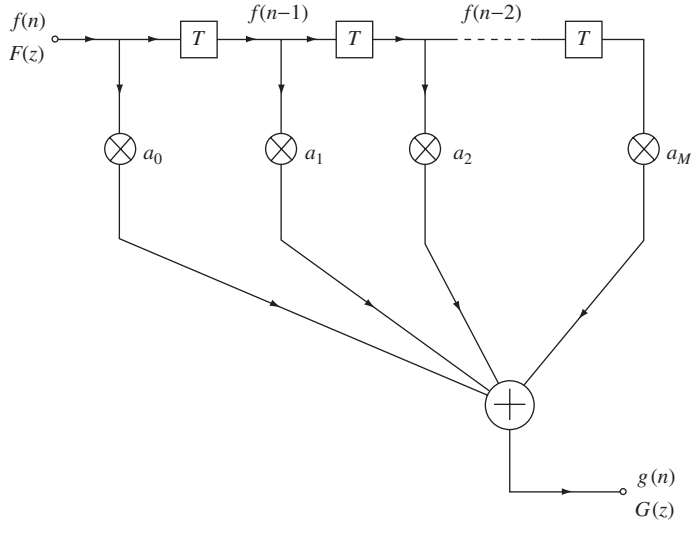


Figure 4.24 Realization of an FIR transfer function

4.5 Conclusion

A review of discrete signals and systems has been given, starting from the process of analog to digital conversion and producing a discrete signal which can be processed by a digital system after quantization and encoding. The z-transformation was used as the

basic analysis tool and the bilinear variable was also discussed due to its importance in the design of digital filters. The class of linear shift-invariant systems was emphasized. This chapter can serve as a concise review for easy reference and to make the treatment in this book self-contained.

Problems

4.1 Find the z-transform of each of the following sequences.
(a) $\{1, 0, 0, 1, 1, 1, 1\}$
(b) $\{1, 1, -1, -1\}$
(c) $\{0, 1, 2, 3, \ldots\} \equiv \{n\}$
(d) $\{0, 1, 4, 9, \ldots\} \equiv \{n^2\}$
(e) $\{1 - e^{-\alpha n}\}$
(f) $\left\{\binom{n}{k}\right\} = \left\{\dfrac{n!}{k!(n-k)!}\right\}$
(g) $\{\sin \alpha n\}$
(h) $\{\cos \alpha n\}$
(i) $\{e^{-\alpha n} \sin \beta n\}$
(j) $\{e^{-\alpha n} \cos \beta n\}$.

4.2 The z-transform can be used to solve difference equations in the same basic manner as the Laplace transform is used to solve differential equations. First, the difference equation is z-transformed and solved for the required variables in the z-domain. Then, the inverse transform is applied to obtain the corresponding sequence. Using this technique, solve the following difference equations.
(a) $f(n) - 4f(n-2) = 0, f(-1) = 0, f(-2) = 2$
(b) $f(n+1) - f(n) = 2n + 3, f(0) = \alpha$
(c) $f(n) + 4f(n-1) + 3f(n-2) = n - 2, f(-1) = f(-2) = 2$.

4.3 Use the z-transform to solve the following set of difference equations

$$3f(n+1) + 2g(n) = 5$$
$$f(n) - g(n+1) = 3$$

with $f(0) = g(0) = 1$.

4.4 Find and plot the discrete convolution of the two sequences

$$\{u_1(n) - u_1(n-4)\} \quad \text{and } \{u_1(n) - u_1(n-9)\}.$$

4.5 Find the inverse z-transform of the following functions, assuming the original sequences to be causal

(a) $\dfrac{z}{z-1}$ (b) $\dfrac{z^2 + 2z}{4z^2 - 5z + 1}$ (c) $\dfrac{1}{z^4(2z-1)}$.

4.6 Write the difference equation describing each of the systems shown in Figure 4.25.

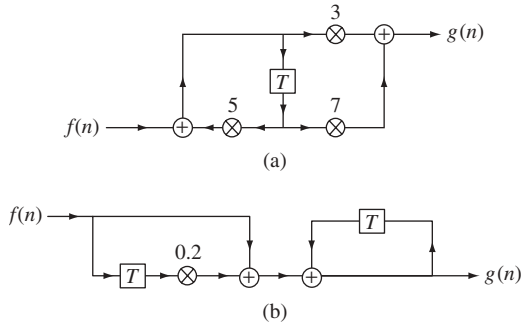


Figure 4.25

4.7 Find the transfer functions, in the z-domain, of the systems of Problem 4.6.

4.8 In each of the following difference equations, $\{f(n)\}$ is the input to a linear shift-invariant system and $\{g(n)\}$ is its output. Obtain the transfer function of each system and test its stability.

(a) $g(n) = 2g(n-1) - g(n-2) + 3f(n) + f(n-1)$

(b) $g(n) = f(n) + 2f(n-1) + g(n-1) + 4g(n-2)$

(c) $g(n) = 0.1f(n) + 0.5f(n-1) - 0.6f(n-2) + 0.3g(n-1) + 0.5g(n-2) + 0.7g(n-3)$.

4.9 Test each of the following transfer functions for stability

(a) $H(z) = \dfrac{z^{-1}(1+z^{-1})}{4 - 2z^{-1} + z^{-2}}$

(b) $H(z) = \dfrac{4z^{-1}(1+z^{-1})}{4 + 3z^{-1} + 2z^{-2} + z^{-3} + z^{-4}}$.

4.10 Find the direct non-recursive realization of the FIR transfer function

$$H(z) = 1 + 2z^{-1} + z^{-2} + 4z^{-3} + 7z^{-4} + 10z^{-5}.$$

4.11 Realize each of the transfer functions of Problem 4.9 in direct canonic form.

4.12 Realize each of the following transfer functions once in direct canonic form, a second time in cascade form and finally in parallel form.

(a) $H(z) = \dfrac{(1+z^{-1})^3}{(1+0.1z^{-1})(1-0.4z^{-1}+0.2z^{-2})}$

(b) $H(z) = \dfrac{(1+z^{-1})^3}{37 + 51z^{-1} + 27z^{-2} + 5z^{-3}}$.

5

Design of Digital Filters

5.1 Introduction

The design techniques of digital filters are discussed in detail. At first, emphasis is laid on the conceptual organization and analytical methods of design. Then the chapter concludes with the details of how to use MATLAB® as a computer-aided design tool. Numerous examples are given throughout the chapter of both analytical and computer-aided design methods.

5.2 General Considerations

As in the case of analog filters studied in Chapter 3, the design of digital filters in the frequency domain can be accomplished according to the following procedure:

(a) The specifications on the filter characteristics are used to obtain a description of the filter in terms of a *stable* transfer function. This is the *approximation problem*.
(b) Once the transfer function of the filter has been determined, it can be realized using the techniques of Chapter 4. This is the *synthesis* or *realization problem* whose solution requires only knowledge of the coefficients of the transfer function.

The transfer function of a general IIR filters is of the form

$$H(z) = \frac{\sum_{n=0}^{M} a_n z^{-n}}{1 + \sum_{n=1}^{N} b_n z^{-n}} = \frac{A_M(z^{-1})}{B_N(z^{-1})} \tag{5.1}$$

and the special case of an FIR filter transfer function is obtained by putting $B_N(z^{-1}) = 1$ so that an FIR digital transfer function is a polynomial of the form

$$H(z) = \sum_{n=0}^{M} a_n z^{-n} \tag{5.2}$$

Signal Processing and Integrated Circuits, First Edition. Hussein Baher.
© 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.

The frequency response of the filter is obtained by letting

$$z \rightarrow \exp(\mathrm{j}\omega T) \tag{5.3}$$

so that

$$H\,(\exp(\mathrm{j}\omega T)) = \frac{\sum_{n=0}^{M} a_n \exp(-\mathrm{j}n\omega T)}{1 + \sum_{n=1}^{N} b_n \exp(-\mathrm{j}n\omega T)} \tag{5.4}$$

which may be put in the form

$$H\,(\exp(\mathrm{j}\omega T)) = |H\,(\exp(j\omega T))|\exp[\mathrm{j}\psi(\omega T)] \tag{5.5}$$

where $|H\,(\exp(\mathrm{j}\omega T))|$ is the amplitude response and $\psi(\omega T)$ is the phase response of the filter.

Clearly

$$|H\,(\exp(\mathrm{j}\omega T))|^2 = H\,(z)H\,(z^{-1})|_{z=\exp(j\omega T)} \tag{5.6}$$

We can obtain expressions for the phase function $\psi(\omega T)$ and group delay $T_{\mathrm{g}}(\omega T)$ in terms of $H\,(z)$. Taking logarithms of both sides of (5.4) we have

$$\begin{aligned}\ln H\,(\exp(\mathrm{j}\omega T)) &= \ln|H\,(\exp(\mathrm{j}\omega T))| + \mathrm{j}\psi(\omega T)\\ &= \tfrac{1}{2}\ln[H\,(\exp(\mathrm{j}\omega T))H\,(\exp(-\mathrm{j}\omega T))] + \mathrm{j}\psi(\omega T)\end{aligned} \tag{5.7}$$

so that if we let

$$\psi(z) \triangleq -\frac{1}{2}\ln\left(\frac{H\,(z)}{H\,(z^{-1})}\right) \tag{5.8}$$

then

$$\psi(\omega T) = -\mathrm{j}\psi(z)|_{z=\exp(\mathrm{j}\omega T)} \tag{5.9}$$

The group-delay is given by

$$\begin{aligned}T_{\mathrm{g}}(\omega T) &= -\frac{\mathrm{d}\psi(\omega T)}{\mathrm{d}\omega}\\ &= \mathrm{j}\frac{\mathrm{d}\psi(z)}{\mathrm{d}z}\bigg|_{z=\exp(\mathrm{j}\omega T)} \cdot \frac{\mathrm{d}\exp(\mathrm{j}\omega T)}{\mathrm{d}\omega}\\ &= T\left(z\frac{\mathrm{d}\psi(z)}{\mathrm{d}z}\right)_{\exp(\mathrm{j}\omega T)}\end{aligned} \tag{5.10}$$

However, (5.8) gives

$$\begin{aligned}-\frac{\mathrm{d}\psi(z)}{\mathrm{d}z} &= \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}z}\left(\ln\frac{H\,(z)}{H\,(z^{-1})}\right)\\ &= \frac{1}{2}\left(\frac{H'(z)}{H\,(z)} + \frac{1}{z^2}\frac{H'(z^{-1})}{H\,(z^{-1})}\right)\end{aligned} \tag{5.11}$$

so that the group-delay in (5.10) becomes

$$T_g(\omega T) = \frac{T}{2}\left(z\frac{H'(z)}{H(z)} + z^{-1}\frac{H'(z^{-1})}{H(z^{-1})}\right)_{z=\exp(j\omega T)} \tag{5.12}$$

or

$$\begin{aligned} T_g(\omega T) &= T\ \mathrm{Re}\ \left(z\frac{H'(z)}{H(z)}\right)_{z=\exp(j\omega T)} \\ &= T\ \mathrm{Re}\ \left(z\frac{\mathrm{d}}{\mathrm{d}z}\ln H(z)\right)_{z=\exp(j\omega T)} \end{aligned} \tag{5.13}$$

It follows that expressions (5.4)–(5.13) can be used for the analysis of any digital filter by evaluating its amplitude, phase and delay responses.

Now, for the realization of a digital filter, its transfer function is expressed in terms of the variable z, as given by (5.1). However, for the solution of the approximation problem, it is much more convenient to employ the variable

$$\begin{aligned} \lambda &= \tanh \tfrac{1}{2}Ts \\ &= \Sigma + j\Omega \end{aligned} \tag{5.14}$$

which was introduced in Chapter 4. It is related to z by

$$\lambda = \frac{1 - z^{-1}}{1 + z^{-1}} \tag{5.15a}$$

or

$$z^{-1} = \frac{1 - \lambda}{1 + \lambda} \tag{5.15b}$$

Therefore a digital transfer function of the form (5.1) can be transformed using (5.15b) into the equivalent form

$$H(\lambda) = \frac{P_m(\lambda)}{Q_n(\lambda)} \quad m \leq n \tag{5.16}$$

where, for stability, $Q_n(\lambda)$ must be strictly Hurwitz in λ.

The mapping between the s-, z- and λ-planes was discussed in detail in Chapter 4 and is shown in Figure 4.14. The main advantage of using the λ-variable is that one can obtain the *amplitude-oriented* digital filter design directly by transforming the prototype function of the analog filters discussed in Chapter 3.

In terms of the *bilinear variable* λ, the frequency response of the filter is obtained by letting $s \to j\omega$ so that

$$\lambda \to j\Omega \tag{5.17}$$

where

$$\begin{aligned} \Omega &= \tan \frac{1}{2}T\omega \\ &= \tan \pi \frac{\omega}{\omega_N} \end{aligned} \tag{5.18}$$

and ω_N is the radian sampling frequency. If *critical sampling* is assumed then ω_N is the radian *Nyquist* frequency, which is chosen as twice the highest frequency of the band of interest. On the jω-axis, (5.16) becomes

$$\begin{aligned} H(\mathrm{j}\Omega) &= \frac{P_m(\mathrm{j}\Omega)}{Q_n(\mathrm{j}\Omega)} \\ &= |H(\mathrm{j}\Omega)| \exp(\mathrm{j}\psi(\Omega)). \end{aligned} \tag{5.19}$$

The group delay is

$$\begin{aligned} T_g(\omega T) &= -\left.\frac{\mathrm{d}\psi(\lambda)}{\mathrm{d}s}\right|_{s=\mathrm{j}\omega} \\ &= -\left.\frac{\mathrm{d}\psi(\lambda)}{\mathrm{d}\lambda}\frac{\mathrm{d}\lambda}{\mathrm{d}s}\right|_{s=\mathrm{j}\omega} \\ &= -\left.\frac{T}{2}(1-\lambda^2)\frac{\mathrm{d}\psi(\lambda)}{\mathrm{d}\lambda}\right|_{s=\mathrm{j}\omega} \end{aligned} \tag{5.20}$$

or

$$T_g(\Omega) = -\left.\frac{T}{2}\mathrm{Re}(1+\Omega^2)\frac{\mathrm{d}\psi(\lambda)}{\mathrm{d}\lambda}\right|_{\lambda=\mathrm{j}\Omega} \tag{5.21}$$

which can be expressed as

$$T_g(\Omega) = \frac{T}{2}(1+\Omega^2)\mathrm{Re}\left[\left(\frac{Q_n'(\lambda)}{Q_n(\lambda)} - \frac{P_m'(\lambda)}{P_m(\lambda)}\right)_{\lambda=\mathrm{j}\Omega}\right] \tag{5.22}$$

Having disposed of the above preliminaries, we now discuss the approximation problem for both IIR and FIR filters. The amplitude-oriented design and phase-oriented design are covered, but filters with simultaneous amplitude selectivity and passband phase linearity are discussed only for the FIR case in which phase linearity can be imposed in a simple manner.

5.3 Amplitude-oriented Design of IIR Filters

5.3.1 Low-pass Filters

The magnitude-squared function of an IIR digital filter can be written from (5.16), (5.17) as

$$\begin{aligned} |H(\mathrm{j}\Omega)|^2 &= \frac{|P_m(\mathrm{j}\Omega)|^2}{|Q_n(\mathrm{j}\Omega)|^2} \\ &= \frac{\sum_{r=0}^{m} c_r \Omega^{2r}}{\sum_{r=0}^{n} d_r \Omega^{2r}} \end{aligned} \tag{5.23}$$

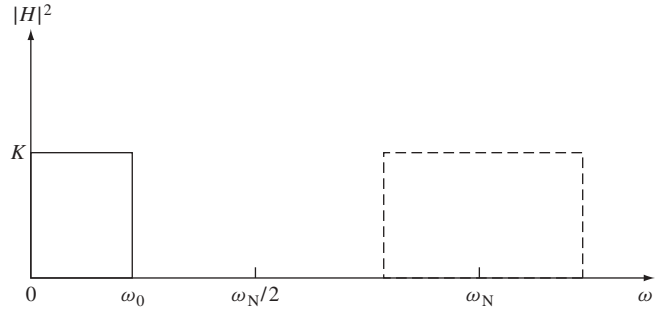


Figure 5.1 The ideal low-pass digital filter characteristic

which, as explained in Chapter 4, is periodic in ω due to the periodicity of the variable $\Omega = \tan(\pi\omega/\omega_N)$. Thus

$$-\infty \leq \Omega \leq \infty \quad \text{for} \quad \frac{-(2r+1)}{2} \leq \frac{\omega}{\omega_N} \leq \frac{2r+1}{2} \quad r = 0, 1, 2, \ldots \tag{5.24}$$

However, the useful band is limited by the sampling theorem to the range

$$0 \leq \frac{|\omega|}{\omega_N} \leq 0.5 \tag{5.25}$$

The ideal low-pass amplitude characteristic is shown in Figure 5.1 where the dotted lines represent the periodic nature of the *hypothetical* response. However, the frequencies above $\omega_N/2$ are excluded if the input signal is bandlimited to below $\omega_N/2$. If not, aliasing will occur.

Now, consider an analog filter low-pass prototype transfer function

$$H(s) = \frac{N_m(s)}{D_n(s)} m \leq n \tag{5.26}$$

whose amplitude response

$$|H(\mathrm{j}\omega)|^2 = \frac{|N_m(\mathrm{j}\omega)|^2}{|D_n(\mathrm{j}\omega)|^2} \tag{5.27}$$

is obtained subject to a certain optimality criterion, with passband edge at $\omega = 1$. In the analog function, let

$$s \rightarrow \frac{\lambda}{\Omega_0} \rightarrow \frac{1}{\Omega_0} \frac{1 - z^{-1}}{1 + z^{-1}} \tag{5.28}$$

where

$$\Omega_0 = \tan \pi \frac{\omega_0}{\omega_N} \tag{5.29}$$

and ω_0 is the actual passband edge of the required digital filter. Expression (5.28) transforms the analog function into a digital one of the form

$$H(\lambda) = \frac{P_m(\lambda)}{Q_n(\lambda)} \tag{5.30}$$

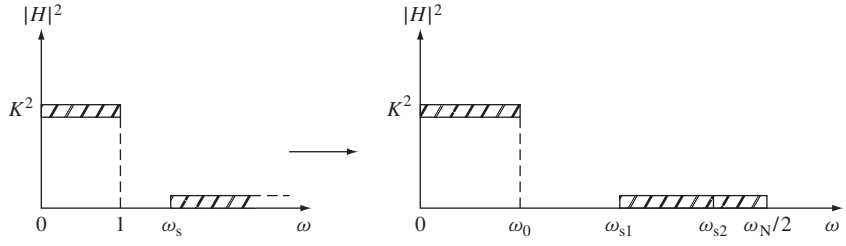


Figure 5.2 Application of the bilinear transformation to (a) the analog low-pass filter specifications, to obtain (b) the digital domain specifications

with amplitude response defined by

$$|H(\mathrm{j}\Omega)|^2 = \frac{|P_m(\mathrm{j}\Omega)|^2}{|Q_n(\mathrm{j}\Omega)|^2}. \tag{5.31}$$

The so-called *bilinear transformation* defined by (5.28) has the following basic features:

(i) The point $\omega = 1$ in the analog prototype is transformed to ω_0 in the digital domain. The point $\omega = 0$ is transformed to the same point $\omega = 0$, while $\omega = \infty$ is transformed to $\omega = \omega_N/2$ since $\Omega = \infty$ corresponds to half the sampling frequency.
(ii) The stability of the analog function is preserved under the transformation, since the resulting denominator $Q_n(\lambda)$ is guaranteed strictly Hurwitz in λ by virtue of (5.28) and stability is assured.
(iii) The properties of the analog filter along the jω-axis are transformed into analogous properties on the jω-axis, which corresponds to the unit circle in the z-plane. However, periodicity applies in this case.
(iv) The resulting transfer function is rational in z^{-1} with real coefficients.

Figure 5.2 shows the transformation of the specifications of an analog low-pass prototype response to the digital domain. Due to the mapping

$$\begin{aligned} &\omega = 0 \rightarrow \omega = 0 \\ &\omega = 1 \rightarrow \omega = \omega_0 \\ &\omega = \infty \rightarrow \omega = \omega_N/2 \end{aligned} \tag{5.32}$$

the stopband now extends from some specified frequency ω_{s1} to half the sampling frequency $\omega_N/2$. It is assumed that the input to the filter is band-limited to $\omega_N/2$ as required by the sampling theorem.

It follows from the above discussion that we can derive the digital counterparts of the maximally flat, Chebyshev or elliptic filters by the transformation in (5.28). Thus, we have the following cases.

5.3.1.1 Maximally Flat Response in Both Bands

Letting $s \rightarrow \lambda/\Omega_0$ in (3.31) we obtain

$$H(s) \rightarrow H(\lambda) = \frac{K}{\prod_{r=1}^{n}\left(\frac{\lambda}{\Omega_0} - \mathrm{j}\,\exp(\mathrm{j}\theta r)\right)} \tag{5.33}$$

where

$$\theta_r = \frac{(2r-1)}{2n}\pi \quad r = 1, 2, \ldots n \tag{5.34}$$

and K is chosen such that the dc-gain is any prescribed value; $K = 1$ for $H(0) = 1$. ω_0 is the point at which the gain falls by 3 dB. In the z-domain, use of (5.28) in (5.33) yields

$$H(z) = \frac{K(1+z^{-1})^n}{\prod_{r=1}^{n}\left[\left(\frac{1}{\Omega_0} - \mathrm{j}\,\exp(\mathrm{j}\theta r)\right) - \left(\frac{1}{\Omega_0} + \mathrm{j}\,\exp(\mathrm{j}\theta r)\right)z^{-1}\right]} \tag{5.35}$$

which can be realized by the techniques of Chapter 4. The resulting response is obtained by letting

$$\omega \rightarrow \Omega/\Omega_0 \tag{5.36}$$

This gives

$$|H(\mathrm{j}\Omega)|^2 = \frac{K^2}{1 + (\Omega/\Omega_0)^{2n}} \tag{5.37}$$

which has $(2n-1)$ zero derivatives around $\Omega = 0$ corresponding to $2n-1$ zero derivatives around $\omega = 0$. It also possesses $(2n-1)$ zero derivatives around $\Omega = \infty$ corresponding to $(2n-1)$ zero derivatives around $\omega_N/2$.

In order to determine the required degree of the filter, let the specifications be given as

$$\begin{aligned} &\text{Passband } 0 \le \omega \le \omega_0, \quad \text{attenuation } \le 3\,\text{dB}. \\ &\text{Stopband } \omega_{s1} \le \omega \le \omega_{s2}, \quad \text{attenuation } \ge \alpha_s\,\text{dB}. \end{aligned} \tag{5.38}$$

We note that the stopband must be finite in the digital domain, ending at a frequency less than half the sampling frequency. We begin by choosing a sampling frequency of at least twice the highest frequency of the band of interest. Thus, we must take

$$\omega_N \ge 2\,\omega_{s2}. \tag{5.39}$$

Let us assume that critical sampling is employed, so that (5.39) is satisfied with equality

$$\omega_N = 2\,\omega_{s2}. \tag{5.40}$$

Then, the passband is defined by

$$0 \leq \frac{\omega}{\omega_N} \leq \frac{\omega_0}{\omega_N} \tag{5.41}$$

and the stopband is in the range

$$\frac{\omega_{s1}}{\omega_N} \leq \frac{\omega}{\omega_N} \leq 0.5. \tag{5.42}$$

Next, the corresponding Ω-values are obtained to define the bands in the Ω-domain as

$$\begin{aligned} &\text{Passband } 0 \leq \Omega \leq \Omega_0 \\ &\text{Stopband } \Omega_{s1} \leq \Omega \leq \Omega_{s2}(=\infty) \end{aligned} \tag{5.43}$$

where

$$\Omega_0 = \tan \pi \frac{\omega_0}{\omega_N} \tag{5.44a}$$

$$\Omega_{s1} = \tan \pi \frac{\omega_{s1}}{\omega_N} \tag{5.44b}$$

$$\Omega_{s2} = \tan \pi \frac{\omega_{s2}}{\omega_N} = \infty. \tag{5.44c}$$

Of course, if the sampling frequency is chosen higher than $2\omega_{s2}$, then the stopband is defined by

$$\Omega_{s1} \leq \Omega \leq \Omega_{s2} \tag{5.45}$$

where, instead of (5.44c) we have

$$\Omega_{s2} = \tan \pi \frac{\omega_{s2}}{\omega_N} \neq \infty. \tag{5.46}$$

Finally, the degree of the required filter is obtained by substitution of (5.44) in (3.20) with $\omega_s \rightarrow \Omega_{s1}/\Omega_0$. This gives

$$n \geq \frac{\log(10^{0.1\alpha_s} - 1)}{2\log(\Omega_{s1}/\Omega_0)}. \tag{5.47}$$

An alternative format for the specifications may be given as

$$\begin{aligned} &\text{Maximum passband attenuation} = \alpha_p \quad \omega \leq \omega_0. \\ &\text{Minimum stopband attenuation} = \alpha_s \quad \omega \geq \omega_{s1}. \end{aligned} \tag{5.48}$$

In this case (5.47) may be used to determine the required degree after transformation to the Ω-domain. This gives

$$n \geq \frac{\log\left(\dfrac{10^{0.1\alpha_s} - 1}{10^{0.1\alpha_p} - 1}\right)}{2\log(\Omega_{s1}/\Omega_0)}. \tag{5.49}$$

Example 5.1 Design an IIR maximally-flat digital filter with the following specifications:

Passband: 0–1 kHz with attenuation $\leq$1 dB.
Stopband: 2–4 kHz with attenuation $\geq$20 dB.

Solution. Assuming critical sampling, the highest frequency in the band of interest is 4 kHz so that we may choose a sampling frequency of twice this value to give

$$\omega_{\mathrm{N}} = (2\pi)8 \times 10^3.$$

Thus, in the Ω-domain we have

$$\Omega_0 = \tan \tfrac{1}{8}\pi = 0.4142$$
$$\Omega_{\mathrm{s1}} = \tan \tfrac{2}{8}\pi = 1.0.$$

Using (5.49) the required filter degree is

$$n \geq 3.373$$

so that we may take

$$n = 4.$$

The filter transfer function is obtained from n and (5.35), and the realization is obtained in direct, parallel or cascade form, as discussed in Chapter 4.

5.3.1.2 Chebyshev Response

Letting $s \to \lambda/\Omega_0$ in the denominator of (3.36) we obtain

$$H(s) \to H(\lambda) = \frac{K}{\prod_{r=1}^{n}\left(\dfrac{\lambda}{\Omega_0} + [\eta \sin\theta_r + \mathrm{j}(1+\eta^2)^{1/2}\cos\theta_r]\right)} \tag{5.50}$$

where θ_r is given by (5.34) and

$$\eta = \sinh\left(\frac{1}{n}\sinh^{-1}\frac{1}{\varepsilon}\right). \tag{5.51}$$

The numerator of $H(\lambda)$ is taken to be an arbitrary constant K, since the digital filter gain can be adjusted arbitrarily, for example at the A/D converter stage, and is not restricted to have a maximum value of unity as in the case of passive filters. Again, the z-domain representation of the Chebyshev filter is obtained as

$$H(z) = \frac{K(1+z^{-1})^n}{\prod_{r=1}^{n}\left[\left(\dfrac{1}{\Omega_0} + \mathrm{j}y_r\right) - \left(\dfrac{1}{\Omega_0} - \mathrm{j}y_r\right)z^{-1}\right]} \tag{5.52}$$

where

$$y_r = \cos(\sin^{-1}\mathrm{j}\eta + \theta_r). \tag{5.53}$$

Evidently, the resulting transfer function has a magnitude squared function given by (3.25) with $\omega \to (\Omega/\Omega_0)$, that is

$$|H(\mathrm{j}\Omega)|^2 = \frac{K^2}{1 + \varepsilon^2 T_n^2(\Omega/\Omega_0)} \tag{5.54}$$

where $T_n(\Omega/\Omega_0)$ is the Chebyshev polynomial defined by (3.26), (3.27) and ε is the ripple factor. $|H(j\Omega)|^2$ has an optimum equiripple response in the passband and $(2n-1)$ zero derivatives around $\Omega = \infty$, that is around $\omega_N/2$.

To determine the required degree of the filter, the specifications are transformed to the Ω-domain as in (5.44), then use is made of (3.24). Let the specifications by given as

$$\begin{aligned} &\text{Passband } 0 \le \omega \le \omega_0, \quad &&\text{attenuation} \le \alpha_p. \\ &\text{Stopband } \omega_{s1} \le \omega \le \omega_{s2}, \quad &&\text{attenuation} \ge \alpha_s. \end{aligned} \tag{5.55}$$

Choosing $\omega_N \ge 2\omega_{s2}$, and evaluating Ω_0 and Ω_{s1} from (5.44), we then use (3.24) to obtain for the degree

$$n \ge \frac{\cosh^{-1}[(10^{0.1\alpha_s}-1)/(10^{0.1\alpha_p}-1)]^{1/2}}{\cosh^{-1}(\Omega_{s1}/\Omega_0)} \tag{5.56}$$

Example 5.2 Design a low-pass Chebyshev digital filter with the following specifications:

Passband: 0–0.5 kHz with 0.1 dB ripple.
Stopband edge: 0.7 kHz with attenuation ≥40 dB.
Sampling frequency: 2 kHz.

Solution. From (5.44) the Ω-domain values are

$$\Omega_0 = \tan\pi\frac{0.5}{2} = 1$$

$$\Omega_{s1} = \tan\pi\frac{0.7}{2} = 1.963$$

and (5.56) gives the required degree as $n = 6$. Also

$$\varepsilon = (10^{0.1\alpha_p}-1)^{1/2} = (10^{0.01}-1)^{1/2} = 0.1526$$

so that the auxiliary parameter is obtained from (5.51) as

$$\eta = \sinh\left(\frac{1}{6}\sinh^{-1}\frac{1}{0.1526}\right) = 0.443.$$

Finally the filter transfer function is obtained from (5.51)–(5.53) as

$$H(z) = H_1(z)H_2(z)H_3(z)$$

with

$$H_1(z) = \frac{0.426 + 0.851z^{-1} + 0.426z^{-2}}{1 + 0.103z^{-1} + 0.805z^{-2}}$$

$$H_2(z) = \frac{0.431 + 0.863z^{-1} + 0.431z^{-2}}{1 - 0.266z^{-1} + 0.459z^{-2}}$$

$$H_3(z) = \frac{0.472 + 0.944z^{-1} + 0.472z^{-2}}{1 - 0.696z^{-1} + 0.192z^{-2}}$$

and for a possible realization, second-order sections of the form shown in Figure 4.18 are used for $H_1(z)$, $H_2(z)$ and $H_3(z)$, then the results are connected in cascade as shown in Figure 4.20.

5.3.1.3 Elliptic Function Response

The low-pass digital elliptic transfer function can be obtained using the same transformation (5.28) and the extensive expressions available for the analog elliptic transfer functions [16]. The procedure is illustrated by the following example.

Example 5.3 The transfer function of a third-order low-pass elliptic analog filter is given by

$$H(s) = \frac{0.314(s^2 + 2.806)}{(s + 0.767)(s^2 + 0.453s + 1.149)}$$

which gives 0.5 dB passband ripple and a minimum stopband attenuation of 21 dB for $\omega_s/\omega_0 \geq 1.5$. Use this prototype function to design a digital filter with passband edge at 500 Hz and a sampling frequency of 3 kHz.

Solution. From the specifications and (5.44)

$$\Omega_0 = \tan \pi \frac{500}{3000} = 0.577.$$

Thus, using the bilinear transformation the digitized transfer function becomes

$$\begin{aligned} H(z) &= \left(\frac{0.126 + 0.126z^{-1}}{1 - 0.386z^{-1}}\right)\left(\frac{1.177 - 0.079z^{-1} + 1.1778z^{-2}}{1 - 0.75z^{-1} + 0.682z^{-2}}\right) \\ &= H_1(z)H_2(z) \end{aligned}$$

which can be realized as a cascade of a first-order section with transfer function $H_1(z)$ of the form shown in Figure 4.19 and a second-order section with transfer function $H_2(z)$ of the form shown in Figure 4.18.

5.3.2 High-pass Filters

The analog low-pass prototype transfer function of Section 3.3 can also be used to derive high-pass digital transfer functions. This can be achieved via the transformation

$$s \rightarrow \frac{\Omega_0}{\lambda} \rightarrow \Omega_0 \left(\frac{1 + z^{-1}}{1 - z^{-1}}\right) \tag{5.57}$$

where

$$\Omega_0 = \tan \pi \frac{\omega_0}{\omega_N} \tag{5.58}$$

in which ω_0 is the high-pass filter passband edge. The transformation is illustrated in Figure 5.3. The point $\omega = 0$ is transformed to $\omega_N/2$ and $\omega = \infty$ is transformed to $\omega = 0$.

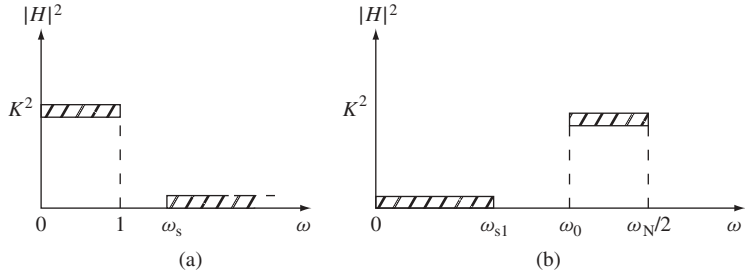


Figure 5.3 Application of the transformation (5.57) to obtain a high-pass digital transfer function. (a) analog low-pass specifications, (b) digital band-pass specifications

The passband now occupies the range

$$\omega_0 \leq |\omega| \leq \omega_N/2. \tag{5.59}$$

The resulting transfer functions are obtained by means of the transformation in (5.57) and any of the expressions in Section 3.3. The maximally flat and Chebyshev amplitude functions are now

$$|H(\mathrm{j}\Omega)|^2 = \frac{K^2}{1 + (\Omega_0/\Omega)^{2n}} \tag{5.60}$$

and

$$|H(\mathrm{j}\Omega)|^2 = \frac{K^2}{1 + \varepsilon^2 T_n^2(\Omega_0/\Omega)}. \tag{5.61}$$

To obtain the required degree of the filter, let the specifications be given as

$$\begin{aligned} &\text{Passband } \omega_0 \leq \omega \leq \omega_N/2, \quad \text{attenuation} \leq \alpha_p \text{ dB} \\ &\text{Stopband } 0 \leq \omega \leq \omega_{s1}, \qquad \text{attenuation} \geq \alpha_s \text{ dB} \end{aligned} \tag{5.62}$$

with

$$\begin{aligned} \Omega_0 &= \tan \pi \frac{\omega_0}{\omega_N} \\ \Omega_{s1} &= \tan \pi \frac{\omega_{s1}}{\omega_N}. \end{aligned} \tag{5.63}$$

The required degree of the maximally flat filter is obtained from (5.48) by letting $(\Omega_s/\Omega_0) \rightarrow \Omega_0/\Omega_s)$ to give

$$n \geq \frac{\log[(10^{0.1\alpha_s} - 1)/(10^{0.1\alpha_p} - 1)]}{2\log(\Omega_0/\Omega_{s1})}. \tag{5.64}$$

Alternatively, if the passband edge is the 3 dB-point then (5.64) gives

$$n \geq \frac{\log(10^{0.1\alpha_s} - 1)}{2\log(\Omega_0/\Omega_{s1})}. \tag{5.65}$$

For the Chebyshev response, the required degree is

$$n \geq \frac{\cosh^{-1}[(10^{0.1\alpha_s} - 1)/(10^{0.1\alpha_p} - 1)]^{1/2}}{\cosh^{-1}(\Omega_0/\Omega_{s1})}. \tag{5.66}$$

Having obtained the required degree, the analog low-pass prototype function is obtained, which upon use of the transformation (5.57) gives the digital high-pass transfer function.

5.3.3 Band-pass Filters

Again, the analog low-pass prototype transfer functions can be used to obtain digital band-pass functions by means of a transformation given by

$$s \rightarrow \frac{\overline{\Omega}}{\Omega_2 - \Omega_1}\left(\frac{\lambda}{\overline{\Omega}} + \frac{\overline{\Omega}}{\lambda}\right) \tag{5.67}$$

where

$$\overline{\Omega} = \tan \pi \frac{\overline{\omega}}{\omega_N} \tag{5.68a}$$

$$\Omega_{1,2} = \tan \pi \frac{\omega_{1,2}}{\omega_N} \tag{5.68b}$$

$$\overline{\Omega} = (\Omega_1\, \Omega_2)^{1/2}. \tag{5.68c}$$

This transformation is illustrated in Figure 5.4. The band centre $\overline{\omega}$ is arbitrary, but the end of the upper stopband is at most equal to $\omega_N/2$. The passband extends from ω_1 to ω_2. Expression (5.15) may be used to transform to the z-domain for the purpose of realization of the transfer function.

The resulting maximally flat and Chebyshev responses are given, respectively, by

$$|H(j\Omega)|^2 = \frac{K^2}{1 + \left[\dfrac{\overline{\Omega}}{\Omega_2 - \Omega_1}\left(\dfrac{\Omega}{\overline{\Omega}} - \dfrac{\overline{\Omega}}{\Omega}\right)\right]^{2n}} \tag{5.69}$$

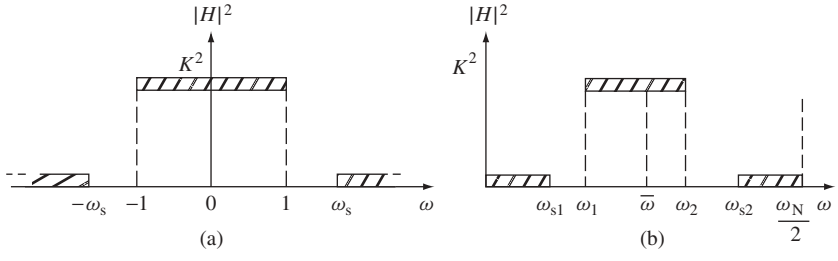


Figure 5.4 Application of the transformation (5.67) to obtain a digital band-pass transfer function (a) analog low-pass specifications. (b) digital high-pass specifications

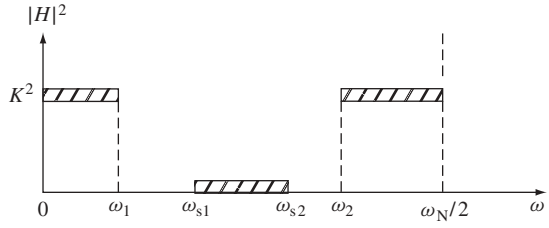


Figure 5.5 Band-stop digital filter specifications obtained from low-pass prototype specifications by the transformation (5.72)

and

$$|H(\mathrm{j}\Omega)|^2 = \frac{K^2}{1 + \varepsilon^2 T_n^2 \left[\dfrac{\overline{\Omega}}{\Omega_2 - \Omega_1} \left(\dfrac{\Omega}{\overline{\Omega}} - \dfrac{\overline{\Omega}}{\Omega} \right) \right]^{2n}} \tag{5.70}$$

The actual filter transfer functions are obtained by the transformation (5.15) together with (5.67) in the appropriate analog functions given, for example, by (3.20) and (3.36).

The required degree of the prototype filter is obtained by substituting in (5.69) or (5.70) for the required attenuation values at the given frequencies and forming the necessary equations. In particular, for the maximally flat response, the two 3 dB points occur at ω_1 and ω_2. Furthermore, there are two frequencies ω_{s1} and ω_{s2} at which the attenuation is the same value α_s, the two frequencies being related by

$$\overline{\Omega} = \left[\left(\tan \pi \frac{\omega_{s1}}{\omega_N} \right) \left(\tan \pi \frac{\omega_{s2}}{\omega_N} \right) \right]^{1/2}. \tag{5.71}$$

The filter is designed according to the more severe of the two requirements, and the response has geometric symmetry about $\overline{\Omega}$ in the Ω domain. These consideration also apply to Chebyshev and elliptic filters.

5.3.4 Band-stop Filters

These can be obtained from an analog prototype low-pass function by the transformation (illustrated in Figure 5.5)

$$s \to \frac{(\Omega_2 - \Omega_1)}{\overline{\Omega}} \left(\frac{\lambda}{\overline{\Omega}} + \frac{\overline{\Omega}}{\lambda} \right)^{-1} \tag{5.72}$$

where $\overline{\Omega}$ and $\Omega_{1,2}$ are also given by (5.68) but in this case ω_1 is the lower passband edge and ω_2 is the upper passband edge. Again, the resulting response exhibits geometric symmetry about $\overline{\Omega}$, so that (5.71) and the preceding discussion are also valid in this case.

5.4 Phase-oriented Design of IIR Filters

5.4.1 General Considerations

A digital transfer function with phase $\psi(\Omega)$ has a group-delay of the form given by (5.22). If we attempt to obtain the digital transfer function from an analog one which approximates

a linear phase in the passband, using the bilinear transformation we see that the delay properties of the analog filter are *not* preserved by this transformation. This is due to the factor $(1 + \Omega^2)$ which multiplies $\mathrm{d}\psi(\Omega)/\mathrm{d}\Omega$ in (5.21). Hence, this non-linear factor renders analog prototypes, such as the Bessel filter, of no use in the phase-oriented design of digital filters. Moreover, frequency scaling by a factor $(\Omega \to k\Omega)$ is not possible here since the factor $(1 + \Omega^2)$ does not scale. This necessitates the incorporation of a bandwidth scaling parameter in the expressions of transfer functions at the outset. We now derive the digital counterpart of the analog low-pass Bessel filter discussed in Section 3.6.2.

5.4.2 Maximally Flat Group-delay Response

Let the digital transfer function be written in the form

$$H(\lambda) = \frac{K}{Q_n(\lambda)} \tag{5.73}$$

where K is a constant and $Q_n(\lambda)$ is a polynomial to be determined such that the group delay is maximally flat around $\Omega = 0$. Write

$$Q_n(\lambda) = M(\lambda) + N(\lambda) \tag{5.74}$$

where $M(\lambda)$ is the even part of $Q(\lambda)$ and $N(\lambda)$ is its odd part. Consider the function

$$\phi(\lambda, \alpha) = \alpha \tanh^{-1} \lambda \tag{5.75}$$

so that

$$\tanh(\phi(\lambda, \alpha)) = \tanh(\alpha \tanh^{-1} \lambda) \tag{5.76}$$

where α is a parameter, whose significance will be clear shortly, and

$$\psi(\lambda) = \tanh^{-1}\left(\frac{N(\lambda)}{M(\lambda)}\right) \tag{5.77}$$

Now, if $\psi(\lambda)$ were to be *identical* to $\phi(\lambda, \alpha)$, then the delay function of the polynomial $Q_n(\lambda)$ is obtained from (5.20) as

$$\begin{aligned} T_{\mathrm{g}}(\lambda) &= -\frac{T}{2}(1 - \lambda^2)\frac{\mathrm{d}\phi(\lambda, \alpha)}{\mathrm{d}\lambda} \\ &= -\frac{T}{2}(1 - \lambda^2)\alpha\frac{\mathrm{d}}{\mathrm{d}\lambda}(\tanh^{-1} \lambda) \\ &= -\frac{T}{2}(1 - \lambda^2)\alpha\frac{1}{1 - \lambda^2} \\ &= -\frac{\alpha T}{2} \text{ a constant.} \end{aligned} \tag{5.78}$$

With the above motivation, we seek a method of approximating $\tanh(\alpha \tanh^{-1}\lambda)$ by the rational function $N(\lambda)/M(\lambda)$, that is

$$\frac{N(\lambda)}{M(\lambda)} \sim \tanh(\alpha \tanh^{-1} \lambda). \tag{5.79}$$

To this end we write

$$\tanh(\alpha \tanh^{-1}\lambda) = \frac{\sinh(\alpha \tanh^{-1}\lambda)}{\cosh(\alpha \tanh^{-1}\lambda)} \tag{5.80}$$

and expand the $\sinh(x)$ and $\cosh(x)$ functions, then find the continued fraction expansion of the right side of (5.80) as

$$\tanh(\alpha \tanh^{-1}\lambda) = \cfrac{\alpha\lambda}{1 + \cfrac{(1-\alpha^2)\lambda^2}{3 + \cfrac{(4-\alpha^2)\lambda^2}{5 + \cfrac{(9-\alpha^2)\lambda^2}{\ddots}}}} \tag{5.81}$$

Finally, the polynomial $Q(\lambda, \alpha)$ is obtained by identifying $N(\lambda)/M(\lambda)$ with the nth approximate in the continued fraction expansion of $\tanh(\alpha \tanh^{-1}\lambda)$. The resulting polynomial can be easily generated using the recurrence formula

$$Q_{n+1}(\lambda, \alpha) = Q_n(\lambda, \alpha) + \frac{\alpha^2 - n^2}{4n^2 - 1}\lambda^2 Q_{n-1}(\lambda, \alpha) \tag{5.82}$$

with

$$Q_0 = 1 \quad Q_1 = 1 + \alpha\lambda. \tag{5.83}$$

It can be shown that the delay response of the resulting filter is maximally flat around $\Omega = 0$. Figure 5.6 shows examples of such responses. However, since all the available degrees of freedom have been used to shape the delay response, the resulting amplitude

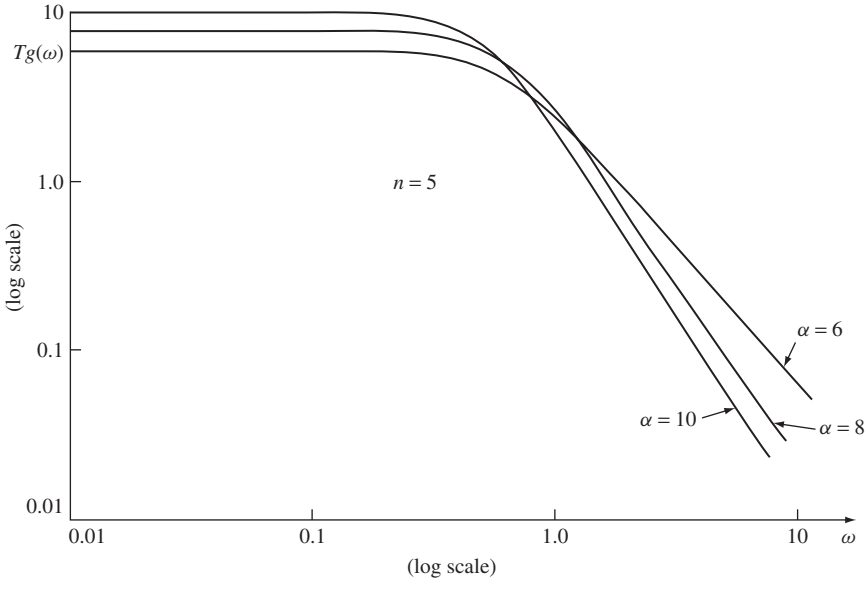


Figure 5.6 Delay response of the maximally flat delay filter

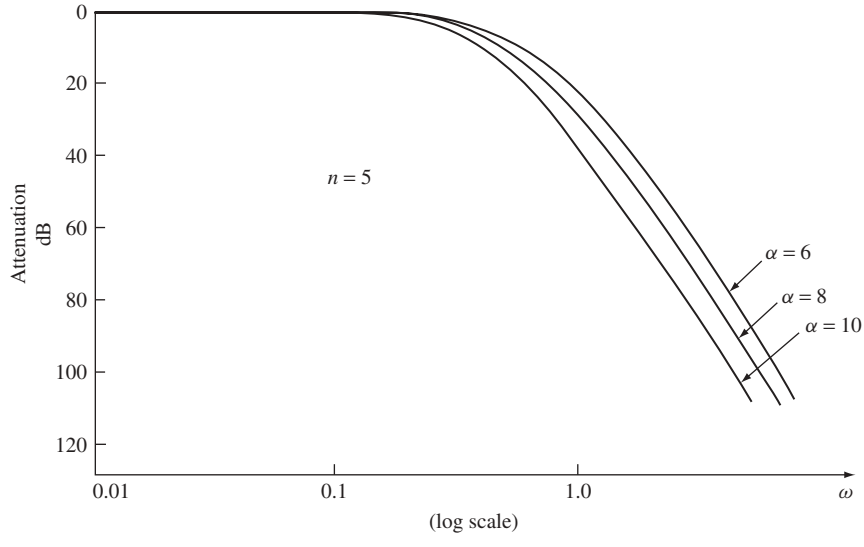


Figure 5.7 Amplitude response of the maximally flat delay filter

response is rather poor as to be expected, examples of which are shown in Figure 5.7. These responses also reveal that the parameter α can be used for bandwidth scaling, placing the passband edge at an arbitrary frequency. For the design of such filters, a simple computer program must be written to vary the degree n and the parameter α so that a combination of these may be found to meet the specifications on the delay response.

In addition to the maximally flat response, there is another near-optimum delay response which will be discussed later in the chapter in relation to a specific application. This is the *equidistant linear phase response filter* [13, 14].

We finally note that a high-pass transfer function can be obtained from the low-pass one in (5.73) by means of the transformation

$$\lambda \rightarrow 1/\lambda \tag{5.84}$$

which results in a maximally-flat delay around $\lambda = \infty$ corresponding to $\omega_N/2$. This can be verified by substitution in (5.21). Note that this result has no counterpart in the lumped analog domain.

5.5 FIR Filters

5.5.1 The Exact Linear Phase Property

It was shown that a digital filter of the FIR type has a transfer function of the form

$$H(z) = \sum_{n=0}^{N} a_n z^{-n} \tag{5.85}$$

That is, it is a polynomial in z^{-1}; therefore it is unconditionally stable (all the poles are at $z = 0$). Its direct non-recursive realization is shown in Figure 4.24. Evidently, the

coefficients a_n constitute the impulse response sequence of the filter; thus

$$\{h(n)\} = a_n \quad n = 0, 1, 2, \ldots, N \tag{5.86}$$

and

$$H(z) = \sum_{n=0}^{N} h(n) z^{-n}. \tag{5.87}$$

and substituting the bilinear variable, it takes the form

$$H(\lambda) = \frac{f_m(\lambda)}{(1+\lambda)^N} \quad m \leq N \tag{5.88}$$

where $f_m(\lambda)$ is a general polynomial and m can be different from N to allow for possible transmission zeros at $\lambda = \infty$ corresponding to $z = -1$ and $\omega = \omega_N/2$.

Now, suppose that $f_m(\lambda)$ is restricted to be an even polynomial, that is it contains only even powers of λ. Then

$$H(\lambda) = \frac{E_m(\lambda)}{(1+\lambda)^N} \tag{5.89}$$

where $E_m(\lambda)$ is even. The delay function of $H(\lambda)$ is obtained from (8.21) as

$$T_g(\lambda) = \frac{T}{2} \mathrm{Ev}\left[(1-\lambda^2)\left(\frac{N(1+\lambda)^{N-1}}{(1+\lambda)^N} - \frac{E_m'(\lambda)}{E_m(\lambda)}\right)\right]. \tag{5.90}$$

Since, however, $E_m(\lambda)$ is assumed even, then $E_m'(\lambda)$ is odd. Hence $\mathrm{Ev}(E_m'(\lambda)/E_m(\lambda)) = 0$, so that

$$T_g(\lambda) = NT/2 \text{ a constant.} \tag{5.91}$$

Therefore, we have the important result that if $f_m(\lambda)$ in (5.89) is an even polynomial, then the filter has a constant group delay (exact linear phase) at all frequencies. The same conclusion is reached if $f_m(\lambda)$ is an odd polynomial. However, due to the zero of transmission at $\lambda = 0$, the function with odd numerator is suitable for the design of high-pass, band-pass filters and differentiators, not for low-pass and band-stop responses which are realizable by $H(\lambda)$ with even numerators.

Let us now examine the constant group delay property (exact linear phase) in terms of the z-domain representation of the FIR transfer function. It is shown below that the constraints on the coefficients a_n (hence the impulse response) of the filter described by (5.87) are those of *even or odd symmetry about some mid-point*.

5.5.1.1 Symmetric Impulse Response

(i) *For N even*

$$h(n) = h(N-n) \quad n = 0, 1, 2, \ldots (N/2 - 1) \tag{5.92a}$$

or, equivalently

$$h\left(\tfrac{1}{2}N - n\right) = h\left(\tfrac{1}{2}N + n\right) \quad n = 1, 2, \ldots \tfrac{1}{2}N. \tag{5.92b}$$

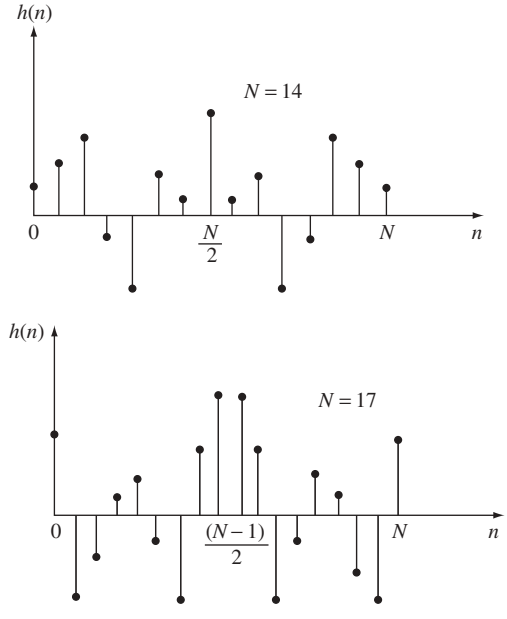


Figure 5.8 The symmetry constraints on the impulse response of an FIR filter

(ii) *For N odd*

$$h(n) = h(N-n) \quad n = 0, 1, 2, \ldots \tfrac{1}{2}(N-1). \tag{5.93a}$$

or, equivalently

$$h\left(\tfrac{1}{2}(N-1) - n\right) = h\left(\tfrac{1}{2}(N+1) + n\right) \quad n = 0, 1, 2, \ldots \tfrac{1}{2}(N-1). \tag{5.93b}$$

The above constraints mean that the impulse response of the filter has even symmetry about a mid-point, as shown in Figure 5.8.

To prove that the symmetry constraints (5.92) or (5.93) lead to a filter with constant group delay at all frequencies, consider (5.92). Write (5.87) as

$$\begin{aligned} H(z) &= \sum_{n=0}^{N} h(n) z^{-n} \\ &= z^{-N/2} \sum_{n=-N/2}^{N/2} h(N/2+n) z^{-n} \end{aligned} \tag{5.94}$$

or

$$H(z) = z^{-N/2} \left(h(N/2) + \sum_{n=0}^{N/2-1} [h(n) z^{N/2-n} + h(N-n) z^{-(N/2-n)}] \right). \tag{5.95}$$

Imposing the constraints in (5.92), the above expression becomes

$$H(z) = z^{-N/2}(h(N/2) + \sum_{n=0}^{N/2-1} h(n)(z^{N/2-n} + z^{-(N/2-n)}) \tag{5.96}$$

which, on the unit circle ($z = \exp(j\omega T)$, reads

$$H(\exp(j\omega T)) = \exp\left(\frac{-j\omega TN}{2}\right)\left(h(N/2) + 2\sum_{n=0}^{N/2-1} h(n)\cos\frac{(N-2n)\omega T}{2}\right)$$

or

$$H(\exp(\mathrm{j}\omega T)) = |H(\exp(\mathrm{j}\omega T))|\exp(\mathrm{j}\psi(\omega)) \tag{5.97}$$

where

$$|H(\exp(\mathrm{j}\omega T))| = h(N/2) + 2\sum_{n=0}^{N/2-1} h(n)\cos\frac{(N-2n)\omega T}{2} \tag{5.98}$$

and

$$\psi(\omega) = -\frac{\omega TN}{2}. \tag{5.99}$$

Therefore, the group delay is

$$\begin{aligned} T_{\mathrm{g}}(\omega) &= -\frac{\mathrm{d}\psi(\omega)}{\mathrm{d}\omega} \\ &= TN/2 \text{ a constant for all } \omega. \end{aligned} \tag{5.100}$$

Hence, we have proved that under the symmetry constraint (5.92), the group delay of the filter is constant at all frequencies. This result is equivalent to the requirement, in the λ-domain, that the numerator of $H(\lambda)$ is an even polynomial.

Similarly, for N odd the symmetry constraints (5.93) leads to the same conclusion with

$$H(\exp(\mathrm{j}\omega T)) = \exp(-\mathrm{j}\omega TN/2)\left(2\sum_{n=0}^{(N-1)/2} h(n)\cos\frac{(N-2n)\omega T}{2}\right). \tag{5.101}$$

At half the sampling frequency ($\omega_{\mathrm{N}}/2$) we have

$$\frac{\omega_{\mathrm{N}}T}{2} = \frac{\omega_{\mathrm{N}}}{2}\cdot\frac{2\pi}{\omega_{\mathrm{N}}} = \pi \tag{5.102}$$

so that expression (5.101) vanishes. Therefore, this case is suitable only for the design of low-pass and band-pass filters, *not* for high-pass or band-stop filters.

5.5.1.2 Antimetric Impulse Response

(i) *For N even*

$$h(n) = -h(N-n) \quad n = 0, 1, \ldots, (N-2)/2 \tag{5.103a}$$

or, equivalently

$$h(N/2-n) = -h(N/2+n) \quad n = 1, 2, \ldots, N/2 \tag{5.103b}$$

with $h(0) = 0$.

(ii) *For N odd*

$$h(n) = -h(N-m) \quad n = 0, 1, 2, \ldots, (N-1)/2 \tag{5.104a}$$

or, equivalently

$$h\left(\frac{N-1}{2} - n\right) = -h\left(\frac{N+1}{2} + n\right) \quad n = 0, 1, 2, \ldots, (N-1)/2. \tag{5.104b}$$

The above constraints imply that the impulse response of the filter has odd symmetry about some midpoint, as illustrated in Figure 5.9. To prove that these constraints lead to a constant group-delay response, consider the case of N even at first. The transfer function, under conditions (5.101), takes the form

$$H(z) = z^{-N/2}\left(0 + \sum_{n=0}^{N/2-1} h(n)(z^{N/2-n} - z^{-(N/2-n)})\right) \tag{5.105}$$

so that

$$\begin{aligned} H(\exp(\mathrm{j}\omega T)) &= \exp(\mathrm{j}\omega TN/2)2\mathrm{j}\sum_{n=0}^{N/2-1} h(n)\sin\frac{(N-2n)\omega T}{2} \\ &= |H(\exp(\mathrm{j}\omega T))|\exp(\mathrm{j}\psi(\omega)) \end{aligned} \tag{5.106}$$

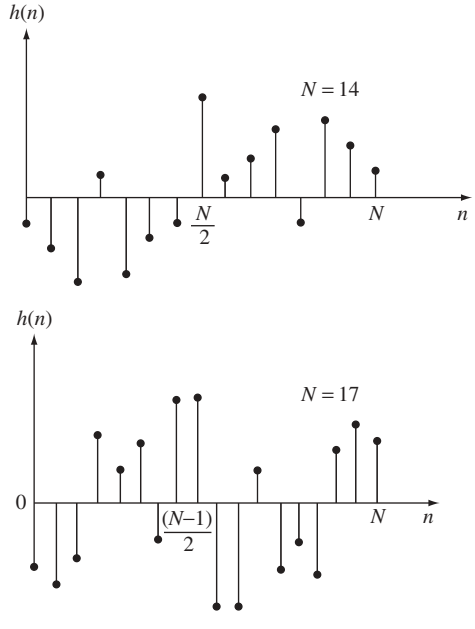


Figure 5.9 The antimetry constraints on the impulse response of an FIR filter

where

$$|H(\exp(\mathrm{j}\omega T)| = 2\sum_{n=0}^{N/2-1} h(n)\sin\frac{(N-2n)\omega T}{2} \tag{5.107}$$

and

$$\psi(\omega) = -\frac{\omega TN}{2} + \frac{\pi}{2}. \tag{5.108}$$

The group delay is given by

$$\begin{aligned} T_{\mathrm{g}}(\omega) &= -\frac{\mathrm{d}\psi(\omega)}{\mathrm{d}\omega} \\ &= \frac{NT}{2} \text{ a constant for all } \omega \end{aligned} \tag{5.109}$$

but the filter introduces a constant phase shift of $\pi/2$. It is also clear that at $\omega = 0$ there is a zero of transmission. Also at half the sampling frequency $(N-2n)\omega T/2 = (N-2n)\pi/2$ and, since N is even, the transfer function is zero at $\omega_{\mathrm{N}}/2$. Therefore this case of antimetric impulse response with N even is suitable only for approximating the ideal characteristics of band-pass filters and differentiators.

Repeating the above analysis for N odd leads to the same conclusion with

$$H(\exp(\mathrm{j}\omega T)) = \exp(-\mathrm{j}\omega TN/2)2\mathrm{j}\sum_{n=0}^{(N-1)/2} h(n)\sin\frac{(N-2n)\omega T}{2}. \tag{5.110}$$

Again, at $\omega = 0$ the above function is zero, hence it is not suitable for low-pass or band-stop filters, but it can be used for the design of high-pass or band-pass filters and differentiators.

Now, let us translate the exact linear phase property constraints into corresponding conditions in terms of the locations of the zeros of the transfer function. We investigate these conditions both in the λ-plane and the z-plane. Consider a transfer function of the form

$$H(\lambda) = \frac{E_m(\lambda)}{(1+\lambda)^N} \tag{5.111}$$

where $E_m(\lambda)$ is an even polynomial for a constant group delay response. But the factors of *any real even* polynomial must be of three types:

(i) $$(\lambda^2 - \Sigma_0^2) \tag{5.112}$$

(ii) $$(\lambda^2 + \Omega_0^2) \tag{5.113}$$

(iii) $$[\lambda + (\Sigma_0 + \mathrm{j}\Omega_0)](\lambda + \Sigma_0 - j\,\Omega_0) \times [\lambda - (\Sigma_0 + \mathrm{j}\Omega_0)][\lambda - (\Sigma_0 + \mathrm{j}\Omega_0)]. \tag{5.114}$$

The complex ones in (5.114) must occur in conjugate pairs to preserve the *realness* of the coefficients of the polynomial. In addition, for every complex zero at $(\Sigma_0 + \mathrm{j}\Omega_0)$ there must be another one at $-(\Sigma_0 + \mathrm{j}\Omega_0)$ so that the four factors should combine to form an

even polynomial. Thus, complex zeros must occur in *quadruplets*. This is another way of stating the rather obvious fact that if $E_m(\lambda)$ is even then

$$E_m(-\lambda) = E_m(\lambda) \tag{5.115}$$

which requires that a zero of $E_m(\lambda)$ must also be a zero of $E_m(-\lambda)$. Typical zero locations for an even polynomial $E_m(\lambda)$ are shown in Figure 5.10. For a function of the form

$$H(\lambda) = \frac{O_m(\lambda)}{(1+\lambda)^N} \tag{5.116}$$

where $O_m(\lambda)$ is an odd polynomial, the same considerations holds true with the additional zero at $\lambda = 0$ because an odd polynomial can be written as λ times an even one. In this case

$$O_m(-\lambda) = -O_m(\lambda). \tag{5.117}$$

Now, in the z-domain, the distribution of the zeros, for the constant group-delay property can be obtained from Figure 5.10 and the mapping between the λ-plane and the z-plane. This leads to the condition that for every zero z_i inside the unit circle we must also have its conjugate z_i^* together with their reciprocals $1/z_i$ and $1/z_i^*$. This should also be clear if we use (5.115) and (5.117) together with (5.111) and (5.116) to write

$$(1+\lambda)^N H(\lambda) = \pm(1-\lambda)^N H(-\lambda) \tag{5.118}$$

where the plus sign holds for an even numerator and the minus sign for the odd numerator case. Then, using (5.115) and noting that replacing λ by $-\lambda$ corresponds to replacing z by z^{-1}, we can write from (5.20)

$$\frac{2^N}{(1+z^{-1})^N} H(z) = \pm \frac{(2z^{-1})^N}{(1+z^{-1})^N} H(z^{-1}) \tag{5.119a}$$

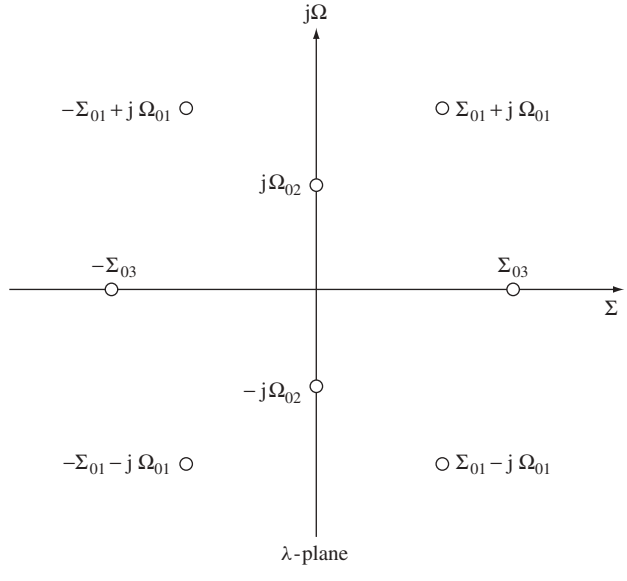


Figure 5.10 Typical zero locations of a constant delay FIR filter in the λ-plane

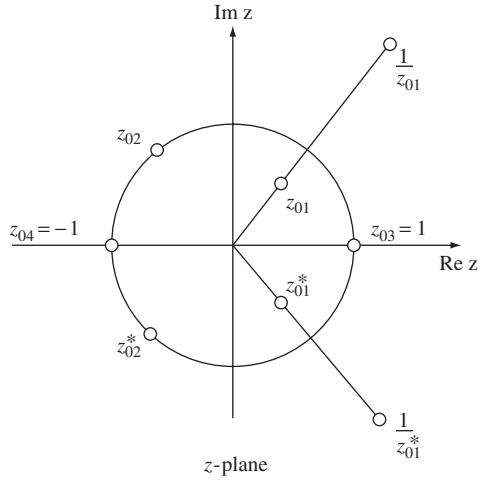


Figure 5.11 Typical zero locations of a constant delay FIR filter in the z-plane

or

$$H(z^{-1}) = \pm z^N H(z) \tag{5.119b}$$

which means that the zeros of $H(z^{-1})$ are the same as those of $H(z)$. This leads to the conclusions given above for the zero locations. In the case of an antimetric impulse response corresponding, in the λ-domain, to a function with an odd numerator, the zero at $\lambda = 0$ corresponds to $z = 1$. The zeros on the unit circle occur in conjugate pairs and are their own reciprocals. Typical zero locations in the z-plane are shown in Figure 5.11.

We now turn to the design techniques of FIR filters. Due to the constrained form of the transfer function, analog prototypes cannot be employed, hence special techniques are required which are sometimes simpler and sometimes more elaborate than those used for the design of IIR filters.

5.5.2 *Fourier-coefficient Filter Design*

We begin by discussing a simple design technique for approximating the ideal filter characteristic [12, 20]. For the sake of brevity, we concentrate on the low-pass even degree case, while the other solutions follow in a similar manner.

The transfer function of the even-degree FIR filter is given by (5.85) and if the filter is required to have a constant group delay at all frequencies, the symmetry constraints in (5.93) must be imposed. To simplify the expressions, put

$$b_n \triangleq h(N/2 + n) \qquad n = -N/2, \ldots N/2 \tag{5.120}$$

so that the transfer function takes the form

$$H(z) = z^{-N/2} \sum_{n=-N/2}^{N/2} b_n z^{-n} \tag{5.121}$$

and the linear phase constraints become

$$b_{-n} = b_n \quad n = 1, 2, \ldots N/2. \tag{5.122}$$

On the unit circle, expression (5.121) becomes

$$H(\exp(\mathrm{j}\omega T)) = \exp(-\mathrm{j}\omega NT/2) \sum_{n=-N/2}^{N/2} b_n \exp(-\mathrm{j}n\omega T). \tag{5.123}$$

The exponential factor in the above expression does not, however, affect the amplitude response; it only introduces a constant delay of $NT/2$ as explained earlier. Therefore, the amplitude response of the filter is determined by the function

$$\hat{H}(\mathrm{j}\omega) = \sum_{n=-N/2}^{N/2} b_n \exp(-\mathrm{j}n\omega T) \tag{5.124}$$

With

$$\theta = \omega T \tag{5.125}$$

we have

$$\hat{H}(\theta) = \sum_{n=-N/2}^{N/2} b_n \exp(-\mathrm{j}n\theta). \tag{5.126}$$

It is now required to obtain the coefficients b_n such that $\hat{H}(\theta)$ approximates an arbitrary desired function $H_\mathrm{d}(\theta)$. Since $\hat{H}(\theta)$ is periodic and looks very much like a truncated Fourier series, an obvious approximation method suggests itself. This consists in expanding the *desired* function $H_\mathrm{d}(\theta)$ in a Fourier series as

$$H_\mathrm{d}(\theta) = \sum_{n=-\infty}^{\infty} c_n \exp(-\mathrm{j}n\theta) \tag{5.127}$$

where

$$c_n = \frac{1}{2\pi} \int_{-\infty}^{\infty} H_\mathrm{d}(\theta) \exp(\mathrm{j}n\theta)\, \mathrm{d}\theta \tag{5.128}$$

then truncate the series and make the identification

$$\begin{aligned} b_n &= c_n \quad |n| = 0, 1, \ldots, N/2 \\ &= 0 \quad |n| > N/2. \end{aligned} \tag{5.129}$$

As an example, consider the approximation of the ideal low-pass characteristic shown in Figure 5.12. This is a real even function defined by

$$\begin{aligned} H_\mathrm{d}(\theta) &= 1 \quad 0 \le |\theta| \le \theta_0 \\ &= 0 \quad \theta_0 < |\theta| < \pi. \end{aligned} \tag{5.130}$$

Calculation of the Fourier coefficients of the desired response function $H_{\mathrm{d}}(\theta)$ gives

$$c_0 = \frac{1}{2\pi}\int_{-\theta_0}^{\theta_0} \mathrm{d}\theta = \frac{\theta_0}{\pi} \tag{5.131}$$

and

$$c_n = \frac{1}{2\pi}\int_{-\theta_0}^{\theta_0} \exp(\mathrm{j}n\theta)\,\mathrm{d}\theta = \frac{1}{n\pi}\sin n\theta_0. \tag{5.132}$$

Therefore, by (5.129) we obtain the filter coefficients as

$$b_n = \frac{1}{n\pi}\sin n\theta_0 \quad n = 1, 2, \ldots N/2 \tag{5.133a}$$

$$b_0 = \theta_0/\pi \tag{5.133b}$$

and to capture the exact linear phase property we impose (5.120). The original impulse response sequence is simply

$$h(N/2+n) = \frac{1}{n\pi}\sin n\omega_0 T \quad n = 1, 2, \ldots (N/2-1) \tag{5.134a}$$

with

$$h(N/2) = \frac{\omega_0 T}{\pi}. \tag{5.134b}$$

This design technique can be used to approximate any desired characteristic, not only the ideal one. Furthermore, for approximation of the ideal high-pass, band-pass and band-stop cases with linear phase responses, the appropriate type of transfer function is chosen.

Now, we have seen in Chapter 2 that representing a desired function by a truncated Fourier series results in the Gibbs' phenomenon in the vicinity of a discontinuity. Since this is the case for the ideal characteristic in Figure 5.12, we expect the response $H(\theta)$ to exhibit such oscillations. It was also explained in Chapter 2 that the Gibbs' phenomenon is due to the non-uniform convergence of the truncated Fourier series in the vicinity of a discontinuity. In the linear phase design, the resulting passband overswing is about 0.8 dB and the first ripple in the stop-band is about 20 dB, as shown in Figure 5.13.

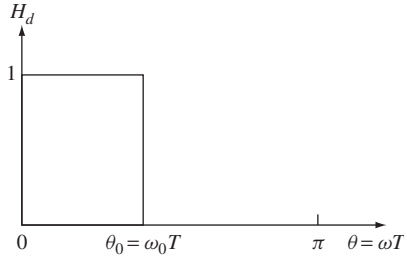


Figure 5.12 The ideal low-pass filter characteristic

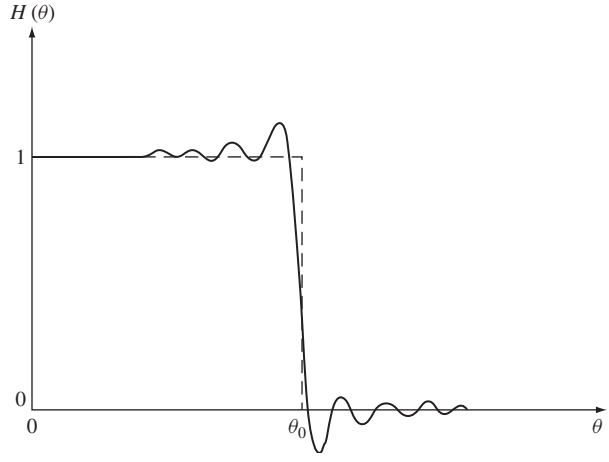


Figure 5.13 The Gibbs' phenomenon in the approximation of the ideal low-pass characteristic by an FIR transfer function

These ripples in $|H(\theta)|^2$ remain fixed in size δ no matter how large the order of the filter is. The ripples are only 'telescoped' to occupy narrower widths as the order of the filter is increased. Thus, for specifications requiring less than 0.8 dB in the passband and/or more than 20 dB in the stop-band simple direct truncation of the Fourier series cannot be employed. As discussed in Section 2.2.5, direct truncation of the Fourier series (5.121) to obtain the filter transfer function is equivalent to multiplying the coefficients c_n by the *rectangular window* shown in Figure 5.14(a) and given by

$$\begin{aligned} w_{\mathrm{R}} &= 1 \quad |n| \leq N/2 \\ &= 0 \quad |n| > N/2 \end{aligned} \tag{5.135}$$

so that

$$b_n = w_{\mathrm{R}}(n) c_n \tag{5.136}$$

and the impulse response of the window $w_{\mathrm{R}}(n)$ is symmetric so that, if required, the linear phase property can be preserved. The frequency response of the window is obtained by taking the Fourier transform of the sampled pulse to give

$$W_{\mathrm{R}}(\omega) = NT \frac{\sin(N\omega T/2)}{(N\omega T/2)} \tag{5.137}$$

which is shown in Figure 5.14(b).

By the frequency convolution relation (2.46), we find that for an arbitrary desired characteristic $H_{\mathrm{d}}(\omega)$, the spectrum of the windowed function is given by

$$H_{\mathrm{w}}(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} H_{\mathrm{d}}(\mu) W_{\mathrm{R}}(\omega - \mu)\, \mathrm{d}\mu \tag{5.138}$$

which is shown in Figure 5.15 in the case where $H_{\mathrm{d}}(\omega)$ is the ideal low-pass characteristic shown in Figure 5.12. The ripples are due to those in the window function $w_{\mathrm{R}}(\omega)$, and in

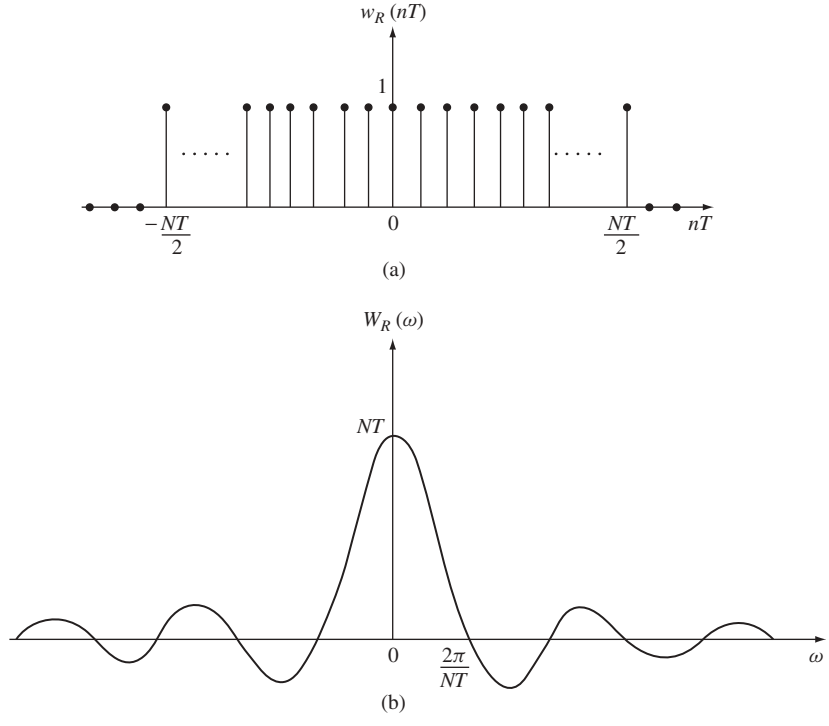


Figure 5.14 (a) Rectangular window and (b) its spectrum

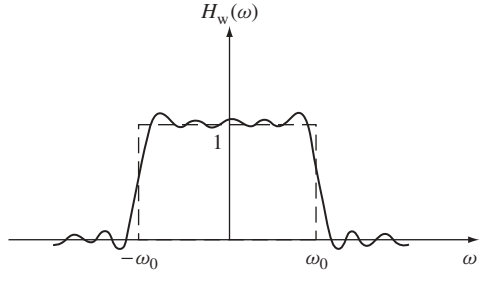


Figure 5.15 Spectrum of windowed ideal characteristic using the rectangular window

order to reduce these ripples we must choose another window whose spectrum introduces smaller ripples than those of the rectangular window.

There is a large number of expressions which may be used to modify the coefficients b_n in order to suppress the Gibbs' phenomenon and improve the convergence of the truncated series. This is generally achieved at the expense of the cut-off rate, that is the side ripples near the discontinuity are reduced, but the transition band becomes wider. The expressions used are precisely the window functions discussed in Chapter 2, but they are repeated here in the context of filter design. The new windowed filter coefficients are given by

$$b_n^w = w(n)b_n \tag{5.139}$$

where $w(n)$ is a window function, examples of which are given below. In the following expressions it is implied that

$$w(n) = 0 \quad |n| > N/2. \tag{5.140}$$

(i) *Fejer Window*

$$w(n) = (N - n + 1)/(N + 1). \tag{5.141}$$

(ii) *Lanczos Window*

$$w(n) = \frac{\sin(2n\pi/N)}{(2n\pi/N)}. \tag{5.142}$$

(iii) *von Hann Window*

$$w(n) = 0.5\left(1 + \cos\frac{2n\pi}{N}\right). \tag{5.143}$$

(iv) *Hamming Window*

$$w(n) = 0.54 + 0.46\cos\frac{2n\pi}{N}. \tag{5.144}$$

(v) *Blackman Window*

$$w(n) = 0.42 + 0.5\cos\left(\frac{2n\pi}{N}\right) + 0.08\cos\left(\frac{4n\pi}{N}\right). \tag{5.145}$$

(vi) *Kaiser Window*

$$w(n) = I_0\left\{\beta\left[1 - \left(\frac{2n}{N}\right)^2\right]^{1/2}\right\} \Big/ I_0(\beta) \tag{5.146}$$

where $I_0(x)$ is the modified zero-order Bessel function of the first kind and β is a parameter. The function $I_0(x)$ can be generated by means of the rapidly convergent series approximation

$$I_0(x) = 1 + \sum_{k=1}^{\infty}\left[\frac{1}{k!}\left(\frac{x}{2}\right)^k\right]^2 \tag{5.147}$$

Figures 5.16 and 5.17 show examples of the responses of filters designed using Fejer and von Hann windows, respectively, for $N = 50$ and $\omega_0 = 0.25\,\omega_N$. The first gives a first ripple of 26 dB in the stopband while the second has a 44 dB first ripple. Table 5.1 gives a comparison between some windows based on the extent to which each one suppresses the Gibbs' phenomenon.

Generally, the more successful the particular window in suppressing the Gibbs' oscillations, the wider the transition bandwidth. Furthermore, each particular window is capable of producing a fixed level of attenuation in the stopband, no matter what the degree is.

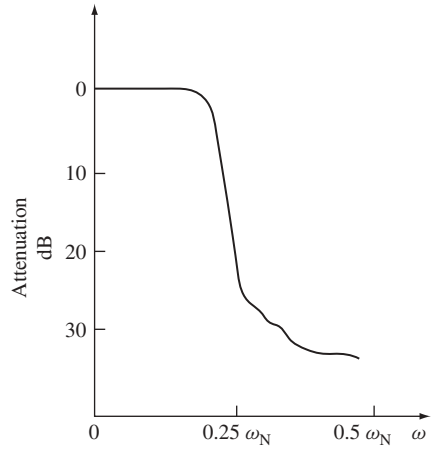


Figure 5.16 Example of the use of the Fejer window in FIR filter design. $N = 50$ and $\omega_0 = 0.25\,\omega_N$

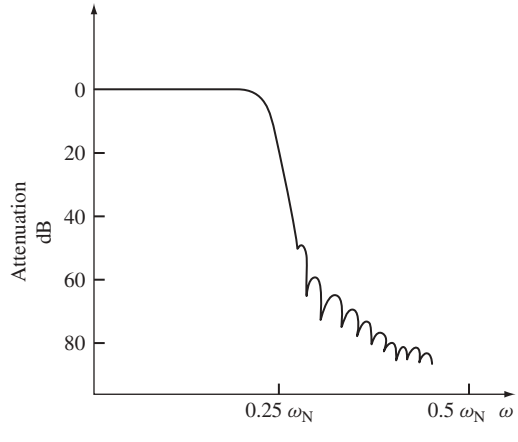


Figure 5.17 Example of the use of the Hann window in FIR filter design. $N = 50$ and $\omega_0 = 0.25\,\omega_N$

Table 5.1 Comparison of the useful available stopband attenuation for various types of window functions

Type of window	Stopband ripple in dB for constant delay designs
Fejer	26
Von Hann	44
Hamming	52
Kaiser, $\beta = 7.8$	80

One exception, however, is the Kaiser window which contains the parameter β, this may be used for trade-off between ripple size and transition bandwidth.

For the Kaiser window, there exist empirical formulae for the determination of the required filter degree. To state these, let the transition bandwidth be normalized with respect to the sampling frequency ω_N to give

$$\Delta\omega = (\omega_s - \omega_0)/\omega_N \tag{5.148}$$

where ω_s is the stopband edge and ω_0 is the passband edge. It is required to have a minimum stopband attenuation of α_s. Then, a formula for the degree is

$$N = \frac{\alpha_s - 7.95}{14.36\Delta\omega} \tag{5.149}$$

The trade-off parameter β is obtained from

$$\beta = 0.1102(\alpha_s - 8.7) \quad \text{for } \alpha_s \geq 50\,\text{dB} \tag{5.150a}$$

or

$$\beta = 0.5942(\alpha_s - 21)^{0.4} + 0.07886(\alpha_s - 21) \quad \text{for } 21 < \alpha_s < 50. \tag{5.150b}$$

As an example, consider the following specifications:

Passband: 0–1 kHz;
Stopband: 1.5–5.0 kHz, $\alpha_s = 50$ dB.

Taking a sampling frequency of twice the highest component in the band of interest, we have $\omega_N = 2\pi \times 10 \times 10^3$. Thus (5.150) gives

$$\Delta\omega = (1.5 - 1)/10 = 0.05 \tag{5.151}$$

so that using (5.149) and (5.150) we obtain

$$N = \frac{50 - 7.95}{14.36 \times 0.05} \sim 60 \tag{5.152}$$

and $\beta = 0.1102(50 - 8.7) = 4.55$.

The above values can be used to evaluate the Kaiser window coefficients in (8.144) using the series approximation in (5.147). Figure 5.18 shows an example of a filter designed using this technique.

5.5.2.1 Fourier-coefficient Design of Differentiators

An ideal analog differentiator is a linear system characterized by the analog transfer function

$$H(s) = s \tag{5.153}$$

so that

$$H(\mathrm{j}\omega) = \mathrm{j}\omega. \tag{5.154}$$

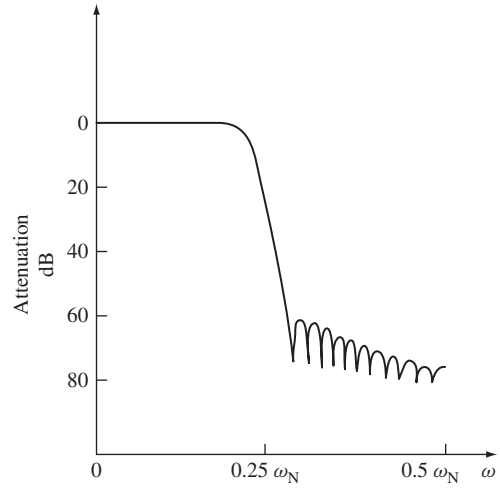


Figure 5.18 Example of the use of the Kaiser window in FIR filter design

Consequently, a digital differentiator with transfer function $H(z)$ is designed such that

$$H(z)|_{z=\exp(j\omega T)} \sim j\omega \quad 0 \le \omega \le \omega_N/2. \tag{5.155}$$

The Fourier coefficient design technique of the previous section can be used to approximate the function in (5.154), which is clearly *odd*. Therefore, if the exact linear phase property is also required, the functions with antimetric impulse response are used. With the same notation used in (5.124) to (5.128) with N changed to $(N-1)$, the Fourier coefficients of the function $j\omega$ are given by

$$b_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} (j\omega)\exp(jn\theta)d\theta \tag{5.156}$$

or, with

$$\theta = \omega T = 2\pi\frac{\omega}{\omega_N} \tag{5.157}$$

$$b_n = \frac{1}{\omega_N}\int_{-\omega_N/2}^{\omega_N/2} (j\omega)\exp(jn\omega T)\,d\omega \tag{5.158}$$

which yields

$$b_0 = 0 \tag{5.159}$$

$$b_n = \frac{1}{nT}\cos n\pi \quad n = 1, 2, \ldots (N-1)/2$$

and the truncated Fourier series of the form (5.126) is used for approximating the differentiator function. Evidently, if we require a constant delay response in addition to approximating the differentiator function, then the antimetry constraints in (5.103) may be imposed because $j\omega$ is an odd function with a transmission zero at $\omega = 0$.

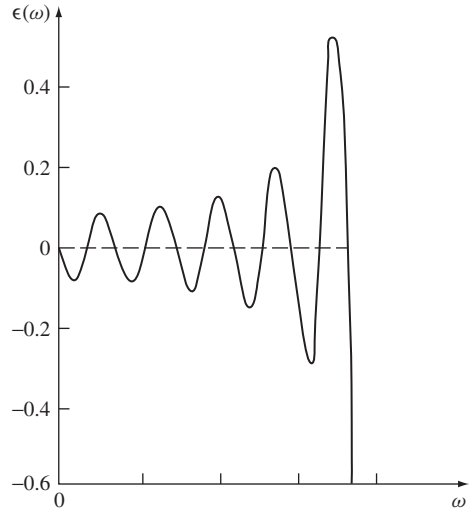


Figure 5.19 Error in approximating the ideal differentiator by direct truncation of the Fourier series

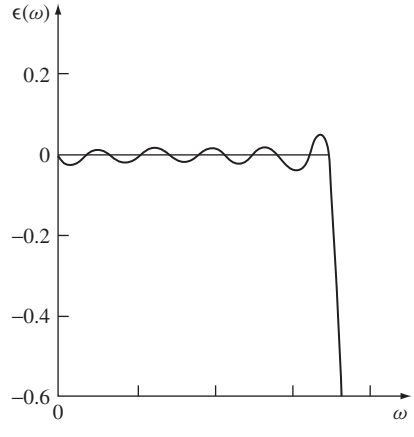


Figure 5.20 Error in the use of the Kaiser window for the design of a differentiator

The error in the amplitude approximation is

$$\varepsilon(\omega) = \left| \sum_{-(N-1)/2}^{(N-1)/2} a_n \exp(-\mathrm{j}n\omega T) \right| - \omega \tag{5.160}$$

which is plotted in Figure 5.19 for $N = 21$. Again the non-uniform convergence of the truncated Fourier series leads to the oscillations shown. These can be reduced using a window such as the Kaiser window. The result for $N = 21$ and $\beta = 3$ is shown in Figure 5.20.

5.5.3 Monotonic Amplitude Response with the Optimum Number of Constraints

We have seen that the transfer function of the FIR filter can be expressed in terms of the bilinear variable λ as

$$H(\lambda) = \frac{f_m(\lambda)}{(1+\lambda)^N} \tag{5.161}$$

and, as shown in Section 5.4.1, if $f_m(\lambda)$ is an even (or odd) polynomial the filter has a constant group-delay at all frequencies. This is equivalent to the symmetry (or antimetry) constraints in the z-domain and the impulse response. Once the exact linear phase property is maintained by choosing f_m as an even polynomial (in the low-pass case), our efforts can be entirely concentrated on the amplitude approximation problem. We now obtain $H(\lambda)$ such that the amplitude response is monotonic with the optimum number of constraints capable of being arbitrarily divided between the passband and stopband. Let N be even and

$$k = N/2. \tag{5.162}$$

We require $H(\lambda)$ to be such that:

(i) $|H(\mathrm{j}\Omega)|^2$ has $(2m-1)$ zero derivatives at $\Omega = 0$ (i.e. $\omega = 0$).
(ii) $|H(\mathrm{j}\Omega)|^2$ has $2(k-m)$ zero derivatives at $\Omega = \infty$ (i.e. at $\omega = \omega_N/2$).

The function with the required properties is given by

$$H(\lambda) = \frac{\sum_{r=0}^{m} (-1)^r \binom{k}{r} \lambda^{2r}}{(1+\lambda)^{2k}} \tag{5.163}$$

which is such that, in $H(\mathrm{j}\Omega)H(-\mathrm{j}\Omega)$, the first m terms of the numerator agree with the corresponding ones in the denominator. Thus, the argument expounded in the derivation of equation (5.72) leads to the conclusion that the first $(2m-1)$ derivatives of $|H(\mathrm{j}\Omega)|^2$ vanish at $\Omega = 0$. Similarly, the degrees of the numerator and denominator of $H(\lambda)$ differ by $2(k-m)$ corresponding to the number of transmission zeros at $\Omega = \infty$ (i.e. $\omega = \omega_N/2$). A combination of n and m must be chosen in order to meet an arbitrary set of specifications. Examples of the responses obtained using this very simple, yet powerful, design technique are shown in Figure 5.21. For the purpose of realization we can transform to the z-domain.

5.5.4 Optimum Equiripple Response in both Passband and Stopband

In the analog domain, and in the case of IIR filters, the optimum solution to the amplitude approximation problem is the one giving rise to equiripple response in both the passband and stopband. It is also possible to arrive at the same conclusion for the case of FIR filters. However, it is unfortunate that due to the constrained form of the FIR transfer function, no analytical solution is known to this problem. Consequently, it has been solved using

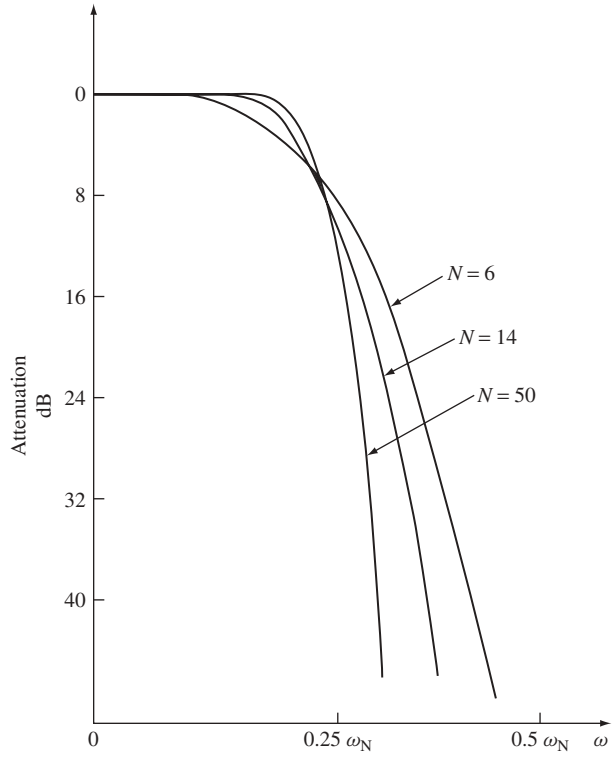


Figure 5.21 Examples of FIR filter responses defined by (5.163)

numerical algorithms. These are based on a number of concepts which are now discussed for the low-pass linear phase design with symmetric impulse response and N even. This is defined by (5.94), which upon the change of variables

$$A(n) = 2h(N/2 - n)$$

$$A(0) = h(N/2) \tag{5.164}$$

becomes

$$H(\exp(\mathrm{j}\omega T) = \exp(-\mathrm{j}\omega TN/2) \sum_{m=0}^{N/2} A(m) \cos m\omega T. \tag{5.165}$$

The amplitude is determined by the summation in the above expression. For convenience we take the sampling frequency $f_\mathrm{N} = 1$ so that $T = 1/f_\mathrm{N} = 1$ and the amplitude response is determined by the function

$$\hat{H}(\omega) = \sum_{m=0}^{N/2} A(m) \cos m\omega. \tag{5.166}$$

The problem is to determine the coefficients $A(m)$ such that the amplitude response is optimum equiripple in both bands. It is to be noted that the problem for all other types of linear phase transfer functions can be reduced to the same problem discussed here, namely the determination of the coefficients of a cosine-series of the general form given by (5.166).

Now, a well known result in the theory of approximation by polynomials asserts that the necessary and sufficient condition that $H(\omega)$ be the unique best weighted Chebyshev approximation to the desired function $D(\omega)$ in the interval $[0,\omega_N/2]$ is that the weighted error function

$$\varepsilon(\omega) = D(\omega) - \hat{H}(\omega) \tag{5.167}$$

exhibit at least $(N/2+1)$ extremal points (maxima and minima) over $[0,\omega_N/2]$. This means that there exist $(N/2+1)$ frequencies ω_i in $[0,\pi]$ such that

$$\omega_0 < \omega_1 < \cdots \omega_N/2 \tag{5.168}$$

$$\varepsilon(\omega_1) = -\varepsilon(\omega_{i+1}) \quad i = 0, 1, \ldots, N/2 \tag{5.169}$$

and

$$|\varepsilon(\omega_i)| = \max|\varepsilon(\omega)| \quad i = 0, 1, \ldots, N/2 \tag{5.170}$$

where the maximization is taken over all ω in $[0,\omega_N/2]$. This result is also valid if a weight function $W(\omega)$ of the error is introduced, so that the weighted error may be written as

$$\varepsilon(\omega) = W(\omega)(D(\omega) - \hat{H}(\omega)). \tag{5.171}$$

In this case $W(\omega)$ allows the designer to specify the relative error in the passband and stopband. For example in the present case of low-pass design with the tolerance scheme shown in Figure 5.22 we have

$$\begin{aligned} D(\omega) &= 1 \quad 0 \le \omega \le \omega_p \\ &= 0 \quad \omega_s \le \omega < \omega_N/2 \end{aligned} \tag{5.172}$$

and we take

$$\begin{aligned} W(\omega) &= (\delta_2/\delta_1) \quad 0 \le \omega \le \omega_p \\ &= 1 \qquad\quad \omega_s \le \omega \le \omega_N/2. \end{aligned} \tag{5.173}$$

The approximation problem can be formulated by writing the set of equations

$$W(\omega_r)(D(\omega_r) - \hat{H}(\omega_r)) = (-1)^r \delta \tag{5.174}$$

or

$$W(\omega_r)\left(D(\omega_r) - \sum_{m=0}^{N/2} A(m)\cos m\omega_r\right) = (-1)^r\delta \quad r = 0, 1, \ldots (N/2+1) \tag{5.175}$$

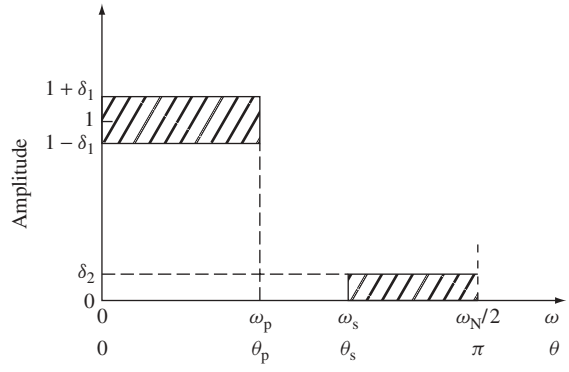


Figure 5.22 Tolerance scheme for low-pass filter design

where the unknowns are the coefficients $A(m)$ and the maximum error is δ. The set of equations (5.175), can be expressed as

$$\begin{bmatrix} 1 & \cos\omega_0 & \cos 2\omega_0 & \cdots & \cos[(N/2)\omega_0] & \dfrac{1}{W(\omega_0)} \\ 1 & \cos\omega_1 & \cos 2\omega_1 & \cdots & \cos[(N/2)\omega_1] & \dfrac{-1}{W(\omega_1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ 1 & \cos\omega_{N/2+1} & \cos 2\omega_{N/2+1} & \cdots & \cos[N/2\,\omega_{N/2+1}] & \dfrac{(-1)^{N/2+1}}{W(\omega_{N/2+1})} \end{bmatrix}$$

$$\times \begin{bmatrix} A(0) \\ A(1) \\ \vdots \\ A(N/2) \\ \delta \end{bmatrix} = \begin{bmatrix} D(\omega_0) \\ D(\omega_1) \\ \vdots \\ D(\omega_{N/2+1}) \end{bmatrix} \tag{5.176}$$

Now, if the extremal frequencies $\omega_r (r = 0, 1, \ldots, N/2 + 1)$ and the filter degree were known, then the solution of (5.174) would give the filter coefficients. However, none of these is known in advance; therefore an iterative procedure is used for the solution for the coefficients. This is based on the *Remez exchange algorithm* which consists of the following stages:

(i) An initial guess is made of the filter degree and the extremal frequencies ω_r.
(ii) The corresponding value of δ is calculated by solving the system of equations (5.176). This gives

$$\delta = \frac{\displaystyle\sum_{i=0}^{N/2+1} c_i D(\omega_i)}{\dfrac{c_0}{W(\omega_0)} - \dfrac{c_1}{W(\omega_1)} + \cdots + \dfrac{(-1)^{N/2+1} c_{N/2+1}}{W(\omega_{N/2+1})}} \tag{5.177}$$

where

$$c_i = \sum_{\substack{i=0 \\ i \neq k}}^{N/2+1} \frac{1}{\cos \omega_k - \cos \omega_i}. \tag{5.178}$$

(iii) The Lagrange interpolation formula is used to interpolate $\hat{H}(\omega)$ on the points ω_r to the values

$$B_k = D(\omega_k) - (-1)^k \frac{\delta}{W(\omega_k)} \quad k = 0, 1, \ldots (N/2 + 1) \tag{5.179}$$

and

$$\hat{H}(\omega) = \frac{\sum_{k=0}^{N/2+1} \left(\frac{\beta_k}{x - x_k} \right) B_k}{\sum_{k=0}^{N/2+1} \left(\frac{\beta_k}{x - x_k} \right)} \tag{5.180}$$

$$\text{where } \beta_k = \sum_{\substack{i=0 \\ i \neq k}}^{N/2+1} \frac{1}{(x_k - x_i)} \tag{5.181}$$

and

$$x = \cos \omega \tag{5.182}$$

(iv) The error $\varepsilon(\omega)$ is calculated on a dense set of frequencies, and is compared with δ. If $\varepsilon(\omega) < \delta$ for all frequencies in the dense set, then the optimum solution has been found. If not, then another set of frequencies ω_r must be chosen as another guess for the extremal frequencies. The new points are chosen as the peaks of the resulting error curve, and the process is iterated until δ converges to its upper bound.

(v) The optimum values of the external frequencies, together with the corresponding δ are used to calculate the filter coefficients $A(m)$ from (5.176). In fact, this can be done without matrix inversion, using another algorithm called the *fast Fourier transform* which will be discussed in a later chapter.

Now, the convergence of the algorithm described above depends on the initialization process, and a computer program is available for implementing all the steps in the design of all types of filter. Furthermore, recent research has led to accelerated procedure in the design, by improving the initialization process.

Since the design of these optimum equiripple filters is always performed using computer-aided design programs, the method itself will not be discussed any further. However, we note that the technique assumes an *a priori* knowledge of the filter degree. This of course is never the case, and the degree N is also to be determined. For the low-pass specifications shown in Figure 5.22 the following empirical estimate can be used

$$N_{\text{est}} = \frac{2}{3} \log \left(\frac{1}{10\delta_1\delta_2} \right) \frac{\omega_{\text{N}}}{\Delta\omega} \tag{5.183}$$

For high-pass filters, the above estimate can also be used. For band-pass designs, (5.183) can be used with $\Delta\omega = \min(\Delta\omega_1, \Delta\omega_2)$, where $\Delta\omega_{1,2}$ is the width of the lower (upper) transition band. It must be noted that (5.183) is only an estimate and more precise formulae may be found in the references. Nevertheless, this estimate can be used as an initial guess in conjunction with the Remez exchange algorithm program. As we shall see, MATLAB® can be used to alleviate the numerical calculations.

5.6 Comparison Between IIR and FIR Filters

In deciding between the two main categories of filters, the designer must keep in mind a number of points:

1. FIR filters can be easily designed to possess a constant group delay response by imposing the trivial constraints of symmetry or antimetry on the impulse response. For IIR filters to possess amplitude selectivity as well as approximating to a constant group delay, much more complex approximation methods are required. It is observed that the constant group delay requirement is important in many applications such as speech processing and data transmission.
2. The non-recursive realizations of FIR filters are inherently stable.
3. Errors due to round-off (to be discussed in a later chapter) can be made small for non-recursive FIR structures.
4. Imposing the symmetry or antimetry constraints to capture the exact linear phase property of FIR filters, also results in simplifications in the realizations.
5. For the same amplitude specifications, a much higher degree FIR filter is required by comparison with an IIR filter.
6. Amplitude-oriented IIR filters can be designed from analog prototypes, hence the wealth of material available in the area of analog filters can be exploited directly. This does not require any optimization or numerical algorithms as in the case of FIR filters with optimum equiripple characteristics.

It follows that the choice of the particular category of filter is determined by the specific application, the number of filters to be designed and the facilities available to the designer.

5.7 Use of MATLAB® for the Design of Digital Filters

The filter transfer functions derived and studied in the previous section, can be easily generated using the *Signal Processing Toolbox* in MATLAB®. The relevant commands and functions are given in this section and may be used to reinforce the readers knowledge and as a labour-saving device. These functions generate the transfer function from the filter specifications and give the result in either of two forms:

1. The poles and zeros of the filter transfer function in the form [z,p,k]. In this case, MATLAB® returns a column matrix [z] with the zero values, a column matrix [p] with the pole values and a value k of the gain of the filter.
2. The coefficients of the numerator and denominator are given in the form [a,b] where [a] is an array of the numerator coefficients in ascending powers of z^{-1}, while [b] is an array of the denominator coefficients in the same order with the first coefficient as 1.

5.7.1 Butterworth IIR Filters

The required degree of the filter is calculated from

$$[\text{n,wn}] = \text{buttord(wp,ws,}\alpha_p, \alpha_s)$$

where for a low-pass and high-pass filters, wp and ws are scalars. For band-pass and band-stop filters each is a vector of two numbers [wp1 wp2] and [ws1 ws2] denoting the two passband edges and the two stopband edges, respectively. Wn is the normalized frequency or frequency vector to be used in the generation of the transfer function. The normalization is with respect to half the sampling frequency.

Low-pass with Normalized Cutoff at wn

```
[z,p,k]=butter(n,wn)
[a,b]=butter(n,wn)
```

High-pass with Normalized Cutoff at wn

```
[z,p,k]=butter(n,wn,'high')
[a,b]=butter(n,wn,'high')
```

Band-pass with wn = [w1 w2] a normalized frequency vector defined by the passband edge frequencies w1 and w2

```
[z,p,k]=butter(n,wn,'bandpass')
[a,b]=butter(n,wn,'bandpass')
```

Band-stop with wn = [w1 w2] a normalized frequency vector defined by the stopband edge frequencies w1 and w2

```
[z,p,k]=butter(n,wn,'stop')
[a,b]=butter(n,wn,'stop')
```

Example 5.4 Design a low-pass Butterworth filter with the following specifications:

1. Passband edge at 5200 Hz with maximum attenuation of 3 dB;
2. Stopband edge at 6000 Hz, with minimum attenuation of 30 dB;
3. Sampling frequency of 16000 Hz.

[n,wn]=buttord(wp,ws,α_p, α_s)

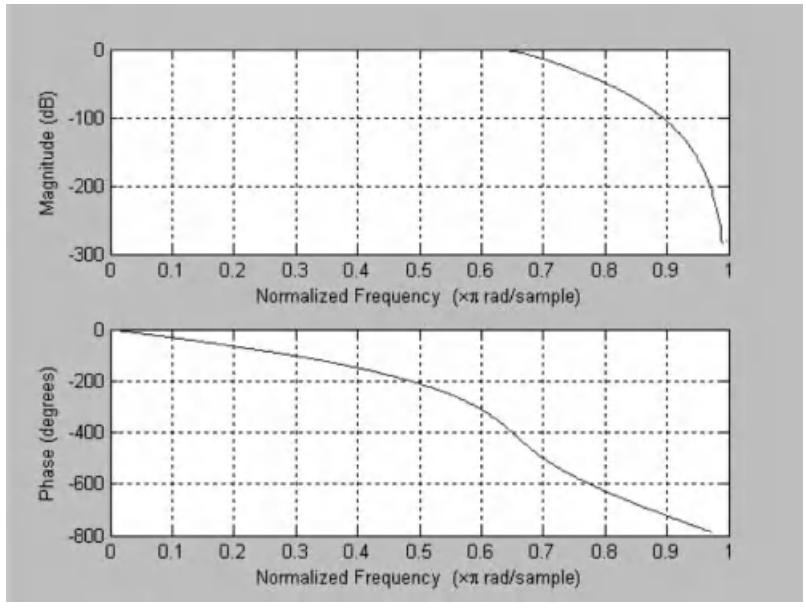
which gives n = 9,wn = 0.6523. Then the coefficients of the transfer function are obtained from

[a,b]=butter(n,wn)

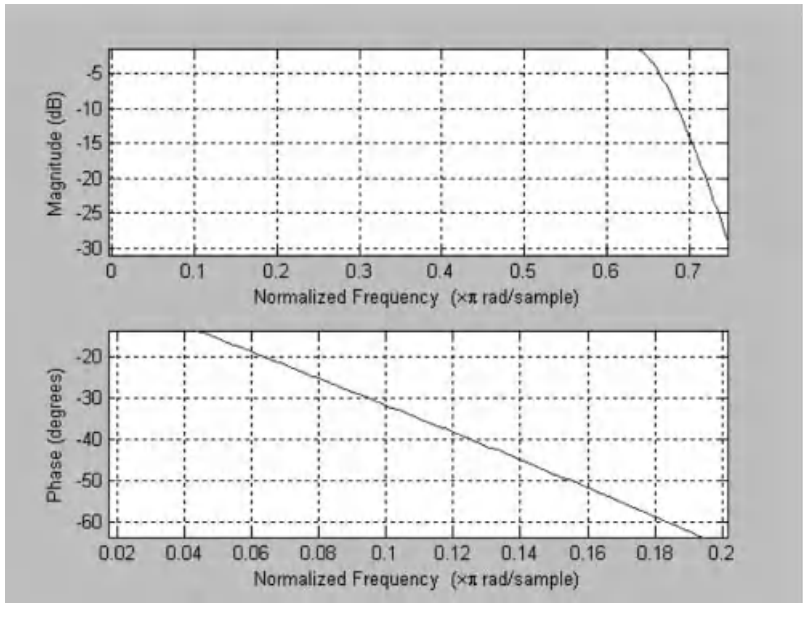
[a] = [0.0344 0.3097 1.2390 2.8910 4.3365 4.3365 2.8910 1.2390 0.3097 0.0344]
[b] = [1.0000 2.7288 4.2948 4.3241 3.0597 1.5299 0.5380 0.1268 0.0181 0.0012]

which are in ascending order in z^{-1}.

The frequency response of the filter can be obtained from freqz(a,b) which gives the plots shown in Figure 5.23.



(a)



(b)

Figure 5.23 Responses of the filter in Example 5.7

5.7.2 Chebyshev IIR Filters

The required degree of the filter is calculated from

[n,wn]=cheb1ord(wp,ws,α_p, α_s)

where for low-pass and high-pass filters, wp and ws are scalars. For band-pass and band-stop filters each is a vector of two numbers [wp1 wp2] and [ws1 ws2] denoting the two passband edges and the two stopband edges, respectively. Wn is the normalized frequency or frequency vector to be used in the generation of the transfer function. In all cases the frequencies are normalized with respect to half the sampling frequency.

Low-pass with Normalized Cutoff at wn

[z,p,k]=cheby1(n,α_p, wn)
[a,b]=cheby1(n,α_p, wn)

High-pass with Normalized Cutoff at wn

[z,p,k]=cheby1(n,α_p, wo,'high')
[a,b]=cheby1(n,α_p, wo,'high')

Band-pass with wn = [w1 w2] a Normalized Frequency Vector Defined by the Pass-band Edge Frequencies w1 and w2

[z,p,k]=cheby1(n,α_p, wn'bandpass')
[a,b]=cheby1(n,α_p, wn,'bandpass')

Band-stop with wn = [w1 w2] a Normalized Frequency Vector Defined by the Stop-band Edge Frequencies w1 and w2

[z,p,k]=cheby1(n,α_p, wn,'stop')
[a,b]=cheby1(n,α_p, wn,'stop')

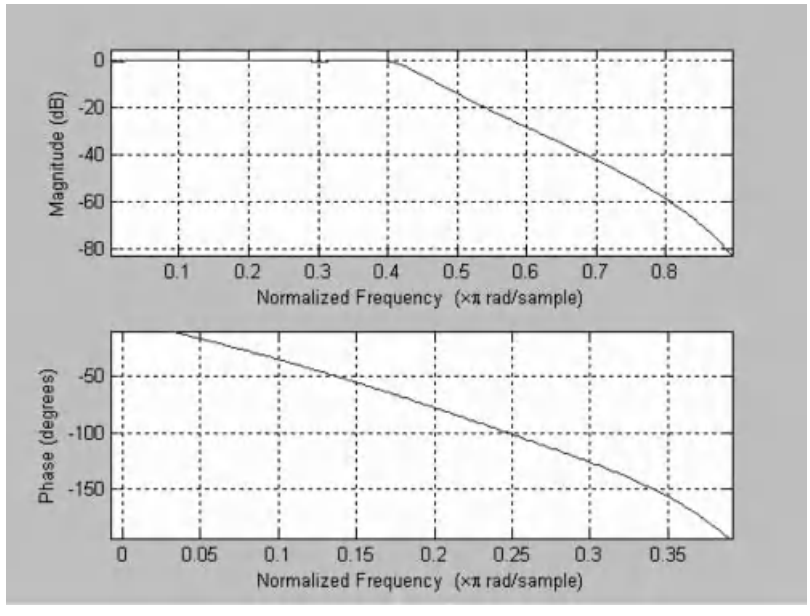
Example 5.5 Design a low-pass Chebyshev filter with the following specifications:

1. Passband edge at 1000 Hz with maximum attenuation of 0.05 dB;
2. Stopband edge at 2000 Hz, with minimum attenuation of 40 dB;
3. Sampling frequency of 16000 Hz.

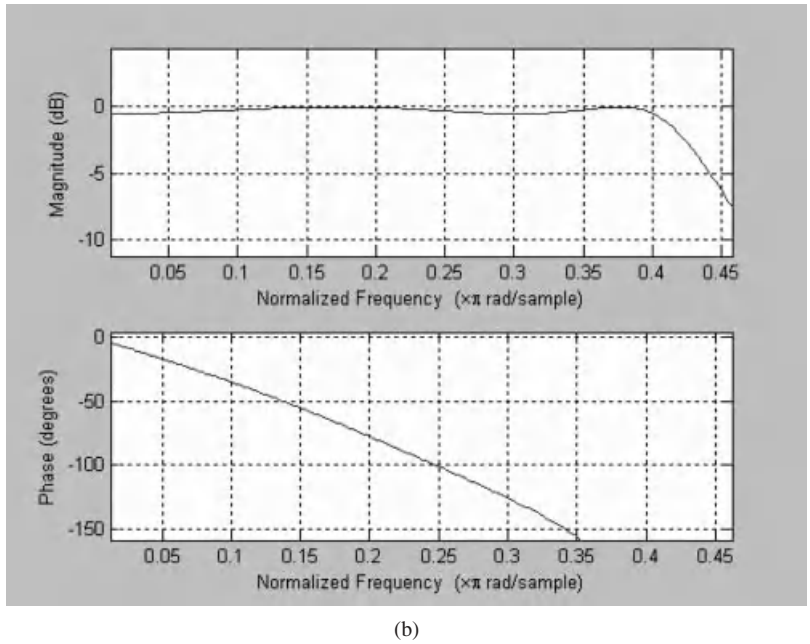
[n,wn]=cheb1ord(wp,ws,α_p, α_s)
n = 6
wn = 0.1200
[a,b]=cheby1(n,α_p, wn)

[a,b]=cheby1(n,0.05,wn)
[a] = [1.0e−003 0.0092 0.0550 0.1376 0.1834 0.1376 0.0550 0.0092]
[b] = [1.0000 −5.0692 10.9243 −12.7889 8.5675 −3.1114 0.4783]

See Figure 5.24 for the responses of the filter.



(a)



(b)

Figure 5.24 Responses of the filter in Example 5.8

5.7.3 Elliptic IIR Filters

The required degree of the filter is calculated from

[n,wn]=ellipord(wo,ws,α_p, α_s)

where for low-pass and high-pass filters, wo and ws are scalars. For band-pass and band-stop filters each is a vector of two numbers [wp1 wp2] and [ws1 ws2] denoting the two passband edges and the two stopband edges, respectively. Wn is the normalized frequency or frequency vector to be used in the generation of the transfer function. In all cases the frequencies are normalized with respect to half the sampling frequency.

Low-pass with Normalized Cutoff at wn

[z,p,k]=ellip(n,α_p, α_s, wn)
[a,b]=ellip(n,α_p, α_s, wn)

High-pass with Normalized Cutoff at wn

[z,p,k]=ellip(n,α_p, α_s, wn,'high')
[a,b]=ellip(n,α_p, α_s, wn,'high')

Band-pass with wn = [w1 w2], a Normalized Frequency Vector Defined by the Passband Edge Frequencies w1 and w2

[z,p,k]=ellip(n,α_p, α_s, wn,'bandpass')
[a,b]=ellip(n,α_p, α_s, wn,'bandpass')

Band-stop with wn = [w1 w2], a Normalized Frequency Vector Defined by the Stopband Edge Frequencies w1 and w2

[z,p,k]=ellip(n,α_p, α_s, wn,'stop')
[a,b]=ellip(n,α_p, α_s, wn,'stop')

Example 5.6 Design an elliptic band-pass IIR filter with the following specifications:

1. Passband: 1000–2000 Hz, maximum attenuation of 0.1 dB;
2. Stopband edges at 500 Hz and 8000 Hz, with minimum attenuation of 30 dB;
3. Sampling frequency = 20 kHz.

The order and normalized frequency vector are obtained from

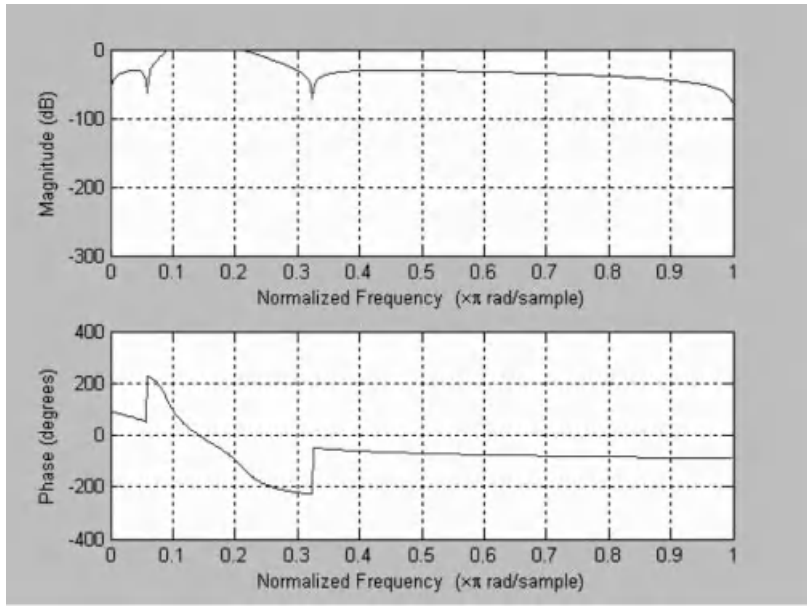
[n,wn]=ellipord(wo,ws,α_p, α_s)
[n,wn]=ellipord([1000 2000]/10000[500 8000]/10000,0.1,30)

n = 3
wn = 0.1000 0.2000
[a,b]=ellip(n,α_p, α_s, wn)
[a] = [0.0308 −0.0929 0.0943 0.0000 −0.0943 0.0929 −0.0308]
[b] = [1.0000 −4.8223 10.1818 −12.0124 8.3481 −3.2415 0.5518]

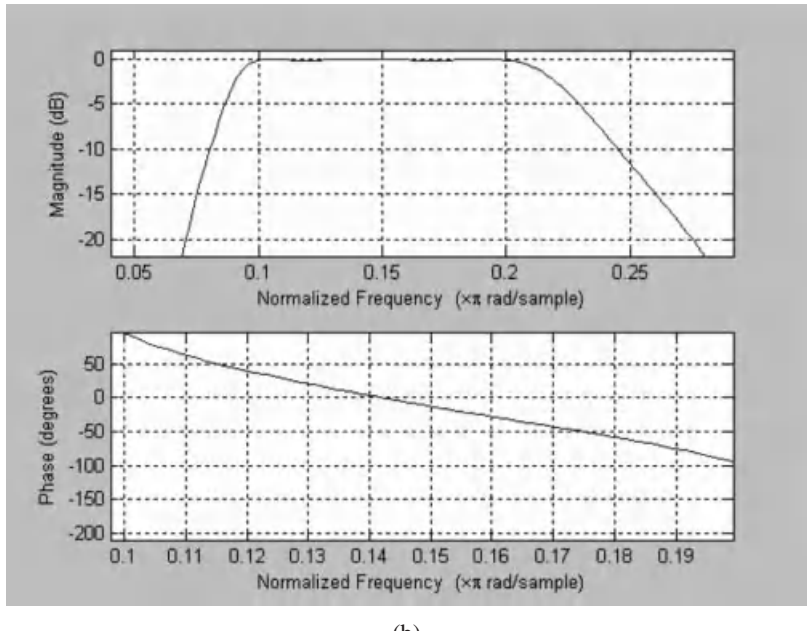
The frequency response of the filter is plotted using

```
Freqz(a,b)
```

and is obtained as in Figure 5.25.

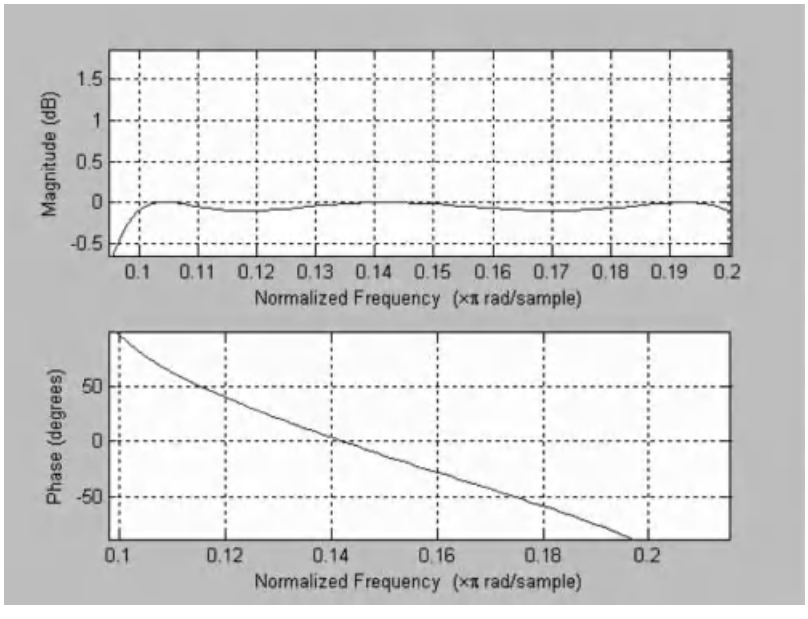


(a)



(b)

Figure 5.25 Responses of the filter in Example 5.9



(c)

Figure 5.25 (*continued*)

5.7.4 Realization of the Filter

From the transfer function a realization in cascade form can be obtained by the function

```
[sos,g]=zp2sos(z,p,k)
```

where

$$sos = \begin{bmatrix} a_{01} & a_{11} & a_{21} & b_{01} & b_{11} & b_{21} \\ a_{02} & a_{12} & a_{13} & b_{03} & b_{12} & b_{22} \\ . & . & . & . & . & . \\ a_{0k} & a_{1k} & a_{2k} & b_{0k} & b_{1k} & b_{2k} \end{bmatrix}$$

in which, each row gives the numerator and denominator of each second-order section. g gives the overall gain of the filter.

5.7.5 Linear Phase FIR Filters

The design using windows can be accomplished using

```
B=fir1(n,wn,'filtertype',window)
```

Where 'filtertype' is either low, high, band or stop. The window is chosen as hamming, Kaiser and so on.

The optimum equiripple design can be obtained using

```
[n,fo,ao,wt]=remezord(f,a,dev,fs)
```

where f is a vector of the edge frequencies defining the pass and stop bands over which the desired values are defined by another vector a and dev is a vector of the maximum deviations δ_1, δ_2 from the desired values in the pass and stop bands. This function returns the required degree, fo, ao and weight wt. Then use the results to find the transfer function vector from

```
a=remez(n,fo,ao,wt)
```

Example 5.7 A low-pass linear-phase FIR Filter with the following specifications:

1. Passband edge at 1000 Hz, with maximum deviation from unity of 0.01;
2. Stopband edge at 2000 Hz, with maximum deviation from zero of 0.001;
3. Sampling frequency of 8000 Hz.

```
[n,fo,ao,wt]=remezord([1000 2000],[1 0],[0.01 0.001],8000)
```

n = 19
fo = 0 0.2500 0.5000 1.0000
ao = 1 1 0 0
wt = 1 10
Then the filter transfer function is obtained from

```
a=remez(n,fo,ao,wt)
```

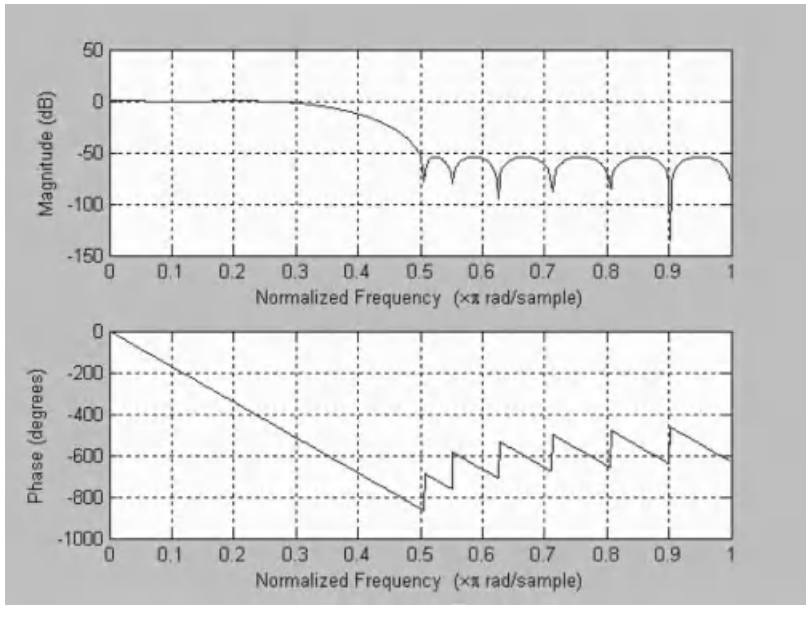


Figure 5.26 Responses of the filter in Example 5.10

as

[a]=[−0.0017 0.0033 0.0150 0.0188 −0.0057 −0.0478 −0.0511 0.0394 0.2033 0.3362 0.3362 0.2033 0.0394 −0.0511 −0.0478 −0.0057 0.0188 0.0150 0.0033 −0.0017]

which are the coefficients of the transfer function in ascending order of z^{-1}.

The amplitude and phase responses are then plotted with the function

freqz(a,128,4000)

to give Figure 5.26.

The detailed step by step design methods discussed so far are instructive and help reinforce the understanding of the material and therefore these are preferred for educational purposes. Having mastered the techniques of filter design, practising filter design engineers can use the much quicker *Filter Analysis and Design Tool* which hides all the detailed steps from the user and produces the design in a single step from the specifications. Once this tool is launched, it will show a tolerance scheme which is easily interpreted in terms of our nomenclature. The key values are entered and the design is obtained directly together with plots of the resulting amplitude and phase responses.

5.8 A Comprehensive Application: Pulse Shaping for Data Transmission

5.8.1 Optimal Design

The derivation of the transfer function of a data transmission IIR digital filter is accomplished such that the same conditions (a) and (b) of Section 3.11 are satisfied [17]. Due to the one to one correspondence between the s-domain and λ-domain, we use the bilinear variable. Let the transfer function be written as

$$H(\lambda) = \frac{P_{2m}(\lambda)}{Q_n(\alpha|\beta|\lambda)} \tag{5.184}$$

where Q_n is the equidistant linear phase polynomial. It can be generated by the recurrence formula

$$Q_{n+1}(\alpha|\beta|\lambda) = Q_n(\alpha|\beta|\lambda) + \gamma_n[\lambda^2 + \tan^2(n\alpha)]Q_{n-1}(\alpha|\beta|\lambda) \tag{5.185}$$

$$\gamma_n = \frac{\cos(n-1)\alpha \cos(n+1)\alpha \sin(\beta - n)\alpha \sin(\beta + n)\alpha \cos^2 n\alpha}{\sin(2n-1)\alpha \sin(2n+1)\alpha \cos^2 \beta\alpha} \tag{5.186}$$

$$Q_0(\alpha|\beta|\lambda) = 1 \tag{5.187}$$

$$Q_1(\alpha|\beta|\lambda) = 1 + \frac{\tan(\alpha\beta)}{\tan\alpha}\lambda \tag{5.188}$$

and

$$\beta\omega_i - \arg\{Q_n(\alpha|\beta|j\tan(\pi\omega_i/\omega_N)\} = 0 \quad \omega_i = i\alpha \text{ for } i = 0,1,\ldots n \tag{5.189}$$

The necessary and sufficient conditions for stability are

$$\alpha < \frac{\pi}{2n}, \beta > n-1, (\beta + n - 1)\alpha < \pi \tag{5.190}$$

Next the numerator of the transfer function must be obtained such that condition (b) is satisfied. A least-square fitting program using MATLAB© can be used to determine the coefficients of the numerator. We can also add an extra constraint σ on the coefficients to force the value of the slope of the magnitude of the transfer function at the midpoint ω_1. This will control the selectivity of the filter.

Example 5.8 As an example, Figure 5.27 shows the impulse and frequency responses for a 10th degree IIR digital impulse-shaping filter with

$$m = 4, \omega_1 = \pi/8, \alpha = \pi/40 = 0.0785, \beta = 16, \sigma = 3.328$$

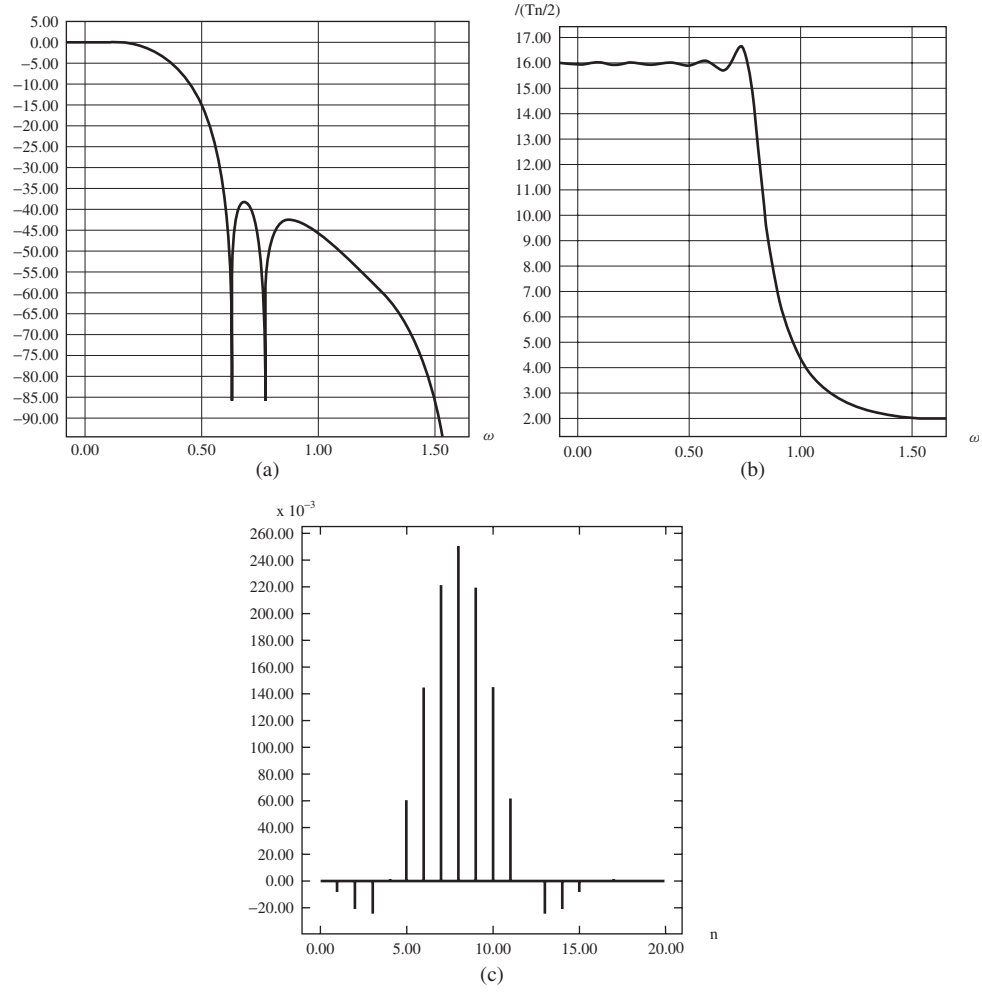


Figure 5.27 Responses of the data transmission filter in Example 5.11: (a) amplitude, (b) group delay, (c) impulse response

The transfer function is given by

$$H(z) = \frac{N(z)}{D(z)}$$

with

$$\begin{aligned} N(z) &= 4.619617 - 281.0891z + 296.0831z^2 + 113.5679z^3 + 44.70272z^4 \\ &\quad + 874.0432z^5 - 44.7.70272z^6 + 113.5679z^7 + 296.0831z^8 \\ &\quad - 281.0891z^9 + 4.619637z^{10} \\ D(z) &= 1489.422 - 12945.68z + 55727.74z^2 - 156511.2z^3 + 318069.5z^4 \\ &\quad - 490121.3z^5 + 582466z^6 - 530424.3z^7 + 357297.9z^8 \\ &\quad - 162687.2z^9 + 38663.07z^{10} \end{aligned}$$

In this example, the maximum overshoot is 0.077 dB, the minimum attenuation is 38 dB, the sampling error is 1.86×10^{-5} and the first side lobe size is equal to 10% of the main lobe.

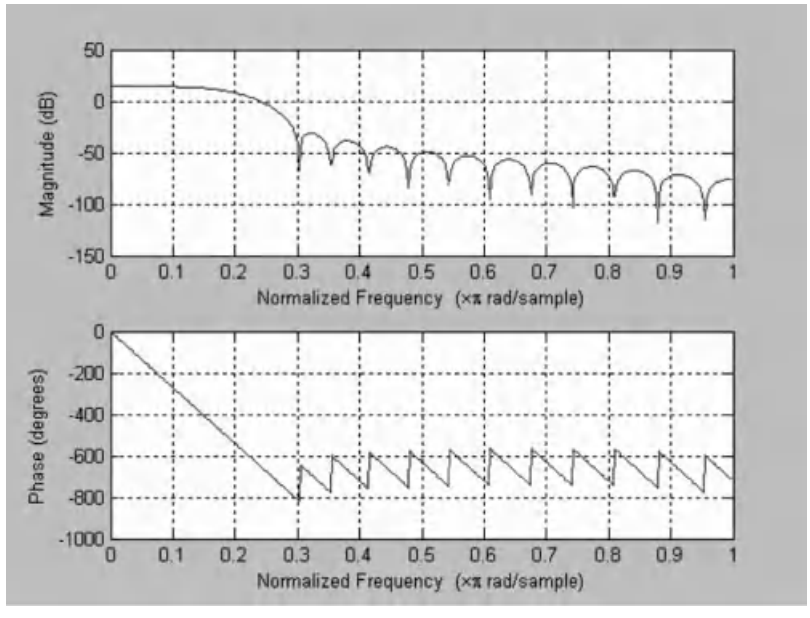
5.8.2 *Use of MATLAB® for the Design of Data Transmission Filters*

The MATLAB® *Communications Toolbox* can also be used to design FIR and IIR data transmission filters based on the raised cosine characteristic. The functions are:

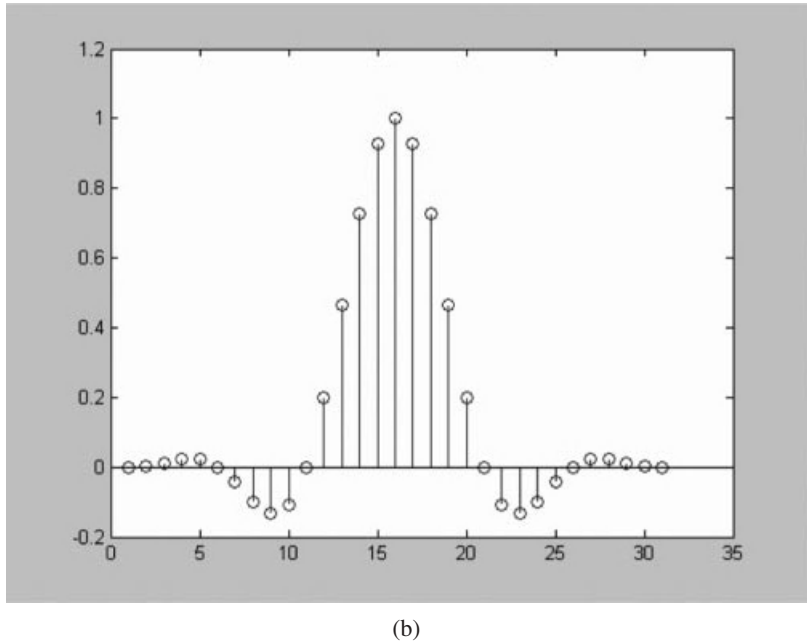
1. num = rcosine(Fd,Fs);
2. [num,den] = rcosine(Fd,Fs,type_flag);
3. [num,den] = rcosine(Fd,Fs,type_flag,r);
4. [num,den] = rcosine(Fd,Fs,type_flag,r,delay);
5. [num,den] = rcosine(Fd,Fs,type_flag,r,delay,tol).

The first function gives an FIR filter. The others allow the specifications of the filter type and input–output delay value. The filter type can be fir/normal, fir/sqrt, iir/normal or iir/sqrt. The "sqrt" category basically designs two filters each with a response approximating the square root of the raised cosine function, one for the transmitter and the second for the receiver end. This is because it has been shown that this configuration gives the best signal to noise ratio. The overall response from transmission to receiver is of course still a raised cosine. r is the roll-off factor and is greater than 1. F_d is the input signal sampling frequency while F_s is the sampling frequency of the filter, the latter must be greater than the former.

Example 5.9 The following is an example of the use of MATLAB® for the design of an FIR data transmission filter, the default order is 30 (the responses are in Figure 5.28).



(a)



(b)

Figure 5.28 Responses of the filter in Example 5.9: (a) frequency response, (b) impulse response

```
num = rcosine(200,1000)
num =
        Columns 1 through 5
        0.0000 0.0030 0.0119 0.0214 0.0211
        Columns 6 through 10
        −0.0000 −0.0441 −0.0981 −0.1324 −0.1095
        Columns 11 through 15
        0.0000 0.2008 0.4634 0.7289 0.9268
        Columns 16 through 20
        1.0000 0.9268 0.7289 0.4634 0.2008
        Columns 21 through 25
        0.0000 −0.1095 −0.1324 −0.0981 −0.0441
        Columns 26 through 30
        −0.0000 0.0211 0.0214 0.0119 0.0030
        Column 31
        0.0000
>> stem(num)
>> freqz(num)
```

5.9 Conclusion

This chapter has dealt with the design of digital filters. These filters have replaced analog continuous-time filters in many applications such as radar, seismic exploration, analysis of vibrations, analysis of biomedical signals and sonar. This is due to many factors such as reliability, reproducibility, high precision and freedom from aging and temperature effects of digital hardware. Furthermore, the decreasing cost and increasing speed of operation of digital hardware, as well as the improvements in computational algorithms and software have contributed to the establishment of digital filters as viable alternatives to analog ones. Other situations which require the use of digital filters are where it is necessary to vary the characteristics of the filter during operation, such as speech processing. Finally a single digital filter can be used to operate in several channels simultaneously; the channels share the same arithmetic elements but each requires its own memory register. MATLAB® has emerged as a powerful tool for computer-aided design of digital filters, and the chapter dealt with the use of this software giving several examples. The chapter concluded with an application in the design of Nyquist data transmission filters of both the FIR and IIR types.

Problems

5.1 Design a low-pass maximally flat IIR filter with the following specifications:
1. Passband: 0–1.5 kHz, attenuation $\leq$0.8 dB;
2. Stopband: 2.0–3.5 kHz, attenuation $\geq$30 dB.

Assume critical sampling, and obtain the realization in parallel form.

5.2 Design a low-pass Chebyshev IIR filter with the following specifications:
1. Passband: 0–3.2 kHz, 0.25 dB ripple;

2. Stopband edge = 4.6 kHz with attenuation ≥50 dB;
3. Sampling frequency: 8 kHz.

Obtain the realization in cascade form.

5.3 Design a band-pass Chebyshev IIR filter with the following specifications:
1. Passband: 300–3200 Hz, 0.25 dB ripple;
2. Stopband edge frequencies = 100 Hz and 4.2 kHz with minimum attenuation of 32 dB;
3. Sampling frequency = 10 kHz.

5.4 Design a band-stop Chebyshev IIR filter with the following specifications:
1. Passband edge frequencies: 1 and 3 kHz with 0.5 dB ripple;
2. Stopband: 1.2–1.5 kHz, with minimum attenuation of 20 dB;
3. Sampling frequency = 8 kHz.

5.5 Use the Fourier-coefficient design method to obtain an FIR low-pass filter approximating the ideal characteristic with a nominal cut off frequency at 3.2 kHz, stopband edge at 4.6 kHz, a sampling frequency of 8 kHz and stopband attenuation of 40 dB. Obtain the design once using the Hamming window and a second time using the Kaiser window. The filters are to possess the constant delay property in addition to satisfying the amplitude specifications.

6

The Fast Fourier Transform and its Applications

6.1 Introduction

In representing an analog periodic signal $f(t)$, with period T, by a Fourier series we write

$$f(t) = \sum_{k=-\infty}^{\infty} c_k \exp(\mathrm{j}k\omega_0 t) \tag{6.1}$$

where

$$c_k = \frac{1}{T}\int_{-T/2}^{T/2} f(t)\exp(-jk\omega_0 t)\,\mathrm{d}t \tag{6.2a}$$

or

$$c_k = \frac{1}{T}\int_{0}^{T} f(t)\ \exp(-jk\omega_0 t)\,\mathrm{d}t \tag{6.2b}$$

and

$$\omega_0 = 2\pi/T. \tag{6.3}$$

In contrast, a non-periodic signal $f(t)$ has a Fourier transform given by

$$F(\omega) = \int_{-\infty}^{\infty} f(t)\ \exp(-\mathrm{j}\omega t)\,\mathrm{d}t. \tag{6.4}$$

In the case of the Fourier series, the evaluation of the integral in (6.2) which defines the coefficients c_k, is usually performed using numerical techniques. This allows the use of efficient high speed computational methods. Therefore, the *integral* in (6.2) must be approximated by a *summation*, since the computer can only process numbers at *discrete* values of the variable t.

Signal Processing and Integrated Circuits, First Edition. Hussein Baher.
© 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.

Furthermore, the representation of $f(t)$ by the Fourier series must be done using only a *finite number of terms* so that we use the nth partial sum, or truncated series

$$f_n(t) = \sum_{k=-n}^{n} c_k \exp(\mathrm{j}k\omega_0 t) \tag{6.5}$$

to approximate the function.

In the case of the Fourier transform $F(\omega)$ in (6.4), we must evaluate the integral between *finite limits* and also approximate it by a *summation* so that it may be evaluated *numerically*.

In this chapter, we discuss concepts which allow the formulation of efficient algorithms for the calculation of the Fourier coefficients and the Fourier transform [11, 12, 17]. These constitute a family of high speed efficient algorithms which have become known collectively as *fast Fourier transform* (FFT) algorithms and are incorporated in software for use with digital computers. Once these algorithms have been incorporated into signal processing techniques, they can be used for a wide variety of other applications such as fast convolution, correlation, filtering and measurement of power spectra. These applications will be also discussed in this chapter.

6.2 Periodic Signals

Consider a periodic signal $f(t)$ and divide the period T into N equally spaced subintervals each of duration T_0, that is

$$T_0 = T/N. \tag{6.6}$$

Suppose we take samples of the function $f(t)$ at the points $0, T_0, 2T_0, \ldots (N-1)T_0$, thus forming a sequence of numbers

$$\{f(mT_0)\} \triangleq \{f(0), f(T_0), \ldots f(mT_0), \ldots f(NT_0\,T_0)\} \tag{6.7}$$

which may be written in the alternative form

$$\{f(m)\} \triangleq \{f(0), f(1), f(2), \ldots, f(m), \ldots f(N-1)\} \tag{6.8}$$

where the curly brackets denote a sequence while $f(mT_0)$ without the brackets denotes the mth sample of $f(t)$. So, we have generated a function of the *discrete* time variable (mT_0) which is defined only at these instants. Using this sequence, we can approximate the *integral* in (6.2) defining the Fourier coefficients c_k by a *summation*. This is achieved by letting

$$\begin{aligned}
&\mathrm{d}t \to T_0 \\
&T \to NT_0 \\
&t \to (mT_0) \quad m = 0, 1, \ldots (N-1) \\
&f(t) \to \{f(mT_0)\} \\
&\int_0^T \to \sum_{m=0}^{(N-1)}
\end{aligned} \tag{6.9}$$

so that the approximate expression of the Fourier coefficient becomes

$$c_k = \frac{1}{NT_0} \sum_{m=0}^{N-1} f(mT_0)[\exp(-\mathrm{j}k\omega_0 m)]T_0 \tag{6.10}$$

or

$$c_k = \frac{1}{N} \sum_{m=0}^{N-1} f(mT_0)\ \exp\left(-\mathrm{j}\frac{2\pi km}{N}\right). \tag{6.11}$$

To simplify the notation we put

$$w = \exp\left(\mathrm{j}\frac{2\pi}{N}\right) \tag{6.12}$$

so that (6.11) takes the form

$$c_k = \frac{1}{N} \sum_{m=0}^{N-1} f(mT_0)\, w^{-km}. \tag{6.13}$$

Using (6.12) in (6.1) we also obtain the values of the function $f(t)$ at the discrete sample points

$$f(mT_0) = \sum_{k=-\infty}^{\infty} c_k w^{km}. \tag{6.14}$$

Noting that

$$w^N = 1, \quad w^{m(k+iN)} = w^{km} \tag{6.15}$$

for any integer i, we write (6.14) explicitly as

$$\begin{aligned} f(mT_0) = {} & (\cdots + c_{-N} + c_0 + c_N + \cdots) \\ & + (\cdots + c_{-N+1} + c_1 + c_{N+1} + \cdots)w^m + \cdots \\ & + (\cdots + c_{-N+k} + c_k + c_{N+k} + \cdots)w^{km} + \cdots \\ & + (\cdots + c_{-1} + c_{N-1} + c_{2_{N-1}} + \cdots)w^{(N-1)m}. \end{aligned} \tag{6.16}$$

A further simplification in the above expressions results if we define the *aliased* coefficients $\hat{c}_k$ as

$$\hat{c}_k = \sum_{i=-\infty}^{\infty} c_{k+iN} \tag{6.17}$$

which are periodic with period N. Therefore (6.14) takes the form

$$f(mT_0) = \sum_{k=0}^{N-1} \hat{c}_k w^{km} \quad m = 0, 1, \ldots (N-1). \tag{6.18}$$

Now, the sample values of $f(t)$ are obtained by putting $m = 0, 1, 2, \ldots (N-1)$ in (6.18). This gives a set of N linear equations of the form

$$\begin{aligned} f(0) &= \hat{c}_0 + \hat{c}_1 \cdots + \hat{c}_k \cdots \hat{c}_{N-1} \\ f(T_0) &= \hat{c}_0 + \hat{c}_1 w + \cdots + \hat{c}_k w^k + \cdots \hat{c}_{N-1} w^{N-1} \\ f(NT_0 - T_0) &= \hat{c}_0 + \hat{c}_1 w^{N-1} + \cdots + \hat{c}_k w^{(N-1)k} + \cdots c_{N-1} w^{(N-1)^2}. \end{aligned} \tag{6.19}$$

In matrix form the above set of equations becomes

$$\begin{bmatrix} f(0) \\ f(T_0) \\ \vdots \\ f(NT_0 - T_0) \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 & \cdots & 1 \\ 1 & w & \cdots & w^k & \cdots & w^{N-1} \\ 1 & w^2 & \cdots & & & \\ 1 & w^3 & \cdots & \vdots & & \vdots \\ \vdots & \vdots & & & & \\ 1 & w^{(N-1)} & \cdots & w^{(N-1)k} & \cdots & w^{(N-1)^2} \end{bmatrix} \begin{bmatrix} \hat{c}_0 \\ \hat{c}_1 \\ \vdots \\ \hat{c}_k \\ \vdots \\ \hat{c}_{N-1} \end{bmatrix} \tag{6.20}$$

or more concisely

$$[f] = [W_N][\hat{c}]. \tag{6.21}$$

The problem now is to determine the coefficients $\hat{c}_k$ from the sample values $f(mT_0)$. To this end we use the dummy variable i instead of k in (6.18), multiply the mth equation by w^{-km} and add all the equations to obtain

$$\begin{aligned} \sum_{m=0}^{N-1} f(mT_0)\, w^{-km} &= \sum_{m=0}^{N-1} w^{-km} \sum_{i=0}^{N-1} \hat{c}_k w^{im} \\ &= \sum_{i=0}^{N-1} \hat{c}_i \sum_{m=0}^{N-1} w^{m(i-k)}. \end{aligned} \tag{6.22}$$

The second sum on the right is, however, a geometric progression with ratio $w^{(i-k)}$, and since $w^N = 1$ we have

$$\begin{aligned} \sum_{m=0}^{N-1} w^{m(i-k)} = \frac{1 - w^{N(i-k)}}{1 - w^{i-k}} &= N \quad \text{for } i = k \\ &= 0 \quad \text{for } i \neq k \end{aligned} \tag{6.23}$$

so that (6.22) becomes

$$\sum_{m=0}^{N-1} f(mT_0)\, w^{-km} = \hat{c}_k N \tag{6.24}$$

or

$$\hat{c}_k \frac{1}{N} \sum_{m=0}^{N-1} f(mT_0)\, w^{-km} k = 0,\ 1, \ldots (N-1) \tag{6.25}$$

expressing the aliased coefficients in terms of the sample values $f(mT_0)$. In matrix form (6.25) reads

$$\begin{bmatrix} \hat{c}_0 \\ \hat{c}_1 \\ \vdots \\ c_{N-1} \end{bmatrix} = \frac{1}{N} \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & w^{-1} & w^{-2} & \cdots & w^{-(N-1)} \\ 1 & w^{-2} & w^{-3} & \cdots & w^{-(N-2)} \\ \vdots & \vdots & & \vdots & \\ 1 & w^{-(N-1)} & & \cdots & w^{-(N-1)^2} \end{bmatrix} \begin{bmatrix} f(0) \\ f(T_0) \\ \vdots \\ f(NT_0 - T_0) \end{bmatrix} \tag{6.26}$$

or

$$[\hat{c}] = \frac{1}{N}[\hat{W}_N][f]. \tag{6.27}$$

Now the aliased coefficients $\hat{c}_k$ were obtained from c_k using (6.17). However, obtaining c_k from $\hat{c}_k$ is not possible in general since the sum in (6.17) is infinite. We usually, however, represent $f(t)$, as in (6.5), by a truncated series (nth partial sum) so that the sample values are approximated, using (6.14), by

$$f(mT_0) = \sum_{k=-n}^{n} c_k w^{km} \tag{6.28}$$

thus *either* assuming that

$$c_k = 0 \quad \text{for } |k| > n \tag{6.29}$$

or *neglecting* those coefficients with $|k| > n$. Moreover, let n be chosen such that

$$N > 2n \tag{6.30}$$

as illustrated in Figure 6.1.

It follows that, under the assumptions of (6.29) and (6.30) the Fourier coefficients c_k can be determined from the aliased coefficients $\hat{c}_k$ using (6.26) as

$$c_k = \begin{cases} \hat{c}_k & |k| \le n \\ 0 & |k| > n\,. \end{cases} \tag{6.31}$$

Therefore, we conclude that if $f(t)$ is represented by a truncated Fourier series of the form (6.5) with n satisfying (6.30), then the coefficients c_k can be determined from the N sample values $f(mT_0)$ using (6.26) and (6.31).

It is to be noted that condition (6.30) can always be met by taking the number of sample points N to be greater than twice the degree of the truncated Fourier series. In terms of sampling the continuous signal $f(t)$, this amounts to stating that the sampling frequency should be taken sufficiently high so that (6.30) is satisfied. This, of course, is a different way of stating the *sampling theorem* discussed in Chapter 4.

6.3 Non-periodic Signals

Next, consider a non-periodic signal $f(t)$ and its Fourier transform $F(\omega)$ as shown in Figure 6.2.

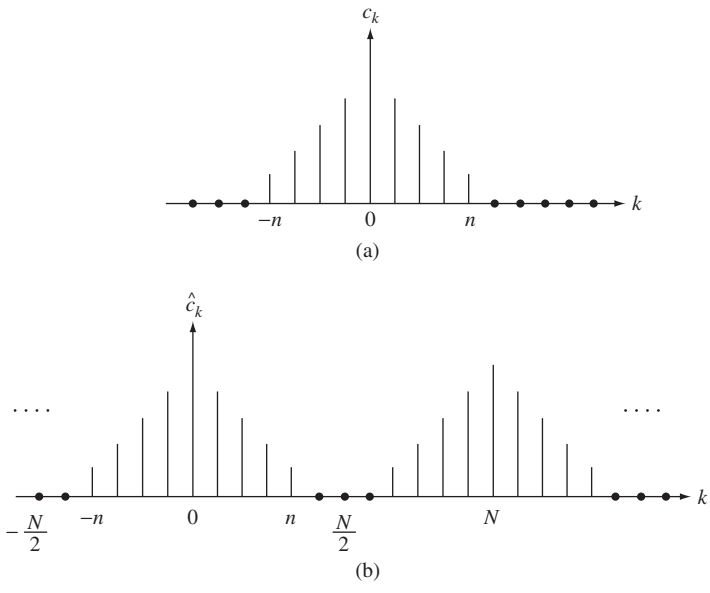


Figure 6.1 (a) The coefficients of the truncated Fourier series of a given function. (b) The aliased coefficients with condition (6.30) satisfied

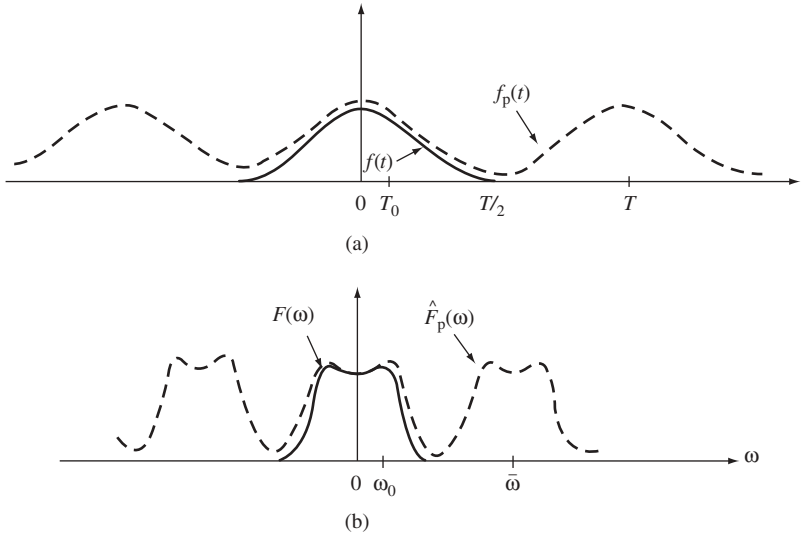


Figure 6.2 (a) A non-periodic function with its extension as defined by (6.32). (b) The spectrum of $f(t)$ and the spectrum defined by (6.43)

Again, we would like to replace the integral in (6.4) by one with finite limits, then approximate the resulting integral with a summation and finally evaluate the sum at a discrete set of ω values. We begin by constructing a periodic function $f_{\mathrm{p}}(t)$ by periodic extension of $f(t)$, as shown in Figure 6.2(a)

$$f_{\mathrm{p}}(t) = \sum_{k=-\infty}^{\infty} f(t + kT) \tag{6.32}$$

and if

$$f(t) \leftrightarrow F(\omega) \tag{6.33}$$

then

$$f_{\mathrm{p}}(t) = \frac{1}{T} \sum_{k=-\infty}^{\infty} F(k\omega_0)\, \exp(\mathrm{j}k\omega_0 t) \tag{6.34}$$

with

$$\omega_0 = 2\pi/T \tag{6.35}$$

and $[(1/T)F(k\omega_0)]$ may be regarded as the Fourier coefficients of the periodic function $f_{\mathrm{p}}(t)$. Consequently, the analysis of the previous section can be applied to $f_{\mathrm{p}}(t)$ with

$$c_k \equiv \frac{1}{T} F(k\omega_0). \tag{6.36}$$

Therefore, let us assume that $F(\omega)$ is band-limited to a certain frequency range, that is

$$F(\omega) = 0 \quad \text{for } |\omega| > \hat{\omega} \tag{6.37}$$

such that

$$F(k\omega_0) = 0 \quad \text{for } |k| > n \tag{6.38}$$

and n is chosen such that

$$n\omega_0 = \hat{\omega}. \tag{6.39}$$

Then, (6.32) is used to approximate $f(t)$ by the truncated series

$$\hat{f}_{\mathrm{p}}(t) = \frac{1}{T} \sum_{k=-n}^{n} F(k\omega_0)\, \exp(\mathrm{j}k\omega_0 t). \tag{6.40}$$

Dividing the interval T into N equally spaced subintervals each of duration T_0, that is

$$T_0 = \frac{T}{N} \tag{6.41}$$

we take N samples of $f_{\mathrm{p}}(t)$ over T and use $w = \exp(j2\pi/N)$ to write (6.40) as

$$\hat{f}_{\mathrm{p}}(mT_0) = \frac{1}{T} \sum_{k=-n}^{n} F(k\omega_0)\, w^{km}. \tag{6.42}$$

Then, we define the function [shown in Figure 6.2(b)]

$$\hat{F}_{\mathrm{p}}(\omega) = \sum_{r=-\infty}^{\infty} F(\omega + r\overline{\omega}) \tag{6.43}$$

which is obtained by periodic extension of $F(\omega)$ every $\overline{\omega}$, where

$$\overline{\omega} = N\omega_0. \tag{6.44}$$

We note that $\hat{F}_{\mathrm{p}}(\omega)$ is *not* the Fourier transform of $f_{\mathrm{p}}(t)$. Next, let us evaluate $F(\omega)$ at a set of frequencies to define the aliased coefficients as

$$\hat{c}_k = \frac{1}{T}\hat{F}_{\mathrm{p}}(k\omega_0). \tag{6.45}$$

If, as in the previous section, we choose

$$N > 2n \tag{6.46}$$

then in this case, (6.39) and (6.44) imply that

$$N\omega_0 > 2n\omega_0 \tag{6.47}$$

or

$$\overline{\omega} > 2\hat{\omega} \tag{6.48}$$

which, in turn, requires $F(\omega)$ to be band-limited to $|\omega| = \hat{\omega} < 2\overline{\omega}$, that is

$$F(\omega) = 0 \quad \text{for } |\omega| > N\omega_0. \tag{6.49}$$

If this condition is satisfied, then the samples $F(k\omega_0)$ can be calculated *exactly* from the samples $\hat{F}_{\mathrm{p}}(k\omega_0)$ by

$$\begin{aligned} F(k\omega_0) &= \hat{F}_{\mathrm{p}}(k\omega_0) \quad \text{for } |k| \le n \\ &= 0 \qquad\qquad\quad \text{for } |k| > n \end{aligned} \tag{6.50}$$

so that we can use (6.45) to write by analogy with (6.28)

$$\hat{f}_{\mathrm{p}} = (mT_0) = \frac{1}{T}\sum_{k=0}^{N-1} \hat{F}_{\mathrm{p}}(k\omega_0)\, w^{km} \tag{6.51}$$

or in matrix form

$$\begin{bmatrix} \hat{f}_{\mathrm{p}}(0) \\ \hat{f}_{\mathrm{p}}(T_0) \\ \vdots \\ \hat{f}_{\mathrm{p}}(NT_0 - T_0) \end{bmatrix} = \frac{1}{T}\left[W_N\right] \begin{bmatrix} \hat{F}_{\mathrm{p}}(0) \\ \hat{F}_{\mathrm{p}}(\omega_0) \\ \vdots \\ \hat{F}_{\mathrm{p}}(N\omega_0 - \omega_0) \end{bmatrix} \tag{6.52}$$

where $[W_N]$ is the same matrix in (6.20).

The above system of equations can be solved for $\hat{F}_{\mathrm{p}}(k\omega_0)$ in terms of $\hat{f}_{\mathrm{p}}(mT_0)$ in the same way as (6.25) was obtained from (6.20) to give

$$\hat{F}_{\mathrm{p}}(k\omega_0) = \frac{T}{N}\sum_{m=0}^{N-1}\hat{f}_{\mathrm{p}}(mT_0)\, w^{-km} \tag{6.53}$$

or

$$\begin{bmatrix} \hat{F}_{\mathrm{p}}(0) \\ \hat{F}_{\mathrm{p}}(\omega_0) \\ \hat{F}_{\mathrm{p}}(\omega_0) \\ \vdots \\ \hat{F}_{\mathrm{p}}(N\omega_0 - \omega_0) \end{bmatrix} = \frac{T}{N}[\hat{W}_N] \begin{bmatrix} \hat{f}_{\mathrm{p}}(0) \\ \hat{f}_{\mathrm{p}}(T_0) \\ \vdots \\ \hat{f}_{\mathrm{p}}(NT_0 - T_0) \end{bmatrix} \tag{6.54}$$

where $[\hat{W}_N]$ is the same matrix in (6.26).

Now, we can make the identification (6.50) only if $F(\omega)$ is band-limited according to (6.49). Then expressions (6.50) and (6.54) yield $F(k\omega_0)$ in terms of $\hat{f}(mT_0)$; that is we obtain the sample values of the Fourier transform of $f(t)$ in terms of the sample values of $\hat{F}_{\mathrm{p}}(t)$. On the other hand if $f(t)$ is not band-limited according to (6.49) but we take N sufficiently large so that $\overline{\omega}$ is, in turn, large enough to allow $F(\omega)$ to be neglected for $|\omega| > \omega_0/2$, that is

$$F(\omega) \approx 0 \quad \text{for } |\omega| > \omega_0/2 \tag{6.55}$$

then

$$F(k\omega_0) \approx \hat{F}_{\mathrm{p}}(k\omega_0) \quad \text{for } |k| < \frac{\overline{\omega}}{2\,\omega_0} \tag{6.56}$$

and we can still use (6.54) to determine $F(k\omega_0)$ from $\hat{f}_{\mathrm{p}}(mT_0)$. However, in this case there will be an error in the approximation

$$\varepsilon = F(k\omega_0) - \hat{F}_{\mathrm{p}}(k\omega_0) \tag{6.57}$$

which is called the *aliasing error*.

6.4 The Discrete Fourier Transform

In the previous two sections, we have seen that the problems of calculating the Fourier coefficients and the Fourier transform, can both be reduced to the problem of calculating a periodic sequence of numbers in terms of another periodic sequence. Therefore, we require efficient computational algorithms for performing the calculations. To this end, we begin by disregarding the origin of the sequences under consideration and introduce some concepts which deal only with the sequences themselves, regardless of how they have been obtained. Consider a periodic sequence of numbers with period N written as

$$\{f(n)\} \triangleq \{f(0), f(1), f(2), \ldots f(n), \ldots f(N-1)\} \tag{6.58}$$

where the curly brackets are used to denote a sequence. We define the *discrete Fourier transform* (DFT) of the sequence $\{f(n)\}$ as another periodic sequence with the same period N

$$\{F(k)\} \triangleq \{F(0), F(1), F(2), \ldots F(n), \ldots F(N-1)\} \tag{6.59}$$

obtained from $\{f(n)\}$ by

$$F(k) = \sum_{n=0}^{N-1} f(n) w^{-nk} \tag{6.60}$$

where

$$w = \exp(\mathrm{j}2\pi/N). \tag{6.61}$$

Conversely, the sequence $\{f(n)\}$ can be expressed in terms of the sequence $\{F(k)\}$ by means of the *inverse discrete Fourier transform* (IDFT) as

$$f(n) = \frac{1}{N} \sum_{k=0}^{N-1} F(k) w^{kn} n = 0, 1, 2, \ldots (N-1). \tag{6.62}$$

The pair of sequences in (6.60) and (6.62) are identical in form to the pair (6.18) and (6.25), so that the proof that (6.62) follows from (6.60) is the same as the proof that (6.25) follows from (6.18). Nevertheless, let us give here an independent proof. Change the index in (6.60) from n to m and substitute into the right-hand side of (6.62)

$$\frac{1}{N} \sum_{k=0}^{N-1} F(k) w^{kn} = \frac{1}{N} \sum_{k=0}^{N-1} w^{kn} \sum_{m=0}^{N-1} f(m) w^{-km} \tag{6.63}$$

that is

$$\frac{1}{N} \sum_{k=0}^{N-1} F(k) w^{kn} = \frac{1}{N} \sum_{m=0}^{N-1} f(m) \left(\sum_{k=0}^{N-1} w^{(n-m)k} \right). \tag{6.64}$$

However

$$\sum_{k=0}^{N-1} w^{(n-m)k} = \frac{1 - w^{(n-m)N}}{1 - w^{n-m}} \tag{6.65}$$

and using

$$w^N = \exp(\mathrm{j}2\pi) = 1 \tag{6.66}$$

we have

$$\frac{1 - w^{(n-m)N}}{1 - w^{(n-m)}} = \begin{cases} 0 & n \neq m \\ N & n = m \end{cases} \tag{6.67}$$

so that (6.64) becomes

$$\begin{aligned} \frac{1}{N} \sum_{k=0}^{N-1} F(k) w^{kn} &= \frac{1}{N} f(n) N \\ &= f(n) \end{aligned} \tag{6.68}$$

as given by (6.62).

Now, the sequences $\{f(n)\}$ and $\{F(k)\}$ are said to form a *discrete Fourier transform* (DFT) *pair of order N*, and this is symbolized by

$$\{f(n)\} \overset{\leftrightarrow}{N} \{F(k)\}. \tag{6.69}$$

The sequences must both be periodic in order that the relations (6.60) and (6.62) should be valid for all k and n, because of the periodicity of w

$$w^{\pm(k+N)n} = \exp\left(\pm \mathrm{j}\frac{2\pi(k+N)n}{N}\right) = w^{\pm kn}. \tag{6.70}$$

Therefore the members of the sequences must be such that

$$F(k+N) = F(k)$$

and

$$f(n+N) = f(n) \tag{6.71}$$

in order for (6.60) to hold for all k and n. Consequently the discrete Fourier transform establishes a one-to-one correspondence between two periodic sequences of period N.

Example 6.1 Consider the sequence

$$\{f(n)\} = \{1, 1, 1, 1, 0, 0, 0, 0, 0, 0\}$$

which has $N = 10$ and is assumed periodic. The discrete Fourier transform of $\{f(n)\}$ is the sequence $\{F(k)\}$ whose members are defined by

$$F(k) = \sum_{n=0}^{9} f(n)w^{-nk} \quad k = 0, 1, \ldots, 9$$

which gives

$$\begin{aligned} F(k) &= 1 \times w^{0k} + 1 \times w^{-k} + 1 \times w^{-2k} + 1 \times w^{-3k} + 0 \times w^{-4k} + 0 \times w^{-5k} \\ &\quad + 0 \times w^{-6k} + 0 \times w^{-7k} + 0 \times w^{-8k} + 0 \times w^{-9k} \\ &= \sum_{n=0}^{4} \exp(-\mathrm{j}2kn\pi/10) = \exp(-\mathrm{j}2k\pi/5)\frac{\sin(k\pi/2)}{\sin(k\pi/10)} \end{aligned}$$

and the sequence $\{F(k)\}$ results as

$$\begin{aligned} \{F(k)\} = \{&5, (1 - \mathrm{j}3.0777), 0, (1 - 0.727), 0, 1, 0, (1 + \mathrm{j}0.727), \\ &0, (1 + 3.0777)\}. \end{aligned}$$

Now, suppose the sequence $\{f(n)\}$ is non-periodic, but instead it is of finite duration N. In this case, we can still represent $\{f(n)\}$ by a periodic sequence of period N, but bearing in mind that the two sequences are equal only over N. This means that we can define the DFT of the finite duration sequence by exactly the same sequence $\{F(k)\}$ as given

by (6.60) with the appropriate interpretation of the expressions, stipulating that they are valid only for

$$n < N$$

and

$$k < N. \tag{6.72}$$

6.5 The Fast Fourier Transform Algorithms

Expressions (6.60) and (6.62) define the DFT of a sequence and the inverse DFT. (6.60) can be written in matrix form as

$$\begin{bmatrix} F(0) \\ F(1) \\ \vdots \\ F(N-1) \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ 1 & w^{-1} & w^{-2} & w^{-3} & \cdots & w^{-(N-1)} \\ 1 & w^{-2} & w^{-4} & w^{-6} & \cdots & w^{-2(N-1)} \\ \vdots & \vdots & \vdots & & \vdots & \\ 1 & w^{-(N-1)} & w^{-2(N-1)} & & \cdots & w^{-(N-1)^2} \end{bmatrix} \begin{bmatrix} f(0) \\ f(1) \\ \vdots \\ f(N-1) \end{bmatrix} \tag{6.73}$$

or

$$[F] = [\hat{w}_N][f]. \tag{6.74}$$

Similarly (6.62) can be put in the form

$$[f] = \frac{1}{N}[W_N][F] \tag{6.75}$$

where $[W_N]$ is obtained from $[\hat{W}_N]$ by changing w^{-k} to w^k. For generality, it can be assumed that the members of the sequences are complex numbers.

We now consider the numerical problems associated with the calculation of the sequence $\{F(k)\}$ from the sequence $\{f(n)\}$, and vice versa, using (6.73) or (6.75). First note that *if we find an efficient computational algorithm for the calculation of the DFT using (6.74), then the IDFT defined by (6.75) can also be calculated using the same algorithm*. This is because (6.74) and (6.75) are identical in form except for replacing w by w^{-1} and the factor N. Therefore, it suffices to consider (6.74).

Now, the matrix $[\hat{W}_N]$ possesses properties that make the calculations of $[F]$ from $[f]$ amenable to considerable simplification. Any algorithm which speeds up these calculations is called a *fast Fourier transform* (FFT) algorithm. In particular, if N is a power of 2, the algorithms become particularly efficient and fast.

It is clear that *direct* calculation of the sequence $\{F(k)\}$ from the sequence $\{f(n)\}$ using (6.73) involves $N(N-1)$ multiplications and a similar number of additions. For N very large, as it is usually the case for an accurate representation of functions, we have approximately $(N-1)^2$ multiplications and a similar number of additions. We now examine the general features of algorithms which reduce the number of multiplications to $(N/2)\log_2(N/2)$ and the number of additions to $N\log_2 N$. The discrete Fourier transform of a sequence of period N is sometimes called an N-point DFT.

6.5.1 Decimation-in-time Fast Fourier Transform

With N a power of 2, decompose the sequence $\{f(n)\}$ into two interleaved sequences, one with even arguments

$$\{f(0), f(2), f(4) \ldots f(N-2)\} \tag{6.76a}$$

and the other with odd arguments

$$\{f(1), f(3), f(5), \ldots f(N-1)\} \tag{6.76b}$$

where N is even, since it is a power of 2.

Using the decomposition in (6.76) we write the first $N/2$ members of the transform sequence $\{F(k)\}$ using (6.73) as

$$\begin{bmatrix} F(0) \\ F(1) \\ F(2) \\ \vdots \\ F\left(\frac{1}{2}N-1\right) \end{bmatrix} = \overbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & w^{-2} & \cdots & w^{-2(N/2-1)} \\ 1 & w^{-4} & \cdots & w^{-4(N/2-1)} \\ \vdots & \vdots & & \vdots \\ 1 & w^{-2(N/2-1)} & \cdots & w^{-2(N/2-1)^2} \end{bmatrix}}^{[\hat{W}_{N/2}]} \begin{bmatrix} f(0) \\ f(2) \\ f(4) \\ \vdots \\ f(N-2) \end{bmatrix}$$
$$+ \underbrace{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ w^{-1} & w^{-3} & \cdots & w^{-(N-1)} \\ w^{-2} & w^{-6} & \cdots & w^{-2(N-1)} \\ \vdots & \vdots & & \\ w^{-(N/2-1)} & w^{-3(N/2-1)} & \cdots & w^{-(N/2-1)(N-1)} \end{bmatrix}}_{[\hat{\hat{W}}_{N/2}]} \begin{bmatrix} f(1) \\ f(3) \\ f(5) \\ \vdots \\ f(N-1) \end{bmatrix}. \tag{6.77}$$

Now, $[\hat{\hat{W}}_{N/2}]$ and $[\hat{W}_{N/2}]$ are related by

$$[\hat{\hat{W}}_{N/2}] = \begin{bmatrix} 1 & & & & \mathrm{O} \\ & w^{-1} & & & \\ & & w^{-2} & & \\ & & & \ddots & \\ \mathrm{O} & & & & w^{-(N/2-1)} \end{bmatrix} [\hat{W}_{N/2}]$$

or

$$[\hat{\hat{W}}_{N/2}] = [\hat{W}_d][\hat{W}_{N/2}] \tag{6.78}$$

so that (6.77) can be written as

$$\begin{bmatrix} F(0) \\ F(1) \\ F(2) \\ \vdots \\ F\left(\frac{1}{2}N-1\right) \end{bmatrix} = [\hat{W}_{N/2}] \begin{bmatrix} f(0) \\ f(2) \\ f(4) \\ \vdots \\ F\left(\frac{1}{2}N-2\right) \end{bmatrix} + [\hat{W}_d][\hat{W}_{N/2}] \begin{bmatrix} f(1) \\ f(3) \\ f(5) \\ \vdots \\ f(N-1) \end{bmatrix}. \tag{6.79}$$

In a similar manner, the last $N/2$ members of the sequence $\{F(k)\}$ can be written as

$$\begin{bmatrix} F\left(\frac{1}{2}N\right) \\ F\left(\frac{1}{2}N+1\right) \\ F\left(\frac{1}{2}N+2\right) \\ \vdots \\ F(N-1) \end{bmatrix} = [\hat{W}_{N/2}] \begin{bmatrix} f(0) \\ f(2) \\ f(4) \\ \vdots \\ f(N-2) \end{bmatrix} - [\hat{W}_d][\hat{W}_{N/2}] \begin{bmatrix} f(1) \\ f(3) \\ f(5) \\ \vdots \\ f(N-1) \end{bmatrix}. \tag{6.80}$$

Clearly, the calculations of $F(k)(k = 0 \rightarrow N/2 - 1)$ from (6.79) are identical to the calculations of $F(k)[k = N/2 \rightarrow (N-1)]$ from (6.80) except for the sign change in the last sum in (6.80). A symbolic representation of the calculations in (6.79) and (6.80) is shown in Figure 6.3.

Thus, we have succeeded in reducing the calculation of an N-point Fourier transform to the calculation of two $(N/2)$-point transforms. If this process is iterated a number of steps equal to

$$\log_2 N - 1 = \log_2 \left(\tfrac{1}{2}N\right) \tag{6.81}$$

then we arrive at transforms of order 2. Each one of these two-point transforms has a matrix

$$[\hat{W}_2] = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \tag{6.82}$$

and no multiplications are needed. Each stage of the reduction requires $(N/2)$ (complex) multiplications; therefore the total number of (complex) multiplications required for the calculation of the complete transform is

$$M_c = \left(\frac{N}{2}\right) \log_2 \left(\frac{N}{2}\right). \tag{6.83}$$

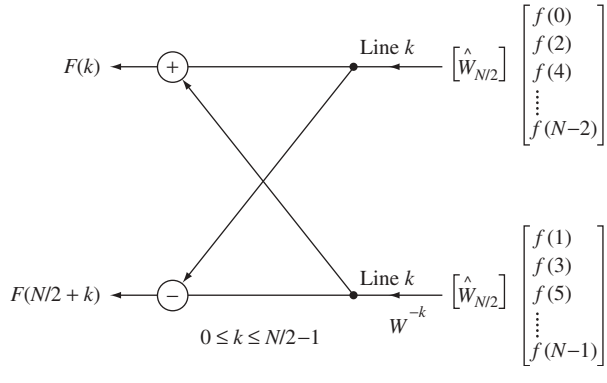


Figure 6.3 Symbolic representation of the calculations in (6.79) and (6.80)

The total number of (complex) additions required is

$$A_c = N \log_2 N. \tag{6.84}$$

These represent great reductions by comparison with direct calculation requiring $(N-1)^2$ multiplications and $N(N-1)$ additions. For example a 64-point transform calculated directly would require

$$(63)^2 = 3639 \text{ multiplications}$$

and

$$64 \times 63 = 4032 \text{ additions.}$$

In contrast, the FFT algorithm requires

$$32 \log_2 32 = 160 \text{ multiplications}$$

and

$$64 \log_2 64 = 384 \text{ additions}$$

which represent a considerable reduction. This saving in computation becomes even more dramatic as N becomes larger.

Example 6.2 Consider the reduction of the matrix of a four-point transform

$$[\hat{W}_4] = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & w^{-1} & w^{-2} & w^{-3} \\ 1 & w^{-2} & w^{-4} & w^{-6} \\ 1 & w^{-3} & w^{-6} & w^{-9} \end{bmatrix}. \tag{6.85}$$

Using

$$\begin{aligned} w^{-1} &= \exp\left(-\mathrm{j}\frac{2\pi}{\mathrm{N}}\right) \\ &= \exp(-\mathrm{j}\pi/2) \\ &= \cos\frac{\pi}{2} - \mathrm{j}\sin\frac{\pi}{2} \\ &= -\mathrm{j} \end{aligned}$$

and noting that $w^{-k} = w^{-(k+N)}$ we have

$$[\hat{W}_4] = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -\mathrm{j} & -1 & \mathrm{j} \\ 1 & -1 & 1 & -1 \\ 1 & \mathrm{j} & -1 & -\mathrm{j} \end{bmatrix} \tag{6.86}$$

Figure 6.4 shows the diagram required for the reduction of $[\hat{W}_4]$.

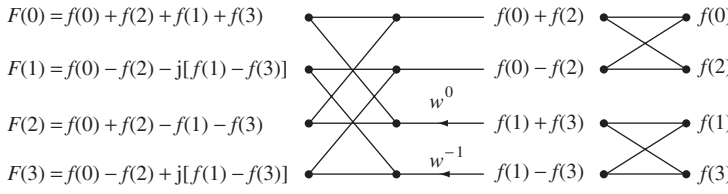


Figure 6.4 Reduction of the matrix needed for the calculation of a four-point FFT

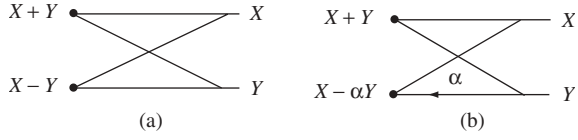


Figure 6.5 The convention in reading a butterfly: (a) addition and subtraction, (b) multiplication, addition and subtraction

These flow graphs are called *butterflies*, and the convention in reading each butterfly, as shown in Figure 6.5, is as follows:

(a) The arrows represent multiplications.
(b) The solid dots at the left of the butterfly represent addition of the two quantities at the right if the dot is above either quantity, or subtraction if the dot is below either quantity.

Example 6.3 Consider the reduction of an eight-point transform with matrix $[\hat{W}_8]$. Using

$$w^{-1} = \exp(-\mathrm{j}2\pi/8)$$

we obtain

$$w^0 = 1,\ w^{-1} = \frac{1-\mathrm{j}}{\sqrt{2}},\ w^{-2} = -\mathrm{j},\ w^{-3} = -\frac{(1+\mathrm{j})}{2^{1/2}},$$

$$w^{-4} = -1,\ w^{-5} = -\frac{(1-\mathrm{j})}{2^{1/2}},\ w^{-6} = \mathrm{j},\ w^{-7} = \frac{1+\mathrm{j}}{\sqrt{2}}.$$

It is also clear that

$$w^{-k} = w^{-(k+8)}$$

and

$$w^{-4} = -w^0,\ w^{-5} = -w^{-1},\ w^{-6} = -w^{-2},\ w^{-7} = -w^{-3}.$$

Figure 6.6 shows the flow graph for the reduction of the eight-point transform.

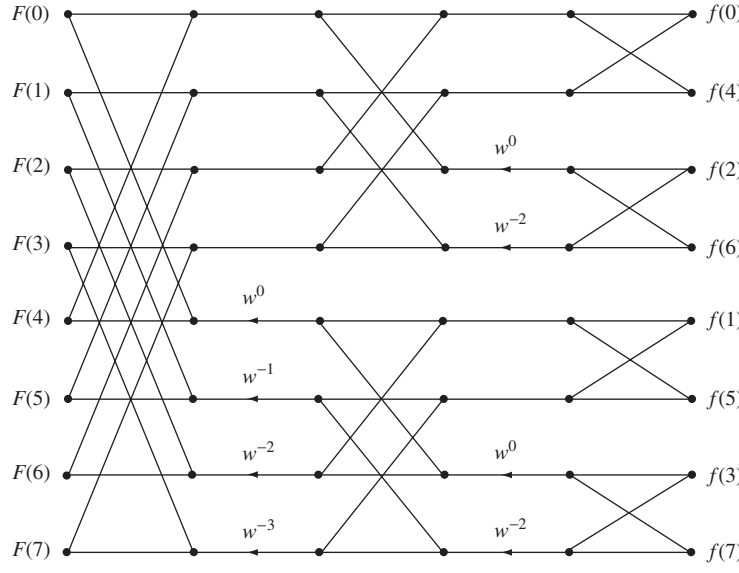


Figure 6.6 Reduction of an eight-point FFT

Bit-Reversal

Examination of Figures 6.4 and 6.6 shows that the members of the sequence $\{F(k)\}$ appear in their natural order. However, the members of the sequence $\{f(n)\}$ appear in permuted order. This is caused by the successive interleaving of the members of the sequence $\{f(n)\}$ necessary for the reduction of the order of the transform. Since the calculations are invariably done on a digital computer, let us represent the arguments of the sequence $\{f(n)\}$ with $N = 8$, by binary numbers. From Figure 6.6 the order in which $\{f(n)\}$ appear obeys the correspondence:

$$
\begin{array}{llcll}
f(0) & [000] & : & f(0) & [000] \\
f(4) & [100] & : & f(1) & [001] \\
f(2) & [010] & : & f(2) & [010] \\
f(6) & [110] & : & f(3) & [011] \\
f(1) & [001] & : & f(4) & [100] \\
f(5) & [101] & : & f(5) & [101] \\
f(3) & [011] & : & f(6) & [110] \\
f(7) & [111] & : & f(7) & [111].
\end{array}
$$

From the above correspondences, it is clear that the order of the sequence $\{f(n)\}$ changes in a manner that may be called *bit reversal*, since it results in a reversal or inversion of the binary representation of the arguments of $\{f(n)\}$.

Next, we note that the *inverse transform* can be performed using exactly the same operations discussed above, if we simply change the sign of the exponent of w. The factor $(1/N)$ in (6.75) can be introduced if we divide by 2 the results of the additions and subtractions made in the butterflies.

Finally, it may be observed that for the calculation of an N-point transform using a computer, we need just enough memory to hold N(complex) positions. This is because the calculations represented by a butterfly are performed on a pair of variables which retain their position in the collection of variables at the end of the operation. Therefore this computation is called *in-place computation*.

6.5.2 Decimation-in-frequency Fast Fourier Transform

The process of decomposition and interleaving which was applied in the previous section to the sequence $\{f(n)\}$ can also be applied to the sequence $\{F(k)\}$. Thus, we partition the members of the sequence $\{F(k)\}$ into those with even arguments and for this set we write from (6.73)

$$\begin{bmatrix} F(0) \\ F(2) \\ F(4) \\ \vdots \\ F(N-2) \end{bmatrix} = \overbrace{\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & w^{-2} & w^{-4} & & w^{-2(N/2-1)} \\ 1 & w^{-4} & w^{-8} & & w^{-4(N/2-1)} \\ \vdots & \vdots & & & \vdots \\ 1 & w^{-2(N/2-1)} & & \cdots & w^{-2(N/2-1)^2} \end{bmatrix}}^{[\hat{W}_{N/2}]} \begin{bmatrix} f(0)+f(N/2) \\ f(1)+f(N/2+1) \\ f(2)+f(N/2+2) \\ \vdots \\ f(N/2)+f(N-1) \end{bmatrix} \tag{6.87}$$

Similarly, for the members with odd arguments

$$\begin{bmatrix} F(1) \\ F(3) \\ F(5) \\ \vdots \\ F(N-1) \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & w^{-1} & \cdots & w^{-(N/2-1)} \\ 1 & w^{-3} & \cdots & w^{-3(N/2-1)} \\ 1 & w^{-5} & \cdots & w^{-5(N/2-1)} \\ \vdots & \vdots & & \vdots \\ 1 & w^{-(N-1)} & \cdots & w^{-(N-1)(N/2-1)} \end{bmatrix}}_{[\hat{\hat{W}}_{N/2}]} \begin{bmatrix} f(0)-f(N/2) \\ f(1)-f(N/2+1) \\ f(2)-f(N/2+2) \\ \vdots \\ f(N/2-1)-f(N-1) \end{bmatrix} \tag{6.88}$$

Comparison of $\left[\hat{W}_{N/2}\right]$ with $\left[\hat{\hat{W}}_{N/2}\right]$ shows that they are related by

$$[\hat{\hat{W}}_{N/2}] = [\hat{W}_{N/2}] \begin{bmatrix} 1 & & & & \\ & w^{-1} & & & \mathrm{O} \\ & & w^{-2} & & \\ & \mathrm{O} & & \ddots & \\ & & & & w^{-(N/2-1)} \end{bmatrix} \tag{6.89}$$

Consequently the matrix $[\hat{W}_{N/2}]$ can be used to calculate $\{F(k)\}$ for k both even and odd, and use is made of (6.89) to obtain the transform of order $N/2$. The diagram in Figure 6.7 is, therefore, obtained to describe the calculations. Using the same convention as in the previous section for the butterflies, we iterate the process until a two-point transform is reached. The number of calculations, in this case, is the same as in the decimation-in-time FFT. Figure 6.8 shows the reduction diagram for an $N = 8$ decimation-in-frequency FFT. It is clear that, here also, the members of the sequence $\{f(n)\}$ appear in their natural order, while the members of the transform sequence $\{F(k)\}$ are permuted appearing with bit reversal.

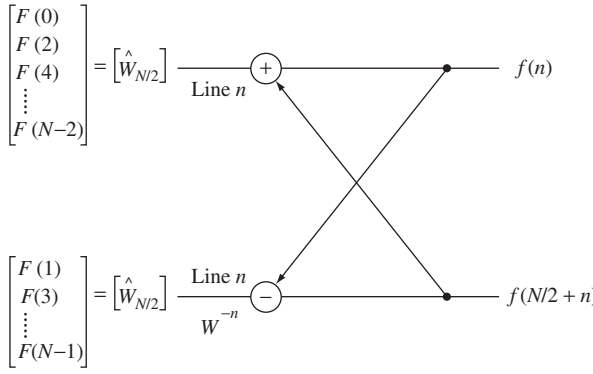


Figure 6.7 Symbolic representation of the calculations in (6.87) and (6.88)

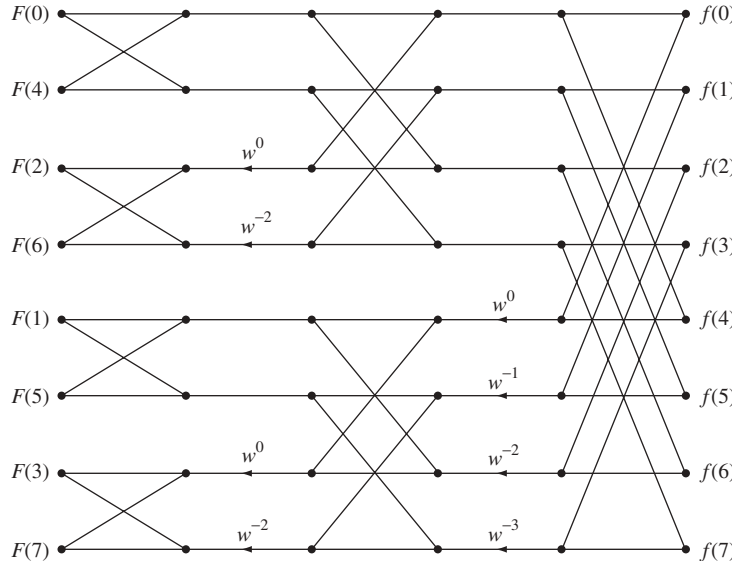


Figure 6.8 Reduction diagram for an eight-point decimation in frequency FFT

6.5.3 Radix 4 Fast Fourier Transform

The algorithms discussed so far are obtained by decomposing an N-point transform (where N is a power of 2) into elementary two-point transforms which require no multiplication. These algorithms are called radix 2 transforms. If N is a power of 4, then it is possible to regard the matrix $[\hat{W}_4]$

$$[\hat{W}_4] = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -\mathrm{j} & -1 & \mathrm{j} \\ 1 & -1 & 1 & -1 \\ 1 & \mathrm{j} & -1 & -\mathrm{j} \end{bmatrix} \tag{6.90}$$

as the elementary matrix of the transform. In this case, the members of the sequence $\{f(n)\}$ are decomposed into four interleaved subsets. Let $[\hat{W}_{N/4}]$ be the square matrix of the $(N/4)$-point transform. Also let

$$[D_i] \triangleq \begin{bmatrix} w^{-1} & & & & \\ & w^{-2} & & & \mathrm{O} \\ & & w^{-3} & & \\ & \mathrm{O} & & \ddots & \\ & & & & w^{-i(N-1)} \end{bmatrix} \quad i = 1, 2, 3, \ldots \tag{6.91}$$

Then the first $N/4$ members of the sequence $\{F(k)\}$ are given by

$$\begin{bmatrix} F(0) \\ F(1) \\ F(2) \\ \vdots \\ F(N/4-1) \end{bmatrix} = [\hat{W}_{N/4}] \begin{bmatrix} f(0) \\ f(4) \\ f(8) \\ \vdots \\ f(N-4) \end{bmatrix} + [D_1][\hat{W}_{N/4}] \begin{bmatrix} f(1) \\ f(5) \\ f(9) \\ \vdots \\ f(N-3) \end{bmatrix}$$
$$+ [D_2][\hat{W}_{N/4}] \begin{bmatrix} f(2) \\ f(6) \\ f(10) \\ \vdots \\ f(N-2) \end{bmatrix} + [D_3][\hat{W}_{N/4}] \begin{bmatrix} f(3) \\ f(7) \\ f(11) \\ \vdots \\ f(N-1) \end{bmatrix} \tag{6.92}$$

The next $N/4$ members of the sequence are given by

$$\begin{bmatrix} F(N/4) \\ F(N/4+1) \\ \vdots \\ F(N/2-1) \end{bmatrix} = [\hat{W}_{N/4}] \begin{bmatrix} f(0) \\ f(4) \\ \vdots \\ f(N-4) \end{bmatrix} - \mathrm{j}[D_1][\hat{W}_{N/4}] \begin{bmatrix} f(1) \\ f(5) \\ \vdots \\ f(n-3) \end{bmatrix}$$
$$- [D_2][\hat{W}_{N/4}] \begin{bmatrix} f(2) \\ f(6) \\ \vdots \\ f(N-2) \end{bmatrix} + \mathrm{j}[D_3][\hat{W}_{N/4}] \begin{bmatrix} f(3) \\ f(7) \\ \vdots \\ f(N-1) \end{bmatrix} \tag{6.93}$$

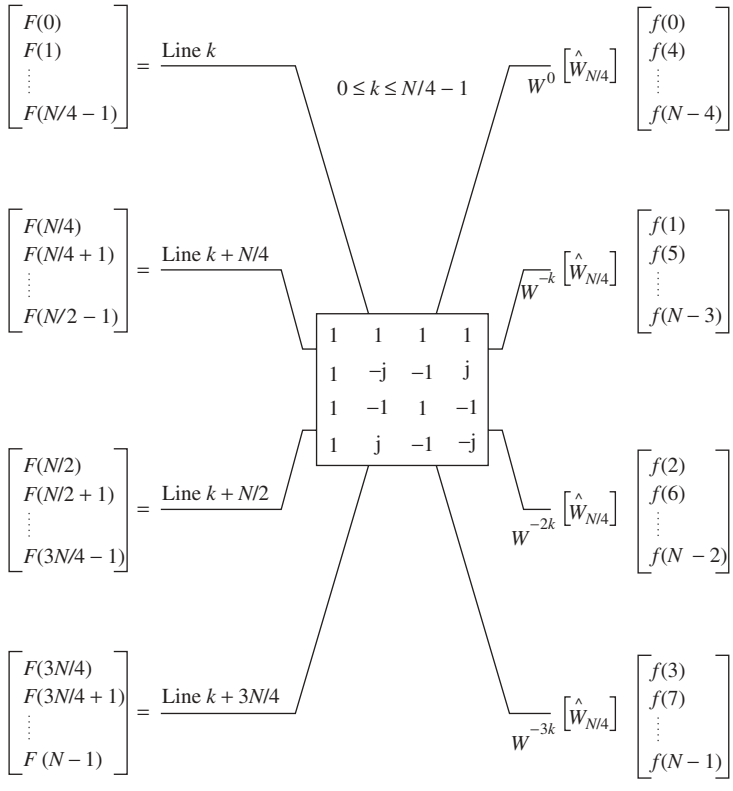


Figure 6.9 Radix 4 FFT

which involves the same calculations as in (6.92) but with the added multiplication by the second row [1 − j − 1j] of $[\hat{W}_4]$ in (4.90). Repeating this process for the third and fourth $(N/4)$ members of the sequence, we obtain the diagram of Figure 6.9.

The number of stages resulting in the reduction of the N-point transform into four-point transform is

$$\log_4 N - 1 = \log_4(N/4) \tag{6.94}$$

and since each stage involves $3(N/4)$ complex multiplications, the total number required is

$$M_c = \frac{3}{4} N \log_4(N/4) \tag{6.95}$$

Similarly, the number of complex additions is

$$A_c = 2N \log_4 N. \tag{6.96}$$

Comparison of (6.96) with (6.84) reveals that the number of additions is the same as in radix 2 transform. However, the radix 4 algorithm reduces the number of multiplications by over 25%, as can be seen by comparing (6.95) with (6.83).

6.6 Properties of the Discrete Fourier Transform

The discrete Fourier transform (DFT) has several important properties which may be viewed as the counterparts of those for the *continuous* case studied in Chapter 2. These properties are now presented since they allow the formulation of a Fourier transform theory for discrete-time signals. These are signals defined only at discrete instants of time, and can therefore be described by sequences.

6.6.1 Linearity

If

$$\{f(n)\} \overset{\leftrightarrow}{N} \{F(k)\} \tag{6.97}$$

and

$$\{h(n)\} \overset{\leftrightarrow}{N} \{H(k)\} \tag{6.98}$$

then

$$a\{f(n)\} + b\{h(n)\} \overset{\leftrightarrow}{N} a\{F(k)\} + b\{H(k)\}. \tag{6.99}$$

This property follows directly from the definition of the DFT in (6.60).

6.6.2 Circular Convolution

Consider two periodic sequences with transforms as in (6.97) and (6.98). Define the *circular convolution* of the two sequences $\{f(n)\}$ and $\{h(n)\}$ as the sequence

$$\begin{aligned} \{g(n)\} &= \{f(n)\} * \{h(n)\} \\ &\triangleq \sum_{m=0}^{N-1} f(m)h(n-m) \end{aligned} \tag{6.100}$$

which is clearly periodic with period N, possessing a DFT. Thus, we can write

$$\{g(n)\} \overset{\leftrightarrow}{N} \{G(k)\} \tag{6.101}$$

where

$$G(k) = \sum_{n=0}^{N-1} g(n)w^{-nk} \tag{6.102}$$

Taking the DFT of both sides of (6.100) and using (6.60) we obtain

$$\begin{aligned} G(k) &= \sum_{n=0}^{N-1} \left(\sum_{m=0}^{N-1} f(m)h(n-m) \right) w^{-nk} \\ &= \sum_{m=0}^{N-1} f(m) \sum_{m=0}^{N-1} h(n-m)w^{-nk} \end{aligned} \tag{6.103}$$

Putting $(n-m)=r$, we have

$$\sum_{n=0}^{N-1} h(n-m)w^{-nk} = w^{-mk} \sum_{r=-m}^{N-1-m} h(r)w^{-rk} \tag{6.104}$$

and since $\{h(r)\}$ and w^{-m} are both periodic sequences with period N, it follows that the sum in (6.104) is the same from $-m$ to $(N-1-m)$ as from 0 to $N-1$. Thus (6.104) becomes

$$w^{-mk} \sum_{r=0}^{N-1} h(r)w^{-rk} = w^{-mk} H(k) \tag{6.105}$$

which, upon substitution in (6.103), gives

$$\begin{aligned} G(k) &= \sum_{m=0}^{N-1} f(m)H(k)w^{-mk} \\ &= H(k) \sum_{m=0}^{N-1} f_m w^{-mk} \\ &= H(k)F(k). \end{aligned} \tag{6.106}$$

Thus, we have proved that the DFT of the convolution of two periodic sequences is the product of the two sequences representing the individual DFTs. Due to its periodicity, the sequence $\{g(n)\}$ is called the periodic or *cirular convolution*. Written explicitly, the convolution relation (6.106), upon use of (6.100) and (6.62), becomes

$$\sum_{m=0}^{N-1} f(m)h(n-m) = \frac{1}{N} \sum_{k=0}^{N-1} F(k)H(k)w^{kn} \tag{6.107}$$

From (6.106), it is clear that the convolution operation is commutative so that

$$\{f(n)\} * \{h(n)\} = \{h(n)\} * \{f(n)\} \tag{6.108}$$

That is

$$\sum_{m=0}^{N-1} f(m)h(n-m) = \sum_{m=0}^{N-1} h(m)f(n-m) \tag{6.109}$$

6.6.3 Shifting

From (6.104) and (6.105) it is observed that if

$$\{f(n)\} \overset{\leftrightarrow}{N} \{F(k)\} \tag{6.110}$$

then we have the shifted pairs

$$\{f(n-m)\} \overset{\leftrightarrow}{N} w^{-mk} \{F(k)\} \tag{6.111}$$

and

$$\{f(n+m)\} \overset{\leftrightarrow}{N} w^{mk}\{F(k)\}. \tag{6.112}$$

This means that shifting the sequences by m amounts to multiplying its DFT by w^{-mk}. It is also clear that, due to the periodicity of the sequences, we obtain the same relations for shifting by $(m \pm N)$.

6.6.4 Symmetry and Conjugate Pairs

Both the sequence $\{f(n)\}$ and its DFT $\{F(k)\}$ are periodic with period N, so that we may write

$$f(-n) = f(N-n) \quad F(-k) = F(N-k) \tag{6.113}$$

From the defining expression of the DFT given by (6.60) we take the complex conjugate of both sides to obtain (noting that $w^* = w^{-1}$)

$$F*(k) = \sum_{n=0}^{N-1} f*(n) w^{kn} \tag{6.114}$$

Putting

$$n = (N-m) \tag{6.115}$$

and using (6.113) we have

$$F*(k) = \sum_{m=0}^{N-1} f*(N-m) w^{(N-m)} \tag{6.116}$$

Also

$$w^m = w^{m \pm N} \tag{6.117}$$

so that (6.116) becomes

$$F*(k) = \sum_{m=0}^{N-1} f*(-m) w^{-km} \tag{6.118}$$

The above sum, however, is the DFT of the sequence $\{f*(-n)\}$. Therefore we have the transform pair

$$\{f^*(-n)\} \overset{\leftrightarrow}{N} \{F^*(k)\} \tag{6.119}$$

Similarly we can write

$$\begin{aligned} F(-k) &= \sum_{n=1}^{N-1} f(n) w^{kn} \\ &= \sum_{m=0}^{N-1} f(N-m)\ \exp[k(N-m)] \end{aligned} \tag{6.120}$$

and use of (6.113) gives

$$F(-k) = \sum_{m=0}^{N-1} f(-m)w^{-km} \tag{6.121}$$

Therefore, we have the transform pair

$$\{f(-n)\} \overset{\leftrightarrow}{N} \{F(-k)\} \tag{6.122}$$

Throughout the above analysis, we have assumed that the sequence $\{f(n)\}$ is complex. If, on the other hand, $\{f(n)\}$ is a sequence of real numbers (as it happens when the sequence is obtained by sampling a real function of time), then

$$\begin{aligned} f*(-n) &= f(-n) \\ F^*(k) &= F(-k) \end{aligned} \tag{6.123}$$

In this case (6.119) reduces to (6.122). Since $F(-k) = F(N-k)$, then if $F(k)$ is known for $k = 0, 1, 2, \ldots N/2$, it is also known for all k.

6.6.5 *Parseval's Relation and Power Spectrum*

As in the case of continuous periodic signals, we are also interested in the power spectrum of discrete signals expressed as periodic sequences.

Consider the transform pairs

$$\{f(n)\} \overset{\leftrightarrow}{N} \{F(k)\}$$

and

$$\{g(n)\} \overset{\leftrightarrow}{N} \{G(k)\} \tag{6.124}$$

By the convolution relation (6.107) we have

$$\sum_{m=0}^{N-1} f(n)g(n-m) = \frac{1}{N} \sum_{k=0}^{N-1} F(k)G(k)w^{kn} \tag{6.125}$$

which, upon setting $n = 0$, gives

$$\sum_{m=0}^{N-1} f(m)g(-m) = \frac{1}{N} \sum_{k=0}^{N-1} F(k)G(k) \tag{6.126}$$

Then, if we replace $g(-m)$ by $g^*(m)$, (6.119) shows that $G(k)$ is replaced by $G^*(k)$. Thus (6.126) becomes

$$\sum_{m=0}^{N-1} f(m)g^*(m) = \frac{1}{N} \sum_{k=0}^{N-1} F(k)G^*(k) \tag{6.127}$$

which is the discrete version of Parseval's relation given in (6.126) for continuous signals. In particular, if we put

$$f(m) = g(m) \tag{6.128}$$

expression (6.127) becomes

$$\sum_{m=0}^{N-1} |f(m)|^2 = \frac{1}{N} \sum_{k=0}^{N-1} |F(k)|^2 \tag{6.129}$$

which is the power of the signal. The above relation is the discrete counterpart of (2.49) and affirms that the power in the parent sequence is equal to the power in the spectrum. The expression $(1/N)|F(k)|^2$ is called the *power spectral density* or simply the *power spectrum* of the signal.

6.6.6 Circular Correlation

The circular correlation between two periodic sequences $\{f(n)\}$ and $\{g(n)\}$ is the sequence $\{R_{fg}(m)\}$ defined by

$$R_{fg}(m) = \frac{1}{N} \sum_{n=0}^{N-1} f(n) g(n+m). \tag{6.130}$$

Taking the DFT of both sides of the above expression, and using (6.60) we obtain

$$\begin{aligned} \text{DFT}\{R_{fg}(m)\} &= \frac{1}{N} \sum_{m=0}^{N-1} \left(\sum_{n=0}^{N-1} f(n) g(n+m) \right) w^{-mk} \\ &= \frac{1}{N} \sum_{n=0}^{N-1} f(n) \sum_{m=0}^{N-1} g(n+m) w^{-mk} \end{aligned} \tag{6.131}$$

Putting $(n+m) = r$ we have

$$\begin{aligned} \sum_{m=0}^{N-1} g(n+m) w^{-nk} &= w^{nk} \frac{1}{N} \sum_{r=n}^{n+N-1} g(r) w^{-rk} \\ &= \frac{1}{N} w^{nk} G(k) \end{aligned} \tag{6.132}$$

due to the periodicity of w and the sequences. Substituting from (6.132) into (6.131) we have

$$\begin{aligned} \text{DFT}\{R_{fg}(m)\} &= \frac{1}{N} \sum_{n=0}^{N-1} f(n) G(k) w^{nk} \\ &= \frac{1}{N} G(k) \sum_{n=0}^{N-1} f(n) w^{nk} \end{aligned} \tag{6.133}$$

or

$$\text{DFT}\{R_{fg}(m)\} = \frac{1}{N} F*(k) G(k) \tag{6.134}$$

The right-hand side of the above expression is called the *cross-power spectrum of the* sequences $\{f(n)\}$ and $\{g(n)\}$. Therefore, we have the transform pair

$$R_{fg}(m) \underset{N}{\leftrightarrow} \frac{1}{N} F*(k)G(k) \tag{6.135}$$

with the explicit use of the frequency variable ω_0 we may write

$$R_{fg}(mT_0) \underset{N}{\leftrightarrow} \frac{1}{N} F*(k\omega_0)G(k\omega_0) \tag{6.136}$$

Now, putting $f(n) = g(n)$ in (6.130) we obtain the *autocorrelation* sequence

$$R_{ff}(m) = \frac{1}{N} \sum_{n=0}^{N-1} f(n)f(n+m) \tag{6.137}$$

so that (6.135) and (6.136) give

$$R_{ff}(m) \underset{N}{\leftrightarrow} \frac{1}{N} F*(k)F(k) \tag{6.138a}$$

or

$$R_{ff}(m) \underset{N}{\leftrightarrow} \frac{1}{N} |F(k)|^2 \tag{6.138b}$$

and

$$R_{ff}(mT_0) \underset{N}{\leftrightarrow} \frac{1}{N} |F(k\omega_0)|^2 \tag{6.139}$$

Comparison of (6.138) with (6.129) shows that the autocorrelation sequence and the power spectrum form a DFT pair.

6.6.7 Relation to the z-transform

Let $F(z)$ be the *one-sided* z-transform of a sequence $\{f(n)\}$ of finite duration N. Thus

$$F(z) = \sum_{n=0}^{N-1} f(n)z^{-n} \tag{6.140}$$

which, on the unit circle ($z = \exp(j\theta)$) becomes

$$F(\exp(\mathrm{j}\theta)) = \sum_{n=0}^{N-1} f(n)\exp(-\mathrm{j}n\theta) \tag{6.141}$$

With $\theta = 2\pi k/N$, the above expression becomes

$$F(k) = \sum_{n=0}^{N-1} f(n)w^{-nk} \tag{6.142}$$

Thus, the DFT of the sequence is obtained from its z-transform by calculating the latter at N equidistant points on the unit circle.

6.7 Spectral Analysis Using the FFT

The efficient algorithms discussed in the previous sections were introduced primarily for the evaluation of Fourier integrals and Fourier series coefficients [11, 12]. This is *spectral analysis*, and the main issues involved in the calculation of the spectrum are discussed in this section.

6.7.1 Evaluation of the Fourier Integral

Given a non-periodic function $f(t)$, we use (6.32) and (6.43) to define two periodic functions

$$f_{\mathrm{p}}(t) = \sum_{k=-\infty}^{\infty} f(t + kT) \tag{6.143}$$

and

$$\hat{F}_{\mathrm{p}}(\omega) = \sum_{r=-\infty}^{\infty} F(\omega + r\overline{\omega}) \tag{6.144}$$

where

$$f(t) \leftrightarrow F(\omega) \tag{6.145}$$

and

$$\overline{\omega} = N\omega_0 = N(2\pi/T) \tag{6.146}$$

Then $f_{\mathrm{p}}(t)$ is sampled over T, every T_0 seconds where

$$T = NT_0 \tag{6.147}$$

and the sample values of $\hat{F}_{\mathrm{p}}(\omega)$ are taken every ω_0. The sample values of (6.143) and (6.144) are then given by

$$\hat{f}(mT_0) = \sum_{k=-\infty}^{\infty} f(mT_0 + kT) \tag{6.148}$$

and

$$\hat{F}(n\omega_0) = \sum_{r=-\infty}^{\infty} F(n\omega_0 + r\overline{\omega}). \tag{6.149}$$

We have seen that $\hat{f}(mT_0)$and$(1/T)\hat{F}(n\omega_0)$ form a discrete Fourier transform pair [see (6.52)], that is

$$\hat{f}(mT_0) \underset{N}{\leftrightarrow} \frac{1}{T}\hat{F}(n\omega_0) \tag{6.150}$$

Hence, the calculation of one sequence from the other can be achieved using the FFT algorithms. It is only left to choose the two parameters required for the calculations; these are T and N.

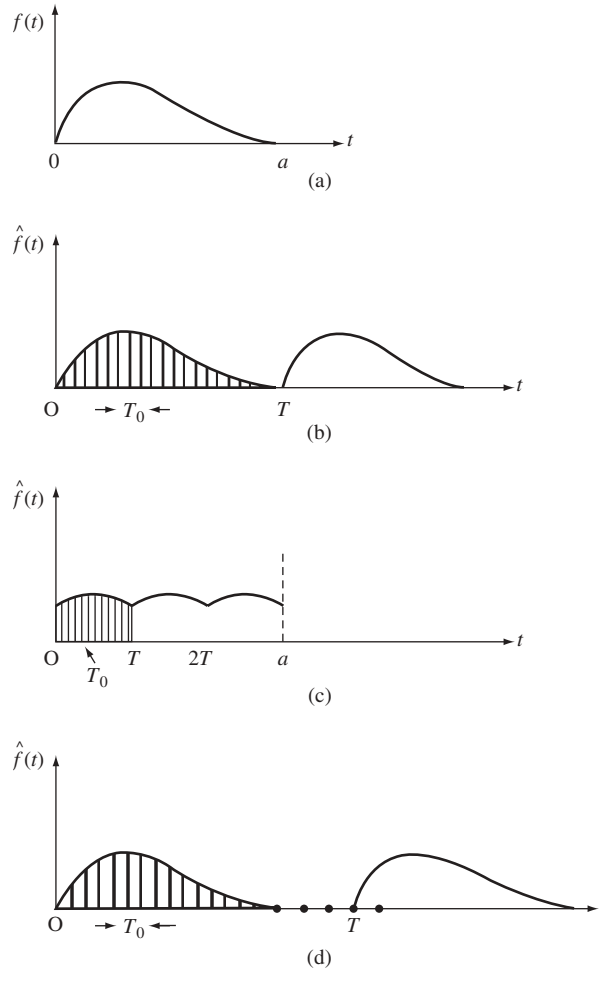


Figure 6.10 (a) A function $f(t)$, (b) choosing $T = a$, (c) $T < a$, (d) $T > a$

Suppose $f(t)$ is specified in the interval $[0, a]$ and is zero outside this interval, as shown in Figure 6.10(a). We have seen in Chapter 3 that a time-limited function cannot be also bandlimited. It follows that the evaluation of the sequence $\hat{F}(mT_0)$, as explained in Section 6.3, can only be approximated giving an aliasing error expressed by (6.57).

Now, if $f(t)$ is specified over $[0, a]$, then we can choose $T = a$, as shown in Figure 6.10(b), and N is the order of the available FFT algorithm.

Therefore we use the algorithm to calculate $\hat{F}(n\omega_0)$ from $\hat{f}(mT_0)$. Then if $F(\omega)$ is negligible for $|\omega| > \hat{\omega}$, with $\hat{\omega} < \overline{\omega}/2$, (this will happen if N is sufficiently large) then we can use (9.56) to make the identification

$$F(n\omega_0) \approx \begin{cases} \hat{F}(n\omega_0) & \text{for } |n| \leq \frac{N}{2} \\ 0 & \text{for } |n| > \frac{N}{2}. \end{cases} \tag{6.151}$$

In these calculations there are two main issues: the *aliasing error* and the *resolution* of the transform.

(i) The Aliasing Error

Naturally, the reduction of the aliasing error should be an important objective. This error is given by

$$\varepsilon = F(n\omega_0) - \hat{F}(n\omega_0) \tag{6.152}$$

which, as explained in Section 6.3, can be reduced by increasing $\overline{\omega}$ as given by (6.44). There are two ways of increasing $\overline{\omega}$, the first is to choose a processor (or software) with large N. If, however, the available processor has a fixed N which yields an unacceptable aliasing error, then another method of increasing $\overline{\omega}$ is to take T smaller than the data interval a, as shown in Figure 4.10(c). However, in this case the samples $\hat{f}(mT_0)$ of $f_p(t)$ are not equal to $f(mT_0)$ but are obtained by summing $f(mT_0)$ as given by (6.148).

(ii) Resolution

If we choose $T = a$, the duration of $f(t)$, and this happens to be large then the samples of $F(\omega)$ will be spaced $\omega_0 = 2\pi/T = 2\pi/a$ apart and are, therefore, closely packed. In this case we say that the *resolution* of $F(\omega)$ is good. Therefore, if it is required to have better resolution, we must increase T, by making it larger than the duration of the signal, that is $T > a$, as shown in Figure 6.10(d). As $f(t) = 0$ for $t > a$, therefore we obtain a function $\hat{f}(t)$ such that

$$\begin{aligned} \hat{f}(t) &= f(t) \quad 0 < t < a \\ &= 0 \quad a < t < T \end{aligned} \tag{6.153}$$

That is we add to $f(mT_0)$ a number of *zero samples* to increase T, as shown in Figure 4.10(d). This process is called *zero padding*.

Now, increasing T results in a decrease in ω_0, which from (6.133) reduces $\overline{\omega}$ for a fixed N, hence increasing the aliasing error. Therefore, to improve the resolution by decreasing ω_0 while keeping the aliasing error the same, we must also increase N, that is we need a longer processor.

6.7.2 *Evaluation of the Fourier Coefficients*

For a periodic function with period T, and Fourier series

$$f(t) = \sum_{k=-\infty}^{\infty} c_k \exp(\mathrm{j}k\omega_0 t)\omega_0 = \frac{2\pi}{T} \tag{6.154}$$

we have seen that if the aliased coefficients are defined as

$$\hat{c}_k = \sum_{r=-\infty}^{\infty} c_{k+rN} \tag{6.155}$$

then the samples $f(mT_0)$ of $f(t)$ and the aliased coefficients are related by (6.20) and (6.26), that is they form a DFT pair

$$f(mT_0) \underset{N}{\leftrightarrow} \hat{c}_k \quad T_0 = T/N \tag{6.156}$$

so that (6.26) gives

$$[\hat{c}_k] = \frac{1}{N}[\hat{W}][f]. \tag{6.157}$$

Again the required calculations can be performed using an FFT algorithm. If we use a truncated Fourier series to represent $f(t)$

$$f(t) \approx \sum_{k=-n}^{n} c_k \exp(\mathrm{j}k\omega_0 t) \tag{6.158}$$

then we have shown in Section 6.2 that choosing

$$N > 2n \tag{6.159}$$

will result in

$$\begin{aligned} c_k &= \hat{c}_k \quad \text{for } |k| \leq \frac{N}{2} \\ &= 0 \quad \text{for } |k| > \frac{N}{2} \end{aligned} \tag{6.160}$$

and the aliasing error is zero. In this case

$$\begin{aligned} c_k &= \frac{1}{T}\int_0^T f(t)\exp(-\mathrm{j}k\omega_0 t)\,\mathrm{d}t \\ &= \frac{1}{N}\sum_{m=0}^{N-1} f(mT_0)\, w^{-mk} \end{aligned} \tag{6.161}$$

That is

$$[c_k] = \frac{1}{N}[\hat{W}][f] \tag{6.162}$$

However, if we do not truncate the Fourier series, but assume that c_k is negligible for $|k| > N/2$ then, as shown in Section 6.2

$$\begin{aligned} c_k &\approx \hat{c}_k \quad \text{for } |k| \leq \frac{N}{2} \\ &\approx 0 \quad \text{for } |k| > \frac{N}{2} \end{aligned} \tag{6.163}$$

and the aliasing error is given by

$$\varepsilon = c_k - \hat{c}_k \tag{6.164}$$

So that if c_k is negligible for $|k| > N/2$, the aliasing error is small. If c_k is not negligible for $|k| > N/2$ then we must increase the order of the FFT algorithm to reduce the aliasing error.

6.8 Spectral Windows

6.8.1 Continuous-time Signals

In Section 6.7, spectrum analysis using the FFT was performed on a signal $f(t)$ which was assumed specified in an interval, and zero outside this interval. An alternative way of looking at the problem is to assume that $f(t)$ does exist for all t but it is known, or observed, only over a finite interval $[-a, a]$, as shown in Figure 6.11. In this section we examine a technique for reducing the errors due to the use of a truncated version of $f(t)$ in spectral analysis.

Given $f(t)$ for every t in an interval $[-a, a]$ we wish to determine the spectrum $F(\omega)$ of $f(t)$. In effect we are given the function

$$f_a(t) = w_r(t) f(t) \tag{6.165}$$

where

$$\begin{aligned} w_r(t) &= 1 \quad |t| < a \\ &= 0 \quad |t| > a. \end{aligned} \tag{6.166}$$

Therefore the spectrum $F(\omega)$ cannot be determined exactly and must be estimated. Let the known segment $f_a(t)$ possess a Fourier transform $F_a(\omega)$. Then

$$w_r(t) f(t) \leftrightarrow F_a(\omega) \tag{6.167}$$

where

$$F_a(\omega) = \int_{-a}^{a} f(t) \exp(-\mathrm{j}\omega t)\, \mathrm{d}t. \tag{6.168}$$

Clearly $w_r(t)$ is a rectangular *window*, which when multiplied by $f(t)$ produces the segment $f_a(t)$ by truncating $f(t)$. However

$$w_r(t) \leftrightarrow \frac{2 \sin a\omega}{\omega} \tag{6.169}$$

so that (6.167) and the frequency convolution theorem gives

$$\begin{aligned} F_a(\omega) &= \frac{1}{2\pi} F(\omega) * \left(\frac{2 \sin a\omega}{\omega} \right) \\ &= \int_{-\infty}^{\infty} F(\mu) \frac{\sin a(\omega - \mu)}{\pi(\omega - \mu)} \mathrm{d}\mu. \end{aligned} \tag{6.170}$$

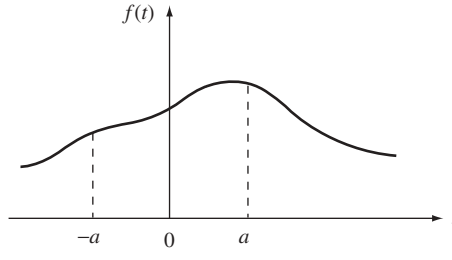


Figure 6.11 A function $f(t)$ observed over a finite interval $[-a, a]$

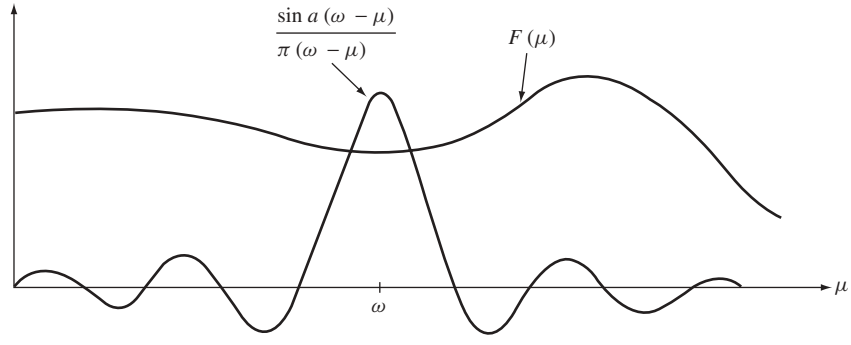


Figure 6.12 The function $\sin a(\omega-\mu)/\pi(\omega-\mu)$ and its use in (6.170)

If this is used as an estimate of $F(\omega)$, as was the case in the previous sections, then clearly this would not be a good estimate for the following reasons:

(a) $F_a(\omega)$ is a weighted average of the values of $F(\omega)$ in an interval of the order of $2\pi/a$. Consequently, rapid variations in $F(\omega)$ in this interval do not show in $F_a(\omega)$.
(b) The weight function $\sin a(\omega-\mu)/\pi(\omega-\mu)$ (see Figure 6.12) has large side lobes which introduce distortion in the estimate.
(c) The function $\sin a(\omega-\mu)/\pi(\omega-\mu)$ can be negative.

Our objective is to improve the estimation of $F(\omega)$ from $f_a(t)$. To this end, we use the windowed function

$$f_{\mathrm{w}}(t) = w(t)f(t) \tag{6.171}$$

where $w(t)$ is real, even and time-limited to $[-a, a]$. Its transform satisfies

$$W(\omega) \geq 0, \quad W(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty} W(\omega)\,\mathrm{d}\omega = 1. \tag{6.172}$$

The *smoothed* spectrum now becomes

$$\begin{aligned} F_{\mathrm{w}}(\omega) &= \int_{-a}^{a} f_{\mathrm{w}}(t)\exp(-\mathrm{j}\omega t)\,\mathrm{d}t \\ &= \int_{-a}^{a} f(t)w(t)\exp(-\mathrm{j}\omega t)\,\mathrm{d}t. \end{aligned} \tag{6.173}$$

Using (6.171) and the frequency convolution theorems we obtain

$$\begin{aligned} F_{\mathrm{w}}(\omega) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\mu)W(\omega-\mu)\mathrm{d}\mu \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega-\mu)W(\mu)\mathrm{d}\mu \end{aligned} \tag{6.174}$$

so that $F_{\mathrm{w}}(\omega)$ is, again the weighted average of $F(\omega)$, but here we can improve the approximation by a proper choice of $w(t)$. The Fourier transform $W(\omega)$ of $w(t)$ is called

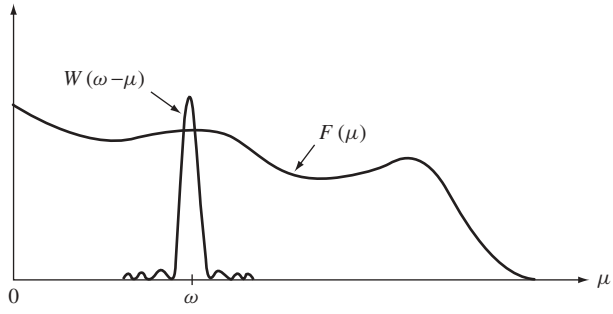


Figure 6.13 Illustration of the use of a general window in spectral analysis as expressed in (6.174)

a *spectral window* whereas $w(t)$ is called a *lag window*. Relation (6.174) is illustrated in Figure 6.13. In selecting a spectral window, two factors are taken into consideration.

(i) Resolution

Suppose the spectrum $F(\omega)$ has peaks at ω_1 and ω_2. In order that these peaks may be detected and *resolved* from the smoothed spectrum $F_w(\omega)$, we must choose $W(\omega)$ with band-width not exceeding the difference $(\omega_2 - \omega_1)$. A reasonable measure of the spectral window band-width is the width of its main lobe. By the uncertainty principle the width of the main lobe is determined by the width of the data interval $2a$.

(ii) Leakage

If one peak $F(\omega_1)$ is small compared with a neighbouring one $F(\omega_2)$, then the smaller peak might not show in $F_a(\omega)$ even if $(\omega_2 - \omega_1)$ exceeds the bandwidth of the window. The extent of this degradation is determined by the size of the *side-lobes* of $W(\omega)$, therefore it can be reduced if $W(\omega)$ decays rapidly as ω increases.

From the above discussion, it follows that, generally speaking, a good window is one whose spectrum is concentrated in its main lobe, this being narrow, and the side lobes are positive and as small as possible. The general properties of windows are based on the following properties of the Fourier transform. Let

$$w(t) \leftrightarrow W(\omega) \tag{6.175}$$

then the following properties [11, 12] can be easily deduced from the discussion in Chapter 2:

(i) If $w(t)$ is an ordinary function containing no impulses, then

$$W(\omega) \to 0 \quad \text{as} \quad \omega \to \infty. \tag{6.176}$$

(ii) If, in addition, $w(t)$, is continuous, then

$$\omega W(\omega) \to 0 \quad \text{as} \quad \omega \to \infty. \tag{6.177}$$

(iii) If the kth derivative of $w(t)$ exists and is continuous then

$$\omega^{k+1} W(\omega) \to 0 \quad \text{as} \quad \omega \to \infty. \tag{6.178}$$

It follows that, if we require $W(\omega)$ to approach zero faster than $1/\omega^k$, than $w(t)$ is selected such that its kth derivative is continuous. $w(t)$ and its first k derivatives must, however, vanish at the end points of the interval $[-a, a]$, that is

$$w(\pm a) = w'(\pm a) = \cdots = w^{(k)}(\pm a) = 0. \tag{6.179}$$

Having chosen a window function, the estimate of the spectrum is obtained using an FFT algorithm applied to $f_{\mathrm{w}}(t)$. Clearly the narrower the bandwidth and the smaller the size of the side-ripples of the spectral window, the better the estimate. In particular, a narrow bandwidth results in a good resolution of the FFT, while small side-ripples minimize the interference between the spectral components of the signal. For example the Hamming window pair (shown in Figure 6.14) is given by

$$\left(0.54 + 0.46 \cos \frac{\pi t}{a}\right) \leftrightarrow \frac{(1.08\pi^2 - 0.16a^2\,\omega^2) \sin a\omega}{\omega(\pi^2 - a^2\,\omega^2)} \tag{6.180}$$

which has the property that $W(\omega) \to 0$ with the speed of $\approx .16/\omega$ as $\omega \to \infty$. Moreover, the spectrum of the Hamming window [shown in Figure 6.14(b)] has 99.96% of its energy

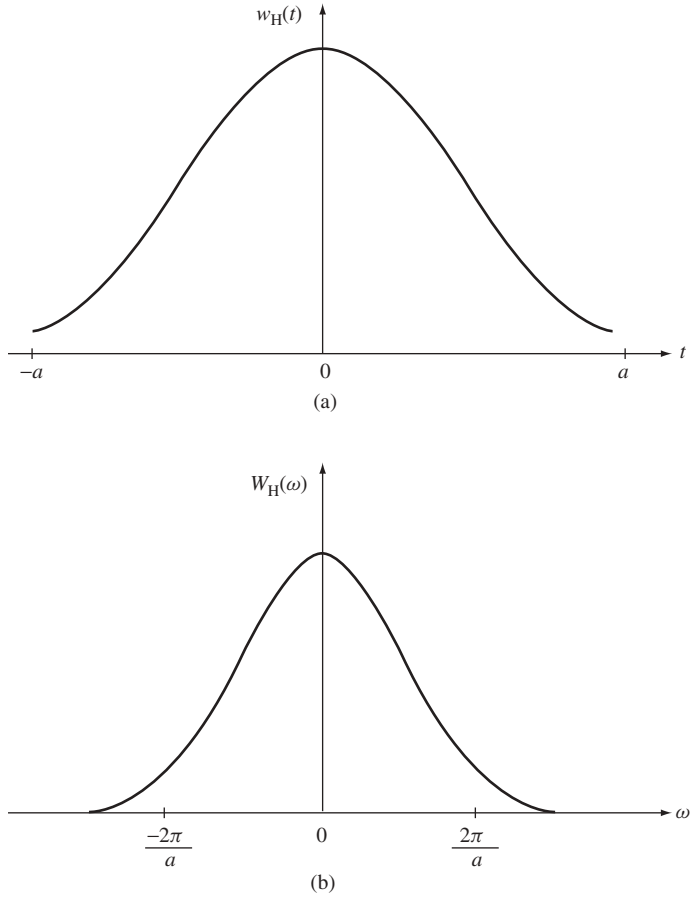


Figure 6.14 (a) The Hamming window and (b) its spectrum

concentrated in its main lobe and the largest side lobe is lower by 43 dB than the main lobe level.

Spectral windows can also be used in conjunction with spectral analysis of periodic signals represented by truncated Fourier series. Thus

$$F_{\mathrm{w}}(\omega) = \sum_{n=-N}^{N} w(n) f(nT) \exp(-\mathrm{j}n\omega T)$$

$$= \frac{1}{2a} \int_{-a}^{a} F(\omega - \mu) W(\mu) \mathrm{d}\mu \tag{6.181}$$

where the window $w(n)$ is one of those expressions given in Chapter 2, with $k \to n$ and $n \to N$.

6.8.2 Discrete-time Signals

The discrete-time version of the analysis in the previous section can be obtained by working with sequences. These are naturally produced when using an FFT algorithm, so that sampling the continuous signal and its spectrum produce a DFT pair in which the signal samples $f(nT)$ form a discrete-time signal. Therefore, if the signal is already a discrete-time one, all we need are discrete-time window functions. These were, in fact, given by (5.141)–(5.147). Therefore, prior to applying an FFT algorithm, the discrete-time function $f(n)$ is windowed by one of these expressions to give

$$f_{\mathrm{w}}(n) = w(n) f(n) \tag{6.182}$$

and the FFT is applied to $f_{\mathrm{w}}(n)$.

6.9 Fast Convolution, Filtering and Correlation Using the FFT

The speed and efficiency of the FFT algorithms have opened many areas for their applications, other than spectral analysis. Three of these applications are the closely related operations of convolution, filtering and correlation.

6.9.1 Circular (Periodic) Convolution

Consider the two DFT pairs

$$\{h(n)\} \underset{N}{\leftrightarrow} \{H(k)\} \tag{6.183}$$

and

$$\{f(n)\} \underset{N}{\leftrightarrow} \{F(k)\} \tag{6.184}$$

where it is assumed, to begin with, that the sequences are periodic of equal duration N. The circular convolution of the two sequences $\{h(n)\}$ and $\{f(n)\}$ is defined by (6.100)

as the sequence

$$\{g(n)\} = \{h(n)\} * \{f(n)\} \triangleq \sum_{m=0}^{N-1} h(m) f(n-m) \tag{6.185}$$

and, as expressed in (6.106), if

$$\{g(n)\} \underset{N}{\leftrightarrow} \{G(k)\} \tag{6.186}$$

then

$$G(k) = H(k)F(k). \tag{6.187}$$

Performing *fast* circlar convolution of the two sequences $\{f(n)\}$ and $\{h(n)\}$ means that we take the DFT of each sequence using an FFT algorithm, then form the product $G(k)$ in (6.187) and finally take the IDFT of $\{G(k)\}$ using an FFT to obtain the required result $\{g(n)\}$.

6.9.2 Non-periodic Convolution

Now, usually $\{f(n)\}$ and $\{h(n)\}$ are non-periodic with finite different durations L and K, respectively. The convolution of the two sequences is given by

$$g(n) = \sum_{m=0}^{n} h(m) f(n-m) \tag{6.188}$$

which is of duration $L + K - 1$. In this case, fast convolution is performed by first choosing a number N which is a power of 2, and also satisfies

$$N \geq L + K - 1. \tag{6.189}$$

Then each sequence is augmented to make it of duration N, by filling it with the appropriate number of zeros. Each of the three sequences is then regarded as one period of a periodic sequence and an FFT algorithm is used as before to compute the cyclic convolution sequence. Although the result is a periodic sequence, only one period is taken as the desired convolution sequence.

6.9.3 Filtering and Sectioned Convolution

Now, it is observed that convolution is essentially a filtering operation in which $\{f(n)\}$ is the input to an FIR filter with impulse response $\{h(n)\}$ and output $\{g(n)\}$. Let us re-examine the convolution operation from this point of view, and show that filtering can be performed using the FFT. Suppose that, as shown in Figure 6.15

$$f(n) = 0 \quad \text{for } 0 > n > L - 1 \tag{6.190}$$

and

$$h(n) = 0 \quad \text{for } 0 > n > K - 1 \tag{6.191}$$

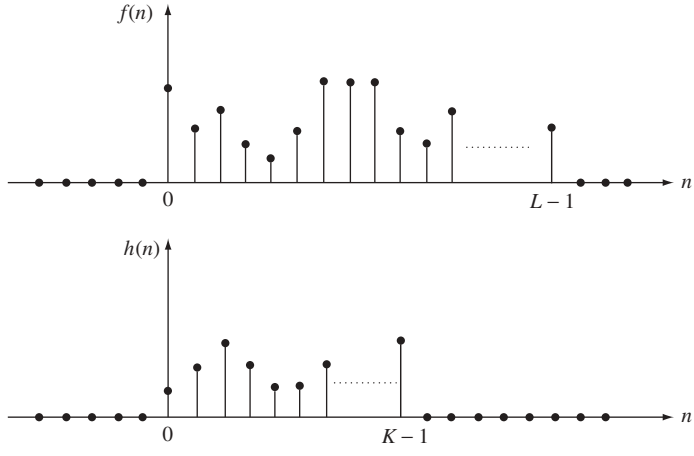


Figure 6.15 The two sequences defined by (6.190) and (6.191)

That is both $\{f(n)\}$ and $\{h(n)\}$ are causal sequences of finite, but unequal durations. The filter output sequence is then given by

$$g(n) = \sum_{m=0}^{K-1} h(m) f(n-m) \quad 0 \le n \le L + K - 2. \tag{6.192}$$

Next, we define N-point DFTs of the sequences, where N satisfies (6.189), say with equality, and zero padding is employed. The response of the filter can be calculated using the following steps:

(i) Calculate the DFT of $\{h(n)\}$ and $\{f(n)\}$ using an FFT algorithm.
(ii) Calculate the product $H(k)F(k)(k = 0, 1, 2, \ldots)$.
(iii) Calculate the IDFT of $G(k)$ using the same FFT algorithm employed in (i).

The above procedure requires $(6 \log_2 N + 4)$ multiplications, whereas direct calculation of (6.192) entails K multiplication. Hence, the usual reduction in computation which characterizes the FFT algorithm makes the evaluation of convolution and the implementation of FIR filters more efficient. For example with $L = K = 256$, the use of the FFT requires 58 multiplications, by contrast with 256 if direct implementation of (6.192) is used.

Now, in following the above procedure, it is assumed that the entire input sequence $\{f(n)\}$ is available before the processing starts. Therefore, if the input sequence $\{f(n)\}$ is much longer than the impulse response sequence $\{h(n)\}$, that is $L \gg K$, a corresponding long delay in the computation ensues, which is objectionable in *real-time* filtering. In these cases the procedure is modified to minimize the delay. To this end we begin by dividing the input sequence into shorter subsequences, each of duration M, which are processed separately. Thus, for $L \gg K$ we write

$$f(n) = \sum_{i=0}^{q} f_i(n) \quad 0 \le n \le (q+1)M - 1 \tag{6.193}$$

where $(q + 1)$ is the number of subsequences. This is related to L and M by

$$q = \text{least integer} \geq L/M\,. \tag{6.194}$$

The order N of the FFT algorithms is chosen to be a power of 2, which also satisfies

$$N \geq M + K - 1. \tag{6.195}$$

The subsequences are defined by

$$\begin{aligned} f_i(n) &= f(n) \quad iM \leq n \leq (i+1)(M-1) \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{6.196}$$

as shown in Figure 6.16 for $M = 4$.

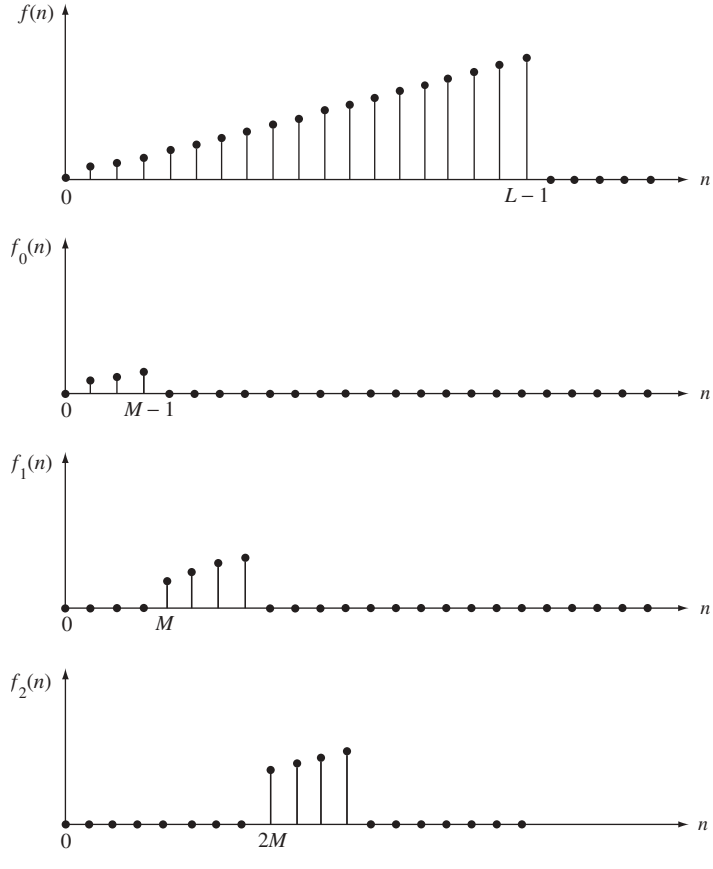


Figure 6.16 The subsequences defined by (9.196)

Thus, the output sequence can be written as

$$g(n) = \sum_{m=0}^{K-1} \sum_{i=0}^{q} h(m) f_i(n-m)$$
$$= \sum_{i=0}^{q} g_i(n) \tag{6.197}$$

where

$$g_i(n) = \sum_{m=0}^{K-1} h(m) f_i(n-m). \tag{6.198}$$

Expressions (6.197) and (6.198) imply that the response sequence $\{g(n)\}$ can be evaluated as the superposition of a number of *partial convolutions*. Each of these is obtained as the non-periodic convolution of a subsequence $f_i(n)$ with the sequence $h(n)$. Consequently, the partial convolution can be calculated by means of the periodic convolution of two augmented sequences, as discussed in Section 6.9.2. With

$$(iM - 1) \leq n \leq (i+1)M + K - 1 \tag{6.199}$$

expression (6.198) gives for the ith partial convolution

$$\begin{aligned}
&g_i(iM-1) = 0\\
&g_i(iM) = h(0) f(iM)\\
&g_i(iM+1) = h(0) f(iM+1) + h(1) f(iM)\\
&g_i[(i+1)M + K - 2] = h(K-1) f[(i+1)M - 1] h(K-1)\\
&g_i[(i+1)M + K - 1] = 0.
\end{aligned} \tag{6.200}$$

The above expressions show that the ith partial convolution sequence has $(M + K - 1)$ non-zero elements, which can be stored in an array g_i as shown in Figure 6.17. The elements of g_i are computed from

$$g_i(n) = \text{IDFT}[H(k) F_i(k)] \tag{6.201}$$

requiring an $(M + K - 1)$-point FFT. Subsequently, we form an array of the values of $g(n)$, by entering the elements of the segments in $g_0, g_1, \ldots$ then adding the elements in overlapping adjacent segments. This is illustrated in Figure 6.17. Therefore, the processing can start when M input samples are available, and the first batch of M output samples becomes available when the first input segment is processed. In this fashion, the delay in the processing is reduced. This technique is called *sectioned convolution*.

6.9.4 Fast Correlation

The periodic cross-correlation between two periodic sequences $\{f(n)\}$ and $\{g(n)\}$ is defined by (6.130) as

$$R_{fg}(m) = \frac{1}{N} \sum_{n=0}^{N-1} f(n) g(n+m) \tag{6.202}$$

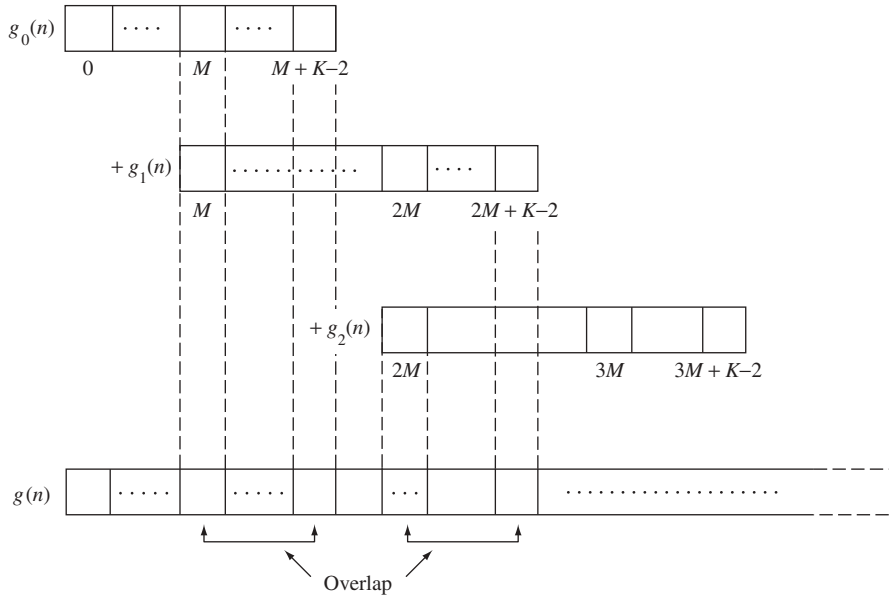


Figure 6.17 Sectioned convolution

and it has been shown that

$$\text{DFT}\{R_{fg}(m)\} = \frac{1}{N}F*(k)G(k). \tag{6.203}$$

It follows that the cross-correlation $R_{fg}(m)$ can be calculated by first taking the DFT of $\{f(n)\}$ and $\{g(n)\}$ using an FFT algorithm then performing the product $F^*(k)G(k)$ and finally taking the IDFT of $(1/N)F^*(k)G(k)$ to obtain the autocorrelation sequence. The resulting reduction in computation as compared with direct evaluation of (6.202) is the same as in the case of convolution. Similarly, the periodic autocorrelation of a sequence is given by

$$R_{ff}(m) = \frac{1}{N}\sum_{n=0}^{N-1} f(n)f(n+m) \tag{6.204}$$

so that (6.137) gives

$$\begin{aligned}\text{DFT}\{R_{ff}(m)\} &= \frac{1}{N}F*(k)F(k)\\ &= \frac{1}{N}|F(k)|^2.\end{aligned} \tag{6.205}$$

Therefore the sequence $R_{ff}(m)$ can be evaluated by first using an FFT algorithm to calculate $F(k)$ from $f(n)$, then $(1/N)|F(k)|^2$ is obtained, whose IDFT is calculated using an FFT to obtain the required sequence $\{R_{ff}(m)\}$.

In the above calculations of the correlation sequences using FFT algorithms, if the sequences are non-periodic, modifications to the procedure are possible along the lines discussed in Section 6.9.2. Furthermore, sectioned correlation can be employed in a manner similar to sectioned convolution discussed in Section 6.9.3. The main applications of

correlation sequences are in the area of statistical signal processing which is the subject of later chapters.

6.10 Use of MATLAB®

The DFT of a sequence can be obtained using the *Signal processing Toolbox* functions in MATLAB©. For a sequence defined by the vector [x], the FFT is obtained by

```
y = fft(x)
```

and the magnitude and phase spectra are obtained by

```
m = abs(y)
p = unwrap(angle(y))
```

It is also possible to specify the number of points n in the transform and use

```
y = fft(x,n)
```

If the sequence is longer than n then it is truncated, whereas if the sequence is shorter than n, zero padding is used.

The convolution of two sequences [a] and [b] is obtained from

```
c = conv(a,b)
```

The inverse DFT is obtained as

```
x = ifft(y)
```

6.11 Conclusion

This chapter has dealt with the algorithms and associated concepts which are commonly used in spectral analysis and filtering of signals with the objective of increasing the speed and efficiency of the calculations. By introducing the *discrete Fourier transform*, two sequences can be related and the calculation of one sequence from another can be accomplished using the computational algorithms known collectively as *fast Fourier transform algorithms*. It must be remembered, however, that the FFT is merely an algorithm, introduced originally for the purpose of calculating the Fourier coefficients and the Fourier integral of functions. However, its applications have extended well beyond these specific roles to perform many operations in signal processing such as filtering, fast convolution and fast correlation.

Problems

6.1 Calculate the 16-point DFT of each of the following sequences:
(a) $\{1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1\}$,
(b) $\{1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0\}$,
(c) $\{0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0\}$.

6.2 Calculate the 128-point DFT of the sequence defined by

$$f(0) = f(1) = f(2) = f(125) = f(126) = f(127) = 1$$
$$f(n) = 0 \quad \text{for} \quad 3 \leq n \leq 124.$$

6.3 Obtain the reduction diagram for a 16-point decimation-in-time-FFT algorithm.

6.4 Calculate the 16-point DFT of the sequence

$$f(n) = \sin(4\pi n/7) + 0.1\ \sin(4\pi \mathrm{n}/13).$$

short. Use the result to compute the convolution of the following sequences

$$\{\sin\ n\} \text{ and } \exp\{-10n\}$$

6.5 Use MATLAB to calculate the power spectrum of the signal in Problem 6.4, assuming only a 16-point FFT and using the Hamming window.

7

Stochastic Signals and Power Spectra

7.1 Introduction

So far in our discussion of signals and systems, it has been tacitly assumed that the signals are defined by analytic expressions, differential equations, difference equations or even arbitrary graphs. Such signals are called *deterministic*. The same is true for deterministic systems described by differential equations, difference equations or any functional. However, most signals are random, or at best contain random components due to factors such as noise introduced by the generating sources or by the channel over which the signals are transmitted. Such signals require the use of statistical methods for their description; this consideration leads to the area of *stochastic signal processing* [11, 12]. This chapter gives an introduction to the concepts and techniques suitable for the description of stochastic (random) signals. The discussion encompasses both analog and digital signals. However the systems which perform the processing of these signals are themselves *deterministic*.

7.2 Random Variables

Consider performing a certain experiment a number of times, and each time an outcome $\zeta_i (i = 1, 2, \ldots)$ results. Thus, we obtain a set Λ of possible outcomes $\zeta_1, \zeta_2, \ldots$ which can be finite or (theoretically) infinite in number. We then assign a number $\mathbf{f}(\zeta)$ to each ζ according to some rule. This way we construct a function $\mathbf{f}(\zeta)$ or simply $\mathbf{f}$ whose domain is the set Λ and whose range is a set of numbers $\mathbf{f}(\zeta_1)$, $\mathbf{f}(\zeta_2) \ldots$. This function is called a *random variable* [11, 12]. *Throughout this chapter random quantities are denoted by boldface characters*.

7.2.1 Probability Distribution Function

Consider a random variable $\mathbf{f}$ which may take real values in a range $[f_1, f_2]$ where f_1 could be as low as $-\infty$ and f_2 as high as $+\infty$. Let us observe this variable over the entire range

Signal Processing and Integrated Circuits, First Edition. Hussein Baher.
© 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.

$[f_1, f_2]$ and define its *probability distribution function* as

$$P(f) \triangleq \text{Prob}[\mathbf{f} < f] \tag{7.1}$$

which is the probability that the random variable **f** assumes a value less than some given number f.

7.2.2 Probability Density Function

If the probability distribution function $P(f)$, of the variable **f**, is differentiable, then we define the probability density function as

$$p(f) = \frac{\mathrm{d}P(f)}{\mathrm{d}f}. \tag{7.2}$$

This has the obvious property that

$$\int_{-\infty}^{\infty} p(f)\,\mathrm{d}f = 1 \tag{7.3}$$

because any value of **f** must lie in the range $[-\infty, \infty]$. Moreover, the probability that **f** lies between f_1 and f_2 is given by

$$\text{Prob}[f_1 < \mathbf{f} < f_2] = \int_{f_1}^{f_2} p(f)\,\mathrm{d}f. \tag{7.4}$$

The shape of the probability density function curve indicates the 'preferred' range of values which **f** assumes. For example, a commonly-occurring probability density function is the Gaussian one given by

$$p(f) = \frac{1}{(2\pi\sigma)^{1/2}} \exp[-(f-\eta)^2/2\sigma^2] \tag{7.5}$$

where σ and η are constants. This is shown in Figure 7.1. Another example is the case shown in Figure 7.2 where there is no preferred range for the random variable **f** between f_1 and f_2. The probability density is said to be *uniform* and is given by

$$\begin{aligned} p(f) &= 1/(f_2 - f_1) \quad f_1 \le \mathbf{f} \le f_2 \\ &= 0 \qquad\qquad\quad \text{otherwise.} \end{aligned} \tag{7.6}$$

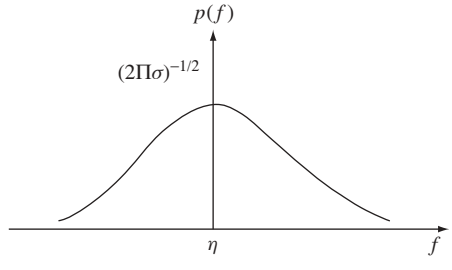


Figure 7.1 Gaussian probability density function defined by (7.5)

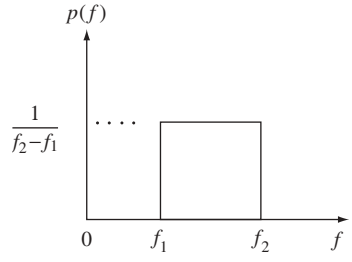


Figure 7.2 Uniform probability density function defined by (7.6)

7.2.3 Joint Distributions

In performing an experiment, we may have two sets of random outcomes

$$\zeta_{f1},\ \zeta_{f2},\ \zeta_{f3},\ \ldots$$

and

$$\zeta_{g2},\ \zeta_{g3},\ \zeta_{g4},\ \ldots \tag{7.7}$$

The probability of observing **f** and **g** below f and g, respectively, is referred to as the *joint distribution function* of **f** and **g**

$$P(f,\ g) = \text{Prob}[\,\mathbf{f} < f,\ \mathbf{g} < g\,]. \tag{7.8}$$

The *joint probability density function* of **f** and **g** is defined by

$$p(f,\ g) = \frac{\partial^2 P(f,\ g)}{\partial f\,\partial g}. \tag{7.9}$$

Again, since the range $[-\infty, \infty]$ includes **f** and **g** we must have

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(f,\ g)\,\mathrm{d}f\,\mathrm{d}g = 1. \tag{7.10}$$

The two random variables **f** and **g** representing the outcomes $(\zeta_{f1}, \zeta_{f2}, \ldots)$ and $(\zeta_{g1}, \zeta_{g2}, \ldots)$ are said to be *statistically independent* if the occurrence of any outcome ζ_g does not affect the occurrence of any outcome ζ_f and vice versa. This is the case if and only if

$$p(f,\ g) = p(f)p(g). \tag{7.11}$$

7.2.4 Statistical Parameters

The description of the properties of random variables can be accomplished by means of a number of parameters. These are now reviewed:

(i) *The mean or first moment, or expectation value* of a random variable **f** is denoted by $E[\mathbf{f}]$ or η_f and is defined by

$$E[\,\mathbf{f}\,] \triangleq \int_{-\infty}^{\infty} fp(f)\,\mathrm{d}f \equiv \eta_f. \tag{7.12}$$

More generally, if a random variable **u** is a function of two other random variables **f** and **g**, that is

$$\mathbf{u} \triangleq \mathbf{u}(\mathbf{f}, \mathbf{g}) \tag{7.13}$$

then

$$E[\mathbf{u}] = \int_{-\infty}^{\infty} up(u)\,\mathrm{d}u \tag{7.14}$$

and

$$\text{Prob}[u < \mathbf{u} < u + \mathrm{d}u] = \text{Prob}\left[f < \mathbf{f} < f + \mathrm{d}f,\ g < \mathbf{g} < g + \mathrm{d}g\right] \tag{7.15}$$

That is

$$p(u)\,\mathrm{d}u = p(f, g)\,\mathrm{d}f\,\mathrm{d}g \tag{7.16}$$

which, upon use of (7.14), gives

$$E[\mathbf{u}] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(f, g)p(f, g)\,\mathrm{d}f\,\mathrm{d}g. \tag{7.17}$$

In the special case of (7.13) with

$$\mathbf{u} = \mathbf{fg} \tag{7.18}$$

we have

$$E[\mathbf{fg}] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} fgp(f, g)\,\mathrm{d}f\,\mathrm{d}g. \tag{7.19}$$

Furthermore, if **f** and **g** are independent variables, then use of (7.11) in (7.19) gives

$$\begin{aligned} E[\mathbf{fg}] &= \int_{-\infty}^{\infty} fp(f)\,\mathrm{d}f \int_{-\infty}^{\infty} gp(g)\,\mathrm{d}g \\ &= E[\mathbf{f}]E[\mathbf{g}]. \end{aligned} \tag{7.20}$$

(ii) The nth moment of the random variable **f** is defined as

$$E[\mathbf{f}^n] = \int_{-\infty}^{\infty} f^n p(f)\,\mathrm{d}f. \tag{7.21}$$

In particular, $n = 2$ gives the second moment as

$$E[\mathbf{f}^2] = \int_{-\infty}^{\infty} f^2 p(f)\,\mathrm{d}f. \tag{7.22}$$

It is always possible to 'centre' the variable by subtracting from it, its mean η; this gives the centred variable $\mathbf{f}_c$ as

$$\begin{aligned} \mathbf{f}_c &= \mathbf{f} - \eta \\ &= \mathbf{f} - E[\mathbf{f}] \end{aligned} \tag{7.23}$$

which is a *zero-mean* variable.

The second central moment of **f** is given by

$$E[\mathbf{f}_c^2] = E[(\mathbf{f}-\eta)^2]. \tag{7.24}$$

Noting that the expectation operator $E[.]$ is linear, we have

$$\begin{aligned} E[(\mathbf{f}-\eta)^2] &= E[\mathbf{f}^2] - E[2\eta\mathbf{f}] + \eta^2 \\ &= E[\mathbf{f}^2] - 2\eta E[\mathbf{f}] + \eta^2 \\ &= E[\mathbf{f}^2] - \eta^2. \end{aligned} \tag{7.25}$$

(iii) *The central second moment* is called the *variance* of **f** and is denoted by σ_f^2. Thus

$$\begin{aligned} \sigma_f^2 &= E[\mathbf{f}^2] - \eta^2 \\ &= E[\mathbf{f}^2] - E^2[\mathbf{f}]. \end{aligned} \tag{7.26}$$

Example 7.1 Find the mean and variance of a variable **f** with uniform probability density defined by (7.6).

Solution.

$$\begin{aligned} E[\mathbf{f}] &= \int_{f_1}^{f_2} \frac{f}{f_2 - f_1}\,\mathrm{d}f = \frac{1}{2}(f_1 + f_2) \\ &= \eta \end{aligned} \tag{7.27}$$

$$E[\mathbf{f}^2] = \int_{f_1}^{f_2} \frac{f^2}{f_2 - f_1}\,\mathrm{d}f = \frac{f_2^3 - f_1^3}{3(f_2 - f_1)}. \tag{7.28}$$

Substituting from the above two expressions into (7.26) we obtain for the variance

$$\sigma_f^2 = \frac{(f_2 - f_1)^2}{12}. \tag{7.29}$$

Example 7.2 Find the mean and variance for the random variable **f** with Gaussian probability density as given by (7.5).

Solution. Direct application of (7.12) and (7.26) shows that for the Gaussian density

$$E[\mathbf{f}] = \eta_f = \eta \tag{7.30}$$

and

$$E[\mathbf{f}^2] = \sigma^2 + \eta^2$$

or

$$\sigma_f^2 = \sigma^2. \tag{7.31}$$

7.3 Analog Stochastic Processes

We have seen that a random variable **f** is a rule for assigning a number $\mathbf{f}(\zeta)$ to each outcome ζ_i of an experiment. We now define a *random* or *stochastic process* (or stochastic signal) $\mathbf{f}(t, \zeta)$ as a rule for assigning a function $\mathbf{f}(t, \zeta)$ to every ζ. That is, a stochastic process is a family (or ensemble) of time-functions depending on the parameter ζ. In other words, **f** is a function of both t and ζ. Figure 7.3 illustrates the construction of a stochastic process. For every ζ_i we have a function of time. The entire set of functions is called an *ensemble*, while each individual function is called a *sample* function, or simply a *sample*.

We denote the stochastic process by $\mathbf{f}(t, \zeta)$, and for simplicity we often drop the parameter ζ and denote the process by $\mathbf{f}(t)$. The stochastic process can assume one of the following interpretations:

(a) If t and ζ are variables, then **f** is an *ensemble* of functions $\mathbf{f}(t, \zeta)$.
(b) If ζ is fixed and t is variable, then $\mathbf{f}(t)$ is a single time function or *sample* of the process.
(c) If t is fixed and ζ is variable, then $\mathbf{f}(t)$ is a random variable.
(d) If both t and ζ are fixed, then $\mathbf{f}(t)$ is a *number*.

7.3.1 Statistics of Stochastic Processes

From the above definitions we see that a stochastic process is an infinite number of random variables: one for every t. For a specific t, $\mathbf{f}(t)$ is, therefore, a random variable with probability distribution function

$$P(f, t) = \text{Prob}[\mathbf{f}(t) < f] \tag{7.32}$$

which depends on t, and it is equal to the probability of the *event* $(\mathbf{f}(t) < f)$ consisting of all outcomes ζ_i such that at the specific time t, the samples $\mathbf{f}(t, \zeta_i)$ of the given process are below the number f. The partial derivative of $P(f, t)$ with respect to f is the probability density

$$p(f, t) = \frac{\partial P(f, t)}{\partial f}. \tag{7.33}$$

The function $P(f, t)$ in (7.33) is called the first-order distribution, and $p(f, t)$ in (10.33) is the first-order density of the process $\mathbf{f}(t)$.

At two specific instants t_1 and t_2, $\mathbf{f}(t_2)$ and $\mathbf{f}(t_2)$ are distinct random variables.

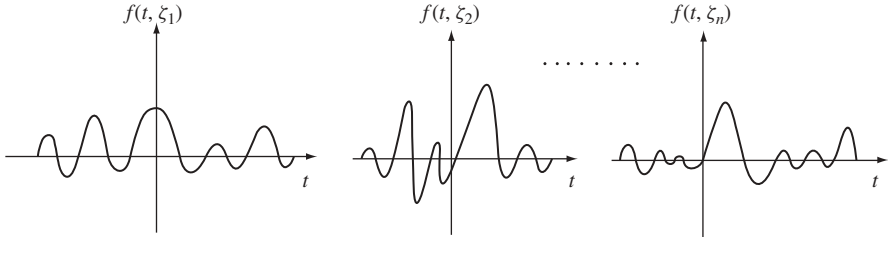


Figure 7.3 Analog stochastic process as an ensemble of samples

Their joint probability distribution is given by

$$P(f_1, f_2;\ t_1,\ t_2) = \text{Prob}[\,\mathbf{f}\,(t_1) < f_1;\ \mathbf{f}\,(t_2) < f_2] \tag{7.34}$$

and their probability density function is

$$p(f_1, f_2;\ t_1,\ t_2) = \frac{\partial^2 P(f_1, f_2;\ t_1,\ t_2)}{\partial f_1 \partial f_2}. \tag{7.35}$$

In order to possess complete information about the properties of a stochastic process, we must know the probability distribution function $P[f_1, f_2, \ldots f_n; t_1, t_2, \ldots t_n]$ for every f_i, t_i and n. However, for many applications only the expected values $E[\mathbf{f}(t)]$ and $E[\mathbf{f}^2(t)]$ are used to characterize the process. These are the second-order properties of the process. For any t, *the mean* $\eta(t)$ of $\mathbf{f}(t)$ is the expected value of the random variable $\mathbf{f}(t)$,

$$\eta(t) = E[\,\mathbf{f}\,(t)] = \int_{-\infty}^{\infty} f p(f,\ t)\,\mathrm{d}f. \tag{7.36}$$

The *mean square* of the process is given by

$$E[\,\mathbf{f}^2(t)] = \int_{-\infty}^{\infty} f^2 p(f,\ t)\,\mathrm{d}f. \tag{7.37}$$

The *autocorrelation* $R_{ff}(t_1, t_2)$ is defined as the expected value (or mean) of the product $\mathbf{f}(t_1)\mathbf{f}(t_2)$, thus

$$\begin{aligned} R_{ff}(t_1,\ t_2) &= E[\,\mathbf{f}\,(t_1)\mathbf{f}\,(t_2)] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P(f_1, f_2;\ t_1,\ t_2)\,\mathrm{d}f_1\,\mathrm{d}f_2 \\ &= R_{ff}(t_2,\ t_1). \end{aligned} \tag{7.38}$$

This parameter is a measure of the inter-relatedness between the instantaneous signal values at t_1 and those at t_2. For $t_1 = t_2 = t$

$$R_{ff}(t,\ t) = E[\,\mathbf{f}^2(t)] \geq 0 \tag{7.39}$$

which is the mean square of the process, and is called the *average power* of $\mathbf{f}(t)$ for reasons to be explained shortly. In fact the autocorrelation is the single most important property of a random process since it leads to a frequency domain representation of the process.

The *cross-correlation* of two processes $\mathbf{f}(t)$ and $\mathbf{g}(t)$ is denoted by $R_{fg}(t_1, t_2)$ and is defined as the expected value of the product $\mathbf{f}(t_1)\mathbf{g}(t_2)$; thus

$$R_{fg}(t_1,\ t_2) = E[\,\mathbf{f}\,(t_1)\mathbf{g}(t_2)]. \tag{7.40}$$

The *cross-covariance* $C_{fg}(t_1, t_2)$ is defined as the expectation of the product

$$\{\mathbf{f}\,(t_1) - \eta_f(t_1)\}\{\mathbf{g}(t_2) - \eta_g(t_2)\}$$

where η_f and η_g are the means of $f(t)$ and $g(t)$ respectively. Thus

$$C_{fg}(t_1,\ t_2) = E[\{\mathbf{f}\,(t_1) - \eta_f(t_1)\}\{\mathbf{g}(t_2) - -\eta_g(t_2)\}]. \tag{7.41}$$

Using the linearity of the expectation operator, expression (7.41) reduces to

$$C_{fg}(t_1, t_2) = E[\mathbf{f}(t_1)\mathbf{g}(t_2)] - \eta_f(t_1)\eta_g(t_2) \tag{7.42}$$

which, upon use of (7.40) becomes

$$C_{fg}(t_1, t_2) = R_{fg}(t_1, t_2) - \eta_f(t_1)\eta_g(t_2). \tag{7.43}$$

The *auto-covariance* of a random process $\mathbf{f}(t)$ is denoted by $C_{ff}(t_1, t_2)$ and is obtained from (7.41) to (7.43) by letting $\mathbf{f} = \mathbf{g}$. Thus

$$\begin{aligned} C_{ff}(t_1, t_2) &= E[\{\mathbf{f}(t_1) - \eta_f(t_1)\}\{\mathbf{f}(t_2) - \eta_f(t_1)\}] \\ &= E[\mathbf{f}^2(t_1)] - \eta_f(t_1)\eta_f(t_2) \\ &= R_{ff}(t_1, t_2) - \eta_f(t_1)\eta_f(t_2). \end{aligned} \tag{7.44}$$

For $t_1 = t_2 = t$ we obtain

$$C_{ff}(t) = R_{ff}(t, t) - \eta_f^2 \tag{7.45}$$

which is the *variance* of $\mathbf{f}(t)$.

7.3.2 Stationary Processes

A stochastic process is said to be *strictly stationary* if all its statistical properties are invariant to a shift of the time origin, that is all its properties are independent of time. However, the process is called *wide-sense stationary* if its mean is independent of time, and its autocorrelation depends only on the difference $\tau = t_1 - t_2$. Thus, a wide-sense stationary process is such that

$$R_{ff}(t_1, t_2) = R_{ff}(\tau) = E[\mathbf{f}(t)\mathbf{f}(t + \tau)]. \tag{7.46}$$

Clearly, if the process is strictly stationary, then it is also wide sense stationary but the converse is not generally true. Henceforth, we shall employ the term *stationary* to mean *wide-sense* stationary.

Two processes $\mathbf{f}(t)$ and $\mathbf{g}(t)$ are said to be *jointly stationary* if both are stationary and their cross-correlation $R_{fg}(t_1, t_2)$ depends only on the difference $(t_1 - t_2) = \tau$. Thus

$$R_{fg}(t_1, t_2) = R_{fg}(\tau) = E\left[f(t)g(t + \tau)\right] \tag{7.47}$$

using (7.43) in the above, it follows that the cross-covariance $C_{fg}(t_1, t_2)$ of two jointly stationary processes depends only on τ, that is

$$C_{fg}(t_1, t_2) = C_{fg}(\tau) = R_{fg}(\tau) - \eta_f\eta_g. \tag{7.48}$$

For a stationary process, the above expression with $\mathbf{f} = \mathbf{g}$ gives

$$C_{ff}(t_1, t_2) = C_{ff}(\tau) = R_{ff}(\tau) - \eta_f^2 \tag{7.49}$$

leading to

$$\begin{aligned} C_{ff}(0) &= E[(\mathbf{f}(t) - \eta_f)^2] \\ &= R_{ff}(0) - \eta_f^2 \end{aligned} \tag{7.50}$$

which is the variance of the stationary random variable $\mathbf{f}(t)$.

7.3.3 Time Averages

Given a stationary random process $\mathbf{f}(t)$ we form a truncated version $\mathbf{f}_{\mathrm{T}}(t)$ as

$$\begin{aligned} \mathbf{f}_{\mathrm{T}}(t) &= \mathbf{f}(t) \quad |t| < \frac{T}{2} \\ &= 0 \quad |t| \geq \frac{T}{2}. \end{aligned} \tag{7.51}$$

Define the integral

$$\begin{aligned} \eta_{\mathrm{T}} &= \frac{1}{T} \int_{-T/2}^{T/2} \mathbf{f}_{\mathrm{T}}(t)\, \mathrm{d}t \\ &= \frac{1}{T} \int_{-T/2}^{T/2} \mathbf{f}(t)\, \mathrm{d}t \end{aligned} \tag{7.52}$$

which varies with each sample of $\mathbf{f}(t)$. Therefore η_{T} is itself a random variable. If the limit

$$\langle \mathbf{f}(t) \rangle \triangleq \lim_{T \to \infty} \eta_{\mathrm{T}} = \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} \mathbf{f}(t)\, \mathrm{d}t \tag{7.53}$$

exists then it is called the *time average of* $\mathbf{f}(t)$.

Similarly, the time average of the product $\{\mathbf{f}(t)\mathbf{f}(t+\tau)\}$ is

$$\langle \mathbf{f}(t)\mathbf{f}(t+\tau) \rangle = \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} \mathbf{f}(t)\mathbf{f}(t+\tau)\, \mathrm{d}t. \tag{7.54}$$

7.3.4 Ergodicity

In the study of stochastic signals, we are interested in estimating their statistical properties such as the mean, mean-square, autocorrelation and so on. As an example, we determine the mean (ensemble average) $\eta(t)$ of the process $\mathbf{f}(t)$ by observing a very large number n of samples $\mathbf{f}(t, \zeta_i)$, thus obtaining an ensemble. The ensemble average is then used as an estimate of $\eta(t)$

$$\eta(t) \approx \frac{1}{n} \sum_i \mathbf{f}(t, \zeta_i). \tag{7.55}$$

However, in many situations we have at our disposal only *one sample* $\mathbf{f}(t, \zeta)$ of the process. Its *time average* is obtained by forming the process in (7.52) and using (7.54). The question is: can we use this *time average* as an estimate of the *ensemble average* $\eta(t)$? Naturally, this would not be possible if $\eta(t)$ depends on t. If $\eta(t) \equiv \eta$, however, a constant then it may be possible to estimate the mean of the process from a single sample. This leads to the subject of *ergodicity*.

7.3.4.1 Definition

A stochastic process is called ergodic if its ensemble averages are equal to the corresponding time averages.

The above definition means that, with probability 'one', any statistic of $\mathbf{f}(t)$ can be determined from a single sample. This definition, however, is too strict, since in most applications we are only interested in *specific* statistics. Therefore, we usually define ergodicity in a limited sense. Thus, if only a certain statistical parameter S can be determined from a single sample, then the process is called *S-ergodic*. For example, we say that the process is *mean-ergodic* or *correlation-ergodic* and so on.

Now, from (7.53) we have

$$E[\eta_{\mathrm{T}}] = \frac{1}{T}\int_{-T/2}^{T/2} E[\mathbf{f}(t)]\,\mathrm{d}t \tag{7.56}$$

where the order of expectation and integration has been reversed. If the process has a constant mean, then

$$\begin{aligned} E[\eta_{\mathrm{T}}] &= \frac{1}{T}\int_{-T/2}^{T/2} \eta_f\,\mathrm{d}t. \\ &= \eta_f. \end{aligned} \tag{7.57}$$

In order that η_{T}, taken over sufficiently large T, to be close to the ensemble mean η_f we must have

$$\lim_{T\to\infty} \eta_{\mathrm{T}} = \eta_f \quad \text{with probability 1} \tag{7.58}$$

and in this case the process is said to be *mean-ergodic*. We express this by the relation

$$E[\mathbf{f}(t)] = \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} \mathbf{f}(t)\,\mathrm{d}t. \tag{7.59}$$

Similarly, *ergodicity of the autocorrelation* for the stationary signal $\mathbf{f}(t)$ gives

$$E[\mathbf{f}(t)\mathbf{f}(t+\tau)] = \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} \mathbf{f}(t)\mathbf{f}(t+\tau)\,\mathrm{d}t \tag{7.60}$$

so that

$$R_{ff}(\tau) = \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} \mathbf{f}(t)\mathbf{f}(t+\tau)\,\mathrm{d}t \tag{7.61}$$

Also, for a correlation-ergodic process, putting $\tau = 0$ in (7.60) we have for the mean square of the process

$$E[\mathbf{f}^2(t)] = \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} \mathbf{f}^2(t)\,\mathrm{d}t. \tag{7.62}$$

Under the same conditions, we also have

$$\begin{aligned} &\lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} [\mathbf{f}(t) + \mathbf{f}(t+\tau)]^2\,\mathrm{d}t \\ &\quad = 2[R(0) + R(\tau)]. \end{aligned} \tag{7.63}$$

7.3.5 Power Spectra of Stochastic Signals

We have seen in Chapter 3 that for a deterministic finite-energy signal $f(t)$, the autocorrelation function $\rho_{ff}(\tau)$ and its energy spectrum $E(\omega)$ constitute a Fourier transform pair, as expressed by the relation

$$\rho_{ff}(\tau) \leftrightarrow E(\omega) \tag{7.64}$$

where

$$\begin{aligned} E(\omega) &= F(\omega)F*(\omega) \\ &= |F(\omega)|^2 \end{aligned} \tag{7.65}$$

with

$$f(t) \leftrightarrow F(\omega) \tag{7.66}$$

That is

$$\rho_{ff}(\tau) \leftrightarrow |F(\omega)|^2. \tag{7.67}$$

Moreover, the cross-correlation $\rho_{fg}(\tau)$ of two finite energy signals $f(t)$ and $g(t)$, together with the cross-energy spectrum $F*(\omega)G(\omega) = E_{fg}(\omega)$ form a Fourier transform pair, that is

$$\rho_{fg}(\tau) \leftrightarrow F*(\omega)G(\omega) \tag{7.68}$$

which is the Wiener–Kintchine relation, a special case of which is (7.65) for $f(t) = g(t)$.

Turning now to stochastic signals, we note that these are not square integrable and, in general, do not possess Fourier transforms. Therefore, we seek an alternative frequency-domain representation of the statistical properties of such signals. This is usually accomplished in terms of their *power spectra*, rather than the energy spectra. We shall concentrate on stationary signals which are also mean-ergodic and correlation-ergodic.

7.3.5.1 Power Spectrum

The power spectral density, or simply the *power spectrum* $P_{ff}(\omega)$ of a stationary process $\mathbf{f}(t)$ is defined as the Fourier transform of its autocorrelation, that is

$$P_{ff}(\omega) = \int_{-\infty}^{\infty} R_{ff}(\tau) \exp(-\mathrm{j}\omega t)\, \mathrm{d}\tau \tag{7.69}$$

with the inverse relation

$$R_{ff}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} P_{ff}(\omega) \exp(\mathrm{j}\omega\tau)\, \mathrm{d}\omega \tag{7.70}$$

so that we have the Fourier transform pair

$$R_{ff}(\tau) \leftrightarrow P_{ff}(\omega) \tag{7.71}$$

which is analogous to (7.64) for finite energy signals. The justification for the use of (7.69) as the definition of the power spectrum stems from the fact that the average power of the process $\mathbf{f}(t)$ is given by

$$E[\mathbf{f}^2(t)] = R(0) \tag{7.72}$$

which is obtained by putting $\tau = 0$ in (7.70) to give

$$E[\mathbf{f}^2(t)] = \frac{1}{2\pi}\int_{-\infty}^{\infty} P_{ff}(\omega)\,\mathrm{d}\omega. \tag{7.73}$$

This reveals that the area under the curve of $P_{ff}(\omega)$ gives the average power of the signal, hence $P_{ff}(\omega)$ is the power spectral density. Clearly, expression (7.73) is the stochastic signal counterpart of Parseval's relation (2.49) for deterministic finite-energy signals. Since (7.69) implies the existence of the power integral, the signals under investigation are called *finite-power* signals.

Now, the stationarity of the random process means that its autocorrelation function

$$R_{ff}(\tau) = E[\mathbf{f}(t)\mathbf{f}(t+\tau)] \tag{7.74}$$

depends only on τ, so that

$$R_{ff}(-\tau) = R_{ff}(\tau) \tag{7.75}$$

Thus it is an even function of τ. From (7.69) it also follows that, since $\mathbf{f}(t)$ is real, the power spectrum $P_{ff}(\omega)$ is real and even

$$P_{ff}(-\omega) = P_{ff}(\omega). \tag{7.76}$$

Therefore, using (7.69) we have

$$P_{ff}(\omega) = \int_{-\infty}^{\infty} R_{ff}(\tau)\cos\,\omega\tau\,\mathrm{d}\tau \tag{7.77}$$

and

$$R_{ff}(\tau) = \frac{1}{2\pi}\int_{-\infty}^{\infty} P_{ff}(\omega)\,\cos\,\omega\tau\,\mathrm{d}\omega. \tag{7.78}$$

At $\omega = 0$

$$P_{ff}(0) = \int_{-\infty}^{\infty} R_{ff}(\tau)\,\mathrm{d}\tau \tag{7.79}$$

which means that the area under the autocorrelation curve equals the power spectrum at zero frequency.

For a correlation-ergodic process, the autocorrelation, hence the power spectrum, can be obtained from time averages as

$$R_{ff}(\tau) = \lim_{T\to\infty}\frac{1}{T}\int_{-T/2}^{T/2}\mathbf{f}(t)\mathbf{f}(t+\tau)\,\mathrm{d}t. \tag{7.80}$$

The above expression together with (7.69) and (7.70) form the basis for the estimation of the power spectrum of a stochastic process. The process is observed over a sufficiently large period and the expression

$$R_{ff}(\tau) \approx \frac{1}{T}\int_{-T/2}^{T/2}\mathbf{f}(t)\mathbf{f}(t+\tau)\,\mathrm{d}t \tag{7.81}$$

is taken as an estimate of the autocorrelation. This estimate can be calculated using the FFT algorithms of Chapter 6 (see Section 6.5). Then from this estimate the power spectrum of the signal is calculated using (7.69) and another FFT step. Naturally, all the preliminary steps of sampling the signal and approximating the integrals by summations are implied in the procedure. Further details of the procedure will be given shortly.

7.3.5.2 Cross-power Spectrum

For two jointly stationary processes $\mathbf{f}(t)$ and $\mathbf{g}(t)$, the cross-power spectrum $P_{fg}(\omega)$ is defined as the Fourier transform of their cross-correlation. Thus

$$P_{fg}(\omega) = \int_{-\infty}^{\infty} R_{fg}(\tau) \exp(-\mathrm{j}\omega\tau)\, \mathrm{d}\tau \tag{7.82}$$

with the inverse relation

$$R_{fg}(\tau) = \frac{1}{2\pi} \int_{-\infty}^{\infty} P_{fg}(\omega) \exp(\mathrm{j}\omega\tau)\, \mathrm{d}\omega \tag{7.83}$$

so that the cross-correlation and the cross-power spectrum (density) form a Fourier transform pair

$$R_{fg}(\tau) \leftrightarrow P_{fg}(\omega). \tag{7.84}$$

Again, for correlation-ergodic jointly stationary processes, the cross-correlation can be obtained from the time average

$$R_{fg}(\tau) = \lim_{T\to\infty} \frac{1}{T} \int_{-T/2}^{T/2} \mathbf{f}(t)\mathbf{g}(t+\tau)\, \mathrm{d}t \tag{7.85}$$

which is equal to the ensemble average given by (7.40). The cross-power spectrum has the property

$$P_{fg}(\omega) = P_{fg}^{*}(\omega). \tag{7.86}$$

Also, using similar analysis to that employed in Chapter 6 together with (7.83) and (7.84), we obtain for stationary correlation-ergodic processes

$$P_{fg}(\omega) = P_{ff}^{*}(\omega) P_{gg}(\omega) \tag{7.87}$$

which is analogous to (2.53) for finite-energy signals.

Again the FFT algorithms of Chapter 6 can be used to estimate the cross-power spectrum. First we observe $\mathbf{f}(t)$ and $\mathbf{g}(t)$ over a sufficiently long T and take

$$R_{fg}(\tau) \approx \frac{1}{T} \int_{-T/2}^{T/2} \mathbf{f}(t)\mathbf{g}(t+\tau)\, \mathrm{d}t \tag{7.88}$$

to be an estimate of the cross-correlation. This estimate can be obtained using an FFT algorithm (see Section 6.5). Then the cross-power spectrum is calculated from (7.82) and the estimate of $R_{fg}(\tau)$ using an FFT algorithm. Of course, all the preliminary steps of sampling the signals and approximating the integrals by sums are implied in the estimation process. The details of this procedure will be given shortly.

7.3.5.3 White Noise

A random process whose power spectrum is constant at all frequencies is called *white noise*. For such a signal

$$P_{\mathrm{WN}}(\omega) = A \quad \text{a constant} \tag{7.89}$$

so that its inverse Fourier transform gives its autocorrelation as

$$R_{\mathrm{WN}}(\tau) = A\,\delta(\tau) \tag{7.90}$$

which is an impulse at $\tau = 0$.

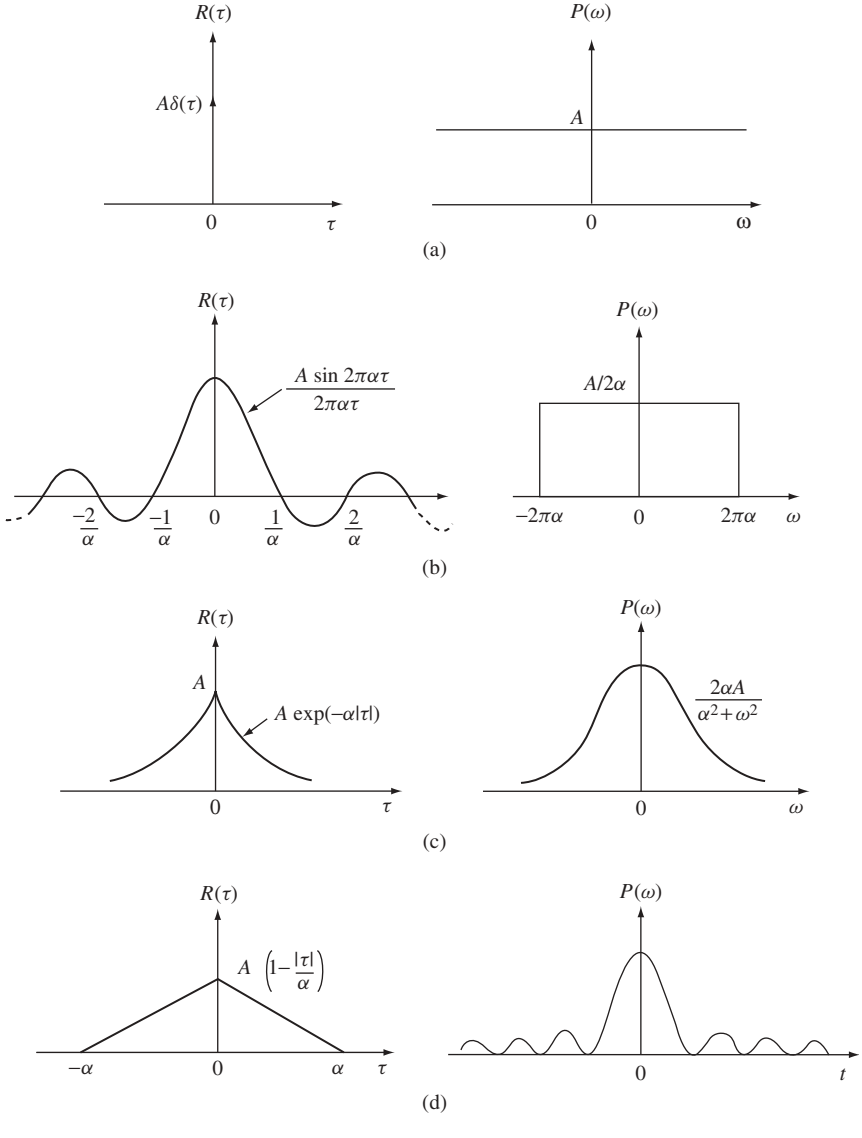


Figure 7.4 Examples of autocorrelation functions and the associated power spectra

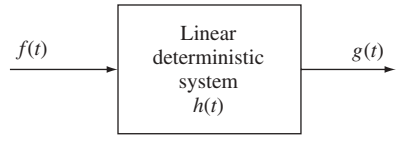


Figure 7.5 A stochastic signal through a linear deterministic system

Figure 7.4 shows some autocorrelation functions and the corresponding power spectra. Figure 7.4(a) is the white noise described by (7.89) and (7.90). Figure 7.4(b) is a band-limited white noise. Figure 7.4(c) represents thermal noise through a resistor.

7.3.6 Signals through Linear Systems

Consider a *deterministic* linear system, as shown in Figure 7.5, with impulse response $h(t)$ and transfer function $H(\omega)$. Thus

$$h(t) \leftrightarrow H(\omega) \tag{7.91}$$

with

$$H(\omega) = \int_{-\infty}^{\infty} h(t) \exp(-\mathrm{j}\omega t)\, \mathrm{d}t. \tag{7.92}$$

Since $h(t)$ is a deterministic function of finite energy, then its energy spectrum is $|H(\omega)|^2$, and its autocorrelation $\rho(\tau)$ is given by

$$\rho(t) = \int_{-\infty}^{\infty} h(\tau) h(\tau + t)\, \mathrm{d}\tau \tag{7.93}$$

and from (7.67) we have the pair

$$\rho(t) \leftrightarrow |H(\omega)|^2. \tag{7.94}$$

Let us now apply a stationary random signal $\mathbf{f}(t)$ as the input to the linear system. The resulting output is given by the convolution

$$\begin{aligned} \mathbf{g}(t) &= \int_{-\infty}^{\infty} \mathbf{f}(t - \alpha) h(\alpha)\, \mathrm{d}\alpha \\ &= \int_{-\infty}^{\infty} \mathbf{f}(\alpha) h(t - \alpha)\, \mathrm{d}\alpha \end{aligned} \tag{7.95}$$

which is also a stochastic process. If the system is *causal*, then $g(t) = 0$ for $t < 0$ and $\mathbf{g}(t)$ becomes

$$\begin{aligned} \mathbf{g}(t) &= \int_{-\infty}^{\infty} \mathbf{f}(t - \alpha) h(\alpha)\, \mathrm{d}\alpha \\ &= \int_{-\infty}^{\infty} \mathbf{f}(\alpha) h(t - \alpha)\, \mathrm{d}\alpha. \end{aligned} \tag{7.96}$$

Next consider the problem of determining the output autocorrelation. R_{gg} and its power spectrum $P_g(\omega)$. Starting from (7.95), we multiply both sides by $\mathbf{f}(t+\tau)$ and take the expectation to obtain

$$E[\mathbf{g}(t)\mathbf{f}(t+\tau)] = \int_{-\infty}^{\infty} E[\mathbf{f}(t+\tau)\mathbf{f}(t-\alpha)]h(\alpha)\,\mathrm{d}\alpha. \tag{7.97}$$

However

$$E[\mathbf{f}(t+\tau)\mathbf{f}(t-\alpha)] = R_{ff}(\tau+\alpha). \tag{7.98}$$

Therefore

$$\begin{aligned} R_{fg}(\tau) &= \int_{-\infty}^{\infty} R_{ff}(\tau+\alpha)h(\alpha)\,\mathrm{d}\alpha \\ &= \int_{-\infty}^{\infty} R_{ff}(\tau-\beta)h(-\beta)\,\mathrm{d}\beta \end{aligned} \tag{7.99}$$

or

$$R_{fg}(\tau) = R_{ff}(\tau) * h(-\tau). \tag{7.100}$$

Similarly, multiplying (7.96) by $\mathbf{g}(t-\tau)$ and taking expectation

$$E[\mathbf{g}(t)\mathbf{g}(t-\tau)] = \int_{-\infty}^{\infty} E[\mathbf{f}(t-\alpha)\mathbf{g}(t-\tau)]h(\alpha)\,\mathrm{d}\alpha. \tag{7.101}$$

Hence

$$R_{gg}(\tau) = \int_{-\infty}^{\infty} R_{fg}(t-\alpha)h(\alpha)\,\mathrm{d}\alpha \tag{7.102}$$

or

$$R_{gg}(\tau) = R_{fg}(\tau) * h(\tau). \tag{7.103}$$

Now, using the convolution theorem and taking the Fourier transform of (7.100) and (7.103) we obtain

$$P_{fg}(\omega) = P_{ff}H*(\omega) \tag{7.104}$$

and

$$P_{gg}(\omega) = P_{fg}(\omega)H(\omega). \tag{7.105}$$

Combining (7.100), (7.103), (7.104) and (7.105) we have

$$R_{gg}(\tau) = R_{ff} * \rho(\tau) \tag{7.106}$$

and

$$P_{gg}(\omega) = P_{ff}|H(\omega)|^2. \tag{7.107}$$

In words, the above expressions mean that the output autocorrelation of a linear system with a stationary input is the convolution of the autocorrelation of the input with the autocorrelation of the (finite energy) impulse response of the systems. Equivalently, in the frequency domain, the output power spectrum is the product of the input power spectrum with the *energy spectrum* of the impulse response function.

7.3.6.1 System Identification

Expression (7.100) provides a means for measuring the impulse response of a system. White noise is fed into the system so that in (7.100) we put $R_{ff}(\tau) = A\delta(\tau)$. Then, the cross-correlation between the input and output is measured. This gives

$$\begin{aligned} R_{fg}(\tau) = R_{fg}(-\tau) &= A\delta(\tau) * h(-\tau) \\ &= Ah(\tau) \end{aligned} \tag{7.108}$$

giving a scaled version of the impulse response of the system.

7.4 Discrete-time Stochastic Processes

We now discuss the discrete counterparts of the definitions and concepts given, so far, for stochastic continuous-time processes. Here also, we concentrate on real processes.

A discrete real process $\mathbf{f}(n)$ is a sequence of real random variables, defined for every integer n. $\mathbf{f}(n)$ may be equal to the time-samples $\mathbf{f}(nT)$, of a continuous-time process $\mathbf{f}(t)$, taken every T seconds. In this case

$$\mathbf{f}(n) \equiv \mathbf{f}(nT). \tag{7.109}$$

For convenience, we drop the sampling period T from the argument. This way, we also accommodate the cases where the argument is any discrete quantity other than time. However, we still refer to n as the discrete time variable.

7.4.1 Statistical Parameters

The mean $\eta(n)$, autocorrelation $R(n_1, n_2)$ and autocovariance $C(n_1, n_2)$ of $\mathbf{f}(n)$ are defined by

$$\eta(n) = E[\mathbf{f}(n)] \tag{7.110}$$

$$R_{ff}(n_1, n_2) = E[\mathbf{f}(n_1)\mathbf{f}(n_2)] \tag{7.111}$$

$$C_{ff}(n_1, n_2) = R_{ff}(n_1, n_2) - \eta(n_1)\eta(n_2). \tag{7.112}$$

For two discrete processes $\mathbf{f}(n)$ and $\mathbf{g}(n)$ we define the cross-correlation as

$$R_{fg}(n_1, n_2) = E[\mathbf{f}(n_1)\mathbf{g}(n_2)] \tag{7.113}$$

and the cross-covariance as

$$C_{fg}(n_1, n_2) = E[\{\mathbf{f}(n_1) - \eta_f(n_1)\}\{\mathbf{g}(n_2) - \eta_g(n_2)\}]. \tag{7.114}$$

7.4.2 Stationary Processes

A discrete-time process $\mathbf{f}(n)$ is said to be (wide-sense) stationary if its mean is constant and its autocorrelation depends only on the difference $m = (n_1 - n_2)$, thus

$$\eta(n) = \eta \quad \text{a constant} \tag{7.115}$$

and

$$R_{\mathbf{ff}}(m) = E[\,\mathbf{f}(n)\mathbf{f}(n+m)]. \tag{7.116}$$

For a stationary process, the mean can be removed giving a zero-mean process. In this case, expressions (7.110) and (7.112) show that the autocorrelation equals the autocovariance.

Two processes $\mathbf{f}(n)$ and $\mathbf{g}(n)$ are said to be *jointly stationary* if both are stationary and their cross-correlation depends on the difference $m = (n_1 - n_2)$ only, thus

$$R_{\mathbf{fg}}(m) = E[\,\mathbf{f}(n)\mathbf{g}(n+m)]. \tag{7.117}$$

For stationary processes with zero means, (7.114) shows that the cross-correlation and cross-covariance are the same.

The *power spectrum* $P_{\mathbf{ff}}(\omega)$ of a stationary process $\mathbf{f}(n)$ is defined as the periodic function with Fourier series coefficients $R(m)$, that is

$$P_{\mathbf{ff}}(\omega) = \sum_{m=-\infty}^{\infty} R_{\mathbf{ff}}(m)\exp(-\mathrm{j}mT\omega) \tag{7.118}$$

and

$$R_{\mathbf{ff}}(m) = \frac{1}{2\alpha}\int_{-\alpha}^{\alpha} P_{\mathbf{ff}}(\omega)\exp(\mathrm{j}mT\omega)\,\mathrm{d}\omega \tag{7.119}$$

where, in general

$$\alpha = \pi/T \tag{7.120}$$

which is an arbitrary constant. However, if $f(n)$ is obtained by sampling a continuous stochastic process, then T is the sampling period. Otherwise we may put

$$T = 1 \;\; \alpha = \pi \tag{7.121}$$

so that

$$P_{\mathbf{ff}}(\omega) = \sum_{m=-\infty}^{\infty} R_{\mathbf{ff}}(m)\exp(-\mathrm{j}m\omega) \tag{7.122}$$

and

$$R_{\mathbf{ff}}(m) = \frac{1}{2\pi}\int_{-\pi}^{\pi} P_{\mathbf{ff}}(\omega)\exp(\mathrm{j}m\omega)\,\mathrm{d}\omega. \tag{7.123}$$

The average power of the signal is its variance obtained by putting $m = 0$ in (7.118) and (7.119) to give

$$\begin{aligned} E[\,\mathbf{f}^2(n)] &= R_{\mathbf{ff}}(0) \\ &= \frac{1}{2\alpha}\int_{-\alpha}^{\alpha} P_{\mathbf{ff}}(\omega)\,\mathrm{d}\omega. \end{aligned} \tag{7.124}$$

From (7.119) it is clear that, since $\mathbf{f}(n)$ is a real process, then $R(m)$ is a real even sequence. It also follows from (7.122) that $R_{\mathbf{ff}}(\omega)$ is a real even function. We also note

that the variance of a stationary signal with zero mean is equal to its average power as given by (7.118).

The *cross-power spectrum* of two jointly stationary process $\mathbf{f}(n)$ and $\mathbf{g}(n)$ is defined by

$$P_{fg}(\omega) = \sum_{m=\infty}^{\infty} R_{fg}(m)\, \exp(-\mathrm{j}mT\omega). \tag{7.125}$$

7.4.2.1 White Noise

A process $\mathbf{w}(n)$ is called white noise, if

$$E[\mathbf{w}(n_1)\mathbf{w}(n_2)] = 0 \;\; n_1 \neq n_2. \tag{7.126}$$

Putting

$$E\{|\mathbf{w}(n)|^2\} = \{I(n)\} \tag{7.127}$$

the autocorrelation of $\mathbf{w}(n)$ is given by

$$R_{ww}(n_1, n_2) = I(n_1)u_0(n_1 - n_2) \tag{7.128}$$

where u_0 is the discrete impulse. If $\mathbf{w}(n)$ is stationary, then $I(n) = I$, a constant so that

$$R_{ww}(m) = Iu_0(m) \tag{7.129}$$

and

$$P_{ww}(\omega) = I. \tag{7.130}$$

With the foregoing definitions, the concept of ergodicity can be also defined in a similar manner to the analog case. Thus $\mathbf{f}(n)$ is mean-ergodic if its ensemble average equals its time average, that is

$$E[\mathbf{f}(n)] = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbf{f}(n). \tag{7.131}$$

Similarly ergodicity of the autocorrelation for the stationary signal $\mathbf{f}(n)$ gives

$$\begin{aligned} E[\mathbf{f}(n)\mathbf{f}(n+m)] &= \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbf{f}(n)\mathbf{f}(n+m) \\ &= R_{ff}(m) \end{aligned} \tag{7.132}$$

and the cross-correlation of two sequences $\mathbf{f}(n)$ and $\mathbf{g}(n)$ satisfies

$$E[\mathbf{f}(n)\mathbf{g}(n+m)] = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbf{f}(n)\mathbf{g}(n+m). \tag{7.133}$$

Again the estimation of power spectra can be accomplished using the FFT algorithms of Chapter 6. First, the cross-correlation or autocorrelation sequence is estimated using a

finite number of samples as

$$R_{fg}(m) \approx \frac{1}{N}\sum_{n=0}^{N-1} \mathbf{f}(n)\mathbf{g}(n+m) \tag{7.134}$$

$$R_{ff}(m) \approx \frac{1}{N}\sum_{n=0}^{N-1} \mathbf{f}(n)\mathbf{f}(n+m). \tag{7.135}$$

Either of the above estimates can be calculated using an FFT algorithm (see Section 6.5). Then the spectra in (7.122) and (7.125) are estimated by finite sums and another FFT step. This procedure will be discussed in detail shortly together with the various methods which may be employed to improve the accuracy of the estimates.

It is also possible to define the (two-sided) z-transform of the autocorrelation sequence $\{R(m)\}$ as

$$P_{ff}(z) = \sum_{m=-\infty}^{\infty} R(m)z^{-m} \tag{7.136}$$

and for the cross-correlation sequence

$$P_{fg}(z) = \sum_{m=-\infty}^{\infty} R_{fg}(m)z^{-m}. \tag{7.137}$$

The one-sided transforms are obtained from the above expressions by changing the lower sum limit to 0. Let a stationary random process $\mathbf{f}(n)$ be applied to a deterministic linear discrete time-invariant system with impulse response sequence $h(n)$ and transfer function $H(z)$. Then the output $\{\mathbf{g}(n)\}$ is also a stationary random process given by

$$\mathbf{g}(n) = \sum_{m=-\infty}^{\infty} \mathbf{f}(m)h(n-m). \tag{7.138}$$

If the system is causal, then $h(n) = 0$ for $n < 0$ and the output becomes

$$\mathbf{g}(n) = \sum_{m=0}^{\infty} \mathbf{f}(m)h(n-m). \tag{7.139}$$

Let

$$H(z) = \sum_{n=-\infty}^{\infty} h(n)z^{-n} \tag{7.140}$$

and the system has a deterministic impulse response $h(n)$ with autocorrelation sequence given by

$$\rho(n) = \sum_{k=-\infty}^{\infty} h(n+k)h(k) \tag{7.141}$$

so that

$$Z\{\rho(n)\} = \sum_{n=-\infty}^{\infty}\left(\sum_{k=-\infty}^{\infty} h(n+k)h(k)\right)z^{-n}. \tag{7.142}$$

Putting $i = n + k$ the above expression becomes

$$Z\{\rho(n)\} = \sum_{i=-\infty}^{\infty} h(i)z^{-i} \sum_{k=-\infty}^{\infty} h(k)z^{k} \tag{7.143}$$

which, upon use of (7.131), gives

$$Z\{\rho(n)\} = H(z)h(z^{-1}). \tag{7.144}$$

Reasoning as in the analog case, we obtain the relations

$$P_{fg}(z) = P_{ff}(z)H(z^{-1}) \tag{7.145}$$

$$P_{gg}(z) = P_{fg}(z)H(z) \tag{7.146}$$

so that

$$P_{gg}(z) = P_{ff}(z)H(z)H(z^{-1}). \tag{7.147}$$

On the unit circle ($z = \exp(j\omega T)$ the above expression becomes

$$P_{gg}(\omega) = P_{ff}(w)|H(\exp(\mathrm{j}wT))|^2. \tag{7.148}$$

The above analysis and conclusions also hold true for the one-sided z-transform and causal sequences.

7.5 Power Spectrum Estimation

7.5.1 Continuous-time Signals

A problem of considerable importance in signal processing is that of the *estimation* of the power spectrum $P_{ff}(\omega)$ of a process $\mathbf{f}(t)$ when only a segment $\mathbf{f}_{\mathrm{T}}(t)$ is available. If the autocorrelation $R_{ff}(\tau)$ is *known* for every τ in the interval $[-T/2, T/2]$ then expression (7.69) can be used together with an FFT algorithm to estimate the power spectrum. In general, however, the function $R_{ff}(\tau)$ is not known exactly and we are only given the value of the process $\mathbf{f}(t)$ for every t in the interval $[-T/2, T/2]$. Therefore, the autocorrelation $R_{ff}(\tau)$ *itself* is estimated from a time average. It follows that the estimation of power spectra consists of two main stages.

Stage 1 Given a segment $\mathbf{f}_{\mathrm{T}}(t)$ defining a single sample of $\mathbf{f}(t)$ for every t in the interval $[-T/2, T/2]$, we evaluate an estimate of the autocorrelation $R_{ff}(\tau)$. It will be shown that the calculations can be performed using an FFT algorithm.

Stage 2 Using the autocorrelation estimate obtained in *Stage 1*, the power spectrum given by (7.69) is calculated using an FFT algorithm.

Now, in estimating any parameter $\mathbf{A}$, it is observed over an interval, and an estimate $\hat{\mathbf{A}}$ is, somehow, made. Two quantities are usually used for measuring the *quality* of the estimate; these are the *estimation bias and variance*. The bias is given by the difference between the *true* value $\mathbf{A}$ and the *expected* value of the estimate, that is

$$\text{Bias} \triangleq \mathbf{A} - E[\hat{\mathbf{A}}]. \tag{7.149}$$

If the bias is zero, the estimate is said to be *unbiased*. The *variance* of the estimate is

$$\sigma_{\mathrm{A}}^2 = E[\{\hat{\mathbf{A}} - E[\hat{\mathbf{A}}]\}^2]. \tag{7.150}$$

The estimate is said to be *consistent* if the bias and variance tend to zero as the observation period increases. In general, a good estimate has small bias and variance.

Let us return to *Stage 1*, which requires the estimation of the autocorrelation $R_{ff}(\tau)$ from the segment $\mathbf{f}_{\mathrm{T}}(t)$. An obvious approach is to use the estimate

$$R_{ff}^{\mathrm{T}}(\tau) = \frac{1}{(T/2) - |\tau|} \int_{-(T/2)+|\tau|/2}^{(T/2)-|\tau|/2} \mathbf{f}(t + \tau/2)\mathbf{f}(t - \tau/2)\,\mathrm{d}t \tag{7.151}$$

which is the same as (7.54) with a shift in the time origin, and taken without the limiting operation. Clearly, expression (7.151) is an *unbiased* estimate of $R_{ff}(\tau)$ since its expectation is $R_{ff}(\tau)$

$$E[R_{ff}^{\mathrm{T}}(\tau)] = R_{ff}(\tau) \quad |\tau| < T. \tag{7.152}$$

However, the Fourier transform of $R_{ff}^{\mathrm{T}}(\tau)$ is

$$P_{ff}^{\mathrm{T}}(\omega) = \int_{-T}^{T} R_{ff}^{\mathrm{T}}(\tau) \exp(-\mathrm{j}\omega\tau)\,\mathrm{d}\tau \tag{7.153}$$

which is a *biased* estimate of $P_{ff}(\omega)$ because its inverse is zero for $|\tau| > T$. Moreover, investigation of the variance of $P_{ff}^{\mathrm{T}}(\omega)$ shows that it is large for *any* T. To minimize this effect a *window* may be used, as discussed in Section 4.8. However, although this reduces the variance, it increases the bias.

An alternative estimate of the autocorrelation is given by

$$\hat{R}_{ff}(\tau) = \frac{1}{T} \int_{-(T/2)+|\tau|/2}^{(T/2)-|\tau|/2} \mathbf{f}(t + \tau/2)\mathbf{f}(t - \tau/2)\,\mathrm{d}t \tag{7.154}$$

which has the distinct advantage over that of (7.151) in that its Fourier transform can be expressed directly in terms of $\mathbf{f}_{\mathrm{T}}(t)$: the available segment. The Fourier transform of (7.154) is

$$\hat{P}_{ff}(\omega) = \int_{-T}^{T} \hat{R}_{ff}(\tau) \exp(-\mathrm{j}\omega\tau)\,\mathrm{d}\tau. \tag{7.155}$$

However, the integral in (7.154) is the convolution of $\mathbf{f}_{\mathrm{T}}(t)$ with $\mathbf{f}_{\mathrm{T}}(-t)$, that is

$$\hat{R}_{ff}(\tau) = \frac{1}{T}\mathbf{f}_{\mathrm{T}}(\tau) * \mathbf{f}_{\mathrm{T}}(-\tau) \tag{7.156}$$

so that, using the convolution theorem with (7.155) we obtain

$$\hat{P}_{ff}(\omega) = \frac{1}{T}|\mathbf{F}_{\mathrm{T}}(\omega)|^2 \tag{7.157}$$

and

$$\mathbf{F}_{\mathrm{T}}(\omega) = \int_{-T/2}^{T/2} \mathbf{f}_{\mathrm{T}}(t) \exp(-\mathrm{j}\omega t)\,\mathrm{d}t. \tag{7.158}$$

Hence, the spectral estimate $\hat{P}_{ff}(\omega)$ in (7.157) is obtained directly in terms of $\mathbf{f}_{\mathrm{T}}(t)$.

$\hat{P}_{ff}(\omega)$ as defined above is called the *sample spectrum* or the *periodogram*. It is a biased spectral estimate because

$$E[\hat{R}_{ff}(\tau)] = \left(1 - \frac{|\tau|}{T}\right) R_{ff}(\tau) \quad |\tau| < T. \tag{7.159}$$

Nevertheless, the error due to the term $(1 - |\tau|/T)$ is small in the region of interest $|\tau| \ll T$. This is also compensated by a reduction in the variance of the estimate. Furthermore, the variance can be reduced further by spectral windows. Thus, we form the windowed autocorrelation

$$\hat{R}^{\mathrm{w}}_{ff}(\omega) = w(\tau)\hat{R}_{ff}(\tau) \tag{7.160}$$

so that the *smoothed spectrum* is obtained as

$$\hat{P}^{\mathrm{w}}_{ff}(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{P}_{ff}(\omega - \mu)W(\mu)\,\mathrm{d}\mu \tag{7.161}$$

with

$$w(t) \leftrightarrow W(\omega) \tag{7.162}$$

in which $W(\omega)$ is a *spectral window*, and its inverse $w(t)$ is a lag window (see Section 6.8). The smoothed spectrum can be written in the alternative form

$$\hat{P}^{\mathrm{w}}_{ff}(\omega) = \int_{-T}^{T} w(\tau)\hat{R}_{ff}(\tau)\exp(-\mathrm{j}\omega t)\,\mathrm{d}\tau. \tag{7.163}$$

The issues involved in using windows are the same as discussed before in Section 6.8 and other locations in the book where windows were used. The computational steps constituting spectral estimation from either (7.151) or (7.154) are as follows:

(i) Compute $\hat{P}_{ff}(\omega)$ from (7.155). This requires one FFT and one multiplication.
(ii) Find the inverse of $\hat{P}_{ff}(\omega)$; this is $R_{ff}(\tau)$. This requires one FFT and one multiplication.
(iii) Chose a window $w(\tau)$ and form the product $\hat{\mathbf{R}}^{\mathrm{w}}_{ff}(\tau) = w(\tau)\mathbf{R}_{ff}(\tau)$. Find the transform $\hat{P}^{\mathrm{w}}_{ff}(\omega)$ of $\hat{R}^{\mathrm{w}}_{ff}(\tau)$. This step requires one multiplication and one FFT. The result is the required spectral estimate.

Finally, the cross-power spectrum $P_{fg}(\omega)$ of two signals $\mathbf{f}(t)$ and $\mathbf{g}(t)$ can be obtained using the same procedure. Instead of (7.154) we have the estimate

$$\hat{R}_{fg}(\tau) = \frac{1}{T}\int_{-(T/2)+|\tau|/2}^{(T/2)-|\tau|/2} \mathbf{f}(t + \tau/2)\mathbf{g}(t - \tau/2)\,\mathrm{d}t \tag{7.164}$$

which leads to

$$\hat{P}_{fg}(\omega) = \frac{1}{T}F*(\omega)G(\omega). \tag{7.165}$$

The windowed cross-correlation is

$$\hat{R}^{\mathrm{w}}_{fg}(\tau) = w(\tau)\hat{R}_{fg}(\tau) \tag{7.166}$$

and the cross-power estimate is obtained by the appropriate rephrasing of the same procedure given in case of the power spectrum in the two stages, and also employing the FFT algorithms.

7.5.2 Discrete-time Signals

A discrete process $\mathbf{f}(n)$ with autocorrelation sequence $R_{ff}(m)$ has a power spectrum given by (7.118) or (7.122). Thus

$$P_{ff}(\omega) = \sum_{m=-\infty}^{\infty} R_{ff}(m) \exp(-\mathrm{j}m\omega T). \tag{7.167}$$

Again, it is assumed that $\mathbf{f}(n)$ is known only for a finite number of samples N. Let this segment be denoted by $\mathbf{f}_N(n)$. The estimation of the power spectrum can be accomplished by following very closely the procedure given before for continuous-time signals, with the added simplification that the signal is already available in discrete form. Thus the FFT algorithms can be used directly without the need for sampling the signal. Therefore, the estimation of the spectrum can be achieved in the following stages.

Stage 1 An estimate $\hat{R}_{ff}(m)$ is found for the autocorrelation as the finite sequence

$$\hat{R}_{ff}(m) = \frac{1}{N} \sum_{n=0}^{N-1} \mathbf{f}_N(n)\mathbf{f}_N(n+m) \tag{7.168}$$

which can be evaluated using an FFT and (7.138). Thus

$$\begin{aligned} \mathrm{DFT}\{R_{ff}(m)\} &= \frac{1}{N} F_N^*(k) F_N(k) \\ &= \frac{1}{N} |F_N(k)|^2. \end{aligned} \tag{7.169}$$

This is the discrete-time version of the *periodogram* given in the continuous-time case by (7.157). Hence, the autocorrelation estimate $\hat{R}_{ff}(m)$ can be found from $\mathbf{f}_N(n)$ by finding its DFT as $F_N(k)$, forming $F_N^*(k)F_N(k)$, then taking the inverse DFT. This involves two DFTs and one multiplication.

Stage 2 The autocorrelation estimate $\hat{R}_{ff}(m)$ obtained in *Stage 1* is multiplied by a window sequence $w(n)$ of the type given by (5.143)–(5.148). The windowed sequence $\hat{R}_{ff}^{\mathrm{w}}(m)$ is used to obtain the power spectrum estimate

$$\hat{P}_{ff}^{w}(k\omega_0) = \sum_{m=0}^{N-1} \hat{R}_{ff}^{w}(m) \exp(-\mathrm{j}mk\omega_0). \tag{7.170}$$

This involves one DFT. The general steps just outlined are depicted schematically in Figure 7.6.

Finally, the cross-power spectrum of two discrete-time signals $\mathbf{f}(n)$ and $\mathbf{g}(n)$ can be estimated in a similar manner. The necessary steps in this case are depicted in Figure 7.7.

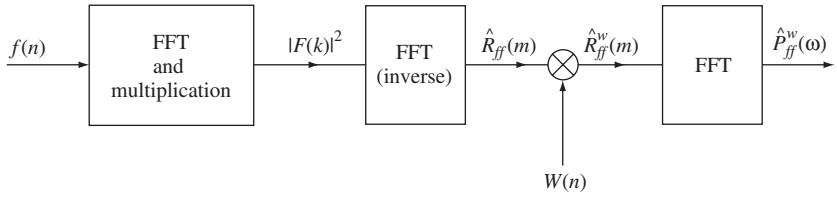


Figure 7.6 Power spectrum estimation using FFT algorithms

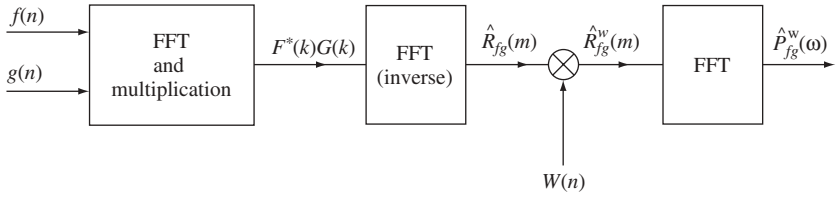


Figure 7.7 Cross-power spectrum estimation using FFT algorithms

7.6 Conclusion

This chapter dealt with the description of stochastic signals and the associated techniques of spectral analysis. Both analog (continuous-time) and discrete-time signals have been treated. In both cases, stochastic signals were shown to be of the *finite-power* type. In this regard, they can be represented in the frequency domain by their *power spectra*. It has also been shown that the autocorrelation and the power spectrum of a stochastic signal constitute a Fourier transform pair; hence the importance of the autocorrelation function. Similarly, the cross-correlation between two signals and their cross-power spectrum constitute a Fourier transform pair. These considerations apply to both analog and digital signals, but in the latter case, the discrete Fourier transform is employed. Finally, it has been shown that the fast Fourier transform algorithms can be used to estimate the power spectra of stochastic signals, and the methods of improving the quality of the estimates were also discussed.

Problems

7.1 Find the probability density $p(\mathbf{f}, t)$, the mean and autocorrelation of the random signal

$$\mathbf{f}(t) = \mathbf{g}(t) - 1$$

where $\mathbf{g}(t)$ is a random variable with probability density

$$P(\mathbf{g}) = \frac{1}{(2\pi)^{1/2}} \exp(-g^2/2) \quad -\infty \leq g \leq \infty.$$

7.2 Show that

$$E\{|\mathbf{f}(t+\tau) + \mathbf{f}(t)|^2\} = 2\mathrm{Re}[R_{ff}(0) + R_{ff}(\tau)].$$

7.3 Let $\mathbf{f}(t)$ be a stochastic signal with constant mean and bounded autocovariance $C(t_1, t_2)$. Prove that the process is mean-ergodic if

$$\lim_{\tau \to 0} C(t + \tau, t) \quad \text{as} \quad \tau \to \infty.$$

7.4 Let $\mathbf{f}(t)$ be a stationary signal with autocorrelation $R_{ff}(\tau)$, and let α, β be two non-commensurate numbers (i.e. α is not related to β by an integer proportionality constant). Suppose that

$$R_{ff}(\alpha) = R_{ff}(\beta) = R_{ff}(0).$$

Show that

$$R_{ff}(\tau) = R_{ff}(0) = \text{a constant}.$$

7.5 Let

$$h(t) = 1/\alpha \quad -\alpha \leq |t| \leq \alpha$$

be the impulse response of a deterministic system.

(a) Find the autocorrelation of the impulse response.

(b) Suppose the input to the system is a stochastic signal with autocovariance $C_{ff}(\tau)$. Obtain the expression for the output covariance and variance $C_{ff}(0)$.

7.6 A linear discrete system with stochastic input $\mathbf{v}(n)$ and output $\mathbf{g}(n)$ is described by the equation

$$\mathbf{g}(n) - \alpha \mathbf{g}(n-1) = \mathbf{v}(n).$$

If $\mathbf{v}(n)$ is white noise with

$$P_{vv}(\omega) = 1$$

find the output power spectrum and autocorrelation.

8

Finite Word-length Effects in Digital Signal Processors

8.1 Introduction

We have seen that a digital filter, or a general digital signal processing system, operates on an input sampled-data signal to produce an output sampled-data signal by means of a computational algorithm. Since the sampled-data signals are represented by number sequences, these are quantized and encoded using binary codes, and the processor algorithm can be implemented either in software using a general purpose computer or in dedicated hardware. The latter approach is becoming increasingly popular due to the advances in *very large scale intergration* (VLSI) and the resulting availability of integrated circuit modules and special purpose hardward with sufficient memory size, complexity and speed so as to render the hardware implementation of digital filters, operating in real time, an attractive technique.

Now, regardless of the type of implementation, the numbers processed by the digital system are ultimately stored in (memory) registers with finite capacity. Therefore, all digital networks operate with only a finite number of binary digits (bits); thus resulting in an inherent limitation on the accuracy of processing. In this chapter, we discuss the affect of using finite word-lengths to represent the numbers and the arithmetic operations, on the accuracy of digital signal processors in general and digital filters in particular [12, 20].

The errors due to the use of finite word-lengths to represent the pertinent numbers can be of the following types:

(i) Input signal quantization effects
These are errors due to the representation of the input signal by a set of values with *discrete amplitudes* and subsequently by binary numbers with finite word-lengths. Typical word-lengths are 32 or 64 bits. Clearly, these errors are inherent in the process of analog to digital conversion discussed in Chapter 4.

Signal Processing and Integrated Circuits, First Edition. Hussein Baher.
© 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.

(ii) Coefficient quantization effects
These result from the representation of each filter coefficient: a_r, b_r in the expression

$$H(z) = \frac{\sum_{r=0}^{M} a_r z^{-r}}{1 + \sum_{r=0}^{N} b_r z^{-r}} \tag{8.1}$$

by a quantized discrete-amplitude value and subsequently in binary form by a finite number of bits. We recall that the solution to the approximation problem leads to a transfer function of the general form (8.1) in which the coefficients (a_r, b_r) are assumed capable of being represented to an arbitrary degree of accuracy. However, prior to implementation, these coefficients are quantized and each represented by a finite number of bits; again typically 32 or 64 bits. Consequently, the filter response may deviate considerably from the desired characteristic. Moreover, the error due to coefficient quantization may cause the poles of the transfer function in (8.1) to change their positions in the z-plane, perhaps moving to points on the unit circle or exterior to it and causing the filter to become unstable.

(iii) Product quantization and accumulation errors
In the practical implementation of digital filters using dedicated hardware or a microprocessor, a fixed length l of shift registers is chosen for use throughout the filter. In the arithmetic operation of multiplication a signal represented by an l-bit number is, however, multiplied by a coefficient represented by another, say, l-bit number also. This results in a product having $2l$ digits. But this number must be *rounded* or *truncated* to l bits, which is the uniform length of each register. This results in errors known as product quantization errors, which accumulate as the operations of multiplication are performed to implement the filter.

(iv) Auto-oscillations due to overflow and limit cycles
In performing arithmetic operations, *overflow* may result, giving rise to self-sustained oscillations, which represent a type of instability of the filter. These may occur if the capacity of the memory is exceeded when logic saturation devices are not used. Furthermore, auto-oscillations may occur even when the filter input is nominally zero, because even in the absence of data an error signal exists which is caused by the quantization of the internal data before storage in the memory. This effect is called *limit cycle oscillations*.

In this chapter we give a discussion of the above types of error, relying on the analysis of random processes given in Chapter 7.

Numbers are represented in either fixed-point or floating-point format. In either case, number quantization is performed by rounding or truncation. In fixed-point arithmetic with l-bit word-length, number quantization is assumed to be a random process with the resulting error having uniform probability density functions in the appropriate range as shown in Figure 8.1. This is a reasonable assumption since it is clear that there is no 'preferred' range of the error. For floating-point arithmetic with l-bit mantissa, the corresponding error probability density functions are shown in Figure 8.2. In our discussion of finite word-length effects, we concentrate on the case of fixed-point arithmetic.

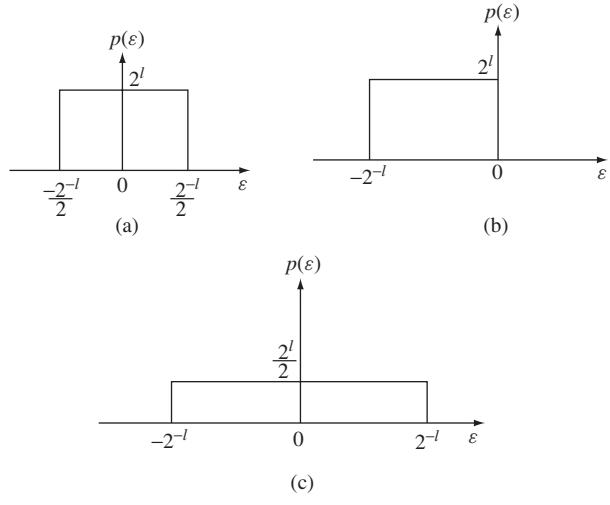


Figure 8.1 Error probability density functions in fixed-point arithmetic: (a) rounding, (b) truncation with two's complement and (c) truncation with one's complement or signed magnitude

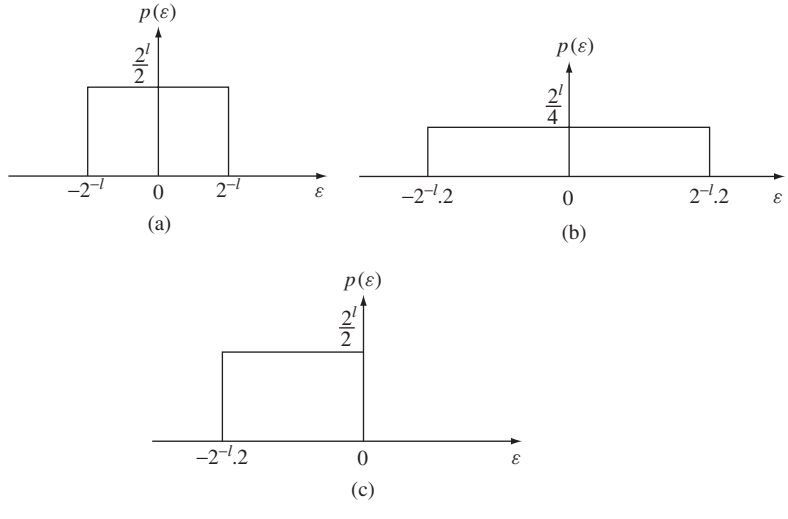


Figure 8.2 Error probability density functions in floating-point arithmetic: (a) rounding, (b) truncation with two's complement and (c) truncation with one's complement or signed-magnitude

8.2 Input Signal Quantization Errors

As we have seen in Chapter 4, the process of analog to digital conversion involves sampling the continuous-time signal $f(t)$, to give the sampled version $f(nT)$, which is quantized in amplitude to give $f_q(nT)$; this is finally encoded to produce the input to the digital system. The process of sampling, if the pertinent rules are followed, does not

produce errors. In contrast, the difference between the actual *analog* input sample and the corresponding binary-coded quantized value, is called *quantization error* or *quantization noise*. This is the first source of degradation in the processing, which is now studied in some detail, concentrating on the use of fixed-point arithmetic.

Typically, an A/D converter used in conjunction with an implementation of a digital filter, has an $(l+1)$-bit fixed-point output, including a sign bit. This corresponds to a resolution of about one part in 2^l. The quantizing step q is, therefore

$$q = 1/2^l. \tag{8.2}$$

If rounding is used in the quantization of the samples $f(nT)$, then the maximum possible quantization error is $\pm q/2$. Figure 8.3(a) shows the output of a quantizer (prior to encoding) with the input as the original continuous signal $f(t)$. Figure 8.3(b) shows the corresponding quantization error, and the amplitude of the error signal lies between $-q/2$ and $q/2$. The amplitudes of the quantized signal are called the *decision amplitudes*. It is generally reasonable to assume that the round-off error ε is white noise with uniform probability density $p(\varepsilon)$ as shown in Figure 8.4(a). The variance (power) of the error signal can be taken as a measure of the degradation suffered by the signal due to quantization. When the quantizing step is very small compared with the signal variations, the error signal can be considered equivalent to the sum of basic error signals, each approximated by a straight line segment as shown in Figure 8.4(b). The average power (variance) of a basic error signal of width τ is

$$\begin{aligned}\sigma_0^2 &= \frac{1}{\tau}\int_{-\tau/2}^{\tau/2} \varepsilon^2(t)\,\mathrm{d}t \\ &= \frac{q^2}{\tau^3}\int_{-\tau/2}^{\tau/2} t^2 \mathrm{d}t\end{aligned}$$

or

$$\sigma_0^2 = \frac{q^2}{12} = \frac{2^{-2l}}{12}. \tag{8.3}$$

Now, after quantization, the input to the filter is

$$f_{\mathrm{q}}(nT) = f(nT) + \varepsilon(nT). \tag{8.4}$$

Since the filter is a linear system, its output is the sum of two components: one due to the signal $f(nT)$ and the other due to the quantization error $\varepsilon(nT)$. Thus, the error is also filtered by the transfer function $H(z)$ of the filter. Since the error $\varepsilon(nT)$ is assumed white noise with zero mean and variance $q^2/12$, then (see Section 7.4) the steady state output component due to $\varepsilon(nT)$ is a zero-mean wide-sense stationary process with power spectral density given, in the z-domain, by

$$P_d(z) = \frac{1}{12}q^2 H(z)H(z^{-1}) \tag{8.5}$$

so that

$$p_d(w) = \frac{1}{12}q^2 H(\exp(-\mathrm{j}\omega T))H(\exp(\mathrm{j}\omega T)). \tag{8.6}$$

Here, we have neglected the effect of coefficient quantization and round-off accumulation since their effect on the response due to $\varepsilon(nT)$ is much smaller than that due to $f(nT)$.

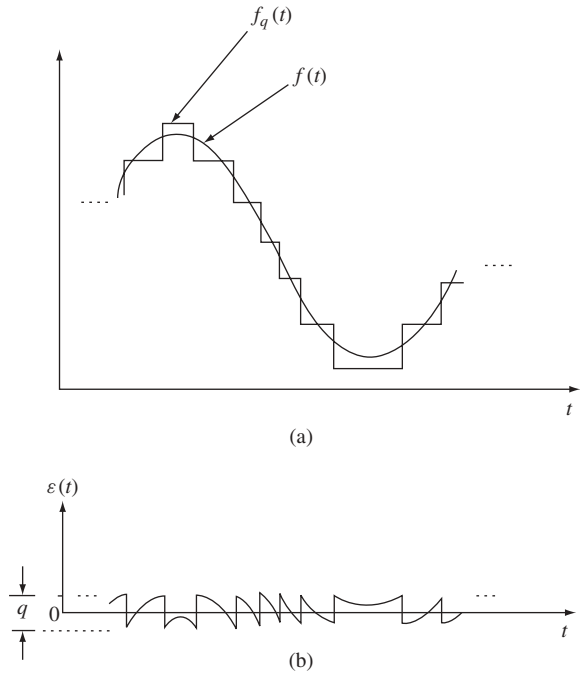


Figure 8.3 (a) The output $f_q(t)$ of a quantizer with input $f(t)$. (b) The quantization error

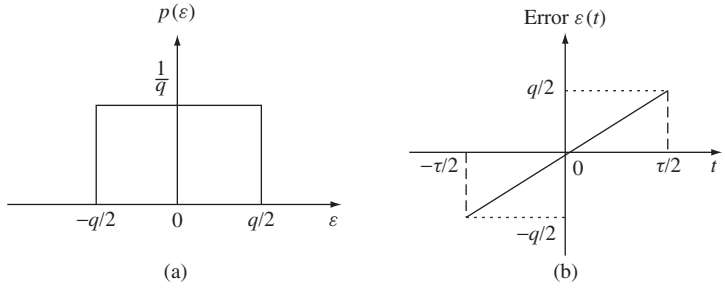


Figure 8.4 (a) Probability density of round-off error. (b) A basic error signal approximation

Thus, the output mean-square error (output noise power) is given by

$$\sigma_{\text{out}}^2 = \frac{q^2}{12} \sum_{n=0}^{\infty} [h(nT)]^2 \tag{8.7}$$

where $\{h(nT)\}$ is the impulse response sequence of the filter, which is assumed causal. Alternatively, the mean square value of the output error due to input quantization may be obtained from (8.5) and Parseval's relation in the z-domain

$$\sigma_{\text{out}}^2 = \frac{q^2}{12} \frac{1}{\text{j}2\pi} \oint_c H(z) H(z^{-1}) \frac{\text{d}z}{z} \tag{8.8}$$

where the integration is carried out over a closed contour c enclosing all the singularities of $H(z)$. The integral can be evaluated using the method of the residues, numerically, algebraically or by a computer programme.

Now, let each member of the quantized sequence $f_q(nT)$ be represented by an l-bit binary number. Then, the maximum number of quantized amplitudes which can be encoded into binary form is 2^l. It follows that the range of amplitudes A which can be encoded lies in the range

$$q \leq A \leq q2^l. \tag{8.9}$$

Therefore, any amplitude value exceeding $q2^l$ cannot be represented, and the signal is *clipped* which results in degradation.

Let the range of signal amplitudes be $[-A_{\mathrm{m}}, A_{\mathrm{m}}]$ so that

$$A_{\mathrm{m}} = \frac{q2^l}{2}. \tag{8.10}$$

If rounding is used in the quantization operation, the error signal is

$$|\varepsilon(nT)| \leq A_{\mathrm{m}}2^{-l}. \tag{8.11}$$

Now, define the *peak power* of a coder as the power of the sinusoidal signal with the maximum possible amplitude A_{m} which the coder can pass without clipping (see Figure 8.5). Thus, the peak power is given by

$$\begin{aligned} P_{\mathrm{c}} &= \frac{1}{2}\left(\frac{q2^l}{2}\right)^2 \\ &= q^2 2^{2l-3.} \end{aligned} \tag{8.12}$$

The coding dynamic range R_{c} is defined as the ratio of the peak power to the quantization noise power. Hence, using (8.3) and (8.12) we have

$$\begin{aligned} R_{\mathrm{c}} &= \frac{P_{\mathrm{c}}}{\sigma_0^2} \\ &= 3(2^{2l-1}) \end{aligned} \tag{8.13}$$

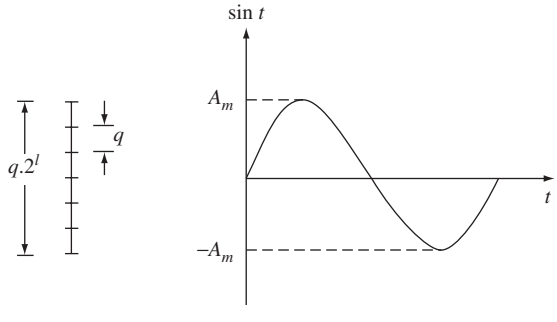


Figure 8.5 Pertinent to the definition of the coding dynamic range of the A/D converter

or

$$10\log\left(\frac{P_c}{\sigma_0^2}\right) = 6.02l + 1.76\,\text{dB}. \tag{8.14}$$

For example, an 8-bit coder has a dynamic range of about 50 dB while a 16-bit coder has a dynamic range of about 98 dB.

If the range of amplitudes of the input signal exceeds the coder dynamic range, then scaling of the signal prior to quantization can be applied to reduce the amplitude range; thus eliminating clipping. Consequently, in the model of quantization expressed by (8.4), a scaling factor K is incorporated such that

$$f_q(nT) = Kf(nT) + \varepsilon(nT) \tag{8.15}$$

where

$$0 < K < 1. \tag{8.16}$$

The signal to noise ratio in this case is given by

$$\text{SNR} = 10\log\left(\frac{k^2\sigma_f^2}{\sigma_0^2}\right)\text{dB} \tag{8.17}$$

where $K^2\sigma_f^2$ is the power of the scaled signal, and σ_0^2 is the quantization noise power. It is found that negligible clipping results with the choice [20]

$$K = 1/5\sigma_f \tag{8.18}$$

so that

$$\begin{aligned}\text{SNR} &= 10\log\left(\frac{1}{25\sigma_0^2}\right) \\ &= 10\log\left(\frac{12}{25q^2}\right)\text{dB}\end{aligned} \tag{8.19}$$

and using (8.2) the above expression becomes

$$\text{SNR} = 6.02l - 3.1876\,\text{dB}. \tag{8.20}$$

Thus, for an 8-bit A/D converter, the signal to noise ratio is about 45 dB, whereas for a 16-bit converter it is about 100 dB.

8.3 Coefficient Quantization Effects

When binary words with finite length are used to represent the filter coefficients a_r and b_r, each coefficient is replaced by its l-bit representation in either fixed-point or floating-point form. Thus, a coefficient a_r is replaced by $(a_r)_q$ where for the fixed-point representation

$$(a_r)_q = a_r + \alpha_r \tag{8.21}$$

while in the floating-point form

$$(a_r)_q = \alpha_r(1 + \alpha_r) \tag{8.22}$$

with

$$|\alpha_r| \leq 2^l. \tag{8.23}$$

Similarly b_r becomes

$$(b_r)_{\mathrm{q}} = b_r + \beta_r \tag{8.24}$$

in the fixed-point form, or

$$(b_r)_{\mathrm{q}} = b_r(1 + \beta_r) \tag{8.25}$$

in the floating-point representation. Again

$$|\beta_r| \leq 2^{-l}. \tag{8.26}$$

It now follows that the filter characteristics deviate from the desired ones. To study this effect we can compute the frequency response of the actual filter with l-bit rounded coefficients, that is using the actual transfer function

$$H_{\mathrm{q}}(z) = \frac{\sum_{r=0}^{M} (\alpha_r)_{\mathrm{q}} z^{-r}}{1 + \sum_{r=1}^{N} (b_r)_{\mathrm{q}} z^{-r}} \tag{8.27}$$

to evaluate

$$H_{\mathrm{q}}(\exp(\mathrm{j}\omega T) = |H_q(\exp(\mathrm{j}\omega T))| \exp(\mathrm{j}\psi_{\mathrm{q}}(\omega T)). \tag{8.28}$$

The result is then compared with the desired theoretical response $H(\exp(\mathrm{j}\omega T))$ of the original design obtained from the solution to the approximation problem. Naturally, the longer the word used to represent the numbers, the closer the actual response to the desired one.

An alternative approach to the study of the above effects is to calculate the movements of the poles and zeros of the transfer function due to coefficient rounding, then apply sensitivity theory to examine the changes in the filter response. Let the poles $H(z)$ be $p_i (i = 1, 2, \ldots N)$ and the poles of $H_{\mathrm{q}}(z)$ be at $(p_i + \Delta p_i)$. Then it has been shown [20] that the variation in the typical pole position is given by

$$\Delta P_i = \sum_{r=1}^{N} \frac{p_i^{r+1}}{\prod\limits_{\substack{m=1 \\ m \neq i}} \left(1 - \dfrac{p_i}{p_m}\right)} \Delta a_r \tag{8.29}$$

where Δa_r is the change in the coefficient due to rounding, which is either α_r or $a_r\,\alpha_r$. Similar results can be obtained for the movement of the zeros. From these perturbations, the deviations in the overall filter response can be examined.

In addition to altering the frequency response, coefficient quantization can also affect the stability of an IIR filter. If the poles happen to be close to the unit circle in the z-plane, coefficient quantization can cause their positions to be sufficiently perturbed so as to move to points on the unit circle or exterior to it, thus producing instability. It has

been shown that for an Nth order low-pass IIR filter operating at a sampling frequency of $1/T$ with distinct poles at $(\cos \omega_r T \pm \mathrm{j}\ \sin \omega_r T)$, stability is guaranteed if the number of bits satisfies the, rather pessimistic, inequality

$$l > -\log_2 \left(\frac{5\sqrt{N}}{2^{N+2}} \prod_{r=1}^{N} \omega_r T \right). \tag{8.30}$$

Example 8.1 Consider the transfer function

$$H(z) = \frac{z}{z^2 - 1.7z + 0.745}$$

and determine the coefficient word-length required to avoid instability.

Solution. The poles occur at $z = (0.85 \pm \mathrm{j}0.15)$, that is

$$\begin{aligned} \omega_1 T &= \tan^{-1}\left(\frac{0.15}{0.85}\right) \\ &= 0.1747 \text{ rad} \\ &= \omega_2 T. \end{aligned}$$

Thus, using the estimate in (8.30) we obtain

$$l > -\log_2 \left(\frac{5\sqrt{2}}{2^4} (0.1747)^2 \right)$$

That is

$$l > -\log_2(0.0135)$$

so that $l > 6.21$ and we take $l = 7.0$. As noted before, the estimate is somewhat pessimistic, particularly since it is based on quantization by truncation rather than rounding.

Generally, the effect of coefficient quantization is more significant for a high degree filter realized in direct form, than the corresponding cascade or parallel realizations. Therefore the parallel or cascade form should be used for high degree filters since the saving in the required word-length is quite significant.

Finally, we note that due to coefficient quantization, the output sequence will differ from the desired one. The analysis of the resulting errors will be undertaken in the next section, in conjunction with round-off accumulation.

8.4 Effect of Round-off Accumulation

To begin with, we shall examine the effect of round-off accumulation due to product quantization without taking into account the effect of coefficient quantization. Then the combined errors due to both effects are studied. In our discussion, we concentrate on the case of fixed-point arithmetic.

8.4.1 Round-off Accumulation without Coefficient Quantization

These errors are dependent on the particular form of realization. Hence each form is treated separately.

8.4.1.1 Direct Form

An IIR filter transfer function

$$H(z) = \frac{\sum_{r=0}^{M} a_r z^{-r}}{1 + \sum_{r=1}^{N} b_r z^{-r}} = \frac{N(z^{-1})}{D(z^{-1})} \tag{8.31}$$

corresponds to the difference equation

$$g(n) = \sum_{r=0}^{M} a_r f(n-r) - \sum_{r=1}^{N} b_r g(n-r) \tag{8.32}$$

where the sampling period T has been dropped for convenience. The direct realization of the filter is shown in Figure 4.17 and requires $(N + M + 1)$ multiplications. In fixed-point arithmetic the result of multiplying two l-bit numbers is a $2l$-bit product; this is rounded to an l-bit word. In dealing with this type of product quantization error we assume that the signal levels throughout the filter are much larger than the quantizing step q. This allows us to treat product quantization errors at the output of the multipliers as uncorrelated (statistically independent) random variables, each being white noise with power spectral density $q^2/12$. Therefore, the total error $\varepsilon(n)$ is the sum of all those due to the multiplications. If μ is the number of coefficients a_r which are neither 0 or 1, and v is the number of coefficients b_r which are neither 0 nor 1, then the total number of multiplications is $(\mu + v)$. Therefore the total error is

$$\varepsilon(n) = \frac{(\mu + v)q^2}{12}. \tag{8.33}$$

For the most general case expressed in (8.31) we may write

$$\mu = M + 1$$
$$v = N \tag{8.34}$$

so that

$$\varepsilon(n) = \frac{(M + N + 1)q^2}{12}. \tag{8.35}$$

Therefore the quantized (rounded) filter output sequence is given by

$$g_{\mathrm{q}}(n) = \sum_{r=0}^{M} a_r f(n-r) - \sum_{r=1}^{N} b_r g_{\mathrm{q}}(n-r) + \varepsilon(n). \tag{8.36}$$

Now, the error in the filter output can be defined as

$$e(n) = g_{\mathrm{q}}(n) - g(n). \tag{8.37}$$

Substituting from (8.32) and (8.36) in (8.37) we have

$$e(n) = \varepsilon(n) - \sum_{r=1}^{N} b_r e(n-r) \tag{8.38}$$

which describes a linear discrete system with input $\varepsilon(n)$ and output $e(n)$. Hence, taking the z-transform of (8.38) we can form the transfer function

$$\begin{aligned} \hat{H}(z) &= \frac{E(z)}{E(z)} = \frac{Z\{e(n)\}}{Z\{\varepsilon(n)\}} \\ &= \frac{1}{1 + \sum_{r=1}^{N} b_r z^{-r}} \end{aligned} \tag{8.39a}$$

or

$$\hat{H}(z) = \frac{1}{D(z^{-1})} \tag{8.39b}$$

where $D(z^{-1})$ is the denominator of the filter transfer function $H(z)$. It follows that $D(z^{-1})$ is the part of the transfer function which contributes to the error noise in the filter output. We can, therefore, construct a model of the filter which takes account of product round-off error accumulation as shown in Figure 8.6.

The average output noise power is

$$\begin{aligned} \sigma^2 &= \frac{(N+M+1)q^2}{12} \sum_{n=0}^{\infty} (\hat{h}(n))^2 \\ &= \frac{(N+M+1)q^2}{12} \frac{1}{\mathrm{j}2\pi} \oint_c \frac{1}{D(z)D(z^{-1})} \frac{\mathrm{d}z}{z}. \end{aligned} \tag{8.40}$$

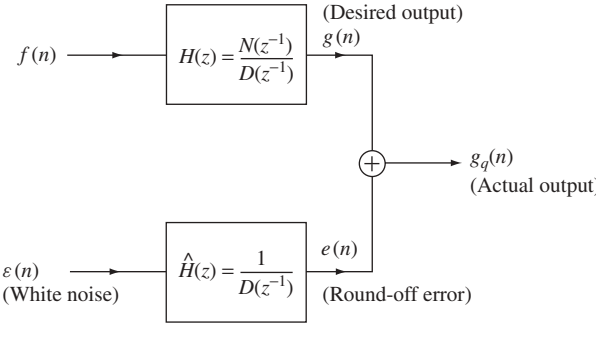


Figure 8.6 A model of an IIR filter, in direct form, taking round-off error accumulation into account

Example 8.2 Calculate the output noise due to round-off accumulation, for the filter described by the transfer function

$$H(z) = \frac{1}{(1 - 0.4z^{-1})(1 - 0.5z^{-1})}.$$

Solution.

$$D(z^{-1}) = (1 - 0.4z^{-1})(1 - 0.5z^{-1})$$
$$D(z) = (1 - 0.4z)(1 - 0.5z).$$

Here, $M = 0$ and $N = 2$. Therefore, (8.40) gives

$$\sigma^2 = \frac{q^2}{8\mathrm{j}\pi} \oint_c \frac{z\,\mathrm{d}z}{(0.4 - z)(0.5 - z)(1 - 0.4)(1 - 0.5z)}.$$

The integral is evaluated by summing the residues due to the poles inside the contour of integration, which in this case is the unit circle. This gives

$$\sigma^2 = \frac{q^2}{8\mathrm{j}\pi} \times \mathrm{j}2\pi \sum \text{residues}.$$

The poles of the integrand inside the unit circle are those at $z = 0.4$ and $z = 0.5$. The corresponding residues are

$$k_1 = \frac{0.4}{(0.4 - 0.5)(1 - 0.4^2)(1 - 0.5 \times 0.4)} = -5.9524$$
$$k_2 = \frac{0.5}{(0.5 - 0.4)(1 - 0.4 \times 0.5)(1 - 0.5^2)} = 8.333.$$

Therefore, the round-off accumulation noise is

$$\sigma^2 = \frac{q^2}{4}(8.333 - 5.924)$$
$$= 0.595q^2.$$

Now, consider the direct non-recursive realization of an FIR filter described by

$$H(z) = \sum_{r=0}^{M} a_r z^{-r}. \tag{8.41}$$

In this case, the round-off accumulation noise can be calculated using (6.40) with $D(z^{-1}) = 1$ and $N = 0$. Thus, the average output noise power is

$$\sigma^2 = \frac{(M + 1)q^2}{12} \frac{1}{\mathrm{j}2\pi} \oint_c \frac{\mathrm{d}z}{z}$$

That is

$$\sigma^2 = \frac{(M + 1)}{12} q^2. \tag{8.42}$$

8.4.1.2 Cascade Realization

As explained in Chapter 4, this method proceeds by expressing the transfer function in the form

$$H(z) = \prod_{k=1}^{m} H_k(z) \tag{8.43}$$

where $H_k(z)$ is a second-order factor of the form

$$\begin{aligned} H_k(z) &= \frac{\alpha_{0k} + \alpha_{1k} z^{-1} + \alpha_{2k} z^{-2}}{1 + \beta_{1k} z^{-1} + \beta_{2k} z^{-2}} \\ &= \frac{A(z^{-1})}{B(z^{-1})} \end{aligned} \tag{8.44}$$

with the possibility of one first-order factor for an odd-degree function; this is of the same form in (8.44) with $\alpha_{2k} = \beta_{2k} = 0$.

The realization is then the cascade connection of the sections described by the second-order terms (and a possible first-order one) which realize each typical transfer function given by (8.44). Without taking round-off errors into account, the realization takes the form shown in Figure 8.5. To include round-off errors in the analysis, the model of Figure 8.6 is used for each transfer function described by (8.44) and the results are connected in cascade. This gives the model of Figure 8.7. The quantization error inputs $\varepsilon_k(n)$ produce noise power components σ_k^2 at the outputs resulting in a total noise power given by

$$\sigma_{\mathrm{T}}^2 = \sum_{k=1}^{m} \sigma_k^2. \tag{8.45}$$

Examination of the model in Figure 8.7 shows that each noise input $\varepsilon_k(n)$ produces an output $e_k(n)$ such that an error transfer function may be formed as

$$\begin{aligned} \hat{H}_k(z) &= \frac{Z[e_k(n)]}{Z[\varepsilon_k(n)]} \\ &= \frac{1}{B_k(z^{-1})} \prod_{r=k+1}^{m} \hat{H}_r \quad k = 1, 2, \ldots (m-1) \end{aligned} \tag{8.46a}$$

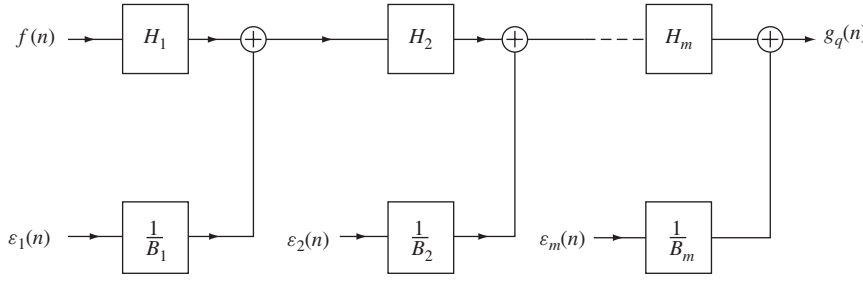


Figure 8.7 A model for an IIR filter, in cascade form, taking round-off error accumulation into account

and

$$\hat{H}_m(z) = \frac{1}{B_m(z^{-1})}. \tag{8.46b}$$

Hence, using Parseval's theorem and a relation similar to (8.40) the kth quantization error component produces a noise power given by

$$\sigma_k^2 = \frac{\mu_k q^2}{12} \frac{1}{\mathrm{j}2\pi} \oint \frac{1}{B_k B_{k^*}} \prod_{r=k+1}^{m} H_r H_{r^*} \frac{\mathrm{d}z}{z} \quad \text{for} \quad k = 1, 2, \ldots (m-1) \tag{8.47a}$$

and

$$\sigma_m^2 = \frac{\mu_m q^2}{12} \frac{1}{\mathrm{j}2\pi} \oint \frac{1}{B_m B_{m^*}} \frac{\mathrm{d}z}{z} \tag{8.47b}$$

where the lower asterisk denotes replacement of z by z^{-1} and μ_k is the number of multiplications in the kth section. $\mu_k = 5$ for a general second-order section and $\mu_k = 3$ for a general first-order one. Finally the total output noise power is obtained from (8.46) and (8.47).

Example 8.3 Find the output noise power due to round-off accumulation in the cascade realization of the transfer function of Example 8.2.

Solution. Let us realize the transfer function as a cascade of two first-order sections. This is possible because both poles are real. Thus

$$H(z) = H_1(z) H_2(z)$$

where

$$H_1(z) \frac{z^{-1}}{(1 - 0.4z^{-1})} = \frac{1}{z - 0.4}$$

$$H_2(z) = \frac{z^{-1}}{(1 - 0.5z^{-1})} = \frac{1}{z - 0.5}.$$

Using (8.47) we obtain with $\mu_1 = 1$

$$\sigma_1^2 = \frac{q^2}{12} \frac{1}{\mathrm{j}2\pi} \oint \frac{5z \, \mathrm{d}z}{(z - 0.4)(z - 0.5)(2.5 - z)(2 - z)}.$$

The two poles inside the unit circle are at $z = 0.4$ and $z = 0.5$. The corresponding residues are -5.9524 and 8.333. Thus

$$\sigma_1^2 = \frac{q^2}{12} \frac{1}{\mathrm{j}2\pi} \mathrm{j}2\pi(-5.9524 + 8.333)$$

$$= 0.1984 q^2.$$

Similarly for $\hat{H}_2$ with $\mu_2 = 1$

$$= \sigma_2^2 = 0.111 q^2.$$

Therefore the total output noise power is

$$\begin{aligned}\sigma_{\mathrm{T}}^2 &= \sigma_1^2 + \sigma_2^2 \\ &= 0.3095q^2.\end{aligned}$$

Comparison with the direct realization in Example 8.2, it is seen that for this transfer function, the cascade form produces lower round-off noise, for the same quantizing step.

8.4.1.3 Parallel Realization

In this method we express the transfer function as

$$H(z) = \sum_{k=1}^{m} H_k(z) \tag{8.48}$$

with

$$\begin{aligned}H_k &= \frac{\alpha_{0k} + \alpha_{1k}z^{-1}}{1 + \beta_{1k}z^{-1} + \beta_{2k}z^{-2}} \\ &= \frac{A_k(z)}{B_k(z)}.\end{aligned} \tag{8.49}$$

As before, we can develop the model of Figure 8.8 which gives the parallel form of realization taking the round-off noise into account.

Each noise error input $\varepsilon_k(n)$ produces a quantization noise power component σ_k^2 at the output. The total output round-off noise power is

$$\sigma_{\mathrm{T}}^2 = \sum_{k=1}^{m} \sigma_k^2 \tag{8.50}$$

where

$$\sigma_k^2 = \frac{\mu_k q^2}{12} \frac{1}{\mathrm{j}2\pi} \oint_c \frac{1}{B_k B_{k^*}} \frac{\mathrm{d}z}{z} \quad k = 1, 2, \ldots m \tag{8.51}$$

and μ_k is the number of multiplications in the kth section.

Example 8.4 Find the output round-off noise power in the parallel realization of the same transfer function of Examples 8.2 and 8.3.

Solution.

$$H(z) = \frac{1}{(z - 0.4)(z - 0.5)}$$

which can be realized as two first-order sections in the parallel form, because both poles are real. Thus

$$H(z) = H_1(z) + H_2(z)$$

where

$$H_1(z) = \frac{10}{z - 0.5}$$

$$H_2(z) = \frac{-10}{z - 0.4}.$$

Thus

$$\sigma_1^2 = \frac{q^2}{12}\frac{1}{\mathrm{j}2\pi}\oint_c \frac{1}{(z - 0.4)(z^{-1} - 0.4)}\frac{\mathrm{d}z}{z}$$

$$= 0.0992q^2$$

and

$$\sigma_2^2 = \frac{q^2}{12}\frac{1}{\mathrm{j}2\pi}\oint_c \frac{1}{(z - 0.5)(z^{-1} - 0.5)}\frac{\mathrm{d}z}{z}$$

$$= 0.111q^2$$

so that

$$\sigma_\mathrm{T}^2 = \sigma_1^2 + \sigma_2^2 = 0.2103q^2$$

which is even smaller than in the cascade form, for this transfer function.

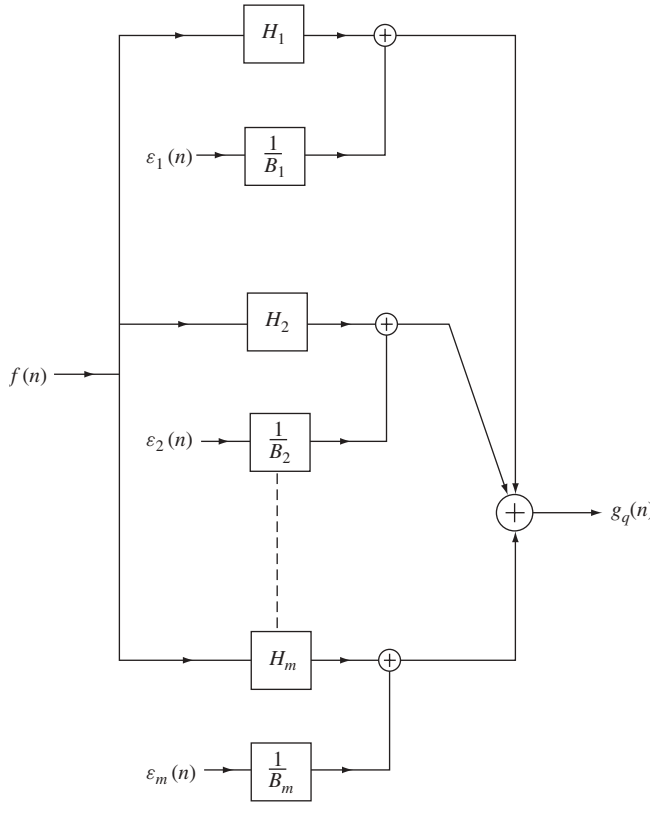


Figure 8.8 A model for an IIR filter, in parallel form, taking round-off error accumulation into account

8.4.2 Round-off Accumulation with Coefficient Quantization

We now repeat the same analysis of the previous section with the additional assumption that the filter *coefficients* are also rounded [12, 20]. Here also, it is assumed that fixed-point arithmetic is used.

8.4.2.1 Direct Realization

With the notation of (8.21)–(8.26) for the rounded coefficients, define

$$\Delta(n) = -\sum_{k=1}^{M} b_k \Delta(n-k) + u(n) \tag{8.52}$$

where

$$u(n) = \sum_{k=0}^{N} \alpha_k f(n-k) - \sum_{k=1}^{M} \beta_k g(n-k) + \varepsilon(n) \tag{8.53}$$

$$A(z) = \sum_{k=0}^{M} \alpha_k z^{-k} \tag{8.54}$$

$$B(z) = \sum_{k=1}^{N} \beta_k z^{-k} \tag{8.55}$$

$$C(z) = A(z) - H(z)B(z) \tag{8.56}$$

$$D(z) = 1 + \sum_{k=1}^{N} b_k z^{-k}. \tag{8.57}$$

Now, with the assumption that the input sequence $\{f(n)\}$ is zero-mean and wide-sense stationary, it has an autocorrelation sequence $\{R_{ff}(n)\}$ and power spectral density $\phi_{ff}(z)$. It follows that the output sequence $\{g(n)\}$ is also zero-mean and wide-sense stationary with power spectral density $\phi_{gg}(z)$ given by (8.147) as

$$\phi_{gg}(z) = H(z)H_*(z)\phi_{ff}(z). \tag{8.58}$$

It can be shown that $\{u(n)\}$, as defined by (8.53), is also zero-mean and wide-sense stationary with an autocorrelation function given by

$$\phi_{uu}(z) = C(z)C_*(z)\phi_{ff}(z) + \frac{(M+N+1)q^2}{12DD_*}. \tag{8.59}$$

The output quantization error $\Delta(n)$ which includes both coefficient quantization and product round-off accumulation, is also zero-mean and wide-sense stationary with auto-correlation

$$\phi_{\Delta\Delta}(z) = \frac{1}{D(z)D_*(z)}\phi_{uu}(z) \tag{8.60}$$

which, upon use of (8.59), becomes

$$\phi_{\Delta\Delta}(z) = \frac{C(z)C_*(z)}{D(z)D_*(z)}\phi_{ff}(z) + \frac{(M+N+1)}{12}q^2. \tag{8.61}$$

Therefore, the mean-square error, or output noise power, is given by

$$\sigma_{\mathrm{T}}^2 = \frac{1}{\mathrm{j}2\pi} \oint \phi_{\Delta\Delta}(z) \frac{\mathrm{d}z}{z}. \tag{8.62}$$

Clearly, if there is no coefficient rounding error, the first term in (8.61) is absent and expressions (8.61) to (8.62) reduce to (8.40) obtained before. However, if there is no round-off error, then the second term in (8.61) is absent. It follows that the total noise at the output, due to internal quantization, is the sum of two components, one is due to round-off accumulation and the other due to coefficient rounding to l bits. The component due to round-off accumulation is uncorrelated with either the input sequence $\{f(n)\}$ or the theoretical output sequence $\{g(n)\}$. A model for the direct realization of the transfer function incorporating these errors can be easily obtained from (8.61) and (8.62). This is shown in Figure 8.9, of which Figure 8.6 is a special case.

8.4.2.2 Cascade Realization

With $H(z)$ factored as in (8.43), the model in Figure 8.9 can be applied to each factor H_k, with $M = N = 2$ (or 1). The notation in (8.21) to (8.36) is used for the quantized coefficients for each second- or first-order section. Thus, with

$$\begin{aligned} \hat{A}_k(z) &= \alpha_{0k} + \alpha_{1k} z^{-1} + \alpha_{2k} z^{-2} \\ \hat{B}_k(z) &= \beta_{1k} z^{-1} + \beta_2 k z^{-2} \\ C_k(z) &= \hat{A}_k(z) - H_k(z) \hat{B}_k(z) \end{aligned} \tag{8.63}$$

we may construct the model in Figure 8.10 of which Figure 8.7 is a special case when coefficient quantization is neglected.

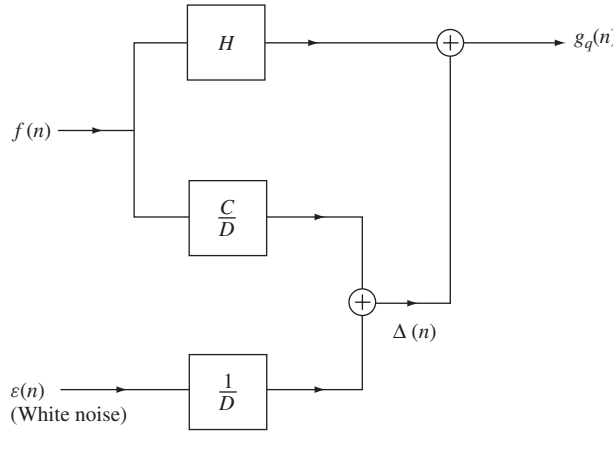


Figure 8.9 A model for an IIR filter in direct form, taking into account both round-off accumulation and coefficient quantization

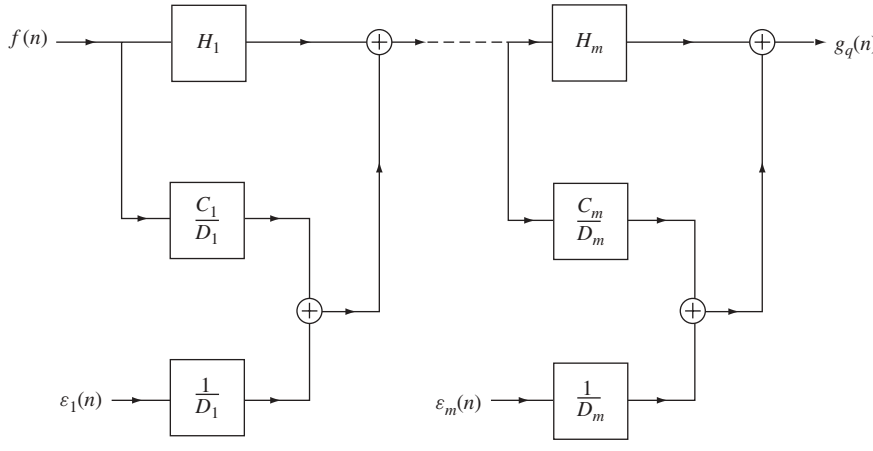


Figure 8.10 A model for an IIR filter in cascade form taking into account both round-off accumulation and coefficient quantization

Thus, for the cascade realization, the power spectral density of the output error is given by

$$\phi_{\Delta\Delta}(z) = \phi_{ff} \sum_{k=1}^{m} \frac{C_k C_{k^*}}{D_k D_{k^*}} \prod_{\substack{r=1 \\ r \neq k}}^{m} H_r H_{r^*}$$

$$+ \frac{q^2}{12} \left(\frac{\mu_m}{D_m D_{m^*}} + \sum_{k=1}^{m-1} \frac{\mu_k}{D_k D_{k^*}} \prod_{\substack{r=1 \\ r \neq k}}^{m} H_r H_{r^*} \right) \tag{8.64}$$

where μ_k is the number of multiplications in the kth section.

Finally, the total output noise power is obtained from (8.62) and (8.64). We also note that for zero coefficient rounding, (8.64) when used in (8.62) reduces to (8.47).

8.4.2.3 Parallel Realization

With $H(z)$ expressed as in (8.48), the model in Figure 8.9 can be applied to each second-order (or first-order) term $H_k(z)$. The notation in (8.21) to (8.26) is used for the quantized coefficients of each term. Thus, with

$$\begin{aligned} \hat{A}_k(z) &= \alpha_{0k} + \alpha_{1k} z^{-1} \\ \hat{B}_k(z) &= \beta_{0k} z^{-1} + \beta_{1k} z^{-2}. \\ C_k(z) &= \hat{A}_k(z) - H_k(z) \hat{B}_k(z) \end{aligned} \tag{8.65}$$

we may construct the model in Figure 8.11 of which Figure 8.8 is a special case when coefficient quantization is absent. Thus, for the parallel form of realization, the power

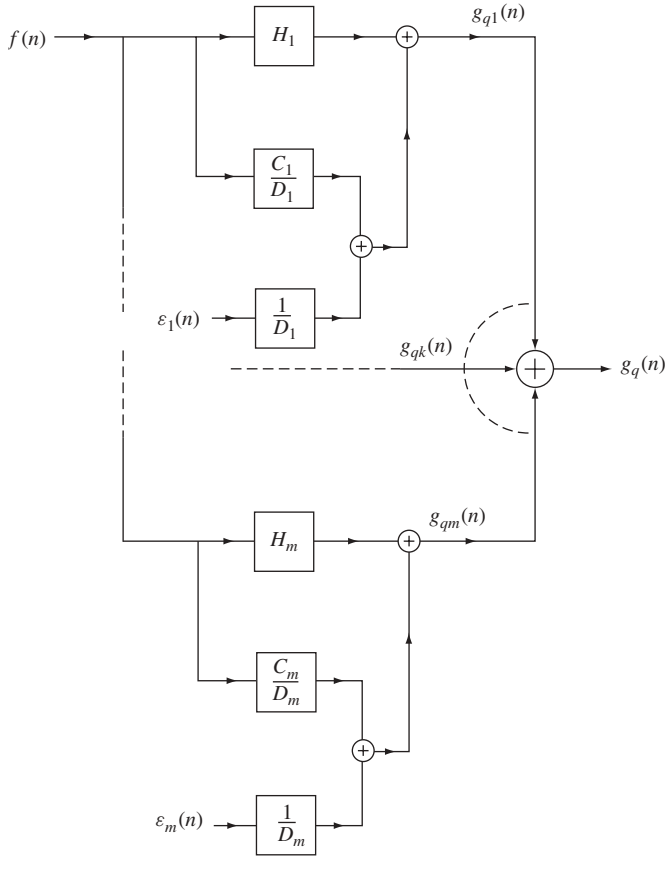


Figure 8.11 A model for an IIR filter in parallel form taking into account both round-off accumulation and coefficient quantization

spectral density of the output error is given by

$$\phi_{\Delta\Delta}(z) = \phi_{ff}(z) \left(\sum_{k=1}^{m} \frac{C_k}{D_k} \right) \left(\sum_{k=1}^{m} \frac{C_{k*}}{D_{k*}} \right) + \frac{q^2}{12} \sum_{k=1}^{m} \frac{\mu_k}{D_k D_{k*}} \tag{8.66}$$

where μ_k is the number of multiplications in the kth section. The total output noise power is obtained from (8.66) and (8.62). For no coefficient quantization, the result reduces to (8.51).

8.5 Auto-oscillations: Overflow and Limit Cycles

8.5.1 Overflow Oscillations

Consider the second-order section as the basic building block in the realization of a digital transfer function [12]. For the section shown in Figure 8.12 described by the difference equation

$$g(n) = f(n) - \beta_1 g(n-1) - \beta_2 g(n-2) \tag{8.67}$$

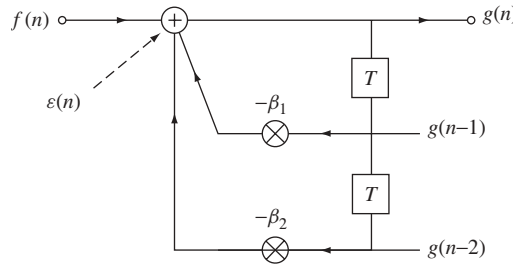


Figure 8.12 A second-order filter section described by (8.67) and (8.68)

the corresponding transfer function is an all-pole one

$$H(z) = \frac{Z\{g(n)\}}{Z\{f(n)\}} = \frac{1}{1 + \beta_1 z^{-1} + \beta_2 z^{-2}}. \tag{8.68}$$

The condition of stability is obtained by finding the poles as

$$P_{1,2} = \frac{-\beta_1 \pm (\beta_1^2 - 4\beta_2)^{1/2}}{2} \tag{8.69}$$

so that for stability $p_{1,2}$ must be inside the unit circle in the z-plane. This condition can be expressed in a (β_1, β_2)-plane as shown in Figure 8.13. For complex poles, the domain of stability is bounded by the parabola

$$\beta_2 = \frac{\beta_1^2}{4} \tag{8.70}$$

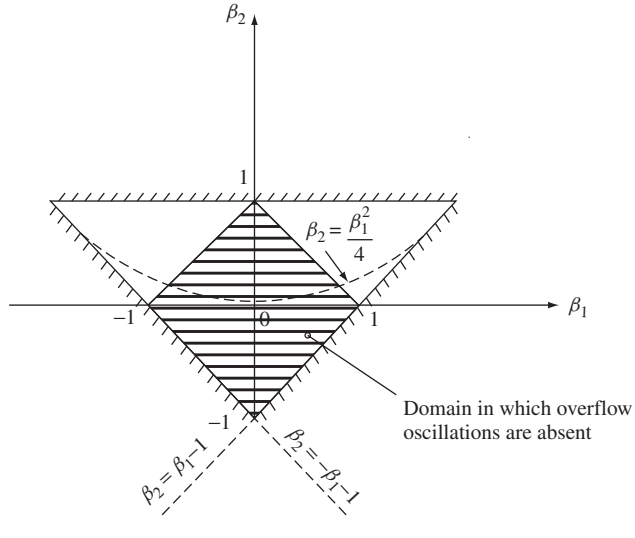


Figure 8.13 Pertinent to the discussion of auto-oscillations

and for stability

$$0 \leq \beta_2 < 1. \tag{8.71}$$

For real poles (8.69) gives

$$-\frac{\beta_1}{2} + \frac{1}{2}(\beta_1^2 - 4\beta_2)^{1/2} < 1$$
$$-1 < \frac{\beta_1}{2} - \frac{1}{2}(\beta_1^2 - 4\beta_2)^{1/2} \tag{8.72}$$

That is

$$\beta_2 > -\beta_1 - 1, \quad \beta_2 > -1 + \beta_1 \tag{8.73}$$

or

$$|\beta_1| < 1 + \beta_2. \tag{8.74}$$

In this case the domain of stability is the triangle defined by the three straight lines

$$\beta_2 = 1, \quad \beta_2 = -\beta_1 - 1, \quad \beta_2 = \beta_1 - 1 \tag{8.75}$$

as shown in Figure 8.13.

Now, even for zero input, and even if the stability conditions are satisfied, self-sustained oscillations may occur. The zero-input difference equation is obtained by setting $f(n) = 0$ in (8.67); thus

$$g(n) + \beta_1 g(n-1) + \beta_2 g(n-2) = 0$$

or

$$g(n) = -[\beta_1 g(n-1) + \beta_2 g(n-2)]. \tag{8.76}$$

When using fixed-point arithmetic in the adder of Figure 8.12 overflow may occur. For example, the transfer characteristic of the two's complement adder is shown in Figure 8.14.

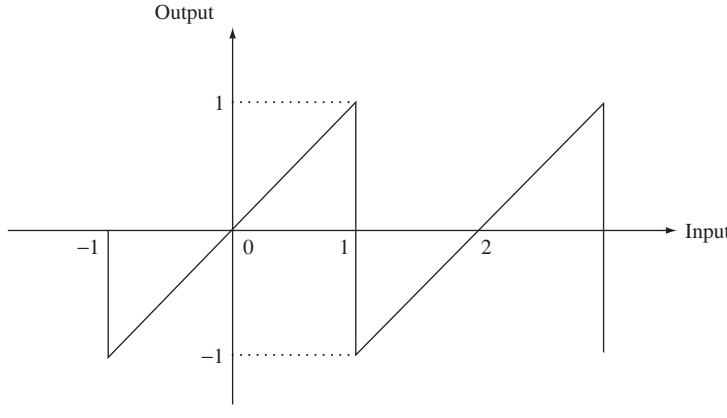


Figure 8.14 Transfer characteristic of a two's complement adder

Clearly overflow will occur if the adder receives at its input numbers whose sum is outside the range $(-1, 1)$. The condition for no overflow is obtained from (8.76) as

$$|g(n)| < 1 \tag{8.77}$$

so that

$$|\beta_1 g(n-1) + \beta_2 g(n-2)| < 1 \tag{8.78}$$

and since $g(n-1)$ and $g(n-2)$ are constrained to be less than unity, then a necessary and sufficient condition for the absence of overflow is

$$|\beta_1| + |\beta_2| < 1. \tag{8.79}$$

This defines a square in the (β_1, β_2)-plane *within* the triangle of stability as shown in Figure 8.13. It has been shown that if (8.79) is not satisfied the adder will operate in a non-linear fashion, producing a filter output even when the input is zero. This output can be either a constant or a periodic signal which is generally called an *overflow oscillation*. The solution to this problem is to use *saturating adders*. The transfer characteristic of such an adder is shown in Figure 8.15 and does not allow the result to exceed the specified dynamic range.

8.5.2 Limit Cycles and the Dead-band Effect

If the input to a digital filter is zero or a constant, the arithmetic round-off errors cannot be treated as uncorrelated random processes in the manner given in Section 8.4. Instead, the round-off noise is dependent on the input signal, and even when the input is switched-off there will be an output determined by this round-off noise. This results in self-sustained oscillations known as *limit* cycles. These are now illustrated by a first-order filter described by the first-order difference equation

$$g(n) = \alpha g(n-1) + f(n) \tag{8.80}$$

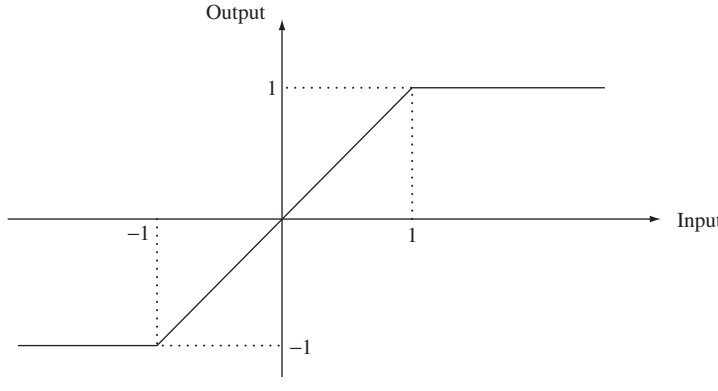


Figure 8.15 Transfer characteristic of a two's complement saturating adder

so that its transfer function is

$$H(z) = \frac{1}{1 - \alpha z^{-1}} \tag{8.81}$$

and stability requires

$$|\alpha| < 1. \tag{8.82}$$

Taking $\alpha = 0.94$, the initial condition $g(-1) = 11$ and assuming that the input is switched-off, that is $f(n) = 0$, we have

$$g(n) = 0.94g(n-1). \tag{8.83}$$

The output is calculated assuming infinite precision arithmetic, then rounded to the nearest integer, giving the values in Table 8.1. These show that although the exact value of $g(n)$ decays exponentially, the rounded value reaches a steady state value of 8. This is what is meant by the *limit cycle* response to zero input. If we repeat the calculations in Table 8.1 with α in (6.80) being -0.94 we obtain Table 8.2 revealing that the rounded value of *g(n) oscillates* between 8 and -8. The range $[-8, 8]$ is called the *dead-band* and the phenomenon is called the *dead-band effect*.

In general, for a first-order filter, the dead-band $[-D, D]$ is obtained from

$$D = \text{integral part of} \left\{ \frac{0.5}{1 - |\alpha|} \right\} \tag{8.84}$$

Table 8.1 Output of (8.83) calculated assuming infinite precision arithmetic and rounded to the nearest integer

n	Exact $g(n)$	Rounded $g(n)$
0	10.34	10
1	9.719 6	9
2	9.136 424	8
3	8.588 238 6	8
4	8.072 944 2	8
5	7.588 567 6	8

Table 8.2 Calculations in Table 8.1 with α in (8.80) being -0.94

n	Exact $g(n)$	Rounded $g(n)$
0	−10.34	−10
1	9.719 6	9
2	−9.136 424	−8
3	8.588 238 6	8
4	−8.072 944 2	−8
5	7.588 567 6	8

where:

(i) for $\alpha > 0$ the limit cycles are constant.
(ii) for $\alpha < 0$ the signs alternate with a frequency of oscillation $\omega_N/2$.

The analysis of limit-cycles in a second-order section shows that it has two modes of auto-oscillations. The first is similar to the first-order case resulting in either a constant output or oscillations with $\omega_N/2$. In the second mode, the filter behaves as though it possessed a pair of conjugate poles on the unit circle.

It must be observed that limit-cycle oscillations occur due to quantization before storage in the registers. They generally have small amplitudes in well-designed systems with a sufficiently large number of bits and a sufficiently small quantizing step. Therefore they are generally smaller in amplitude than possible overflow oscillations in the absence of logic saturation devices. An upper bound on limit cycle amplitudes can be easily obtained by noting that the quantization error signal is bounded by

$$\varepsilon(n) \leq \frac{q}{2}. \tag{8.85}$$

Therefore, if the filter impulse response sequence is $\{h(n)\}$, then

$$|g(n)| \leq \frac{q}{2} \sum_n |h(n)| \tag{8.86}$$

which is, in fact, too pessimistic. A more realistic estimate of the limit cycle amplitudes is given by

$$A_{1c} = \frac{q}{2} \max |H(\exp(\mathrm{j}\omega T))| \tag{8.87}$$

where $H(\exp(\mathrm{j}\omega T))$ is the transfer function of the filter section. For example, a second-order section described by

$$H(z) = \frac{1}{1 + \beta_1 z^{-1} + \beta_2 z^{-2}} \tag{8.88}$$

has poles as

$$P_{1,2} = |r| \exp(\mathrm{j} \pm \theta) \tag{8.89}$$

so that (8.86) gives

$$|g(n)| \leq \frac{q}{2} \frac{1}{(1-r)\sin\theta}$$

while (8.87) gives

$$A_{1\mathrm{c}} = \frac{q}{2} \frac{1}{(1-r^2)\sin\theta}. \tag{8.90}$$

Finally, it is observed that in order to minimize limit-cycle oscillations, the register length must be chosen sufficiently large and the quantizing step must be sufficiently small. They can also be eliminated by quantization using *truncation* rather than rounding. However, the price is an increased quantization noise *in the presence of a signal*.

8.6 Conclusion

This chapter dealt with the various sources of degradation in performance of digital signal processors, which are inherent in such systems due to the use of finite digital words to represent signal quantities and the coefficients of the transfer function. The study of these finite word-length effects is essential, since they must be taken into consideration before the design of a digital system is completed and implemented. These limitations on the performance of a digital signal processing system, such as a digital filter, constitute a fundamental difference between these systems and their analog counterparts which do not suffer from such limitations.

Problems

8.1 For each of the following filter transfer functions:

(a) $H(z) = \dfrac{(1+z^{-1})^3}{(1+0.1z^{-1})(1-0.4z^{-1}+0.2z^{-2})}$

(b) $H(z) = \dfrac{z^{-1}(1+z^{-1})}{(4-2z^{-1}+z^{-2})(37+51z^{-1}+27z^{-2}+5z^{-3})}$

calculate the output error due to input signal quantization with 32-bit fixed-point number representation.

8.2 For the transfer functions of Problem 8.1, calculate the coefficient word-length required to avoid instability.

8.3 For the transfer functions of Problem 8.1, calculate the output error due to round-off accumulation without coefficient quantization, in each of the cases of direct realization, cascade realization and parallel realization.

8.4 Repeat the calculations of Problem 8.3 taking into account the effect of coefficient quantization. Assume 16-bit word-lengths throughout.

9

Linear Estimation, System Modelling and Adaptive Filters

9.1 Introduction

In this chapter, a central problem in signal processing is addressed, namely: the estimation of some signal of interest from a set of received noisy data signals [11, 12, 20, 21]. If the signal is deterministic with known spectrum and this spectrum does not overlap with that of the noise, then the signal can be recovered by the conventional filtering techniques discussed earlier. However, this situation is very rare. Instead, we are often faced with the problem of estimating an *unknown random* signal in the presence of noise, and this is usually accomplished so as to minimize the error in the estimation according to a certain criterion. This leads to the area of adaptive filtering [21]. A closely related area is that of the *modelling* or *simulation* of the behaviour of an unknown system (or process) by a linear system. Initially, the principles of linear estimation and modelling are discussed, then it is shown how these can be implemented using adaptive algorithms. In linear estimation theory, we use techniques that are derived from the classical mean-square approximation method for deterministic functions. Therefore this chapter begins by a discussion of such methods, in preparation for extension to stochastic signals. The exposition in this chapter is intended to serve as an introduction to linear estimation by analog and digital techniques. Emphasis is laid on the Wiener filter and the associated adaptive algorithms in the discrete domain. We rely very heavily on the results and notation of Chapter 7.

9.2 Mean-square Approximation

9.2.1 Analog Signals

Suppose we are given a function $f(t)$ and it is required to approximate $f(t)$ by a linear combination of N independent signals $x_k(t)$ of the form

$$\hat{f}(t) = \sum_{k=0}^{N-1} a_k x_k(t). \tag{9.1}$$

Signal Processing and Integrated Circuits, First Edition. Hussein Baher.
© 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.

The error in the approximation is given by

$$\begin{aligned} e(t) &= f(t) - \hat{f}(t) \\ &= f(t) - \sum_{k=0}^{N-1} a_k x_k(t). \end{aligned} \tag{9.2}$$

The constants a_k are to be determined so as to minimize the resulting mean square error

$$\varepsilon = \int_{-\infty}^{\infty} e^2(t)\,\mathrm{d}t \tag{9.3}$$

That is

$$\varepsilon = \int_{-\infty}^{\infty} \left(f(t) - \sum_{k=0}^{N-1} a_k x_k(t) \right)^2 \mathrm{d}t \tag{9.4}$$

We have encountered an example of this type of approximation in the case of representing a function by a truncated Fourier series. In that case the signals $x_k(t)$ were of the form $\sin k\omega_0 t$, $\cos k\omega_0 t$ or $\exp(\mathrm{j}\omega_0 t)$. Here we study the problem in its most general form where the signals $x_k(t)$ are not restricted to be of a particular type. We confine ourselves, however, to real signals.

9.2.1.1 The Orthogonality Principle

First we note that two functions $x(t)$ and $y(t)$ are said to be *orthogonal* over an interval $[a, b]$, if

$$\int_a^b x(t)y(t)\,\mathrm{d}t = 0. \tag{9.5}$$

The interval may be $[-\infty, \infty]$ and in this case orthogonality requires

$$\int_{-\infty}^{\infty} x(t)y(t)\,\mathrm{d}t = 0. \tag{9.6}$$

Let us now consider the mean square error ε as given by (9.4) and note that it is a function of the coefficients a_k. In order that ε be minimum, we differentiate it with respect to a typical coefficient a_i and equate to zero

$$\frac{\partial \varepsilon}{\partial a_i} = 0 \quad i = 0, 1, \ldots (N-1) \tag{9.7}$$

or

$$\frac{\partial}{\partial a_i} \left[\int_{-\infty}^{\infty} \left(f(t) - \sum_{k=0}^{N-1} a_k x_k(t) \right)^2 \right] = 0 \tag{9.8}$$

That is

$$-2 \int_{-\infty}^{\infty} \left(f(t) - \sum_{k=0}^{N-1} a_k x_k(t) \right) x_i(t)\,\mathrm{d}t = 0 \tag{9.9}$$

or

$$\int_{-\infty}^{\infty} (f(t) - \hat{f}(t)) x_i(t)\, \mathrm{d}t = 0 \quad i = 0, 1, \ldots (N-1) \tag{9.10}$$

so that

$$\int_{-\infty}^{\infty} e(t) x_i(t)\, \mathrm{d}t = 0 \quad i = 0, 1, \ldots (N-1) \tag{9.11}$$

which means that the error, $e(t) = (f(t) - \hat{f}(t))$, must be *orthogonal to the signals* $x_i(t)$.

Writing (9.11) explicitly we have

$$a_0 \int_{-\infty}^{\infty} x_0(t) x_i(t)\, \mathrm{d}t + a_1 \int_{-\infty}^{\infty} x_1(t) x_i(t) + \cdots + a_{N-1} \int_{-\infty}^{\infty} x_{N-1}(t) x_i(t)$$

$$= \int_{-\infty}^{\infty} f(t) x_i(t)\, \mathrm{d}t \quad i = 0, 1, \ldots (N-1) \tag{9.12}$$

which is a system of N equations in the coefficients $a_0, a_1 \cdots a_{N-1}$ whose solution gives the optimum values which minimize the mean square error.

Clearly, if we impose the extra condition that the signals $x_k(t)$ *themselves* form an orthogonal set, then

$$\int_{-\infty}^{\infty} x_i(t) x_k(t)\, \mathrm{d}t = E_i \quad i = k$$

$$= 0 \quad i \neq k. \tag{9.13}$$

If this is the case, then (9.12) and (9.13) give

$$a_i = \frac{1}{E_i} \int_{-\infty}^{\infty} f(t) x_i(t)\, \mathrm{d}t. \tag{9.14}$$

9.2.2 Discrete Signals

Two sequences $\{x(n)\}$ and $\{y(n)\}$ are said to be orthogonal over the interval $[-\infty, \infty]$ if

$$\sum_{n=-\infty}^{\infty} x(n) y(n) = 0. \tag{9.15}$$

Now, given an arbitrary sequence $\{f(n)\}$ it is required to approximate it by a linear combination of N linearly independent sequences $\{x_k(n)\}$ as

$$\hat{f}(n) = \sum_{k=0}^{N-1} a_k x_k(n). \tag{9.16}$$

The error in the approximation is

$$e(n) = f(n) - \hat{f}(n)$$

$$= f(n) - \sum_{k=0}^{N-1} a_k x_k(n). \tag{9.17}$$

Again we wish to determine the coefficients a_k so as to minimize the mean square error

$$\varepsilon = \sum_{n=-\infty}^{\infty} e^2(n)$$
$$= \sum_{n=-\infty}^{\infty} \left(f(n) - \sum_{k=0}^{N-1} a_k x_k(n) \right)^2. \tag{9.18}$$

To this end, the above expression is differentiated with respect to a typical coefficient and equated to zero. This gives

$$\frac{\partial \varepsilon}{\partial a_i} = -2 \sum_{n=-\infty}^{\infty} (f(n) - \hat{f}(n)) x_i(n) = 0. \tag{9.19}$$

Thus

$$\sum_{n=-\infty}^{\infty} (f(n) - \hat{f}(n)) x_i(n) = 0 \quad i = 0, 1, \ldots (N-1) \tag{9.20}$$

or

$$\sum_{n=-\infty}^{\infty} e(n) x_i(n) = 0 \quad i = 0, 1, \ldots (N-1) \tag{9.21}$$

which means that the error is orthogonal to the signals $x_i(n)$. Therefore, the orthogonality principle holds also for discrete signals. To determine the coefficients, we write (9.19) explicitly as

$$a_0 \sum_{n=-\infty}^{\infty} x_0(n) x_i(n) + a_1 \sum_{n=-\infty}^{\infty} x_1(n) x_i(n) + \cdots + a_{N-1} \sum_{n=-\infty}^{\infty} x_{N-1}(n) x_i(n)$$
$$= \sum_{n=-\infty}^{\infty} f(n) x_i(n) \quad i = 0, 1, \ldots (N-1). \tag{9.22}$$

The solution of the above N equations gives the optimum values for the coefficients a_k.

In the rest of this chapter, we shall adapt the mean-square approximation principle to the use in stochastic signal estimation. The term *estimation*, rather than *approximation* is used since the signals are random or at best contain random components.

9.3 Linear Estimation, Modelling and Optimum Filters

Consider a stochastic signal $\mathbf{f}(t)$ which can only be observed in the presence of additive noise [12]. The noise is also a stochastic signal $v(t)$ and the received data signal is given by

$$\mathbf{x}(t) = \mathbf{f}(t) + v(t) \tag{9.23}$$

which is represented symbolically in Figure 9.1(a). The problem considered here is that of the estimation of the signal of interest $\mathbf{f}(t)$ when the available data is $\mathbf{x}(t)$. In (9.23) it is assumed that $\mathbf{f}(t)$ is an unknown sample of the stochastic process and that $\mathbf{f}(t)$ and $v(t)$ are jointly stationary with known power spectra.

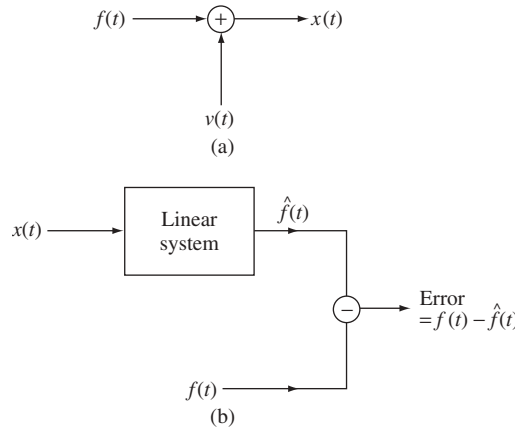


Figure 9.1 (a) Symbolic representation of (9.23). (b) Linear estimation

We distinguish between three types of estimation:

(a) The estimation of $\mathbf{f}(t)$ at time t from data $\mathbf{x}(t)$ available up to time t. This is usually referred to as *filtering*.
(b) The estimation of $\mathbf{f}(t)$ at time t from data available *up* to time t *and* those measured *later* than t. This is called *smoothing*. In this case there is a delay in estimating the result.
(c) The signal is estimated at time $(t + \tau)$ in the *future* using data observed up to time t. This is called *prediction*.

Generally speaking, the linear estimation problem consists in processing the available noisy signal by passing it through a linear system [see Figure 9.1(b)] which will be, loosely speaking, called a *filter* whether the processing is filtering, smoothing or prediction. The output of the filter $\hat{\mathbf{f}}(t)$ must be an estimate of the signal $\mathbf{f}(t)$ with the requirement of suppressing the noise $v(t)$.

The pioneering work in estimation theory was accomplished by Kolmogorov and Wiener in the 1940s in the case of stationary processes. The solution uses the *minimum mean square error* as an optimality criterion, and the resulting optimum filter is known as the *Wiener filter*. However, the Kolmogorov–Wiener theory was not suitable for dealing with non-stationary signals and noise. This triggered the work of Kalman in the 1960s, who developed a new theory applicable to non-stationary processes. The solution gives the optimum design using the method of least squares, and the result is a variable-coefficient system called the *Kalman filter*.

Another problem, which is very closely related to the optimum filtering one, is that of *modelling* or *simulating* the behaviour of an unknown system by a linear system (filter). This situation is depicted in Figure 9.2, in which the error between the output of the filter and the desired output is minimized. The solution to the optimum modelling problem is identical in form to that of the estimation problem. Therefore, the discussion in this chapter applies to both cases. However, we shall concentrate on either one or the other for the sake of being specific.

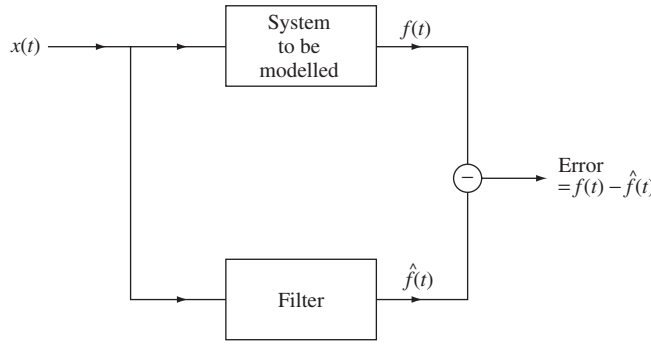


Figure 9.2 Modelling of a system by a linear filter

9.4 Optimum Minimum Mean-square Error Analog Estimation

Let us return to expression (9.1) with the corresponding representation or model in Figure 9.1 or Figure 9.2. We now wish to process the signal $\mathbf{x}(t)$ by passing it through an analog (continuous-time) filter which produces an output $\hat{\mathbf{f}}(t)$. This output must be an estimate of the signal $\mathbf{f}(t)$ with the requirement of suppressing the noise $\nu(t)$. Assuming that the received data signal $\mathbf{x}(t)$ is (wide-sense) stationary, then its mean is

$$E[\mathbf{x}(t)] = \eta \text{ a constant} \tag{9.24}$$

and its autocorrelation is

$$R_{xx}(\tau) = E[\mathbf{x}(t)\mathbf{x}(t+\tau)] \tag{9.25}$$

depending only on the difference τ. Moreover, since we can always subtract the mean η, it is assumed that the process has zero-mean.

At first, we shall consider data smoothing by *non-causal* filters, then discuss the problem of *causal* estimation.

9.4.1 Smoothing by Non-causal Wiener Filters

Consider Figure 9.3 in which the data $\mathbf{x}(t)$ is applied to a filter with impulse response $h(t)$. We wish to find $h(t)$ or its Fourier transform $H(\omega)$ such that the output $\hat{\mathbf{f}}(t)$ of the filter is an estimate of $\mathbf{f}(t)$. In this case we estimate $\mathbf{f}(t)$ by a linear combination of *past and future* values of the data $\mathbf{x}(t)$. The output $\hat{\mathbf{f}}(t)$ of the (non-causal) estimator is related to its input by the convolution

$$\hat{\mathbf{f}}(t) = \int_{-\infty}^{\infty} h(\alpha)\mathbf{x}(t-\alpha)\,\mathrm{d}\alpha. \tag{9.26}$$

Thus, $\hat{\mathbf{f}}(t)$ is a linear combination of the data $\mathbf{x}(t-\tau)$ where τ takes all values from $-\infty$ to ∞. The error in the estimation is given by

$$\mathbf{e}(t) = \mathbf{f}(t) - \hat{\mathbf{f}}(t) = \mathbf{f}(t) - \int_{-\infty}^{\infty} h(\alpha)\mathbf{x}(t-\alpha)\,\mathrm{d}\alpha. \tag{9.27}$$

It is now required to determine the impulse response of the filter such that the error signal is minimum in the mean-square sense. This can be achieved by a generalization of the

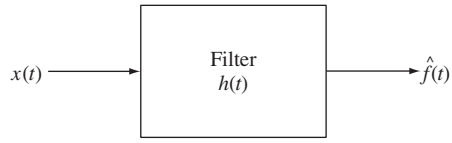


Figure 9.3 A filter with input as the received noisy data signal and output as the estimate of the desired signal

orthogonality principle discussed in Section 9.2.1. To this end, we first extend expressions (9.16) and (9.17) to infinite summations then perform limiting operations on the resulting summations, converting them into integrals. This gives for the mean square error

$$\begin{aligned}\varepsilon &= E[\mathbf{e}^2(t)] = E[(\mathbf{f}(t) - \hat{\mathbf{f}}(t))^2] \\ &= E\left[\left(\mathbf{f}(t) - \int_{-\infty}^{\infty} h(\alpha)\mathbf{x}(t-\alpha)\,\mathrm{d}\alpha\right)^2\right]\end{aligned} \tag{9.28}$$

which, by the orthogonality principle, is minimum for the error $\mathbf{e}(t)$ orthogonal to the data. Thus

$$E\left[\left(\mathbf{f}(t) - \int_{-\infty}^{\infty} h(\alpha)\mathbf{x}(t-\alpha)\,\mathrm{d}\tau\right)\mathbf{x}(t-\tau)\right] = 0 \tag{9.29}$$

Using the linearity property of the expectation operator we obtain

$$E[\mathbf{f}(t)\mathbf{x}(t-\tau)] = \int_{-\infty}^{\infty} h(\alpha)E[\mathbf{x}(t-\alpha)\mathbf{x}(t-\tau)]\,\mathrm{d}\alpha = \int_{-\infty}^{\infty} h(\alpha)E[\mathbf{x}(\tau-\alpha)]\,\mathrm{d}\alpha \tag{9.30}$$

That is

$$R_{fx}(\tau) = \int_{-\infty}^{\infty} h(\alpha)R_{xx}(\tau-\alpha)\,\mathrm{d}\alpha \tag{9.31}$$

which is called the *Wiener–Hopf condition for optimum non-casual estimation*. Hence, the impulse response of the optimum non-causal filter is obtained by solving equation (9.31) for $h(t)$. This can be achieved by first taking the Fourier transform of both sides of (9.31). With the transform pairs

$$R_{fx}(\tau) \leftrightarrow P_{fx}(\omega) \tag{9.32}$$

$$R_{xx}(\tau) \leftrightarrow P_{xx}(\omega) \tag{9.33}$$

$$h(t) \leftrightarrow H(\omega) \tag{9.34}$$

the transformed version of (9.31) becomes

$$P_{fx}(\omega) = H(\omega)P_{xx}(\omega). \tag{9.35}$$

It follows that the transfer function of the optimum filter is given by

$$H(\omega) = \frac{P_{fx}(\omega)}{P_{xx}(\omega)} \tag{9.36}$$

which is determined by the cross-power spectrum of the signal and data, as well as the power spectrum of the data. A filter whose transfer function is given by (9.36) is called a

non-causal Wiener filter. If t is real time, then the response defined by (9.36) cannot be realized and must, therefore, be approximated as it is always the case with unrealizable idealized characteristics.

Now, under the condition of minimum mean-square error, the error $e(t)$ is orthogonal to the data $\mathbf{x}(t-\tau)$ for all τ. Hence, (9.27) and (9.29) give

$$E[\{\mathbf{f}(t)-\hat{\mathbf{f}}(t)\}\hat{\mathbf{f}}(t)]=0 \tag{9.37}$$

so that

$$E[\mathbf{f}(t)\hat{\mathbf{f}}(t)]=E[\hat{\mathbf{f}}^2]. \tag{9.38}$$

However, (9.28) gives

$$\varepsilon_{\min}=E[\mathbf{f}(t)-\hat{\mathbf{f}}(t))^2]=E[\mathbf{f}^2(t)]-2E[\mathbf{f}(t)\hat{\mathbf{f}}(t)]+E[\hat{\mathbf{f}}^2(t)]$$

which, upon use of (9.38), gives

$$\varepsilon_{\min}=E[\mathbf{f}^2(t)]-E[\mathbf{f}(t)\hat{\mathbf{f}}(t)]=R_{ff}(0)-\int_{-\infty}^{\infty}h(\alpha)R_{fx}(\alpha)\,\mathrm{d}\alpha. \tag{9.39}$$

The above minimum error can be expressed in the frequency domain using Parseval's relation (7.73) together with (7.84). This gives

$$\varepsilon_{\min}=\frac{1}{2\pi}\int_{-\infty}^{\infty}P_{ff}(\omega)\,\mathrm{d}\omega-\frac{1}{2\pi}\int_{-\infty}^{\infty}P_{fx}^{*}(\omega)H(\omega)\,\mathrm{d}\omega. \tag{9.40}$$

An important special case of the Wiener filter occurs when the signal $\mathbf{f}(t)$ and the noise $v(t)$ are *uncorrelated*; they are called *orthogonal*. In this case $R_{fv}(\tau)=0$ and we have

$$P_{fx}(\omega)=P_{ff}(\omega) \tag{9.41}$$

and

$$P_{xx}(\omega)=P_{ff}(\omega)+P_{vv}(\omega). \tag{9.42}$$

Therefore, the transfer function of the required filter becomes

$$H(\omega)=\frac{P_{ff}(\omega)}{P_{ff}(\omega)+P_{vv}(\omega)} \tag{9.43}$$

with the associated minimum error

$$\varepsilon_{\min}=\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{P_{ff}(\omega)P_{vv}(\omega)}{P_{ff}(\omega)+P_{vv}(\omega)}\,\mathrm{d}\omega. \tag{9.44}$$

Example 9.1 Suppose the noise is white, so that

$$P_{vv}(\omega)=A \quad \text{a constant}$$

and the signal is known to have the spectrum

$$P_{ff}(\omega)=\frac{1}{\omega^2+\gamma^2}.$$

Then, it is reasonable to assume that the signal and noise are uncorrelated, so that the transfer function of the required filter is given by (9.43) as

$$H(\omega) = \frac{1}{1 + A(\gamma^2 + \omega^2)} = \frac{1/A}{\omega^2 + (1 + A\gamma^2)/A}$$

so that

$$h(t) = \frac{1}{A} \exp[-(1 + A\gamma^2)/A]|t|$$

which is clearly non-causal.

9.4.2 Causal Wiener Filters

When the processing of signals is performed in real time, only past values of the data are available at the present time t. Thus, we require an estimate $\hat{\mathbf{f}}(t)$ of the signal $\mathbf{f}(t)$, which is obtained as a linear combination of the data $\mathbf{x}(t - \tau)$ in which τ takes values from 0 to ∞. Therefore we need a *causal* filter with impulse response satisfying

$$h(t) = 0 \quad t < 0. \tag{9.45}$$

The general analysis of the previous section is still valid, but instead of (9.26) we now have

$$\hat{\mathbf{f}}(t) = \int_0^\infty h(\alpha)\mathbf{x}(t - \alpha)\, \mathrm{d}\alpha. \tag{9.46}$$

To minimize the mean square error, the orthogonality principle in (9.29) is used with the lower limit of the integral changed to 0. This gives

$$E\left[\left(\mathbf{f}(t) - \int_0^\infty h(\alpha)\mathbf{x}(t - \alpha)\, \mathrm{d}\alpha\right) \mathbf{x}(t - \tau)\right] = 0 \tag{9.47}$$

so that the equivalent of (9.31) for a causal Wiener filter is given by

$$R_{fx}(\tau) = \int_0^\infty h(\alpha)R_{xx}(\tau - \alpha)\, \mathrm{d}\alpha \quad \text{for } \tau > 0 \tag{9.48}$$

which is the *Wiener–Hopf condition* for optimum causal estimation. Since (9.48) is valid only for $\tau > 0$, it cannot be solved in a manner similar to the solution of (9.31) which is valid for all τ. We shall not pursue the solution of (9.48) to obtain $h(t)$ of the required filter. Instead, we move, shortly, to the *discrete-time* version of the estimation problem where digital techniques have overwhelming advantages over analog ones. However, before doing so, a special type of analog processors is introduced due to its importance in communication systems.

9.5 The Matched Filter

An important special processor is the *matched filter* [11]. The optimality criterion in this case is different from that for the Wiener filter. Returning to (9.23) we assume that the signal $f(t)$ is deterministic and known but corrupted by additive random noise, so that the received signal is

$$\mathbf{x}(t) = f(t) + \nu(t). \tag{9.49}$$

Now, applying the data $\mathbf{x}(t)$ to a filter with impulse response $h(t)$ and transfer function $H(\omega)$ the output is the random signal

$$\mathbf{y}(t) = \mathbf{x}(t) * h(t) = y_f(t) + \mathbf{y}_v(t) \tag{9.50}$$

where

$$y_f(t) = f(t) * h(t) \tag{9.51}$$

which is the component of the output due to the signal $f(t)$, and

$$\mathbf{y}_v(t) = v(t) * h(t) \tag{9.52}$$

which is the component of the output due to the random noise $v(t)$. At any time τ the signal to noise ratio at the output is

$$\text{SNR} = \frac{|\, y_f(\tau)\,|}{\{E[|\, y_v(\tau)\,|^2\}^{1/2}}. \tag{9.53}$$

It is now required to determine the filter transfer function such that the output SNR is maximum. The output $y_f(t)$ due to $f(t)$ can be written from (9.51) as

$$y_f(t) = \Im^{-1}[H(\omega)F(\omega)] = \frac{1}{2\pi}\int_{-\infty}^{\infty} H(\omega)F(\omega)\exp(\mathrm{j}\omega t)\,\mathrm{d}\omega \tag{9.54}$$

whereas for the output due to the input noise we have

$$E[\mathbf{y}_v^2(t)] = \frac{1}{2\pi}\int_{-\infty}^{\infty} P_v(\omega)|\, H(\omega))\,|^2 \tag{9.55}$$

where $P_{vv}(\omega)$ is the power spectrum of the stationary noise $\boldsymbol{v}(t)$.

Now, consider the two functions

$$G_1(\omega) = (P_{vv}(\omega))^{1/2} H(\omega) \tag{9.56}$$

and

$$G_2(\omega) = \frac{F(\omega)}{(P_{vv}(\omega))^{1/2}}\exp(\mathrm{j}\omega t) \tag{9.57}$$

so that the product $G_1(\omega)G_2(\omega)$ is the integrand in (9.54). Applying Schwartz's inequality [11, 12] to $G_1(\omega)$ and $G_2(\omega)$ we find

$$\left|\int_{-\infty}^{\infty} F(\omega)H(\omega)\exp(\mathrm{j}\omega t)\,\mathrm{d}\omega\right|^2 \le \frac{|\, F(\omega)\,|^2}{P_{vv}(\omega)}\,\mathrm{d}\omega \int_{-\infty}^{\infty} P_{vv}(\omega)/H(\omega)\,|^2\,\mathrm{d}\omega. \tag{9.58}$$

Using (9.53) in the above we obtain

$$(\text{SNR})^2 = \frac{y_f^2(\tau)}{E[\mathbf{y}_v^2(\tau)]} \le \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{|F(\omega)\,|^2}{P_{vv}(\omega)}\,\mathrm{d}\omega. \tag{9.59}$$

This attains its maximum value with equality [11, 12], which occurs when $G_1(\omega)$ is proportional to $G_2^*(\omega)$

$$G_1(\omega) = cG_2^*(\omega) \quad c \text{ is a constant} \tag{9.60}$$

or from (9.56) and (9.57)

$$(P_{\nu\nu}(\omega))^{1/2}H(\omega) = c\frac{F*(\omega)}{(P_{\nu\nu}(\omega))^{1/2}}\exp(\mathrm{j}\omega t) \tag{9.61}$$

so that the transfer function of the required filter is

$$H(\omega) = c\frac{F*(\omega)}{P_{\nu\nu}(\omega)}\exp(-\mathrm{j}\omega t) \tag{9.62}$$

which yields the maximum SNR given by

$$(\mathrm{SNR})^2_{\max} = \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{|F(\omega)|^2}{P_{\nu\nu}(\omega)}\,\mathrm{d}\omega. \tag{9.63}$$

In the special case of white noise

$$P_{\nu\nu}(\omega) = A \quad \text{a constant} \tag{9.64}$$

and (9.62) becomes

$$H(\omega) = \frac{c}{A}F*(\omega)\exp(-\mathrm{j}\omega t) \tag{9.65}$$

so that

$$h(t) = \frac{c}{A}f(\tau - t) \tag{9.66}$$

with

$$\begin{aligned}(\mathrm{SNR})^2_{\max} &= \frac{1}{2\pi A}\int_{-\infty}^{\infty}|F(\omega)|^2\,\mathrm{d}\omega \\ &= \frac{E}{A}\end{aligned} \tag{9.67}$$

or

$$E(\mathrm{SNR})_{\max} = (E/A)^{1/2} \tag{9.68}$$

where E is the energy of the signal $f(t)$.

9.6 Discrete-time Linear Estimation

Again, consider a stochastic signal $\mathbf{f}(t)$ which can only be observed in the presence of additive noise. The noise is also a stochastic signal $\nu(t)$ and the received data signal is $\mathbf{x}(t)$ given by (9.23) (see Figure 9.1). The problem considered here is exactly the same as that of Section 9.4 namely: the estimation of the signal of interest $\mathbf{f}(t)$ when the available information is $\mathbf{x}(t)$. However we wish to use discrete-time techniques, in particular digital filters. The digital filter is required to operate in *real time* and we concentrate solely on causal estimation. In order to use discrete-time filters for the processing, we first produce a sampled version of $\mathbf{x}(t)$ by taking samples of the available signal every T seconds. Thus (9.23) becomes

$$\mathbf{x}(nT) = f(nT) + \nu(nT) \tag{9.69}$$

which, for convenience, is written as

$$\mathbf{x}(n) = \mathbf{f}(n) + \nu(n) \tag{9.70}$$

upon dropping the constant T. A symbolic representation of (9.70) is shown in Figure 9.4. The set of available data is the sequence $\{\mathbf{x}(n)\}$ and is called a *random time-series*.

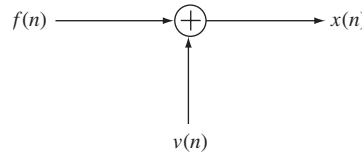


Figure 9.4 Symbolic representation of (9.70)

9.6.1 Non-recursive Wiener Filtering

We now wish to process the signal $\mathbf{x}(n)$ by passing it through a discrete-time linear system, as shown in Figure 9.5, which produces an output $\hat{\mathbf{f}}(n)$ as an *estimate* of the signal $\mathbf{f}(n)$. The requirement placed on the system is to suppress the effect of the noise $v(n)$. The estimator has an impulse response sequence $\{h(n)\}$ which is assumed here to be of finite duration, that is we use an FIR filter for the processing. The duration of the filter is taken to be equal to the number of samples of the input signal.

Consider the FIR filter shown in Figure 9.6 whose transfer function is given by

$$H(z) = \sum_{n=0}^{M-1} h(n) z^{-n} \tag{9.71}$$

and let the input to the filter be the available signal $\mathbf{x}(n)$. Assuming that the random signal $\mathbf{x}(n)$ is (wide-sense) stationary, then

$$\begin{aligned} &\text{(i)} \;\; E[\mathbf{x}(n)] = \eta, \quad \text{a constant} \\ &\text{(ii)} \;\; R_{xx}(n, m) = E[\mathbf{x}(n)\mathbf{x}(n-m)] \end{aligned} \tag{9.72}$$

depending only on the difference m. Note that we are using $-m$ instead of m in the above definitions. Moreover, we can subtract the mean η so that it is assumed that the process has zero mean.

Now, if the filter output is $\hat{\mathbf{f}}(n)$, *then*

$$\hat{\mathbf{f}}(n) = \sum_{k=0}^{M-1} h(k)\mathbf{x}(n-k) \tag{9.73}$$

which is a linear combination of the received data $\mathbf{x}(n-k)$. The above expression can be put in matrix form as

$$\hat{\mathbf{f}}(n) = [h(n)]'[\mathbf{x}(n)] \tag{9.74}$$

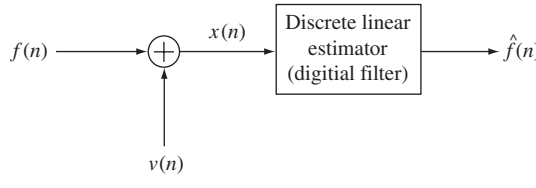


Figure 9.5 Discrete-time linear estimation

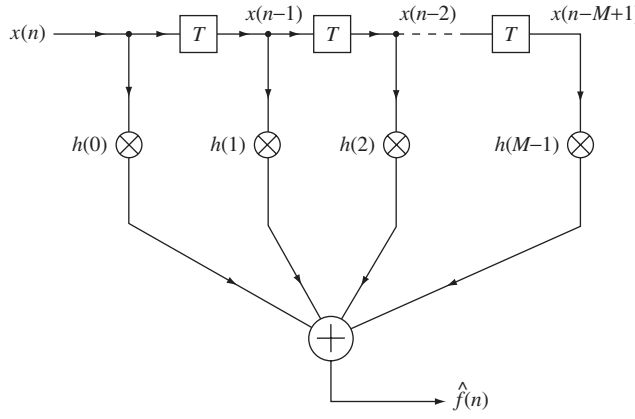


Figure 9.6 An FIR filter as a non-recursive linear estimator

where

$$[h(n)] = \begin{bmatrix} h(0) \\ h(1) \\ \vdots \\ h(M-1) \end{bmatrix} \tag{9.75}$$

$$[\mathbf{x}(n)] = \begin{bmatrix} \mathbf{x}(n) \\ \mathbf{x}(n-1) \\ \vdots \\ \mathbf{x}(n-M+1) \end{bmatrix}. \tag{9.76}$$

Since the *desired* filter output is $\mathbf{f}(n)$, then the error in the estimation is

$$\mathbf{e}(n) = \mathbf{f}(n) - \hat{\mathbf{f}}(n). \tag{9.77}$$

It is now required to determine the filter coefficients $\{h(n)\}$ such that the error signal is minimum in the mean-square sense. This can be easily achieved by a generalization of the orthogonality principle discussed in Section 9.2.2, to encompass stochastic processes. Thus, using (9.73) and (9.77), the mean square error is given by

$$\varepsilon(n) = E[\mathbf{e}^2(n)] = E\left[\left(\mathbf{f}(n) - \sum_{k=0}^{M-1} h(k)\mathbf{x}(n-k)\right)^2\right]. \tag{9.78}$$

The variable parameters here are the filter coefficients $h(k)$. Therefore, in order to minimize the mean square error with respect to the typical coefficient $h(m)$ we differentiate (9.78) to obtain

$$\frac{\partial \varepsilon(n)}{\partial h(m)} = 2E\left[\left(\mathbf{f}(n) - \sum_{k=0}^{M-1} h(k)\mathbf{x}(n-k)\right)\mathbf{x}(n-m)\right]. \tag{9.79}$$

In order that $\varepsilon(n)$ be a minimum, the above expression must be zero for all m. Thus

$$E\left[\left(\mathbf{f}(n) - \sum_{k=0}^{M-1} h(k)\mathbf{x}(n-k)\right)\mathbf{x}(n-m)\right] = 0 \quad m = 0, 1, \ldots (M-1) \tag{9.80}$$

which means that the error signal (in the curved brackets above) is orthogonal to the data. Using the linearity property of the expectation operator, (9.80) becomes

$$E[\mathbf{f}(n)\mathbf{x}(n-m)] - \sum_{k=0}^{M-1} h(k)E[\mathbf{x}(n-h)\mathbf{x}(n-m)] = 0. \tag{9.81}$$

Assuming stationary and jointly stationary processes we can write

$$E[\mathbf{f}(n)\mathbf{x}(n-m)] = R_{fx}(m) \tag{9.82}$$

which is the cross-correlation between $\mathbf{f}(n)$ and $\mathbf{x}(n)$. Also

$$E[\mathbf{x}(n-k)\mathbf{x}(n-m)] = R_{xx}(m-k) \tag{9.83}$$

the autocorrelation of the process $\mathbf{x}(n)$. Thus the orthogonality principle (9.80) assumes the form

$$R_{fx}(m) = \sum_{k=0}^{M-1} h_{\text{op}}(n)R_{xx}(m-k) \tag{9.84}$$

where the subscript 'op' stands for 'optimum'. This is a set of M simultaneous equations in the optimum filter coefficients $h_{\text{op}}(n)$ as the unknowns. The known quantities are: (i) the autocorrelation sequence $\{R_{xx}(m-k)\}$ of the filter input $\mathbf{x}(n)$ and (ii) the cross-correlation sequence $\{R_{fx}(m)\}$ between the desired signal sequence and the filter input sequence. Expression (9.84) is the discrete version of the Wiener–Hopf condition (9.48) given earlier for continuous-time causal Wiener filters.

Now, the M equations expressed by (9.84) can be written in matrix form. Put

$$[h_{\text{op}}(n)] = \begin{bmatrix} h_{\text{op}}(0) \\ h_{\text{op}}(1) \\ h_{\text{op}}(2) \\ \vdots \\ h_{\text{op}}(M-1) \end{bmatrix} \tag{9.85}$$

$$[R_{fx}(n)] = \begin{bmatrix} R_{fx}(0) \\ R_{fx}(1) \\ \vdots \\ R_{fx}(M-1) \end{bmatrix}. \tag{9.86}$$

Noting that, for a stationary process $\mathbf{x}(n)$

$$R_{xx}(n-m) = R_{xx}(m-n) \tag{9.87}$$

then we write

$$[R_{xx}(n)] = \begin{bmatrix} R_{xx}(0) & R_{xx}(1) & \ldots & R_{xx}(M-1) \\ R_{xx}(1) & R_{xx}(0) & & R_{xx}(M-2) \\ \vdots & & & \vdots \\ R_{xx}(M-1) & \ldots & & R_{xx}(0) \end{bmatrix} \tag{9.88}$$

which is called the *autocorrelation matrix* of the *input* signal and is clearly symmetric. Thus (9.84) takes the form

$$[R_{fx}] = [R_{xx}]\,[h_{\text{op}}] \tag{9.89}$$

in which $[h_{\text{op}}]$ defines the coefficients of the optimum estimator called a *non-recursive Wiener filter*. Consequently, the evaluation of the optimum filter coefficient matrix $[h_{\text{op}}]$ requires the inversion of the autocorrelation matrix $[R_{xx}]$ to obtain

$$[h_{\text{op}}] = [R_{xx}]^{-1}[R_{fx}]. \tag{9.90}$$

However, the autocorrelation matrix $[R_{xx}]$ has many interesting properties which simplify the calculations. These are now examined [21]:

(i) $[R_{xx}]$ is symmetric about *both* diagonals. This follows from condition (9.87) for stationary signals.

(ii) $[R_{xx}]$ is *Toeplitz*. This means that the entries on every diagonal are equal. That is: the main diagonal entries are the same, and also the entries on any diagonal parallel to the main diagonal are the same and so on.

(iii) $[R_{xx}]$ is *positive semidefinite* and almost always *positive definite*. To see this let $[A]$ be any column matrix of order M and define the random variable

$$\begin{aligned}\boldsymbol{\alpha} &= [A]'[\mathbf{x}(n)] \\ &= [\mathbf{x}(n)]'[A]\end{aligned} \tag{9.91}$$

where $[\mathbf{x}(n)]$ is the input signal column matrix of order M. The mean-square value of α is

$$\begin{aligned}E[\boldsymbol{\alpha}^2] &= E[[A]'[\mathbf{x}(n)][\mathbf{x}(n)]'[A]] \\ &= [A]'E\{[\mathbf{x}(n)][\mathbf{x}(n)]'\}[A] \\ &= [A]'\,[R_{xx}]\,[A]\end{aligned} \tag{9.92}$$

which is a quadratic form and, since $E[\boldsymbol{\alpha}^2]$ must be non-negative, then $[R_{xx}]$ is positive semidefinite. However, in practical problems we always find that $[R_{xx}]$ is positive definite and hence non-singular.

Now, returning to expression (9.78), it is found that the estimation error, under optimum conditions, can be obtained by substituting the optimum filter coefficients, as given by (9.84), in (9.78). This gives

$$\begin{aligned}\varepsilon_{\min} &= E\left[\left(\mathbf{f}(n) - \sum_{k=0}^{M-1} h_{\text{op}}(k)\mathbf{x}(n-k)\right)\mathbf{f}(n)\right] \\ &= R_{ff}(0) - \sum_{k=0}^{M-1} h_{\text{op}}(k)R_{fx}(n-k) \\ &= R_{ff}(0) - [R_{fx}]'\,[h_{\text{op}}]\end{aligned} \tag{9.93}$$

which, upon use of (9.89), becomes

$$\varepsilon_{\min} = R_{ff}(0) - [R_{fx}]'\,[R_{xx}]^{-1}\,[R_{fx}]. \tag{9.94}$$

9.6.2 *Adaptive Filtering Using the Minimum Mean Square Error Gradient Algorithm*

We have shown that, in order to optimize the coefficients of the FIR filter in the minimum mean-square sense, we require knowledge of the following quantities:

(i) The autocorrelation matrix $[R_{xx}]$ of the input signal $\mathbf{x}(n)$;
(ii) The cross-correlation matrix $[R_{fx}]$ between the desired signal $\mathbf{f}(n)$ and the data $\mathbf{x}(n)$.

Access to the above quantities allows us to evaluate the optimum coefficients vector $[h_{\text{op}}]$ by means of expression (9.90). This involves the inversion of $[R_{xx}]$, then multiplying the result by $[R_{fg}]$. Although matrix inversion, for a high order case, involves a large amount of computation, the double symmetry and Toeplitz character of $[R_{xx}]$ allow considerable simplifications in the computation.

However, there are situations where the filter operates in an environment for which the autocorrelation matrix $[R_{xx}]$ and the cross-correlation matrix $[R_{fx}]$ are unknown. In such a case we can use the data obtained up to time n, to evaluate *estimates* $[\hat{R}_{xx}]$ and $[\hat{R}_{fx}]$ of the matrices $[R_{xx}]$ and $[R_{fx}]$, then use the results to obtain the filter coefficients using (9.90). These correlation calculations can be performed using an FFT algorithm along the lines explained in Section 7.5. However, for high degree filters, this procedure becomes very inefficient. For this reason and to provide an alternative to the inversion of $[R_{xx}]$ we seek a different approach to the problem of determining the optimum values of the filter coefficients. It must be stressed, however, that we still seek the optimization in the minimum mean-square error sense.

This alternative approach is *iterative* in nature, incorporating into the filter an *adaptive algorithm*, as shown in Figure 9.7, which allows automatic adjustment of the filter coefficients. We start with an initial *guess* of the coefficient vector $[h]$. Then the coefficients are changed in a certain iterative procedure so as to guarantee that after a finite number of iterations we reach a set of values close enough to the optimum Wiener solution for the coefficient vector as defined by (9.90). Naturally, we may only reach an approximate solution, but this is the price we pay for speed of computation.

The adaptive algorithm discussed here is called the minimum mean-square (MMS) error algorithm, or the method of *steepest descent*. This is applied according to the following procedure:

1. We start by making an initial guess of the filter coefficient vector $[h]$.
2. The MMS error *gradient vector* is calculated. This is a column matrix whose entries are equal to the first derivatives of the mean square error $\varepsilon(n)$ with respect to the filter coefficients.
3. The coefficients are changed in the direction *opposite to that of the gradient vector*.
4. The new error gradient vector is calculated and the process is iterated until these successive corrections to the gradient vector lead to the minimum mean square error, thus reaching the optimum values of the coefficients $[h_{\text{op}}]$.

It has been shown that this method of *steepest descent* or the MMS error *gradient algorithm* converges to the optimum Wiener filter coefficients $[h_{\text{op}}]$ irrespective of the initial guess. Let us examine this gradient algorithm in some detail. Figure 9.7 shows a schematic diagram of the FIR filter incorporating the *adaptive algorithm*. Here, it is

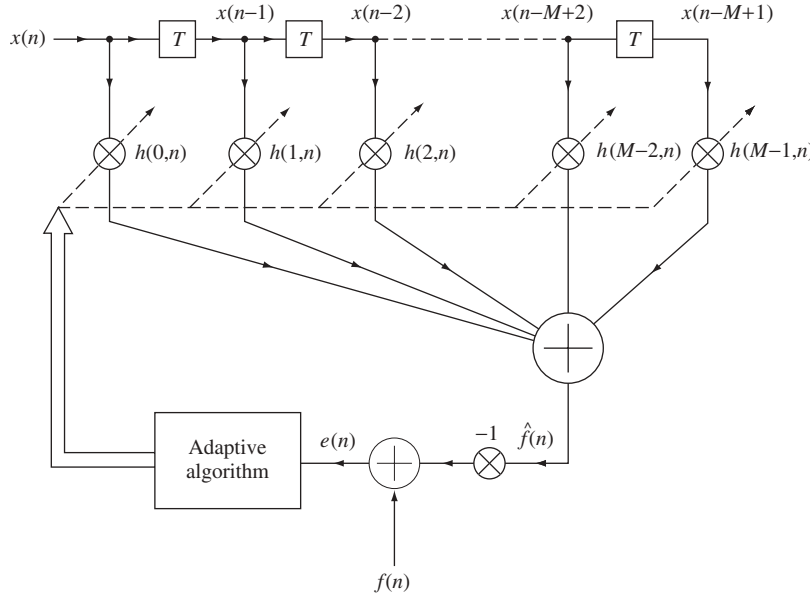


Figure 9.7 An FIR filter incorporating an adaptive algorithm

assumed that the filter coefficients are adjustable according to the adaptive algorithm. Naturally, as discussed in Chapter 5, the realization of the filter including the algorithm can be conveniently achieved in software form. Let the filter coefficients at time n be denoted by

$$h(0,n), h(1,n), h(2,n), \ldots h(M-1,n). \tag{9.95}$$

In the *adaptive filter*, two processes occur. The first is the *adaptation* or automatic adjustment of the filter coefficients according to the algorithm. The second process is the *filtering process*, which produces the output signal calculated from the set of coefficients obtained in the adaptation process. During the second process, a desired response is fed into the adaptive algorithm in order to provide a guide for adjusting the values of the coefficients. At time n, the output of the filter is given by

$$\hat{\mathbf{f}}(n) = \sum_{k=0}^{M-1} h(k,n)\mathbf{x}(n-k) \tag{9.96}$$

which is compared with the desired response $\mathbf{f}(n)$ to produce an error signal

$$\mathbf{e}(n) = \mathbf{f}(n) - \hat{\mathbf{f}}(n). \tag{9.97}$$

The adaptive algorithm is designed such that it uses the error signal $\mathbf{e}(n)$ to produce *corrections* of the filter coefficients such that the optimum Wiener solution defined by (9.90) is approached. From (9.79), the derivative of the error signal with respect to a typical coefficient $h(k,n)$ can be written as

$$\frac{\partial \varepsilon(n)}{\partial h(k,n)} = -2E[\mathbf{e}(n)\mathbf{x}(n-k)] = -2R_{ex}(k) \quad k = 0, 1, 2, \ldots \tag{9.98}$$

where $R_{ex}(n)$ is the cross-correlation between the error signal and the signal at the kth tap input. Expression (9.98) should reach zero at the minimum error point. Defining the error *gradient vector* as

$$\nabla(n) = \begin{bmatrix} \partial\varepsilon(n)/\partial h(0,n) \\ \partial\varepsilon(n)/\partial h(1,n) \\ \vdots \\ \partial\varepsilon(n)/\partial h(M-1,n) \end{bmatrix} \tag{9.99}$$

expression (9.98) can be written in the form

$$\nabla(n) = -2E\{\mathbf{e}(n)[\mathbf{X}(n)]\} \tag{9.100}$$

where

$$[\mathbf{x}(n)] = \begin{bmatrix} \mathbf{x}(n) \\ \mathbf{x}(n-1) \\ \vdots \\ \mathbf{x}(n-M+1) \end{bmatrix} \tag{9.101}$$

Also let the filter coefficients at time n be put in the form

$$[h(n)] = \begin{bmatrix} h(0,n) \\ h(1,n) \\ \vdots \\ h(M-1,n) \end{bmatrix}. \tag{9.102}$$

We now state the gradient (steepest-descent) algorithm. The updated coefficient vector at time $(n+1)$ is obtained according to

$$[h(n+1)] = [h(n)] + \frac{1}{2}\mu[-\nabla(n)] \tag{9.103}$$

where μ is a positive number which determines the *correction step*. Intuitively, (9.103) should eventually lead to the optimum Wiener solution since the filter coefficients are changed in the direction *opposite* to the MMS error gradient. Use of (9.100) in (9.103) gives

$$[h(n+1)] = [h(n)] + \mu E[\mathbf{e}(n)\mathbf{x}(n)] = [h(n)] + \mu[\mathbf{R}_{ex}(n)]. \tag{9.104}$$

It follows that the coefficient vector is updated by applying to the old values a correction equal to μ times the cross-correlation between the error signal $\mathbf{e}(n)$ and the input vector $[\mathbf{x}(n)]$. Hence μ controls the size of the correction and is called the *step-size parameter*. The error signal $\mathbf{e}(n)$ is given by

$$\mathbf{e}(n) = \mathbf{f}(n) - [\mathbf{x}(n)]'\,[h(n)]. \tag{9.105}$$

Expressions (9.104) and (9.105) together define the MMS *error gradient (steepest-descent) algorithm*. To perform adaptive filtering, the algorithm is started with an initial guess of $[h(n)]$, which is usually taken to be a column of zeros, that is we start with zero values of the filter coefficients. The coefficients are updated according to the algorithm to obtain $[h(1)]$, $[h(2)]$, ... until the optimum Wiener solution defined by (9.90) is reached.

9.6.3 The Least Mean Square Error Gradient Algorithm

Now, the main disadvantage of the steepest-descent (MMS) gradient algorithm, lies in the fact that exact measurement of the gradient vector is required at each step in the iteration process. This is not practical and the need arises for an algorithm for deriving *estimates of the gradient vector* from the available data. This can be achieved by means of the least mean square (LMS) error gradient algorithm discussed now. Its advantages are: its simplicity, it does not require matrix inversion, and it does not require correlation measurements.

Instead of using *expectation* values for calculation of the gradient vector given by (9.100), the LMS algorithm uses *instantaneous* estimates of the vector based on *sample* values of the input $[\mathbf{x}(n)]$ and the error $e(n)$. From (9.100), the instantaneous estimate of the gradient vector is

$$\hat{\nabla}(n) = -2\mathbf{e}(n)[\mathbf{x}(n)]. \tag{9.106}$$

Such an estimate is obviously *unbiased* since its expected value is the same as the value in (9.100). In the LMS algorithm, the filter coefficients are changed along the direction of the gradient vector estimate according to the relation

$$[h(n+1)] = [h(n)] + \tfrac{1}{2}\mu[-\hat{\nabla}(n)] = [h(n)] + \mu\mathbf{e}(n)[\mathbf{x}(n)] \tag{9.107}$$

which is much simpler than (9.104) requiring knowledge only of the data $[\mathbf{x}(n)]$ and no cross-correlation estimate. The error signal is still defined by (9.105).

The adaptive filtering process, is then, the same as with the steepest-descent algorithm. Thus we begin by initializing the coefficient values as $[h(0)]$ at $n = 0$, these may all be set to zero as an initial guess.

At any time n, the coefficients are updated as follows:

(a) From $[h(n)]$, the input vector $[\mathbf{x}(n)]$ and the desired response $\mathbf{f}(n)$ we calculate the error signal as

$$\mathbf{e}(n) = \mathbf{f}(n) - [\mathbf{x}(n)]'\,[h(n)]. \tag{9.108}$$

(b) The new estimate $[h(n+1)]$ is obtained from (9.107).
(c) The time index n is incremented by one, and the process is iterated until we reach a steady state.

It is to be noted, however, that the LMS algorithm provides only an approximation to the optimum Wiener solution. Further, it circumvents the difficulty inherent in estimating the mean-square gradient vector in (9.99), basically by replacing ensemble averages by time averages. Moreover, the LMS algorithm has been shown capable of operation in a slowly varying non-stationary environment.

9.7 Adaptive IIR Filtering and System Modelling

It has been observed that the problems of linear estimation and system modelling have, basically, the same solution. In our discussion of linear estimation we have concentrated on the use of FIR filters with the associated adaptive algorithms. We now consider the same problem using IIR filters, but the formulation is changed into that of the modelling of an unknown linear system.

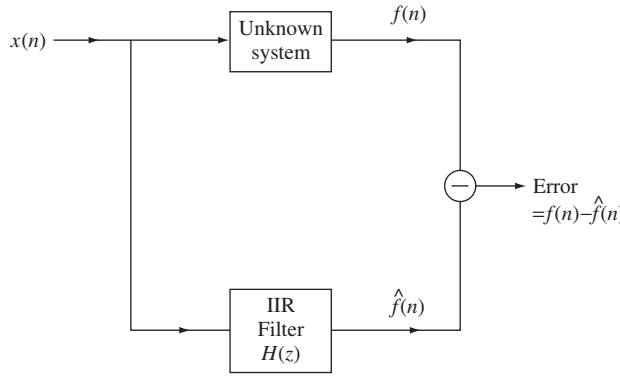


Figure 9.8 System modelling by an IIR filter

With reference to Figure 9.8, which is the discrete counterpart to Figure 9.2, given an unknown system with an input signal $x(n)$, we take a set of measurements of the output $f(n)$. It is now required to design an IIR filter model of the unknown system with $f(n)$ as the *desired response*. The filter output is $\hat{f}(n)$ which differs from the desired output by an error signal $e(n)$.

A possible IIR model has a transfer function of the form

$$H(z) = \frac{1}{1 + \sum_{r=1}^{N} b_r z^{-r}} \tag{9.109}$$

corresponding to a difference equation of the form

$$\hat{f}(n) = x(n) - \sum_{r=1}^{N} b_r \hat{f}(n-r). \tag{9.110}$$

The above expression describes a system in which the output sample at time n is *regressed* on the N past samples. Therefore, (9.109) describes what is called an *autoregressive* (AR) model.

It is also possible to model the system by a more general IIR filter with a minimum-phase transfer function

$$H(z) = \frac{P(z^{-1})}{Q(z^{-1})} \tag{9.111}$$

where $P(z^{-1})$ has all its zeros inside the unit circle in the z-plane. For such a function, it is possible to write

$$1/P(z^{-1}) \approx 1 + \sum_{r=1}^{M} c_r z^{-r} \tag{9.112}$$

so that

$$H(z) \approx \frac{1}{\left(1 + \sum_{r=1}^{N} b_r z^{-r}\right)\left(1 + \sum_{r=1}^{M} c_r z^{-r}\right)} \tag{9.113}$$

and we have a situation that can be treated as the case described by (9.109). Now, the most general IIR transfer function is of the form

$$H(z) = \frac{P_M(z^{-1})}{Q_N(z^{-1})} = \frac{\sum_{r=1}^{M} a_r z^{-r}}{1 + \sum_{r=1}^{N} b_r z^{-r}} \tag{9.114}$$

and is not restricted to be of the minimum-phase type. It contains the factor $[1/Q_N(z^{-1})]$ which corresponds to an autoregressive (AR) model, as well as the factor $P_M(z^{-1})$ which corresponds to a *moving average* (MA) model. Therefore, the model described by (9.114) is said to be *autoregressive with moving average* or an ARMA model. This model can be used in conjunction with the gradient algorithm for adaptive filtering or system modelling. The mean-square error in the case of system modelling is given by

$$\varepsilon(n) = \frac{1}{N} \sum_{n=0}^{N-1} (f(n) - \hat{f}(n))^2 \tag{9.115}$$

which can be minimized using the gradient algorithm. The error gradients are obtained as

$$\frac{\partial \varepsilon}{\partial a_r} = \frac{2}{N} \sum_{n=0}^{N-1} (f(n) - \hat{f}(n)) \frac{\partial \hat{f}(n)}{\partial q_r} \quad 0 \le r \le M$$

$$\frac{\partial \varepsilon}{\partial b_r} = \frac{2}{N} \sum_{n=0}^{N-1} (f(n) - \hat{f}(n)) \frac{\partial \hat{f}(n)}{\partial b_r} \quad 1 \le r \le N \tag{9.116}$$

where

$$\frac{\partial \hat{f}(n)}{\partial a_r} = x(n-r) - \sum_{r=1}^{N} b_r \frac{\partial \hat{f}(n-r)}{\partial a_r}$$

$$\frac{\partial \hat{f}(n)}{\partial b_r} = -\hat{f}(n-r) - \sum_{r=1}^{N} b_r \frac{\partial \hat{f}(n-r)}{\partial b_r}. \tag{9.117}$$

For implementation of the above two expressions we note that

$$\hat{f}(n) = \frac{1}{2\pi \mathrm{j}} \oint_c z^{n-1} H(z) X(z)\, \mathrm{d}z \tag{9.118}$$

so that

$$\frac{\partial \hat{f}(n)}{\partial a_r} = \frac{1}{2\pi \mathrm{j}} \oint_c z^{n-1} z^{-r} \frac{X(z)}{Q(z)}\, \mathrm{d}z$$

$$\frac{\partial \hat{f}(n)}{\partial b_r} = \frac{-1}{2\pi \mathrm{j}} \oint_c z^{n-1} z^{-r} \frac{H(z)X(z)}{Q(z)}\, \mathrm{d}z. \tag{9.119}$$

Hence, the gradient functions are obtained by applying $x(n)$ and $\hat{f}(n)$ to a filter with transfer function $1/Q_{\mathrm{N}}(z^{-1})$.

The steps necessary for the application of the gradient algorithm are generally similar to those discussed in the case of FIR filters. Thus the filter coefficients are initialized, then they are changed in a direction opposite to that of the error gradient. Iteration of the process, by making successive corrections to the gradient, leads to the coefficient values which minimize the mean-square error between the desired system output and its model. However, there are other algorithms which are particularly suitable for adaptive IIR filters. These, together with the further ramifications of the technique, may be found in the references [21].

9.8 An Application of Adaptive Filters: Echo Cancellers for Satellite Transmission of Speech Signals

The applications of adaptive filtering span a wide range of areas, particularly in telecommunications. System modelling discussed earlier may be considered a broad category of the applications of adaptive filters and the associated algorithms. As a further application we now discuss *echo cancellation* on telephone lines.

Consider Figure 9.9, in which the transmission links of telephone lines between two speakers employ both *four-wire* and *two-wire* networks. At each end the connection between the two types of network is effected by means of a *hybrid transformer* for the conversion between four-wire and two-wire transmission. At this hybrid, mismatching and imperfections occur, resulting in an *echo signal* which is reflected back to the speaker.

For long distance transmission, for example using satellites, this is an objectionable feature from the subjective viewpoint. Specifically, the high altitude of the satellite results in a delay of 270 ms in each four-wire path. The ideal situation is that the speech from speaker A should follow path A to hybrid B which converts it to a two-wire connection. However due to the mismatch at the hybrid, some of the speech energy is returned to be heard by the speaker himself. This takes the form of an echo occurring 540 ms later than the moment at which speaker A starts to talk. A possible solution to this problem is to employ *echo cancellers* as shown in Figure 9.10. The function of each canceller is to estimate the echo and subtract it from the return signal. Hence an adaptive FIR filter with the associated algorithm can be used for this purpose. Naturally, A/D and D/A converters are employed if the transmission path is to operate in the digital model.

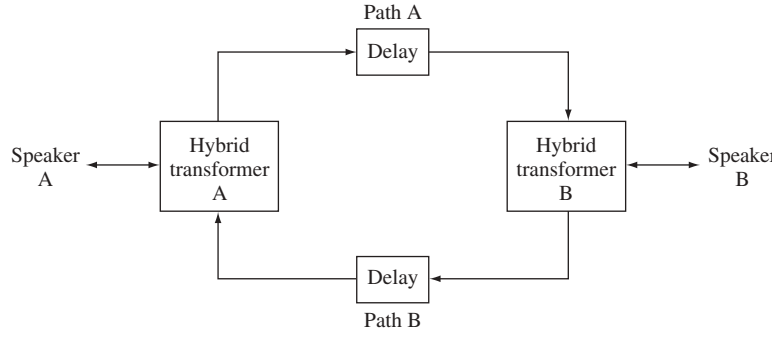


Figure 9.9 Satellite transmission of telephone conversation

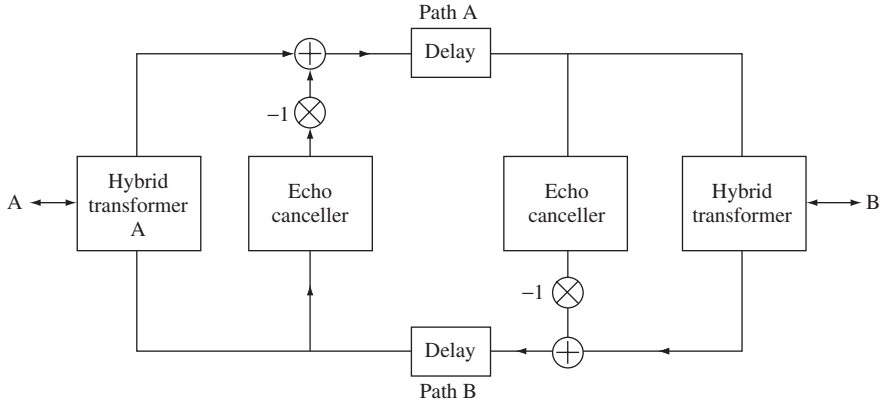


Figure 9.10 Employing echo cancellers in the network of Figure 9.9

9.9 Conclusion

This chapter dealt with the processing of stochastic signals by means of linear systems. Emphasis has been laid on digital methods which are now almost exclusively used for adaptive filtering, system modelling, linear estimation and prediction. The chapter concluded with an application of adaptive filters in a satellite system used for the transmission of speech signals.

Part III

Analog MOS Integrated Circuits for Signal Processing

'In science, the man of real genius is the man who invents a new method. The notable discoveries are often made by his successors, who can apply the method with fresh vigour, unimpaired by the previous labour of perfecting it; but the mental calibre of the thought required for their work, however brilliant is not as great as that required by the first inventor of the method.'

Bertrand Russell
'The Place of Science in a Liberal Education'

10

MOS Transistor Operation and Integrated Circuit Fabrication

10.1 Introduction

This chapter gives a review of the operation of the metal oxide semiconductor (MOS) transistor [22–24] together with the technological processes employed in the fabrication of MOS transistors and integrated circuits. Knowledge of device physics is assumed, including the operation of the pn junction and basic transistor operation. The chapter begins with a review of the MOS transistor operation equations and introduces the complementary (CMOS) circuits. A description is next given of the fabrication of MOS devices which is necessary for understanding the performance and limitations of MOS integrated circuits used in signal processing. Layout rules and area requirements of integrated circuits are described. The chapter concludes by a discussion of the subject of noise in MOSFETs.

10.2 The MOS Transistor

Figure 10.1 shows the physical structure of an n- channel enhancement-type MOSFET. The device is fabricated on a p-type substrate consisting of a single-crystal silicon wafer. The n^+ regions are heavily-doped n-type silicon, constituting the source and drain regions. A thin silicon dioxide (SiO_2) layer is grown on the substrate, extending over the area between the source and drain. For the electrodes, metal can be used as contacts to the gate, source, drain and substrate. The gate electrode can also be made from poly-crystalline silicon (polysilicon) in the process of silicon-gate technology. The oxide layer results in the current in the gate terminal being very small ($\approx 10^{-15}$ A.)

The normal operation of the MOS transistor requires the pn junctions formed between the substrate and each of the drain and source to be reverse-biased. Generally the drain voltage is higher than that of the source, and the two pn junctions referred to above would be reverse biased if the substrate is connected to the source. We shall assume this to be the case in the following analysis, then examine the effect of the substrate shortly afterwards.

Signal Processing and Integrated Circuits, First Edition. Hussein Baher.
© 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.

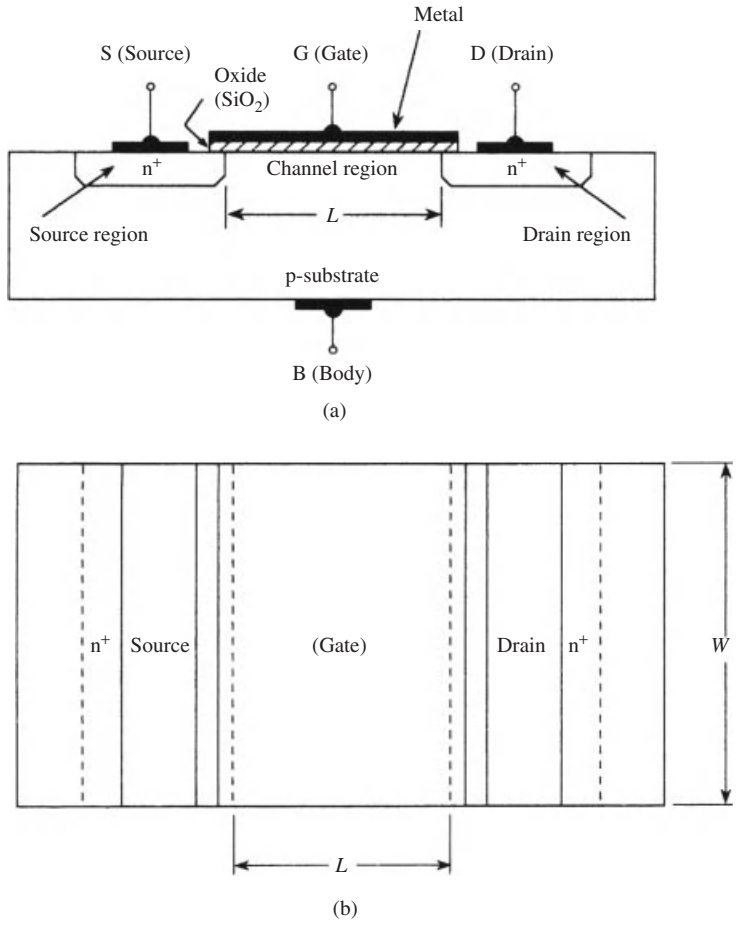


Figure 10.1 The enhancement-type MOSFET: (a) cross-section, (b) top view

10.2.1 Operation

1. For $v_{GS} = 0$ the drain-substrate and source-substrate pn-junctions form two diodes back-to-back. Thus no current flows between the source and drain, since the path between these regions has a resistance of the order of $10^{12}\ \Omega$.
2. With $v_{GS} > 0$, the free holes in the substrate region near the gate are repelled into the substrate, thus creating a depletion region, as illustrated in Figure 10.2, which contains bound negative charges. In addition, the positive v_{GS} attracts electrons from the source and drain n^+ regions upwards towards the gate. These conditions result in the formation of a negatively charged induced channel connecting the source and drain regions. Consequently, if a voltage is applied between the drain and source, a current flows between the two regions via the induced channel. This n-type channel results from the *inversion* of the p-type region near the gate into an n-type region. Hence, the channel is called an *inversion layer* and the resulting structure is called an n-channel MOSFET, or an NMOS transistor.
   Now, there is a critical value of v_{GS} at which a conducting channel is formed. This value is called the *threshold voltage* V_t which is determined during the fabrication process.

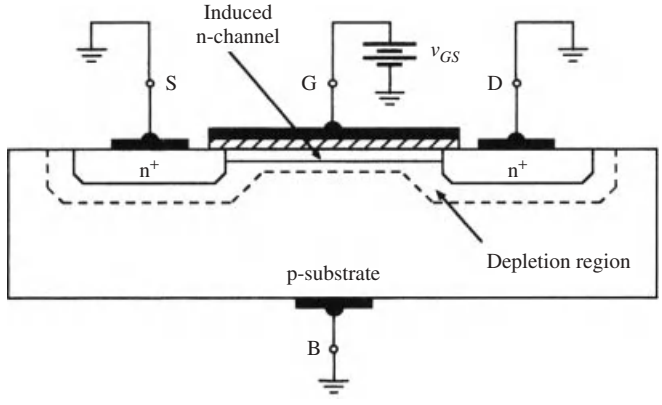


Figure 10.2 Enhancement-type NMOS device showing the induced n-channel as a result of applying a positive $v_{GS} > V_t$

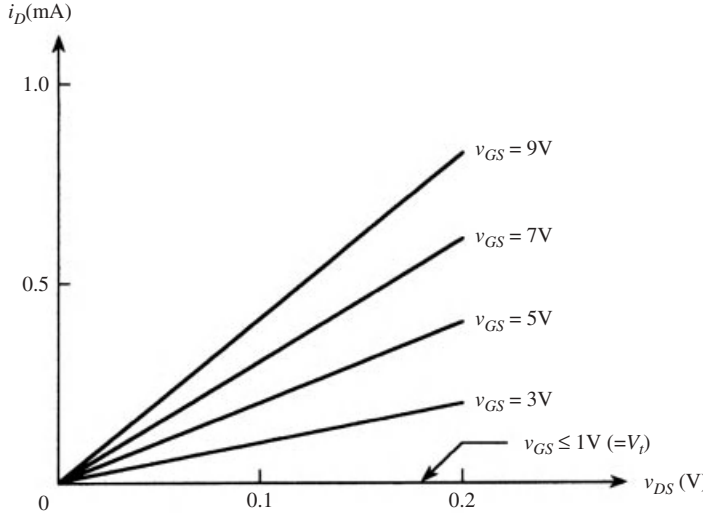


Figure 10.3 The $i_D - v_{DS}$ characteristic of a MOSFET with $V_t = 1\,\text{V}$ for different v_{GS}

3. With $v_{GS} > V_t$ the n-channel is formed and applying a voltage v_{DS} results in the $i_D - v_{DS}$ characteristic shown in Figure 10.3 for values of v_{DS} in the range 0–0.2 V. Thus, the induced channel is enhanced by taking $v_{GS} > V_t$ and this gives the device its name of *enhancement-type* MOSFET.
4. Again, with $v_{GS} > V_t, v_{DS}$ is increased further. Clearly, there is a voltage gradient across the n-channel since the voltage between the source and drain varies from v_{GS} at the source to $v_{GS} - v_{DS}$ at the drain. Thus, the channel depth varies in a tapered shape as shown in Figure 10.4. At a value $v_{DS} = (v_{GS} - V_t)$, the channel depth at the drain becomes zero and the channel is said to be *pinched-off*. Then, the channel shape is unaffected by a further increase in v_{DS}. Beyond pinch-off the device enters saturation resulting in the $i_D - v_{DS}$ characteristic shown in Figure 10.5.

Figure 10.6 shows the circuit symbols for the NMOS device.

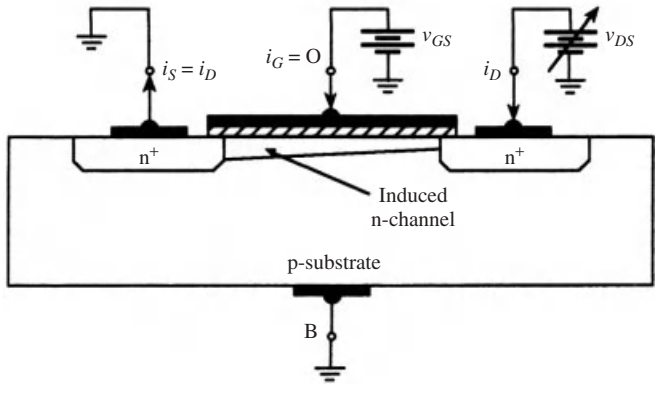


Figure 10.4 Effect of increasing v_{DS} on the shape of the induced channel

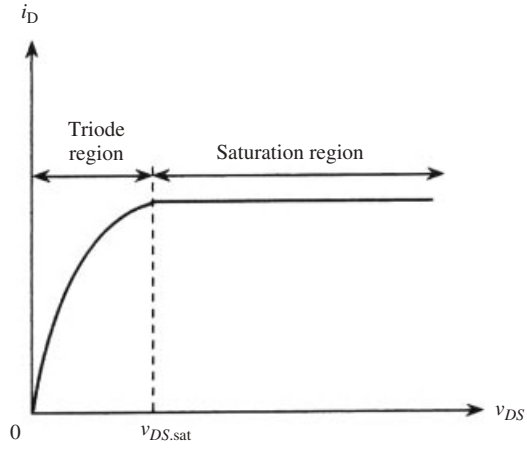


Figure 10.5 Typical characteristic for enhancement NMOS transistor with $v_{GS} > V_t$

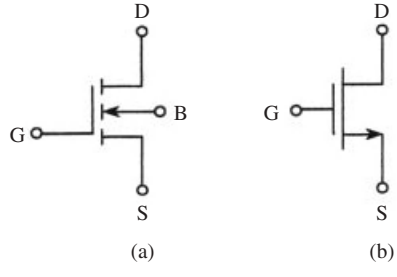


Figure 10.6 Symbols of the NMOSFET: (a) showing the substrate, (b) simplified symbol when B is connected to S

We now consider quantitatively the various regions of operation of the MOS transistor. The device is cut off when

$$v_{GS} < V_t \tag{10.1}$$

where V_t is the threshold voltage of the device.

To operate in the triode region we must have

$$v_{GS} \geq V_t \tag{10.2}$$

and keep v_{DS} small enough for the channel to remain continuous, that is

$$v_{GD} > V_t \tag{10.3}$$

or noting that

$$v_{GD} = v_{GS} + v_{SD} = v_{GS} - v_{DS} \tag{10.4}$$

we have

$$v_{DS} < v_{GS} - V_t \tag{10.5}$$

In this region of operation, an approximate relation is given by

$$i_D = K[2(v_{GS} - V_t)v_{DS} - {v_{DS}}^2] \tag{10.6}$$

with

$$K = \frac{1}{2}\mu_n C_{ox}\left(\frac{W}{L}\right) \tag{10.7}$$

where μ_n is the electron mobility in the induced channel, C_{ox} is the oxide capacitance per unit area of the gate to body capacitor

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}} \tag{10.8}$$

with ε_{ox} being the permittivity of the oxide and t_{ox} is its thickness. L is the length of the channel and W is its width. Usually the quantity $\frac{1}{2}\mu_n C_{ox}$ is determined by the fabrication process. Thus, K is determined by the *aspect ratio* (W/L) of the device. The transconductance parameter K' is defined as

$$K' = \mu_n C_{ox} \tag{10.9}$$

so that

$$K = K'(W/L)/2 \tag{10.10}$$

Near the origin, for very small v_{DS}, we can neglect v_{DS}^2 and (10.6) becomes

$$i_D \cong 2K(v_{GS} - V_t)v_{DS} \tag{10.11}$$

This allows the use of the transistor as a voltage-controlled linear resistor of value

$$r_{DS} = v_{DS}/i_D = 1/[2K(v_{GS} - V_t)] \tag{10.12}$$

which is controlled by v_{GS}.

In the saturation region we must have

$$v_{GS} \geq V_t \tag{10.13}$$

and

$$v_{GD} \leq V_t \tag{10.14}$$

or

$$v_{DS} \geq v_{GS} - V_t \tag{10.15}$$

The three regions of operation are shown in Figure 10.7 for different v_{GS} values.

From (10.5) and (10.15), the boundary between the triode and saturation regions is defined by

$$v_{DS} = v_{GS} - V_t \tag{10.16}$$

which upon substitution in (10.6) gives the saturation value

$$i_D = K(v_{GS} - V_t)^2 \tag{10.17}$$

which is independent of v_{DS}. This relationship is shown in Figure 10.8 and shows that in saturation the device behaves as an ideal voltage-controlled current source whose strength is determined by v_{GS} in a non-linear fashion as shown in Figure 10.8(c). The large signal model of the n-channel device in saturation is shown in Figure 10.8(b).

10.2.2 The Transconductance

For a MOSFET in saturation we can use (10.17) with v_{GS} having the dc component V_{GS} (corresponding to the bias point) and the signal component v_{gs} to write

$$\begin{aligned} i_D &= K(V_{GS} + v_{gs} - V_t)^2 \\ &= K(V_{GS} - V_t)^2 + 2K(V_{GS} - V_t)v_{gs} + Kv_{gs}^2 \end{aligned} \tag{10.18}$$

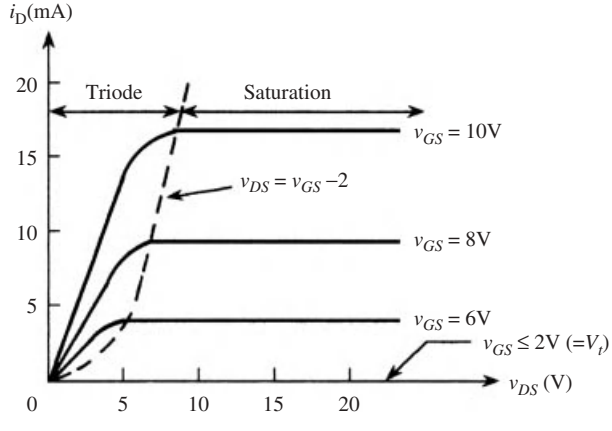


Figure 10.7 Typical $i_D - v_{DS}$ characteristics of an NMOSFET with varying v_{GS}, for which $V_t = 2\,\text{V}$

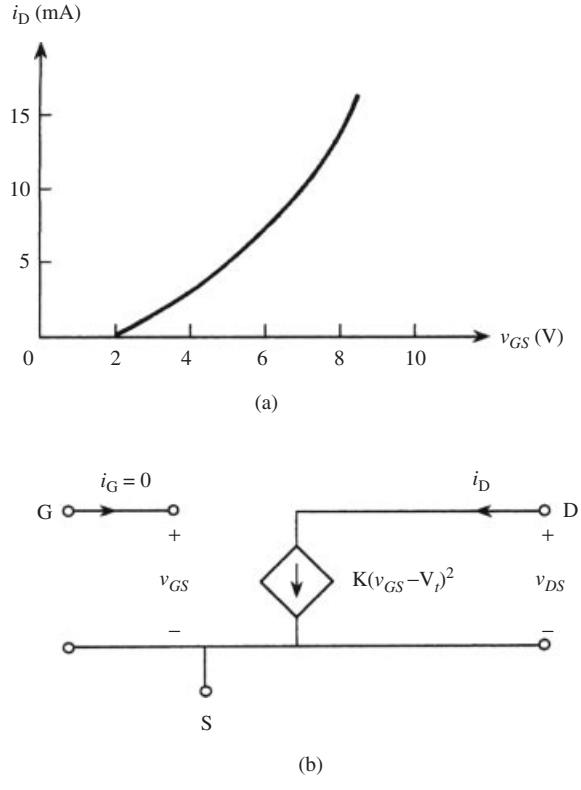


Figure 10.8 The enhancement NMOSFET in saturation: (a) $i_D - v_{DS}$ characteristic, (b) large signal equivalent circuit

and for small signal operation the quadratic term can be neglected. The first term is a dc or quiescent current. Therefore, the signal current is

$$i_d = 2K(V_{GS} - V_t)v_{gs} \tag{10.19}$$

The transconductance of a MOSFET is given by

$$\begin{aligned} g_m &= \frac{i_d}{v_{gs}} \\ &= 2K(V_{GS} - V_t) \end{aligned} \tag{10.20}$$

so that substitution for K from (10.7) gives

$$g_m = (\mu_n C_{ox})(W/L)(V_{GS} - V_t) \tag{10.21}$$

More specifically,

$$g_m = \left.\frac{\partial i_D}{\partial v_{GS}}\right|_{v_{GS}=V_{GS}} \tag{10.22}$$

For the dc value I_D we also have

$$I_D = K(V_{GS} - V_t)^2 \tag{10.23}$$

which upon use in (10.20) and (10.21) gives the alternative expression for the transconductance as

$$\begin{aligned} g_m &= \sqrt{2\mu_n C_{ox}(W/L)I_D} \\ &= \sqrt{2K'(W/L)I_D} \\ &= 2\sqrt{KI_D} \end{aligned} \tag{10.24}$$

10.2.3 Channel Length Modulation

The idealized behaviour of the MOSFET in saturation shown in Figure 10.7 assumes that the channel shape is not affected by an increase of the drain to source voltage beyond pinch off. This leads to the horizontal portion of the characteristic which implies that the output resistance in saturation

$$r_{outsat} = \frac{\partial v_{DS}}{\partial i_D} \tag{10.25}$$

is infinite. However, in practice, increasing v_{DS} beyond $v_{DS,sat}$ has an effect on the channel length, which actually decreases. This is called the *channel-length modulation* phenomenon. But from (10.7), K is inversely proportional to the channel length and therefore, i_D increases with v_{DS} resulting in the modified set of curves shown in Figure 10.9. The linear dependence of i_D on v_{DS} is taken into account by multiplying (10.17) by the factor $(1 + \lambda v_{DS})$ to give

$$i_D = K(v_{GS} - V_t)^2(1 + \lambda v_{DS}) \tag{10.26}$$

where, by reference to Figure 10.9

$$\lambda = 1/V_A \tag{10.27}$$

which is a device parameter called the *channel length modulation parameter* and V_A is the early voltage.

As a consequence of the relationship in (10.26) the output resistance of the device in saturation is now finite and is given by

$$\begin{aligned} r_o &= \left[\frac{\partial i_D}{\partial v_{DS}}\right]^{-1}, v_{GS} = \text{constant} \\ &= 1/\lambda K(V_{GS} - V_t)^2 \end{aligned} \tag{10.28}$$

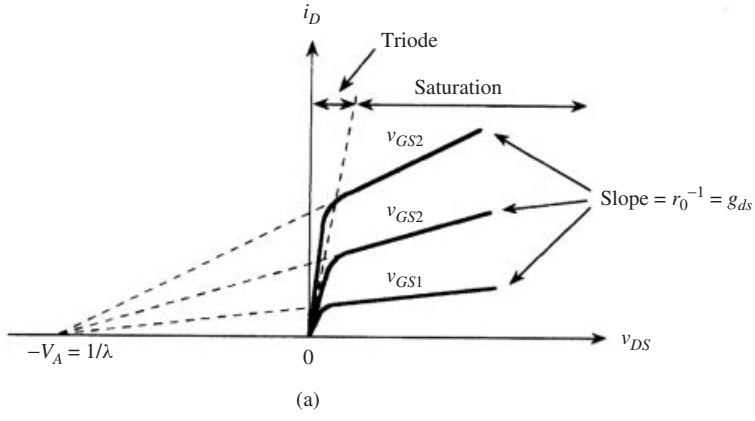
or

$$\begin{aligned} r_o &\cong 1/\lambda I_D \\ &\cong V_A/I_D \end{aligned} \tag{10.29}$$

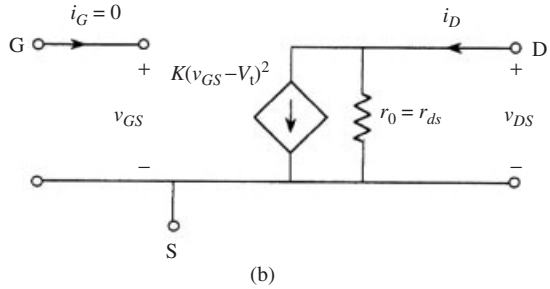
with I_D as the drain current for a specific V_{GS}. An alternative notation is

$$\begin{aligned} g_{ds} = g_o &= 1/r_o \\ &= I_D/(1 + \lambda V_{DS}) \\ &\cong \lambda I_D \end{aligned} \tag{10.30}$$

which is also called the *small signal channel conductance*.



(a)



(b)

Figure 10.9 Effect of finite output resistance of the MOSFET in saturation: (a) characteristic, (b) large signal equivalent circuit

10.2.4 PMOS Transistors and CMOS Circuits

A MOS transistor can be fabricated with an n-type substrate and p^+ type drain and source regions. The created channel becomes of the p-type. In this case the operation is similar to the NMOS case, but V_t and V_A are negative. Figure 10.10 shows the symbols of the PMOS transistor. The most important use of this PMOS transistor lies in employing it together with an NMOS device in a complementary manner, resulting in a CMOS circuit. Figure 10.11 shows such an IC.

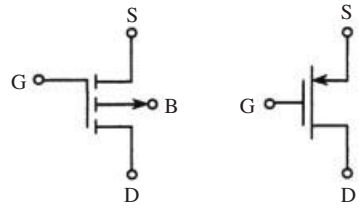


Figure 10.10 Symbols of PMOSFET: (a) showing the substrate, (b) simplified symbol when B is connected to S

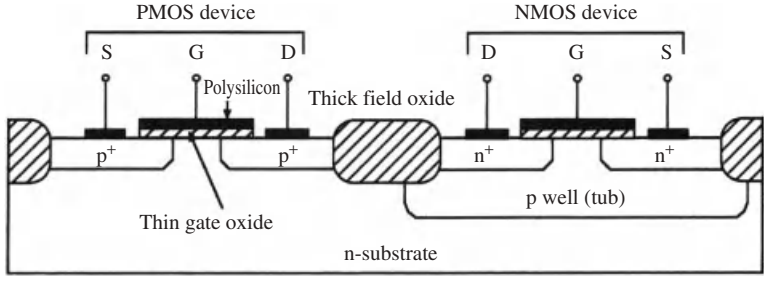


Figure 10.11 Cross-section of a CMOS circuit

10.2.5 The Depletion-type MOSFET

The structure of this device is identical to that of the enhancement type with one important difference: it has a physically implanted channel, and there is no need to create one. In this case a positive v_{GS} *depletes* the channel from its charge carriers; hence the name. The circuit symbol of the depletion type n-channel MOSFET is shown in Figure 10.12.

10.3 Integrated Circuit Fabrication

All the non-ideal effects in integrated circuits have their origins in the fabrication process of the devices. Therefore, a reasonable understanding of this process is quite helpful in predicting and possibly reducing these non-ideal effects. In this section, an outline of the fabrication process of MOS integrated circuits is given.

The basic processing steps in the production of MOS transistors, and subsequently larger circuits composed of a number of these transistors, are the following:

(a) Diffusion and ion implantation,
(b) Oxidation,
(c) Photolithography,
(d) Chemical vapour deposition,
(e) Metallization.

These steps are applied to a silicon crystal which must be produced and prepared first. This is outlined next before the processing itself is discussed.

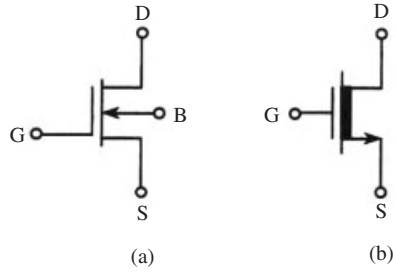


Figure 10.12 Symbols of the depletion- type NMOSFET: (a) showing the substrate, (b) simplified symbol when B is connected to S

10.3.1 Wafer Preparation

The starting point in silicon processing according to the above steps is a single crystal silicon wafer of appropriate conductivity and doping. The starting material for silicon growth is very pure polycrytstalline silicon referred to as *semiconductor-grade* silicon corresponding to an impurity concentration of less than one part per billion silicon atoms. The number of silicon atoms is usually 5×10^{22} atoms/cm^3. The resistivity corresponding to the number of impurities, if they are of the acceptor type, is about 300 Ωcm. Actually, polycrystalline silicon with impurity concentrations down to 0.1 parts per billion is also available.

The semiconductor-grade silicon is used to grow single crystals in the form of ingots having a diameter of about 10–15 cm and length of the order of 1 m. A crystallographic orientation flat is also ground along the length of the ingot. Then the extreme top and bottom of the ingot are cut off and the ingot surface is ground to produce a constant and precise diameter. The ingot is next sliced producing circular slices or wafers about 0.5–1.0 mm thick. Then the wafer is subjected to polishing and cleaning processes in order to remove the silicon damaged from the slicing operation and to produce a highly planar flat surface which is necessary for fine line device geometries and for improving the parallelism of the two surfaces in preparation for photolithography. Usually, one side of the wafer is given a mirror-smooth finish while the other (backside) is only treated for an acceptable degree of flatness.

10.3.2 Diffusion and Ion Implantation

Diffusion is a process by means of which dopants are introduced into the surface of the silicon wafer. The most commonly used difusants are *substitutional dopants*. These have atoms which are too large to fit in the interstices between the silicon atoms and therefore the only way they can enter the silicon crystal structure is by replacing silicon atoms. Those of the donor type are phosphorus, arsenic and antimony, while boron is practically the only acceptor type dopant used.

For the process of diffusion to occur, the presence of vacancies in non-ideal crystals is necessary. The vacancy density can be increased by raising the temperature. The vacancies can be generated at the crystal surface or interior to it. The diffusion process relies on the flow of atoms caused by a concentration gradient. The particle flow or flux F is proportional to the concentration gradient, that is

$$F = -D\frac{\partial N}{\partial y} \tag{10.31}$$

where F is the rate of flow of atoms per unit area (per second) and the derivative indicates the concentration gradient. D is the diffusion coefficient. Also the diffusion rate obeys the continuity equation

$$\frac{\partial N}{\partial t} = -\frac{\partial F}{\partial y} \tag{10.32}$$

Combining (10.31) with (10.32) we obtain

$$\frac{\partial N(y,t)}{\partial t} = D\frac{\partial^2 N(y.t)}{\partial y^2} \tag{10.33}$$

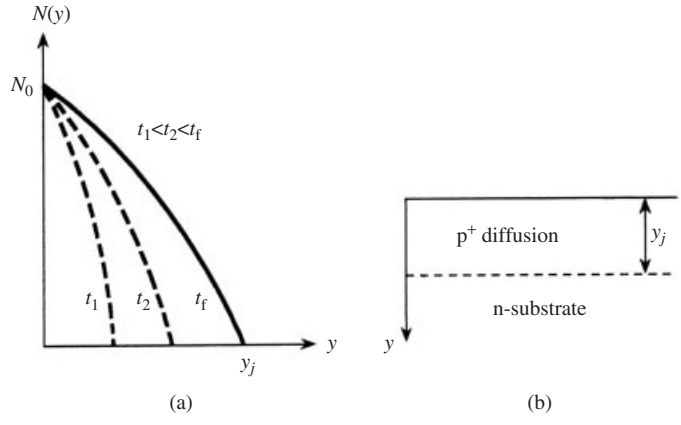


Figure 10.13 Deposition diffusion: (a) profile, (b) junction depth

The solution of the above equation, subject to the boundary conditions, gives the profile of the distribution of the diffusant particles $N(y,t)$. This is considered below, according to the type of diffusion employed.

There are two types of diffusion. The first occurs under constant surface concentration or infinite source conditions and is referred to as deposition diffusion. The diffusion concentration at the surface of the silicon crystal ($y = 0$) is assumed constant so that $N(0,t) = N_0$, a constant. The other boundary condition is that the concentration should tend to zero as $y \to \infty$. The solution to the diffusion equation under these conditions is given by

$$N(y,t) = N_0 \operatorname{erfc}\left(\frac{y}{2\sqrt{Dt}}\right) \tag{10.34}$$

where erfc is the complementary error function. Figure 10.13 shows an example of the diffusion profile of p-type boron diffusion into an n-type (phosphorus-doped) substrate.

Deposition diffusion of the type described above is usually followed by a second diffusion process called *drive-in diffusion* in which the external source is removed. There are no more external dopants entering the silicon, but those already there move further inside and are redistributed. The impurity profile is obtained in this case by applying the boundary condition that the total diffusant density (atoms per unit area) remains constant. Alternatively, this implies that the net flow of diffusant atoms in or out of the silicon is zero at the surface (y = 0). Thus, solution of (10.33) with the boundary condition that (10.33) is zero at $y = 0$ gives

$$N(y,t) = \frac{Q}{\sqrt{Dt}} \exp\left(-\frac{y^2}{4Dt}\right) \tag{10.35}$$

which is a Gaussian distribution with typical behaviour as shown in Figure 10.14.

10.3.2.1 Sheet Resistance

Two main parameters are used to characterize diffusion layers. These are the junction depth y_j and the sheet resistance. The former has been described, while the latter is given

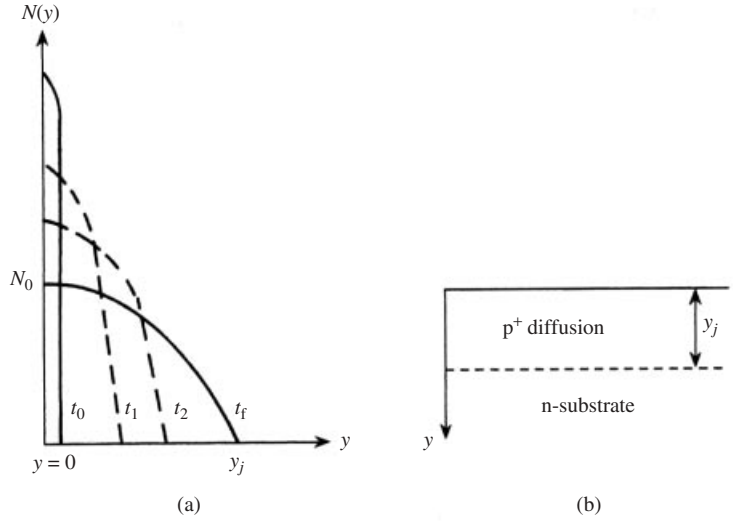


Figure 10.14 Drive-in diffusion: (a) profile, (b) junction depth

by $R = \rho\ell/A = \rho\ell/wt$ where ρ is the resistivity of the layer, ℓ is its length, t its thickness and w its width. In the case of a square shape, $\ell = w$ and we have $R_s = \rho/t$ which is independent of the lateral dimension. This is referred to as the sheet resistance expressed in Ω/square. Thus the resistance of a layer is expressed as

$$R = R_s \ell/w \tag{10.36}$$

10.3.2.2 Ion Implantation

This can be used as an alternative to deposition diffusion to produce a shallow region of dopant atoms. It has the advantage of more precise control over the sheet resistance and therefore more commonly used than deposition diffusion. In this technique, silicon wafers are placed in vacuum and scanned by a beam of high energy dopant atoms which have been accelerated by high voltages. The atoms impinge on the surface of the wafer and penetrate into a small layer. The depth of penetration is called the *projected range*.

The distribution of the implanted ions is Gaussian, and a typical profile is shown in Figure 10.15. An ion implantation pattern on a wafer is defined by a mask made up of SiO_2 or Si_3N_4 and photoresist for low energy ions (<100 kV). Heavy metals such a gold or titanium deposited on SiO_2 are used as masks with higher energy ions.

10.3.3 Oxidation

A native oxide layer about 20–30 Å thick would naturally form on a silicon surface if exposed to air. But this will inhibit further growth of oxide to a desired thickness, which is of the order of 5000–10 000 Å. This is used as either a diffusion or ion implantation mask or as a passivation layer for the protection of the devices. For this purpose, thermal oxidation is used, which consists in heating the wafers to about 1000 °C and simultaneously exposing them to a gas containing oxygen or water vapour. The diffusion rate of the

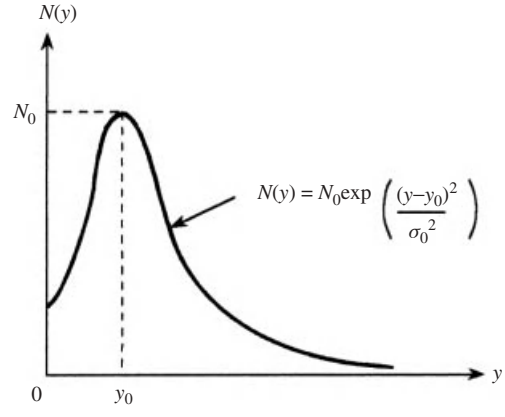


Figure 10.15 Ion implantation impurity profile, y_o = range with σ_o its standard deviation and N_o the peak concentration

gas increases with temperature so that oxidation will continue until the desired thickness is reached. In this thermal oxide growth, some of the silicon will be consumed.

10.3.3.1 Oxide Masking

Oxide layers are used for producing patterns on a wafer by creating openings or windows through which ion implantation or diffusion may be applied while protecting the other parts of the wafer from the process. Thus, the dopant pattern coincides with that of the windows in the oxide layers. The technique is essential for the production of microscopic size devices. Figure 10.16 shows a pn junction produced by diffusion through an oxide window. Diffusion will occur vertically and also horizontally as shown. The junction intersects the silicon surface well below the protective thermally grown oxide layer. Thus, this layer protects the junction against the surrounding effects and is called a *passivation junction*. The production of the oxide mask requires the use of photolithography which is discussed next.

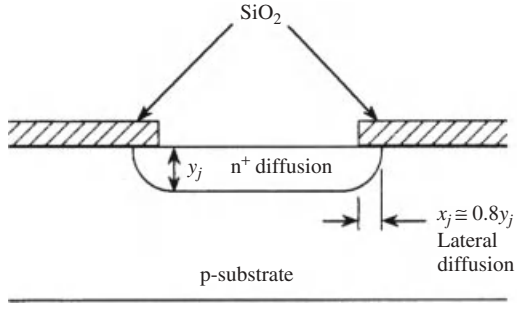


Figure 10.16 The use of an oxide layer for diffusion masking

10.3.4 Photolithography

This is the process by means of which a pattern is created on the wafer defining the locations of the devices and circuits to be produced. Device dimensions of the order of 1.5 μm are possible with conventional ultraviolet (UV) light exposure techniques. Using electron beam or X-ray methods, submicron dimensions are achievable. Photolithography involves a number of steps which are illustrated in Figure 10.17. First, a photoresist which is a light-sensitive liquid, is applied to the surface of the wafer. A thin layer is produced by placing a drop of photoresist at the centre of the wafer then spinning it at high speed. The wafer is then prebaked by heating to expel solvents from the photoresist and harden it to a semi-solid state. The coated wafer is then aligned with a photomask.

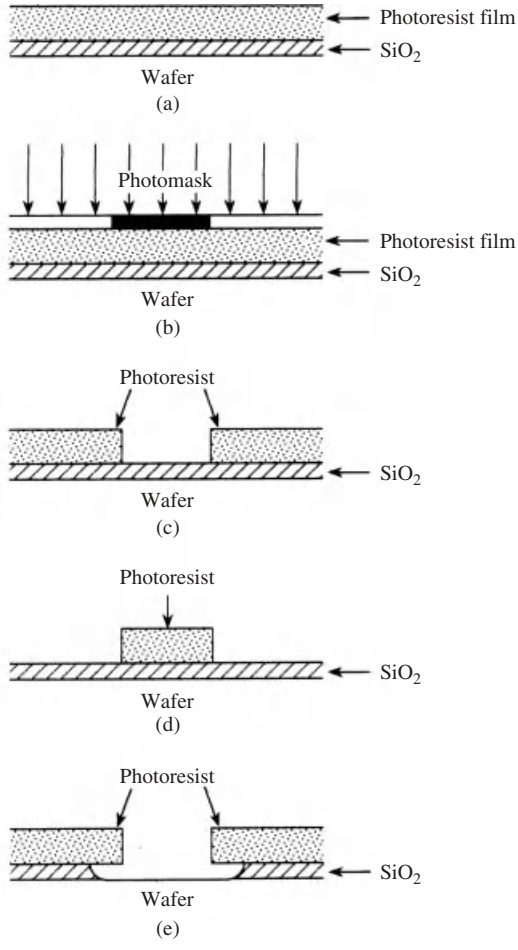


Figure 10.17 Photolithography: (a) film spinning, (b) exposure, (c) negative photoresist development, (d) positive photoresist development, (e) oxide etching

This is a glass plate which has a photographic emulsion or thin metal pattern on one side. The pattern consists of clear and opaque areas. The photomask is produced by computer-aided design tools which define the pattern by a computer connected to the mask-making machine. This pattern is produced such that the wafer will be divided into a number of chips each containing an integrated circuit made up of a large number of transistors. Then follows the process of the development of the photoresist by UV radiation. There are two types of photoresist: negative and positive. If the former is used, the areas exposed to UV light becomes polymerized, hardened and virtually insoluble in the developer solution. This will develop into a copy of the mask patterns as shown in Figure 10.17(c) where the clear areas on the mask coincide with the areas where the photresist remains on the wafer. If positive photoresist is used, the opposite situation results in which exposure to UV radiation depolymerizes the photoresist making the exposed areas soluble in the developer solution while the unexposed areas are rendered virtually insoluble. In this case, the clear areas on the photomask will coincide with the areas where the photoresist was removed. After development, the wafer is post-baked to harden the photoresist for better adhesion to the wafer and increase its resistance to acids used for etching the oxide. This oxide etching process is next and can be either wet or dry. In the former method, the wafer is exposed to an etching solution to remove the oxide layer in the areas where there is no photoresist as shown in Figure 10.17(e). The outcome is the pattern of windows coinciding with the corresponding photomask. In dry etching, gaseous plasma is used instead of the chemical solution. The plasma is produced by an RF (radio frequency) field and this type of etching results in the ability to produce smaller windows. The final step in photolithography is the removal of the photoresist. Positive photoresists can be easily removed by organic solvents such as acetone. Negative ones require a more elaborate process such as immersion in sulphuric acid together with mechanical scrubbing.

It must be noted that the principles illustrated in Figure 10.17 would be the same if electron-beam or X-ray methods are used instead of UV light, except that the materials used are modified accordingly.

10.3.5 Chemical Vapour Deposition

Chemical vapour deposition (CVD) is the process by means of which thin films of material are deposited on a substrate. Examples of materials deposited in IC processing are silicon dioxide, silicon nitride, silicon epitaxial layers and silicon hetero-epitaxial films. The material to be deposited is in gaseous form and is placed in a reaction chamber with the substrate where the chemical reaction causes the atoms that are produced to be deposited on the substrate.

SiO_2 CVD can be achieved at relatively lower temperatures by comparison with the thermal oxide growth discussed earlier. The process is also much faster, being completed in a few minutes, but the resulting oxide film is generally of lower quality giving lower dielectric strength than a thermally grown one. The oxide, however, can be deposited on a wafer for post-metallization passivation without seriously affecting the existing condition, while providing a protective layer after all the processes, including metallization, have taken place. CVD oxide can also be used for isolating metallization levels.

Silicon nitride can also be deposited using CVD for device protection and passivation against penetration of certain contaminants against which SiO_2 is less effective. It can also be used as a diffusion or ion implantation mask.

10.3.6 Metallization

This is the final step in the wafer processing sequence and consists in producing a thin metal layer serving as the conducting medium linking the devices and circuits on a chip. It is also used to produce bonding pads around the periphery of the chip which are needed to bond the wire pads from an IC package to the chip. These bonding wires are usually made of gold. The pads are square in shape to allow for the flattened ends of the wires and for some placement errors.

The metallization process usually uses aluminium. First, the thin film is deposited on the wafer using vacuum evaporation. Then the required pattern of the metal is created using photolithography. Etching uses phosphoric acid or plasma. Alternatively, metallization can be accomplished using a lift-off process. A positive photoresist is deposited and patterned using standard photolithography, then the metal film is deposited on the remaining photoresist. Next the photoresist is dissolved and lifted off the wafer, taking those parts of the metal on its top with it. This process is capable of producing very fine width patterns even for a larger film thickness.

10.3.7 MOSFET Processing Steps

Based on the previous description of the fabrication process, we may now state the sequence of processing steps necessary to produce an n-channel aluminium-gate MOSFET of the type shown in Figure 10.1:

1. The starting material is p-type silicon.
2. A thermal oxide layer is grown on the silicon.
3. A first photolithography process produces windows for the source and drain diffusions.
4. n^+ phosphorous diffusion is used to produce the source and drain regions.
5. A second photolithography process is used to remove the oxide from the channel region between the source and drain.
6. A very thin oxide layer is grown over the channel region.
7. A third photolithography step produces the contact windows.
8. Metallization is employed using an aluminium thin film.
9. A fourth photolithography step produces metallization patterns for the gate electrode, as well as the source and drain contact areas.
10. Contact sintering is applied followed by backside metallization.

Self-aligned Gate Structures

As described in Section 10.2, the inversion layer producing the channel must cover the entire region between the source and drain. Hence, the gate electrode must extend throughout this region. To allow for possible mask errors, the gate electrode is designed to overlap the source and drain regions by small distances. This will create small *overlap capacitances* C_{gs} between the gate and source, and C_{gd} between the gate and drain. C_{gd} represents a feedback capacitance from drain to gate and will affect the frequency response of the transistor adding to the Miller effect.

To reduce the overlap capacitances a self-aligned gate structure is used as shown in Figure 10.18. This is produced using the same steps as for a p-channel device, except

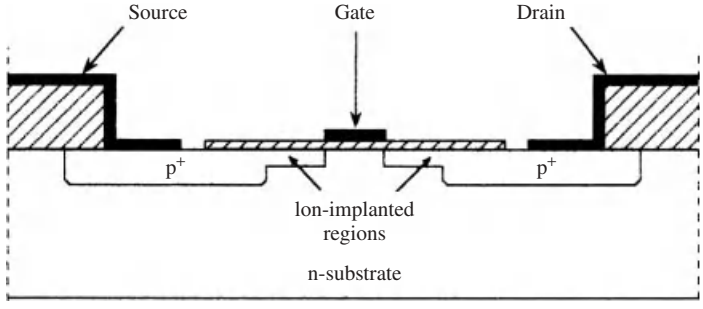


Figure 10.18 Self-aligned gate PMOSFET using ion implementation

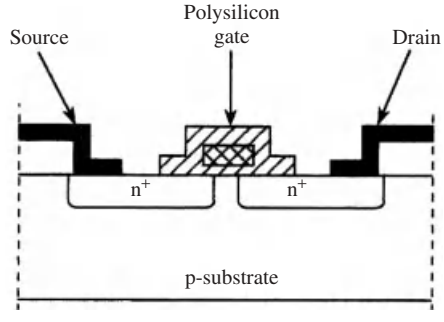


Figure 10.19 Polysilicon gate MOSFET

that the source and drain regions do not extend under the gate. Boron ion implantation is used to produce p-type extensions of the source and drain regions up to the edge of the gate. The high-energy boron ions penetrate the thin gate oxide but are blocked by the thick field oxide and by the gate. Therefore, the gate electrode itself also acts as an implantation mask and the source and drain regions effectively end just under the gate edges. This reduces the overlap capacitances.

An alternative self-aligned gate structure is shown in Figure 10.19, which uses polysilicon as the gate. This can withstand the high temperatures of the diffusion process and as a consequence the gate can also serve as a diffusion mask. Naturally, there will be lateral diffusion under the gate, but the overlap capacitances are still much smaller than those created in the conventional structure.

10.4 Layout and Area Considerations for IC MOSFETs

In integrated circuits, minimum clearance values are required between the various diffused regions, contact windows and metallized contact to allow for mask registration errors and the minimum line resolution determined by the photolithographic process. The minimum dimension resolution is specified as λ (not to be confused with the channel modulation parameter). The minimum clearances and dimensions are usually given as multiples of this quantity resulting in a set of dimensions called the *design rules*.

MOSFETs produced on the IC have two important attributes which allow high density of devices to be produced on the same chip. The first is the simple geometry and the

second is the self-isolating property. Figure 10.20 shows PMOSFETs sharing the same n-type substrate. The oxide thickness under the gate is usually less than one- tenth of the thickness of the field oxide. The channel is created at the threshold voltage V_t which is much smaller than necessary to produce an inversion layer under the thick field oxide. Therefore, no n-channels will be created between adjacent MOSFETs, and hence they are self isolating requiring no special isolation regions.

Consider the NMOS device and its layout shown in Figure 10.21. For illustration we take a uniform design rule dimension of 10λ (corresponding to $\alpha = 10$ in Figure 10.21) assumed for all clearances and spacings. The overall dimensions of the PMOS transistor are approximately 75λ and 30λ so that the area is $225\lambda^2$ per transistor.

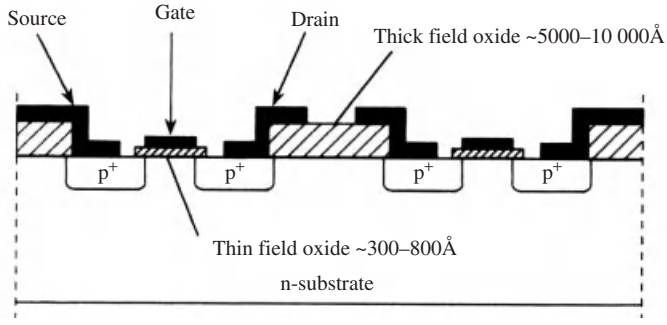


Figure 10.20 Illustrating the self-isolating properties of integrated MOSFETs

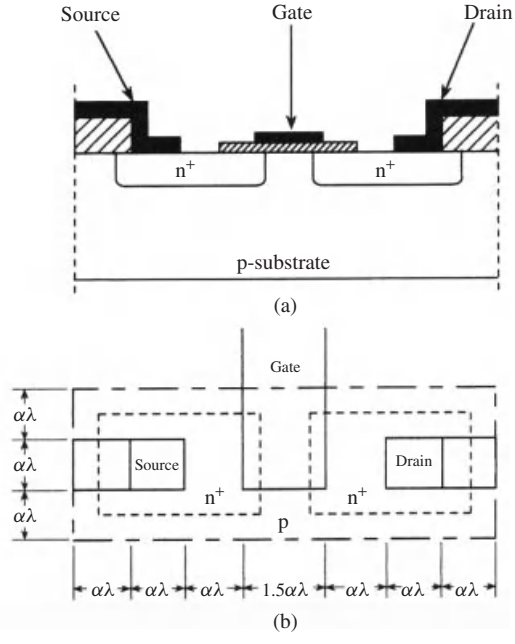


Figure 10.21 NMOSFET: (a) cross section, (b) top-surface layout

Generally, the design rules are specified by the wafer manufacturer and must be observed by the designer in the layout of the integrated circuit.

10.5 Noise In MOSFETs

MOS components suffer from noise generated by small fluctuations of analog signals within the components. The various types of noise are now discussed.

10.5.1 Shot Noise

This results from the dc current flowing across a pn-junction. It has the mean square value

$$\langle i^2 \rangle = 2qI_D \Delta f \tag{10.37}$$

where q is the electronic charge, I_D is the average dc current of the pn-junction and Δf is the bandwidth. The noise current spectral density is given by

$$\langle i^2 \rangle / \Delta f = 2qI_D \tag{10.38}$$

10.5.2 Thermal Noise

This is a result of random electron motion, with a mean square value having the typical form

$$\langle v^2 \rangle = 4kTR \Delta f \tag{10.39}$$

where R is the equivalent resistance of the noise source, T is the absolute temperature and k is Boltzmann's constant. For the device in saturation, the channel is tapered and R can be approximated by $R = 3/g_m$.

10.5.3 Flicker (1/f) Noise

This results from charge carrier traps, which capture and release carriers in a random manner. The time dependence of this phenomenon results in noise which, for a specific device and process, has a spectral density of the form

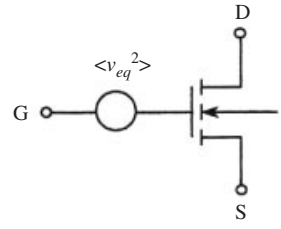
$$\langle i^2 \rangle / \Delta f = K_f (2K'/C_{ox}L^2)[I^a/f] \tag{10.40}$$

where K_f is a constant with a typical value of 10^{-24} V^2. Farads and $a = 0.5$–2.0 taken here to be $= 1$.

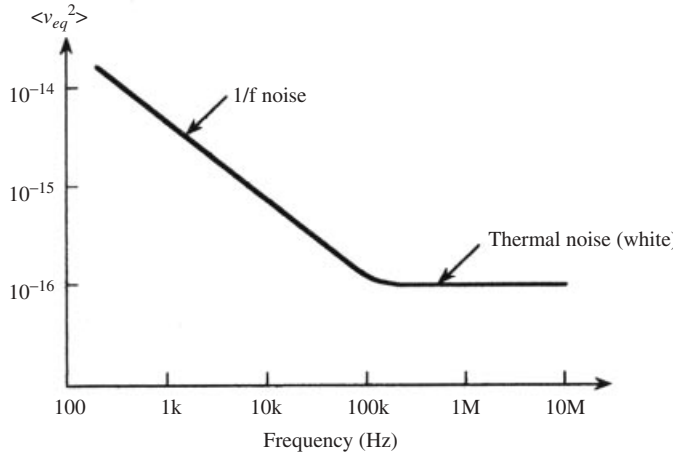
10.5.4 Modelling of Noise

In evaluating the performance of CMOS circuits, taking into account the above types of noise, they can be incorporated in the equivalent circuits as sources. Both thermal and $1/f$ noise can be modelled by a current source in parallel with $i_{D.}$ in the large signal model of the MOS transistor. The mean square current noise value is

$$\langle i_N^2 \rangle = [(8/3)kTg_m + (2K'K_f I_D)/fC_{ox}L^2]\Delta f \tag{10.41}$$



(a)



(b)

Figure 10.22 (a) Input-referred noise source model. (b) Noise spectrum of a typical MOSFET

This current can be referred to the gate of the MOS transistor, as shown in Figure 10.22(a) by dividing the above expression by g_m^2 $(=4KI_D)$ which leads to

$$\langle v_{eq}^2 \rangle = [(8/3g_m)kT + (K_f)/fC_{ox}WL]\Delta f \tag{10.42}$$

For frequencies below 1 kHz, the $1/f$ noise is the dominant source and for many practical cases, the above expression reduces to

$$\langle v_{eq}^2 \rangle = [(K_f)/fC_{ox}WL]\Delta f \tag{10.43}$$

Figure 10.22(b) shows the noise spectrum of a typical MOSFET.

Problems

10.1 An NMOS transistor has a drain current of 6.5 mA at $V_{GS} = V_{DS} = 10V$. The drain current decreases to 2 mA for $V_{GS} = V_{DS} = 6V$. Calculate the values of K and V_t for this transistor.

10.2 In a certain fabrication process, the transconductance parameter $K' = 20\,\mu A/V^2$ and the threshold voltage $V_t = 1V$. It is required to operate an NMOS transistor

device over a range in which $v_{GS} = v_{DS} = 5V$ and producing a drain current of 1 mA for the device having a minimum length of $10\,\mu m$. Find the required value of channel width.

10.3 A MOSFET has $V_t = 1V$ and $K = 500\,\mu A/V^2$. If the device is to operate in saturation with $i_D = 10\,mA$, calculate the required value of v_{GS} and the minimum required value of v_{DS}.

10.4 A MOSFET operating in saturation at a constant v_{Gs} has $i_D = 1mA$ and $v_{DS} = 2V$. When $v_{DS} = 7V, i_D$ becomes $1.1mA$. Calculate the corresponding values of r_o, V_A and λ.

11

Basic Integrated Circuits Building Blocks

11.1 Introduction

This chapter gives a concise presentation of the elementary circuits, employing the metal oxide semiconductor (MOS) structure, which form the basis for the design of the more composite building blocks of analog integrated circuits [22–24]. The use of MOS transistors as load devices is first introduced followed by the design principles of MOS amplifiers. Next, parasitic capacitances are discussed due to their importance to the operation of signal processing systems at high frequencies. In this context, the cascode amplifier is introduced with the objective of reducing the Miller effect. Finally, the current mirror is introduced followed by a discussion of the CMOS amplifier.

11.2 MOS Active Resistors and Load Devices

Active loads in MOS technology can be constructed from either enhancement-type or depletion-type devices. In the former case, the drain is connected to the gate, while in the latter the source is connected to the gate.

Figure 11.1 shows an NMOS diode-connected enhancement-type transistor together with its v-i characteristic, which is defined by

$$i = K(v - V_t)^2 \tag{11.1}$$

and it always operates in saturation.

For a transistor biased at a voltage V, we can write

$$\begin{aligned} i &= K(V + v_{gs} - V_t)^2 \\ &= K(V - V_t)^2 + 2K(V - V_t)v_{gs} + Kv_{gs}^2 \end{aligned} \tag{11.2}$$

and for small signal operation the last term can be neglected. The first term is a dc or quiescent current, therefore the signal current is

$$i = 2K(V - V_t)v_{gs} \tag{11.3}$$

and the terminal resistance of the device is $1/g_m$.

Signal Processing and Integrated Circuits, First Edition. Hussein Baher.
© 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.

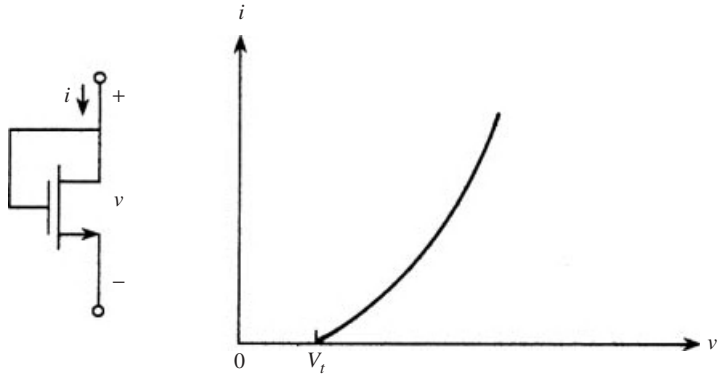


Figure 11.1 A diode-connected MOS transistor

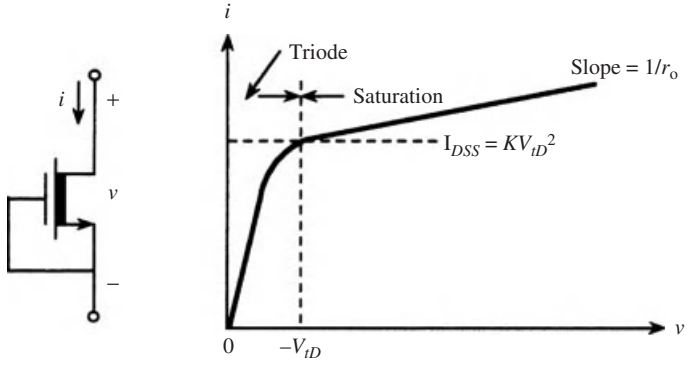


Figure 11.2 A diode-connected depletion type MOSFET and its terminal characteristic

Similarly the diode connected depletion-type MOS transistor shown in Figure 11.2 can be used as an active load. For the device to operate in saturation, the voltage across the two terminals must exceed $-V_{tD}$ (the threshold voltage). In this case

$$i \cong KV_{tD}{}^2 \left(1 + \frac{v}{V_A}\right) \tag{11.4}$$

and the device can be used to provide large resistance load values as usually required for high-gain amplifiers.

An alternative use of the MOS transistor as a resistor is shown in Figure 11.3. In the vicinity of $v_{DS} \approx 0$, the resistance of the device is given by $r_{ds} = r_o$.

A number of diode-connected transistors can be stacked to form simple CMOS voltage references as shown in Figure 11.4 for two devices.

Analysis of the circuit gives

$$V_{ref} = \frac{V_{SS} + V_{tn} + \sqrt{K_2/K_1}(V_{DD} - |V_{tp}|)}{1 + \sqrt{K_2/K_1}} \tag{11.5}$$

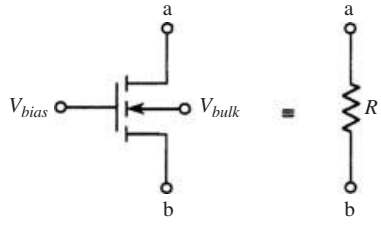


Figure 11.3 A MOSFET as an active resistor

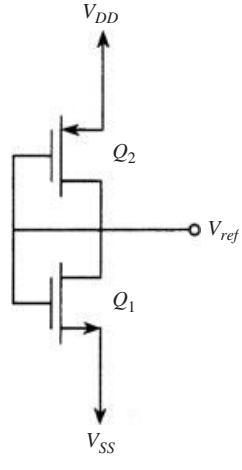


Figure 11.4 Active resistors for voltage division

11.3 MOS Amplifiers

11.3.1 NMOS Amplifier with Enhancement Load

The simplest amplifier circuit employing a diode-connected enhancement-type NMOS transistor as an active load is shown in Figure 11.5 together with its small signal equivalent circuit.

Assume that Q_1 and Q_2 have infinite resistances in saturation and identical threshold voltages V_t but different K-values. Then

$$v_{out} = \left(V_{DD} - V_t + \sqrt{\frac{K_1}{K_2}} V_t \right) - \sqrt{\frac{K_1}{K_2}} v_{in} \tag{11.6}$$

signifying a linear transfer characteristic. The large signal gain is, therefore

$$\begin{aligned} A_V &= \sqrt{\frac{K_1}{K_2}} \\ &= \sqrt{(W_1/L_1)/(W_2/L_2)} \end{aligned} \tag{11.7}$$

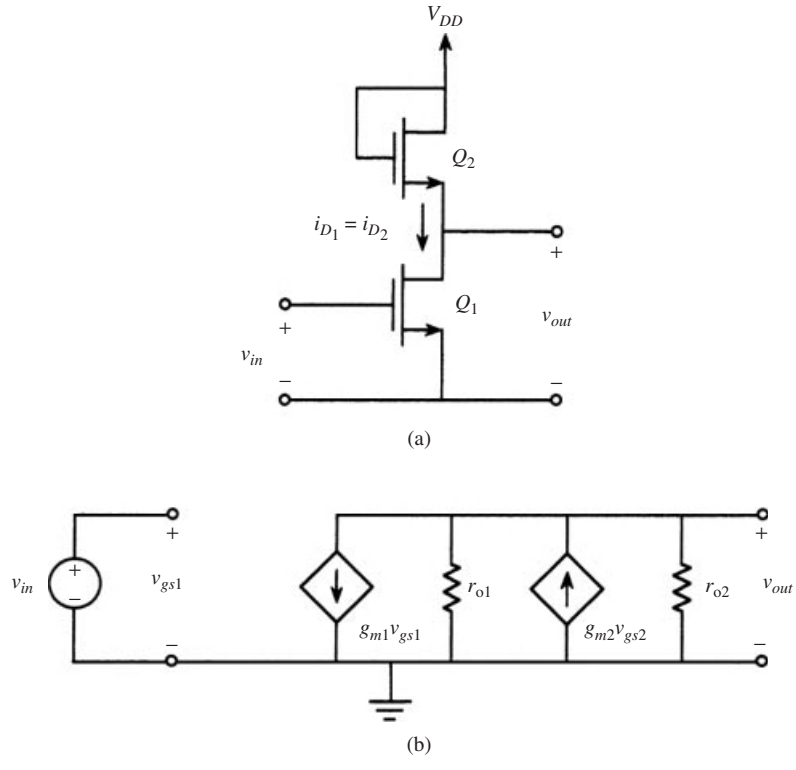


Figure 11.5 Enhancement-load amplifier: (a) circuit, (b) equivalent circuit

In order to determine the small signal gain of the amplifier, we use the equivalent circuit shown in Figure 11.5(b). This gives

$$A_v = \frac{v_{out}}{v_{in}} = \frac{-g_{m1}}{g_{m2} + (1/r_{o1}) + (1/r_{o2})} \tag{11.8}$$

Usually, $r_{o1}, r_{o1} \gg 1/g_{m2}$ and the above expression simplifies to

$$A_v = -g_{m1}/g_{m2} \tag{11.9}$$

11.3.2 Effect of the Substrate

In the above analysis we have tacitly assumed that the substrate of each transistor is connected to the source. However, in integrated circuit realizations, the two transistors share the same substrate, which is grounded. Thus, for Q_2 the substrate is grounded while the source is not, and a signal voltage v_{bs} develops between the body and the source, giving rise to a drain current component, represented by the additional current source in the equivalent circuit of Figure 11.6.

In Figure 11.6(b) the body transconductance is

$$g_{mb} = \chi g_m \tag{11.10}$$

where χ is the body factor, with typical values in the range 0.1–0.3.

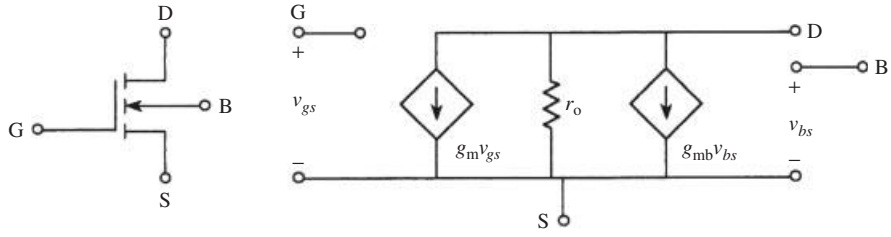


Figure 11.6 Effect of the substrate: (a) MOSFET with B not connected to S, (b) small signal equivalent circuit

Thus the enhancement-load amplifier incorporating the body effect shown in Fig. 10.6(a) has the equivalent circuit shown in Figure 11.6(b). Direct analysis gives for the gain

$$A_v = \frac{-g_{m1}}{g_{m2} + g_{mb2} + (1/r_{o1}) + (1/r_{o2})} \tag{11.11}$$

Usually, $r_{o1}, r_{o2} \gg 1/g_{m2}$, and we have

$$A_v \cong \frac{-g_{m1}}{g_{m2} + g_{mb2}} \tag{11.12}$$

which upon use of (11.10) gives

$$A_v = -\frac{g_{m1}}{g_{m2}} \frac{1}{1 + \chi} \tag{11.13}$$

Thus, the body effect reduces the gain by a factor of $1/(1 + \chi)$.

We note that the enhancement-load amplifier has a limited output signal swing which cannot exceed $V_{DD} - V_t$.

11.3.3 NMOS Amplifier with Depletion Load

Next, consider Figure 11.7(a) which shows an amplifier using an enhancement transistor together with a depletion-type diode-connected load transistor. It is stipulated that the amplifier is biased to operate in the region where both transistors are in saturation. Figure 11.7(b) shows the equivalent circuit of the amplifier incorporating the body effect.

Direct analysis under similar assumptions to those made in the previously discussed amplifier gives

$$A_v \cong \frac{-g_{m1}}{g_{mb2}} \cong \frac{-g_{m1}}{g_{m2}} \left(\frac{1}{\chi}\right) \tag{11.14}$$

or

$$A_v \cong \sqrt{\frac{(W_1/L_1)}{(W_2/L_2)}} \left(\frac{1}{\chi}\right) \tag{11.15}$$

By comparison with the gain of the enhancement-load amplifier, it is seen that the depletion-load amplifier provides a higher gain by a factor of $(1 + \chi)/\chi$.

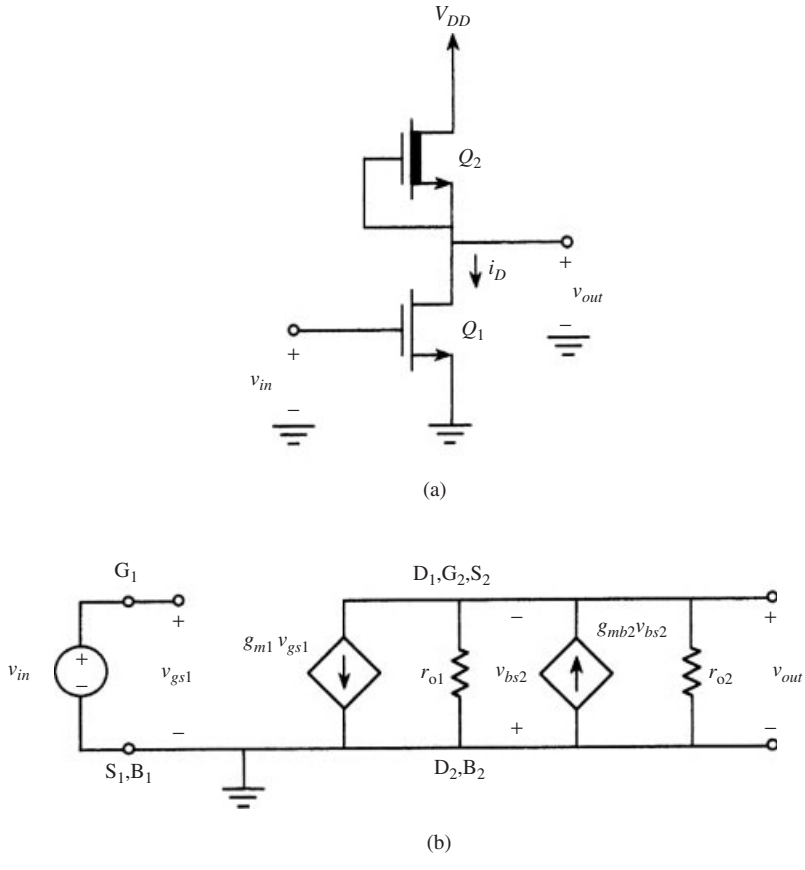


Figure 11.7 (a) Depletion-load amplifier. (b) Equivalent circuit

11.3.4 The Source Follower

The basic amplifiers discussed above are of the *common-source type*, since a signal ground is established at the source. They provide a high input impedance, a large negative voltage gain and a large output resistance. The latter is not a desirable property for voltage amplifiers. Therefore, it is often required to design an output buffer stage for obtaining a low output resistance without affecting the gain of the previous stage. This can be achieved by using the common-drain or source follower configuration shown in Figure 11.8.

The voltage gain is obtained as

$$\frac{v_{out}}{v_{in}} = \frac{[(1/g_{mb})||r_o]}{(1/g_m) + [(1/g_{mb})||r_0]} \tag{11.16}$$

which for $r_o \gg 1/g_m$, becomes

$$\frac{v_{out}}{v_{in}} \cong \frac{g_m}{g_m + g_{mb}} \tag{11.17}$$

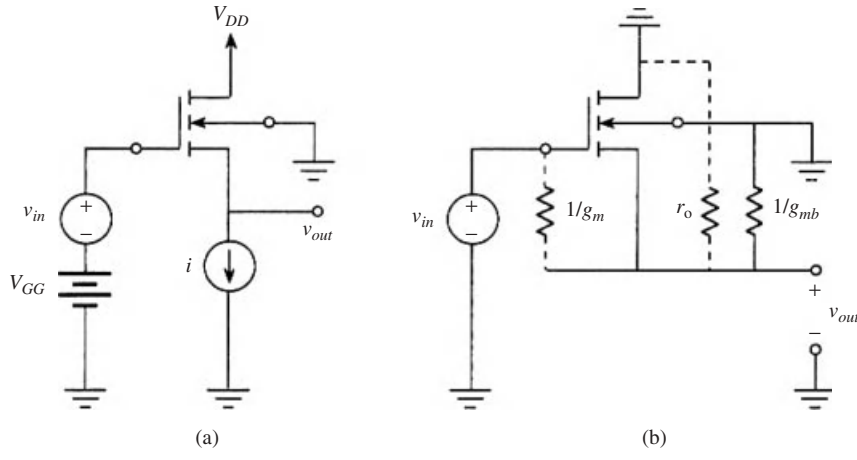


Figure 11.8 (a) Source follower configuration. (b) Finding its output resistance

and with

$$g_{mb} = \chi g_m \tag{11.18}$$

$$\frac{v_{out}}{v_{in}} \cong \frac{1}{1+\chi} \tag{11.19}$$

which is almost unity if the body effect is neglected. This is the no-load (open-circuit) voltage gain. The output resistance of the source follower is given from Figure 11.8(b) by

$$R_o = (1/g_m)||(1/g_{mb})||r_o \tag{11.20}$$

For example, with a transistor with

$$W/L = 10, \mu_n C_{ox} = 100\,\mu\text{A/V}^2, V_t = 1\,\text{V}, V_A = 100\,\text{V}, I = 20\,\text{mA}, 0.1,$$

we have the voltage gain as 0.9 and the output resistance as 140 Ω.

Now, we consider the configuration shown in Figure 11.9 which is known as the common- gate amplifier and gives high input and output conductances.

Using the equivalent circuit of Figure 11.9(b) incorporating the body effect we obtain

$$A_v = \frac{g_m(1+\chi)R_L}{1+g_m(1+\chi)R_s} \tag{11.21}$$

$$g_{in} = g_m(1+\chi) \tag{11.22}$$

$$g_{out} = \frac{g_o}{1+g_m(1+\chi)R_s} \tag{11.23}$$

It follows that unlike other configurations, the body effect does not degrade the performance of the amplifier; on the contrary it increases the effective transconductance. The

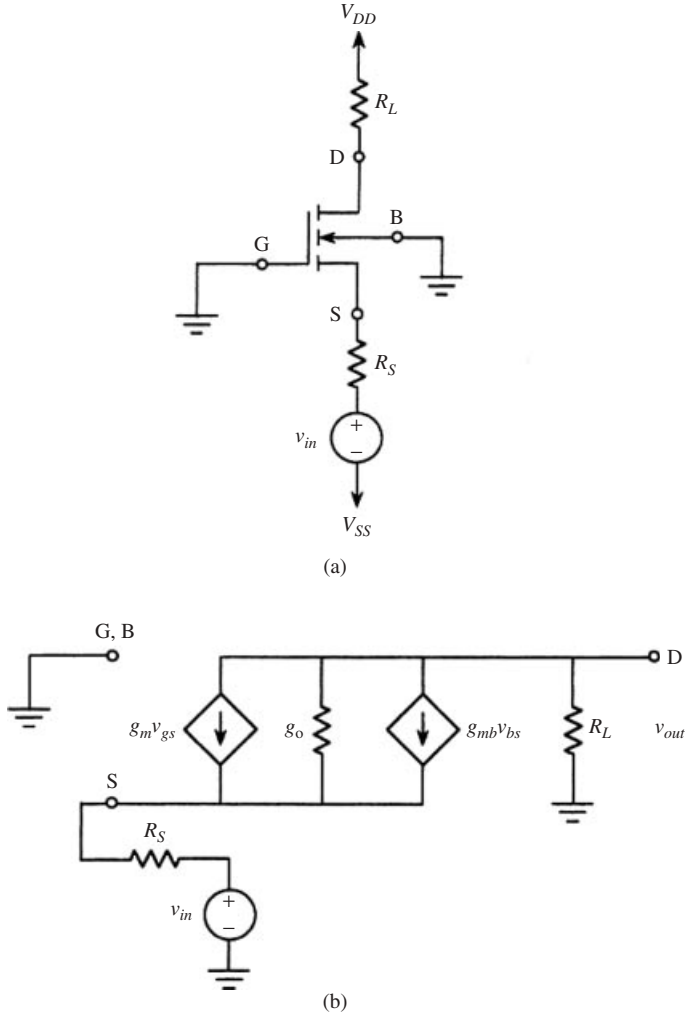


Figure 11.9 The common-gate amplifier: (a) circuit with the substrate shown, (b) equivalent circuit

main advantage of the configuration, however, lies in its wider bandwidth as we shall see in the next section.

11.4 High Frequency Considerations

11.4.1 Parasitic Capacitances

Figure 11.10 shows the high-frequency equivalent circuit of a MOSFET containing the intrinsic components of the terminal capacitances as well as the extrinsic ones. Those of the former type are associated with reverse-biased p-n junctions, channel and depletion regions. These are strongly dependent on the region of operation of the devices. Those of the latter type (extrinsic) are made up of components which are largely constant and are due to layout parasitics and overlapping regions.

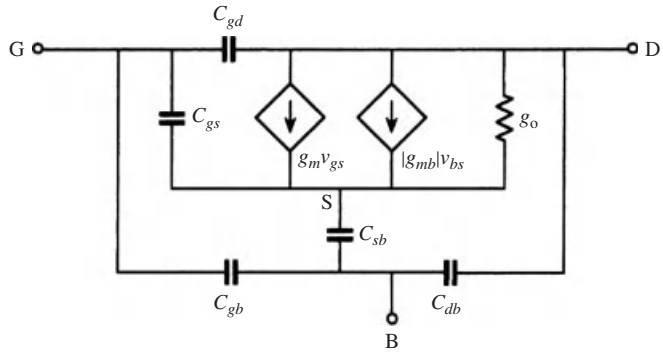


Figure 11.10 High-frequency equivalent circuit of MOSFET

In saturation, the most significant capacitances are the following:

1. C_{gd}: Gate to drain capacitance. This is a thin oxide capacitance due to the overlap of the gate and drain diffusion, and as such can be assumed to be voltage independent.
2. C_{gs}: Gate to source capacitance, which has two components:
   (a) C_{gs1} which is a thin oxide capacitance due to the gate to source overlap.
   (b) C_{gs2} which is the gate to channel capacitance. In saturation, this is around $\frac{2}{3}C_{ox}$. Here, C_{ox} is the total thin oxide capacitance between the gate and the surface of the substrate.

   In saturation, C_{gs} is almost voltage-independent.
3. C_{sb}: Source to substrate capacitance. This has two components:
   (a) C_{sdpn} which is the p-n junction capacitance between the source diffusion and substrate.
   (b) C'_{sb} which is about 2/3 of the capacitance of the depletion region below the channel, C_{sb} has a voltage dependence similar to that of an abrupt p-n junction.
4. C_{db}: Drain to substrate capacitance, which is a voltage-dependent capacitance of a p-n junction.
5. C_{gb}: Gate to substrate capacitance, which is normally small in saturation, typically around $0.1C_{ox}$.

Now, for the common source amplifier of Figure 11.11(a), the capacitances discussed above come into play at high frequencies and dictate the frequency response of the amplifier. Figure 11.11(b) shows the equivalent circuit of such an amplifier including the parasitic capacitances, when the amplifier is loaded with C_L, that is a capacitive load. This equivalent circuit simplifies to that of Figure 11.11(c) by straightforward addition of capacitors and sources, in which

$$\begin{aligned} G_{Leq} &= g_{d1} + g_{d2} + g_{m2} + |g_{mb2}| \\ C_{Leq} &= C_{db1} + C_{gs2} + C_{sb2} + C_L \end{aligned} \tag{11.24}$$

Direct analysis of this circuit gives for the gain

$$A_v(s) = \frac{G_s(C_{gd1} - g_{m1})}{[s(C_{gs1} + C_{gd1}) + G_s][s(C_{gd1} + C_{Leq}) + G_{Leq}] - sC_{gd1}(sC_{gs1} - g_{m1})} \tag{11.25}$$

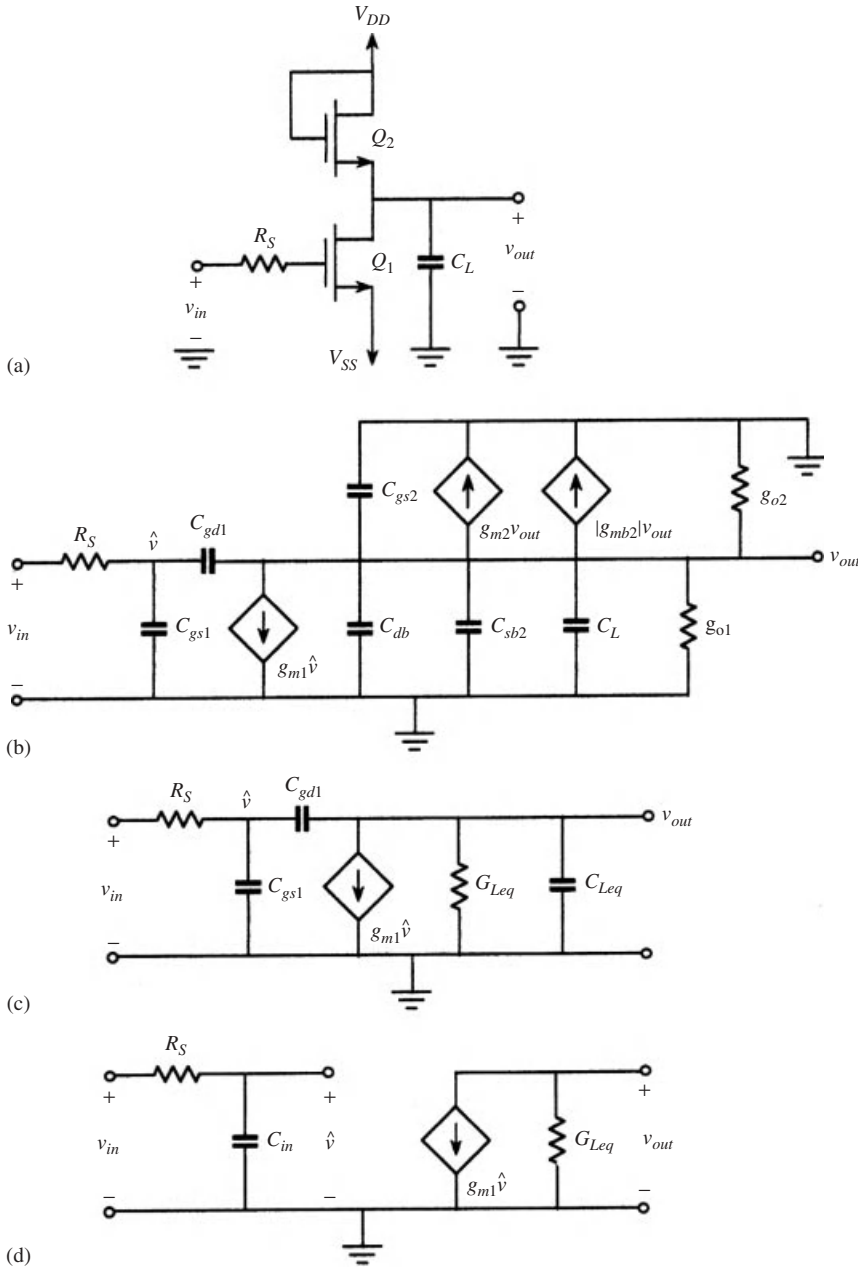


Figure 11.11 Capacitively loaded NMOS amplifier at high frequencies: (a) circuit, (b) equivalent circuit, (c) simplified equivalent circuit, (d) approximate equivalent circuit

As $s = j\omega$ and for moderate frequencies we can take

$$\begin{aligned} g_{m1} &\gg \omega C_{gd1} \\ G_{Leq} &\gg \omega(C_{gd1} + C_{Leq}) \end{aligned} \tag{11.26}$$

and the gain

$$\begin{aligned} A_v(j\omega) &= \frac{-g_{m1}G_s}{G_s G_{Leq} - j\omega[G_{Leq}(C_{gs1}) + g_{m1}C_{gd1}]} \\ &= \frac{A(0)}{1 + j\omega R_s C_{in}} \end{aligned} \tag{11.27}$$

where

$$A(0) = -g_{m1}/G_{Leq} \tag{11.28}$$

and

$$\begin{aligned} C_{in} &= C_{gs1} + C_{gd1}(1 + g_{m1}/G_{Leq}) \\ &= C_{gs1} + C_{gd1}[1 + |A_v(0)|] \end{aligned} \tag{11.29}$$

Therefore, the gain is seen to be the gain of the approximate equivalent circuit of Figure 11.11(d). In particular the input capacitor C_{in} results from the gate to source capacitance C_{gs1} plus the gate to drain capacitor C_{gd1} magnified by the factor $[1 + |A_v(0)|]$. The latter is the familiar Miller effect and, since $|A_v(0)| \gg 1$, the Miller effect results in a serious reduction in the bandwidth.

11.4.2 The Cascode Amplifier

The Miller effect can be eliminated or at least reduced by using a common-gate MOSFET Q_2 together with the common-source amplifier resulting in the cascode configuration shown in Figure 11.12(a). Q_2 isolates the input and output nodes. It provides a low input resistance $1/g_{m2}$ at its source and a high one at its drain to drive Q_2.

The low-frequency small signal equivalent circuit is shown in Figure 11.12(b) which gives

$$g_{m1} = -g_{m2}\hat{v} = -g_{m3}v_{out} \tag{11.30}$$

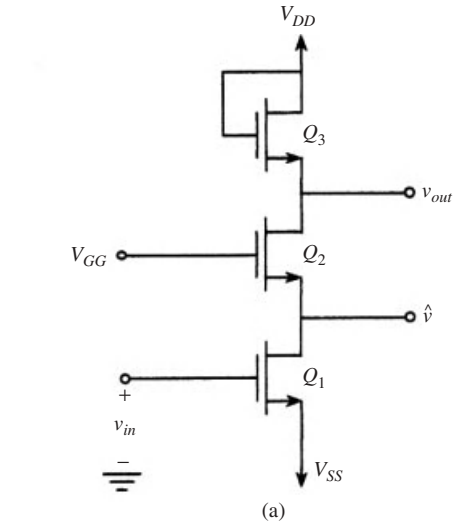
or

$$\hat{v} \cong -\frac{g_{m1}}{g_{m2}}v_{in} \tag{11.31}$$

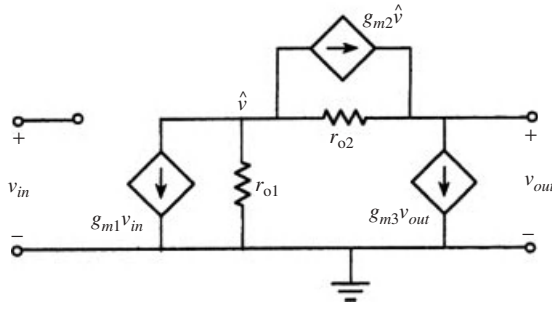
and

$$v_{out} \cong \frac{g_{m2}}{g_{m3}}\hat{v} \cong -\frac{g_{m1}}{g_{m3}}v_{in} \tag{11.32}$$

Hence, the gate to drain gain of Q_1 is $-g_{m1}/g_{m2}$ and C_{gd1} of the driver transistor is multiplied by $(1 + g_{m1}/g_{m2})$. If we choose $g_{m1} = g_{m2}$ then this factor is 2, and the Miller effect is reduced considerably.



(a)



(b)

Figure 11.12 (a) Cascode amplifier. (b) Low-frequency equivalent circuit

11.5 The Current Mirror

In integrated circuits, a stable dc reference current source is designed, then used to generate other dc currents which are multiples or fractions of this source at other points of the circuit for biasing the transistors. The current mirror shown in Figure 11.13 is a universal circuit for producing a current I_o which is proportional to a reference current I_{ref}. It comprises two enhancement MOSFETs Q_1 and Q_2 with the same V_t but possibly different aspect ratios.

In the circuit of Figure 11.13, both Q_1 and Q_2 operate in saturation and due to the parallel connection both transistors have the same V_{GS}. Thus, we can write

$$I_{ref} = K_1(V_{Gs} - V_t)^2 \tag{11.33}$$

$$I_o = K_2(V_{Gs} - V_t)^2 \tag{11.34}$$

assuming the output resistance of Q_2 to be infinite. Hence

$$I_o = I_{ref}\left(\frac{K_2}{K_1}\right) = I_{ref}\frac{(W_2/L_2)}{(W_1/L_1)} \tag{11.35}$$

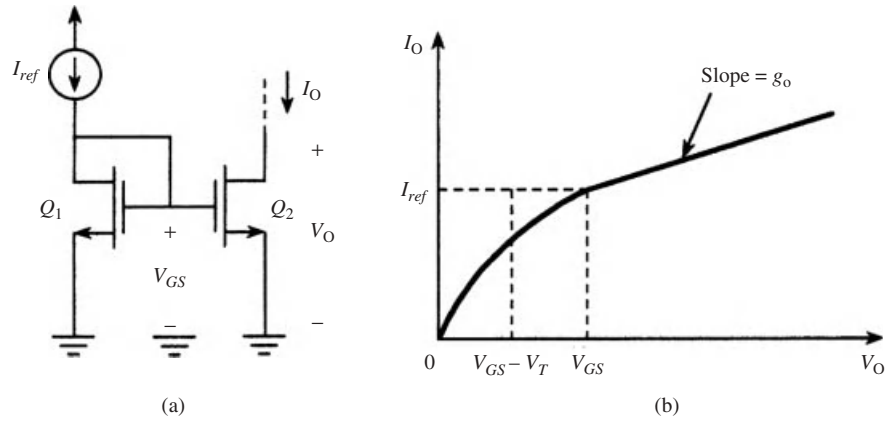


Figure 11.13 (a) Basic current mirror. (b) Its output characteristic for matched Q_1 and Q_2

However, due to the finite output resistance (= r_o of the transistor Q_2) the above expression is only approximate. This output resistance of the current mirror can be increased by using the cascode mirror of Figure 11.14(a). For this circuit the incremental resistance of each of the diode-connected Q_1 and Q_4 is $1/g_m$ which is relatively small. Replacing Q_2 by r_{o2} and using the equivalent circuit of Q_3 we obtain the circuit of Figure 11.14(c) which gives for the output resistance

$$R_o = v/i = r_{o3} + r_{o2} + g_{m3} r_{o3} r_{o2} \tag{11.36}$$

With $r_{o2} = r_{o3} = r_o$, we have

$$R_o = r_o(2 + g_m r_o) \tag{11.37}$$

which is larger than that of the simple mirror of Figure 11.13 by a factor $\approx g_m r_o$.

Another circuit is the Wilson current mirror shown in Figure 11.15 for which

$$r_o = (g_{m1} r_{o1}) \left(\frac{g_{m3}}{g_{m2}} \right) r_{o3} \tag{11.38}$$

with a typical value

$$r_o \approx (100)(1)(10^5) \approx 10\, M\Omega$$

However, this circuit suffers from the fact that the drain-to-source voltage drops of Q_1 and Q_2 are unequal; these can be equalized by adding the diode-connected transistor Q_4 as shown in Figure 11.16. The output resistance of this improved Wilson current source is still the same as that of the Wilson mirror.

11.6 The CMOS Amplifier

Complementary n-channel and p-channel devices are used together in integrated circuit CMOS technology yielding greater design flexibility and in addition, they are fabricated in a manner that eliminates the body effect. They are also used with great effect in

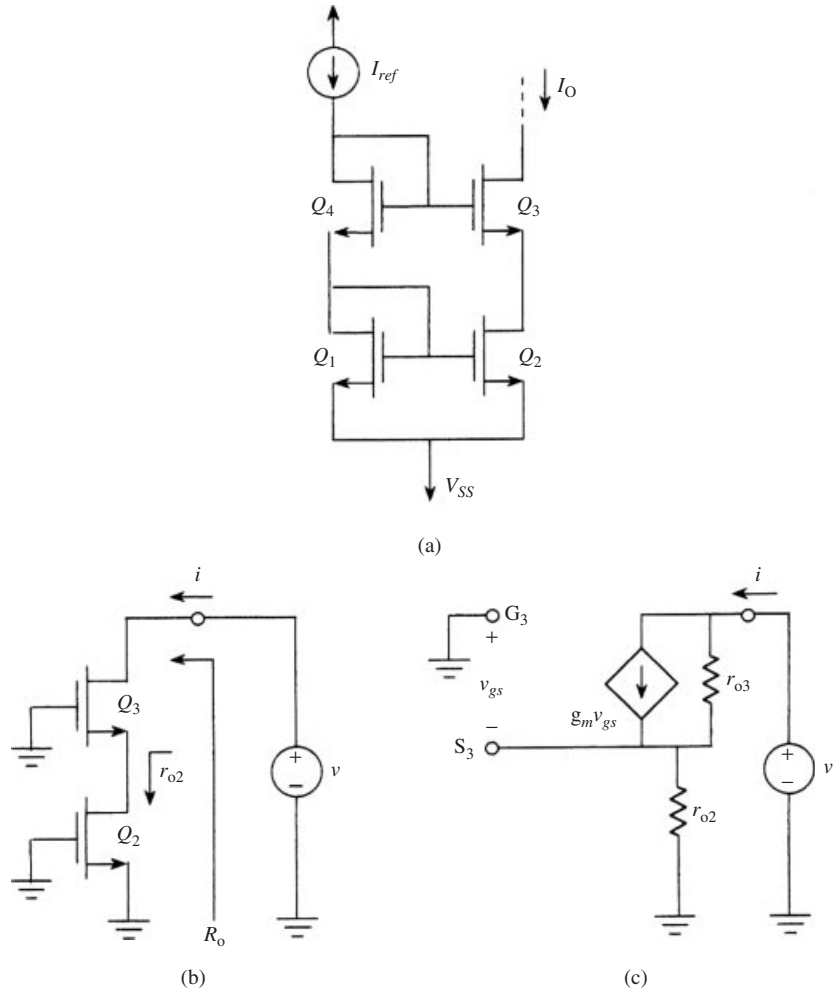


Figure 11.14 (a) Cascode current mirror. (b, c) Finding its output resistance

the design of switches to minimize an undesirable effect called clock feed-through in switched-capacitor circuits, as we shall see in a later chapter.

The basic CMOS amplifier is shown in Figure 11.17(a). It has the following features:

1. Q_2 and Q_3 are matched p-channel devices forming a current source with the v-i characteristic shown in Figure 11.17(b). Q_2 is forced to operate in saturation by ensuring that its drain voltage is lower than its source voltage V_{DD} by at least $V_{SG} - |V_{tp}|$, where V_{SG} is the dc bias voltage corresponding to a drain current of I_{ref}. In saturation, Q_2 has the high output resistance

$$r_{o2} = \frac{|V_A|}{I_{ref}} \tag{11.39}$$

2. Q_2 is an active load for the amplifying transistor Q_1.

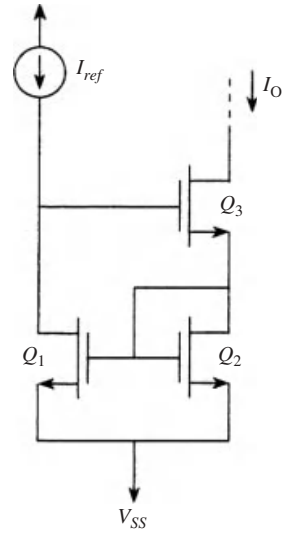


Figure 11.15 Wilson current mirror

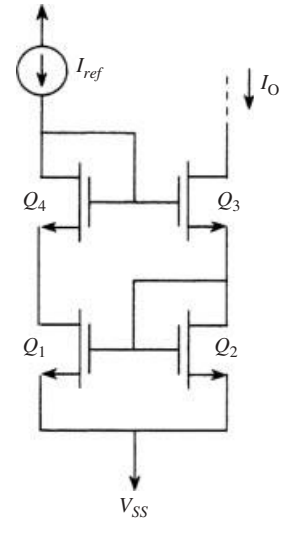


Figure 11.16 Improved Wilson current mirror

3. When Q_1 is in saturation, the small signal voltage gain is given by

$$A_v = -(g_{m1})(r_{o1}||r_{o2}) \tag{11.40}$$

and since Q_1 operates at the dc bias current I_{ref}, then using (10.24) we have

$$g_{m1} = \sqrt{2(\mu_n C_{ox})(W/L)I_{ref}} \tag{11.41}$$

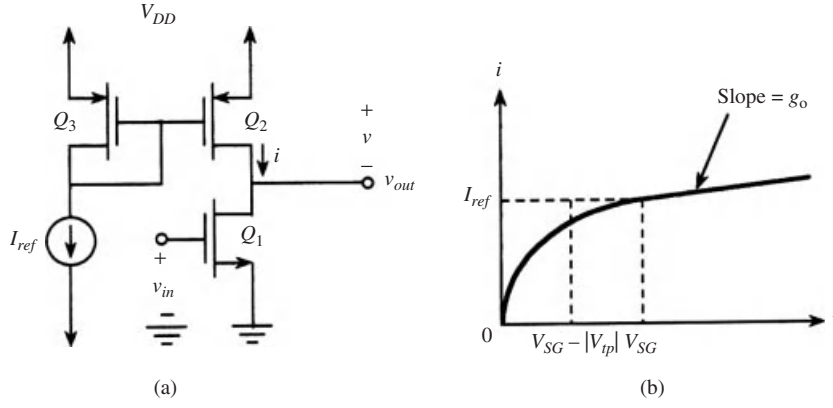


Figure 11.17 (a) CMOS amplifier. (b) Terminal characteristic of the active load

Thus with $r_{o1} = r_{o2}$

$$A_v = -\frac{\sqrt{K_n}|V_A|}{\sqrt{I_{ref}}} \tag{11.42}$$

11.7 Conclusion

This chapter dealt with elementary integrated circuit components which are necessary for the construction of more elaborate circuits. In particular, the use of MOS transistors as active resistors and loads and as simple amplifying or gain stages were reviewed. The effect of parasitic capacitances on the high frequency performance of the circuits was discussed. The various types of current mirror were also given and used in the design of simple CMOS amplifiers.

Problems

11.1 Design an NMOS common source amplifier with an enhancement-type load device, to have a voltage gain of 10 and an output resistance of 1 kΩ taking the effect of the substrate into account. The minimum channel length in the process is 10 μm and for both devices $\chi = 0.2$, $K' = 10\,\mu A$, $V_t = 1\,V$. Assume a bias value of 2 V.

11.2 Design an NMOS common-source amplifier with a depletion-load transistor as a load, to have a gain of 100 taking the effect of the substrate into account. The minimum channel length in the process is 3 μm and $\chi = 0.1$.

11.3 Design a cascode amplifier to have a gain of 100, taking into account the effect of the substrate for which $\chi = 0.2$. The minimum channel length for the process is 5 μm.

11.4 The Wilson current mirror of Figure 11.15 employs transistors with $K = 100\,\mu A$, $V_t = 1\,V$, $V_A = 50\,A$. The reference current is $50\,\mu A$ and $V_{SS} = 0$. Find I_0 and the output resistance of the mirror.

11.5 Design a CMOS amplifier of the type shown in Figure 11.17 with a voltage gain of 100. Take the transistor parameters to be any reasonable set, based on the experience gained in solving the problems of Chapter p10, and use $I_{ref} = 100\,\mu A$.

12

Two-stage CMOS Operational Amplifiers

12.1 Introduction

A critical building block in the design of analog signal processing systems is the CMOS operational amplifier (Op Amp). It may be regarded as a composite building block employing the elementary circuits and fundamental concepts given in the previous chapter. This chapter deals with the design of integrated CMOS Op Amps [22–24] and gives complete design examples. The chapter begins by a summary of Op Amp performance parameters and the fundamentals of feedback amplifier characteristics, then proceeds to give a discussion of the CMOS differential pair which is the first stage in the Op Amp. Next, the very popular two-stage CMOS Op Amp architecture is developed and a detailed account of the design considerations is given together with a summary of the design equations for easy reference and use by the reader. This is followed by a complete design example starting from the design specifications and showing in detail how to arrive at the structure and element values of the final design of an integrated CMOS Op Amp.

12.2 Op Amp Performance Parameters

Ideally, an Op Amp is a voltage-controlled voltage source as shown in Figure 12.1 whose output v_o and two inputs v_1 and v_2 are related by

$$v_o = A(v_1 - v_2) \tag{12.1}$$

where A is a constant, independent of frequency, and is very large (ideally $= \infty$). In practice, however, A is frequency-dependent and there are a number of non-ideal effects in real MOS Op Amps.

The performance and specifications of an Op Amp are usually stated in terms of the following parameters:

1. *The open loop dc-gain* A_o. This is the value of the voltage gain at zero frequency, that is when the input is a constant.
2. *Output voltage swing*. This is the output range over the linear range of operation, for example for $\pm 5V$ supplies, the linear range could be $-4V < v_o < 4V$.

Signal Processing and Integrated Circuits, First Edition. Hussein Baher.
© 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.

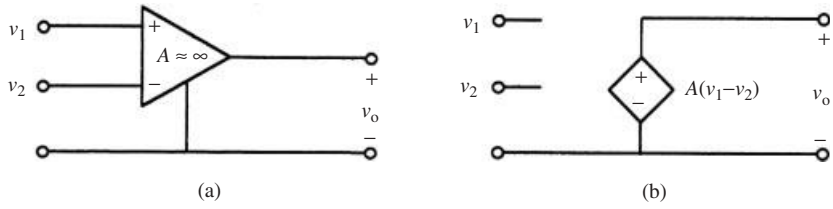


Figure 12.1 The ideal Op Amp: (a) symbol, (b) equivalent circuit

3. *Input offset voltage*. Ideally, for $v_1 = v_2$, $v_o = 0$. However, in practice, $v_0 \neq 0$ for zero differential input. The voltage at the input which is required to reduce the output to zero, is called the *input offset voltage*.
4. *Common–mode rejection ratio* (CMRR). Ideally, v_0 depends only on the differential input $(v_1 - v_2)$. In reality, v_0 is also affected by the average or common-mode voltage

$$v_{CM} = \frac{v_1 + v_2}{2} \tag{12.2}$$

as well as the differential voltage

$$v_d = (v_1 - v_2) \tag{12.3}$$

and we can write for the output

$$v_0 = A_d v_d + A_{CM} v_{CM} \tag{12.4}$$

where A_d is the differential gain and A_{CM} is the common-mode gain. The common-mode rejection ratio is defined as

$$CMRR = 20 \log \frac{A_d}{A_{CM}} \text{dB} \tag{12.5}$$

which, for an ideal Op Amp should be infinite.
5. *Common-mode range* (CMR). This is the range of the common-mode voltage over which the CMRR remains acceptably high.
6. *Power supply rejection ratio*. Any additive noise at either terminal of the voltage supply appears at the output v_0 with a gain of A_{ps}. The power supply rejection ratio is defined by

$$PSRR = 20 \log \frac{A_d}{A_{ps}} \tag{12.6}$$

7. *The unity-gain bandwidth*. As the frequency increases, the gain of the Op Amp rolls off. The frequency at which the gain reaches unity (or 0 dB) is the unity-gain bandwidth.
8. *Settling time*. This is the time required for the output to reach its final value (with an error of typically 0.1–1.0%) when the input voltage changes over the linear range of operation. The settling time is related to the unity-gain bandwidth, the output impedance and the load.

9. *Slew rate*. As the input varies by a jump discontinuity, the output cannot follow the difference instantaneously. The maximum rate of change dv_0/dt is called the slew rate (SR). It is basically a non-linear effect.
10. *Output resistance*. Ideally, this should be zero. Actual Op Amps have a finite resistive output impedance, which may be too large, thus affecting the charging time of a capacitor connected to the output and limiting the highest signal frequency.
11. *Dynamic range*. The real Op Amp has a finite linear range in its transfer characteristic. Therefore, there is a maximum signal amplitude $v_{in.max} \cong V_{cc}/A$ where A is the open loop gain. Also, due to spurious signals such as noise, there is a minimum signal value $v_{in.min}$ that is not lost in noise. The dynamic range of the Op Amp is defined as

$$\text{Dynamic range} = 20\log(v_{in.\max}/v_{in.\min})\,\text{dB} \tag{12.7}$$

12. *dc power dissipation*. The general structure of a two-stage Op Amp is shown in Figure 12.2. The input differential stage is designed to provide a high input impedance, large CMRR, a large PSRR, low noise, low offset voltage and high gain. The next stage can be designed to perform one or more of several functions: (a) level shifting to compensate for the dc voltage change in the input stage, ensuring proper dc bias for the following stages, (b) additional gain and (c) differential to single-ended conversion. The output buffer may be needed in some applications. In some cases, most Op Amps used are required to drive on-chip capacitances of low values and the output stage is not needed. However, if the Op Amps are required to drive large capacitive or resistive loads, the output buffer stage is needed, which provides the Op Amp with a low output impedance.

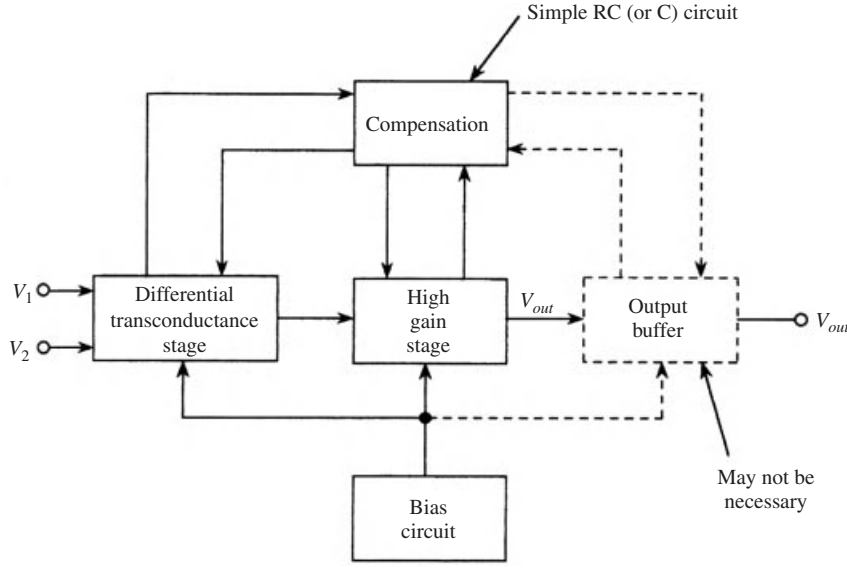


Figure 12.2 General structure of a two-stage Op Amp

12.3 Feedback Amplifier Fundamentals

Negative feedback is employed almost invariably in the design of operational amplifiers to achieve greater control over the frequency and time responses of the amplifier. As expected, this accomplished at the expense of reduced gain. Figure 12.3 shows the general feedback amplifier topology, where $G(s)$ is the open loop gain (without feedback) and $\beta(s)$ is the transfer function of the feedback network. The latter could be as simple as an RC network.

The transfer function of the feedback amplifier is given by

$$A(s) = \frac{G(s)}{1 + \beta(s)G(s)} \tag{12.8}$$

and if we define the loop gain L(s) as

$$L(s) = -\beta(s)G(s) \tag{12.9}$$

then

$$A(s) = \frac{G(s)}{1 - L(s)} \tag{12.10}$$

If ω_0 is the frequency at which the argument of the loop gain is zero, then stability (no sustained oscillations) requires

$$|L(\omega_0)| < 1 \tag{12.11}$$

Alternatively, if the 3 dB frequency ω_{03dB} is defined by

$$|L(\omega_{03dB})| = 1 \tag{12.12}$$

then stability requires

$$\arg[L(\omega_{03dB}] > 0 \tag{12.13}$$

These relations are illustrated in Figure 12.4 for a second-order amplifier transfer function which shows the amplitude and phase of the loop gain.

Stability requires that the magnitude crosses the 0 dB point before the phase reaches zero. A standard measure of stability is the phase margin ϕ_M defined as the value of the phase at the frequency where the magnitude reaches unity. Thus

$$\phi_M = \arg |L(j\omega_{03dB})| \tag{12.14}$$

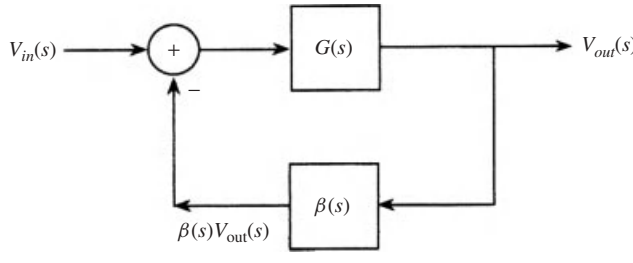


Figure 12.3 Feedback amplifier topology

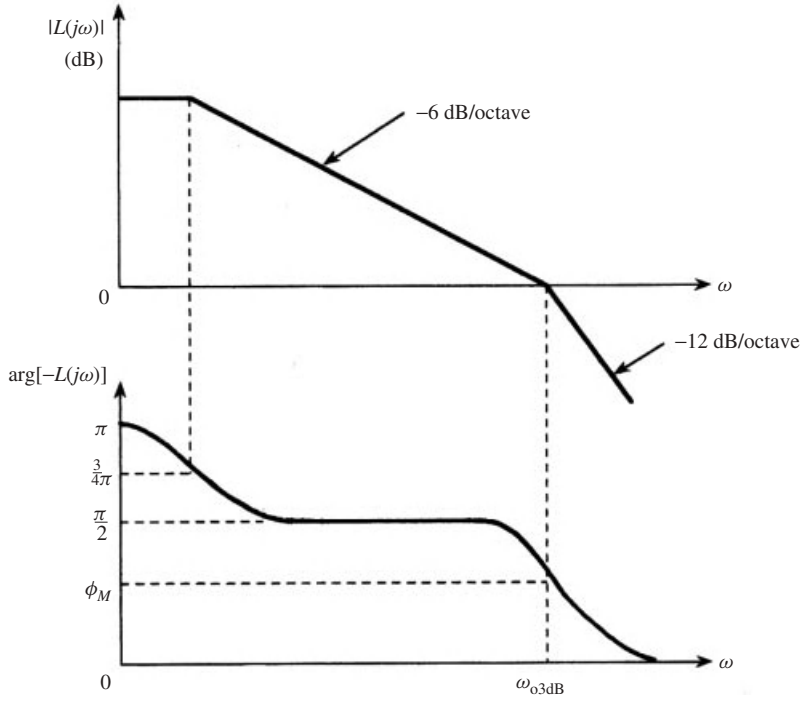


Figure 12.4 Typical amplitude and phase of the loop gain of a second order transfer function

Since the phase margin is determined by the pole-zero pattern of the transfer function, it also determines the time response of the amplifier. Figure 12.5 shows examples for a second-order function for several values of the phase margin. Since too much ringing is undesirable, a phase margin higher that 45° is usually sought. A value of 60° is generally considered acceptable for most applications.

In determining the frequency response of operational amplifiers, one usually reaches a transfer function of the form

$$A(s) = A(0)\frac{\prod(1 + s/\omega_{zi})}{\prod(1 + s/\omega_{pi})} \tag{12.15}$$

Usually the zeros ω_{zi} are at frequencies so high that they do not affect the 3 dB frequency. If, in addition, one of the poles ω_{pk}, say, is at a frequency much lower that than all the others, this is called the *dominant pole* and we can write

$$A(s) \cong \frac{\omega_{pk}A(0)}{(s + \omega_{pk})} \tag{12.16}$$

so that

$$\omega_{3dB} \cong \omega_{pk} \tag{12.17}$$

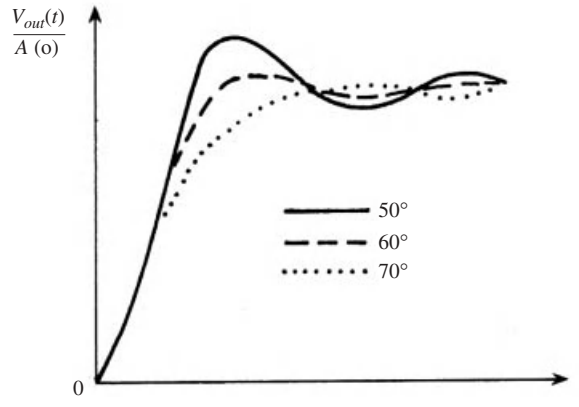


Figure 12.5 Step response of a typical second-order feedback amplifier for different values of phase margin

The unity gain frequency is given by

$$\omega_t = A(0)\omega_{3dB} \tag{12.18}$$

12.4 The CMOS Differential Amplifier

The basic MOSFET differential pair is shown in Figure 12.6 as composed of two NMOS transistors Q_1 and Q_2. Ideally, the transistors are perfectly matched.

Neglecting the output resistances of the two devices, we can write in saturation

$$\begin{aligned} i_{D1} &= K(v_{GS1} - V_t)^2 \\ i_{D2} &= K(v_{GS2} - V_t)^2 \end{aligned} \tag{12.19}$$

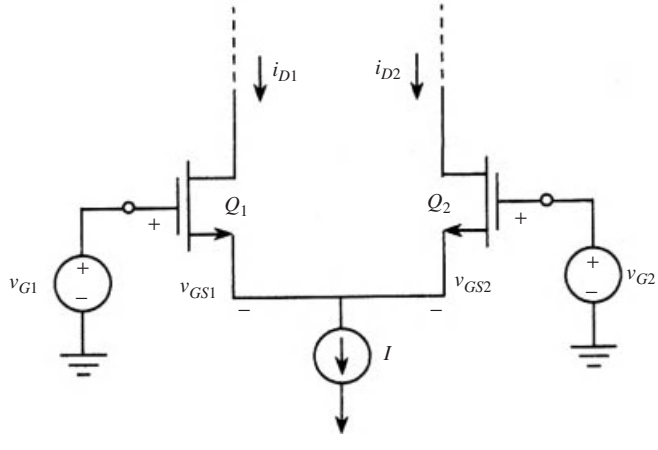


Figure 12.6 Basic MOSFET differential pair

where

$$K = \tfrac{1}{2}\mu_n C_{ox}(W/L) \tag{12.20}$$

Solving for v_{GS1} and v_{GS2} and subtracting we have

$$\sqrt{i_{D1}} - \sqrt{i_{D2}} = \sqrt{K}v_{iD} \tag{12.21}$$

with

$$v_{iD} = v_{GS1} - v_{GS2} \tag{12.22}$$

But

$$i_{D1} + i_{D2} = I \tag{12.23}$$

so that

$$\begin{aligned} i_{D1} &= \frac{I}{2} + \sqrt{2KI}\left(\frac{v_{id}}{2}\right)\sqrt{1 - \frac{(v_{id}/2)^2}{(I/2K)}} \\ i_{D1} &= \frac{I}{2} - \sqrt{2KI}\left(\frac{v_{id}}{2}\right)\sqrt{1 - \frac{(v_{id}/2)^2}{(I/2K)}} \end{aligned} \tag{12.24}$$

But at the bias point $v_{id} = 0$, so that

$$i_{D1} = i_{D2} = I/2 \tag{12.25}$$

and

$$v_{GS1} = v_{GS2} = V_{GS} \tag{12.26}$$

with

$$I/2 = K(V_{GS} - V_t)^2 \tag{12.27}$$

which when used in (12.25) and (12.26) give

$$\begin{aligned} i_{D1} &= \frac{I}{2} + \left(\frac{I}{V_{GS} - V_t}\right)\left(\frac{v_{id}}{2}\right)\sqrt{1 - \left(\frac{v_{id}/2}{V_{GS} - V_t}\right)^2} \\ i_{D1} &= \frac{I}{2} - \left(\frac{I}{V_{GS} - V_t}\right)\left(\frac{v_{id}}{2}\right)\sqrt{1 - \left(\frac{v_{id}/2}{V_{GS} - V_t}\right)^2} \end{aligned} \tag{12.28}$$

For small signal operation, $v_{id} \ll V_{GS} - V_t$

$$\begin{aligned} i_{D1} &\cong \frac{I}{2} + \left(\frac{I}{V_{GS} - V_t}\right)\left(\frac{v_{id}}{2}\right) \\ i_{D1} &\cong \frac{I}{2} - \left(\frac{I}{V_{GS} - V_t}\right)\left(\frac{v_{id}}{2}\right) \end{aligned} \tag{12.29}$$

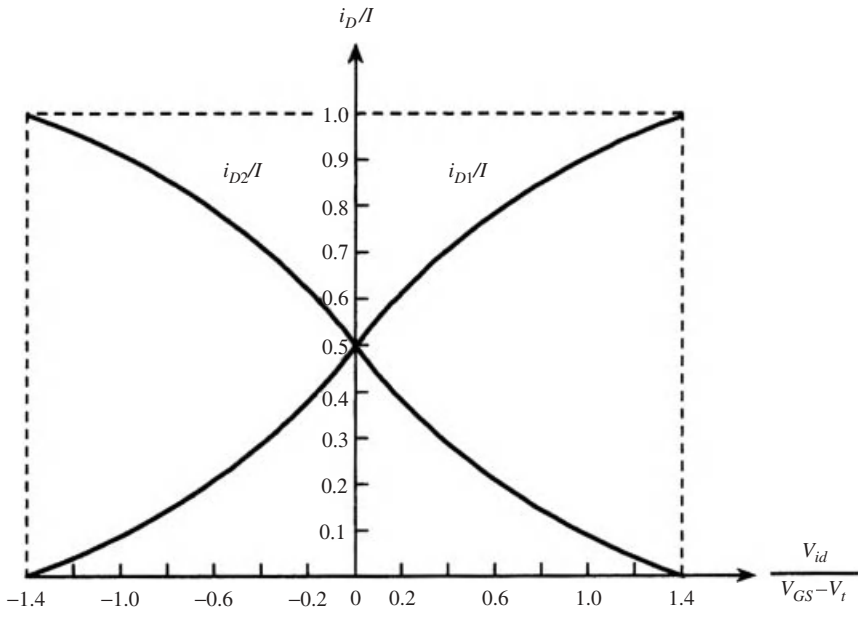


Figure 12.7 Illustration of Equation (12.29) for the MOSFET differential pair

But a MOSFET biased at I_D has $g_m = 2I_D/(V_{GS} - V_t)$. Thus, for Q_1 or Q_2

$$g_m = \frac{2(I/2)}{V_{GS} - V_t} = \frac{I}{V_{GS} - V_t} \tag{12.30}$$

Figure 12.7 shows a graphical representation of equations (12.29). We also note that for differential input signals, each of the two transistors has an output resistance of r_0.

Next, we consider the use of a current mirror composed of the transistors Q_3 and Q_4 as a load for the differential pair, as shown in Figure 12.8. The mirror uses PMOS transistors while the differential pair uses NMOS types; therefore the result is a simple yet popular CMOS differential amplifier configuration.

Analysis of the circuit gives

$$i = g_m(v_{id}/2) \tag{12.31}$$

with

$$g_m = \frac{I}{V_{GS} - V_t} \tag{12.32}$$

Also

$$v_{out} = 2i(r_{ot} \| r_{o4}) \tag{12.33}$$

with

$$r_{o2} = r_{o4} = r_o \tag{12.34}$$

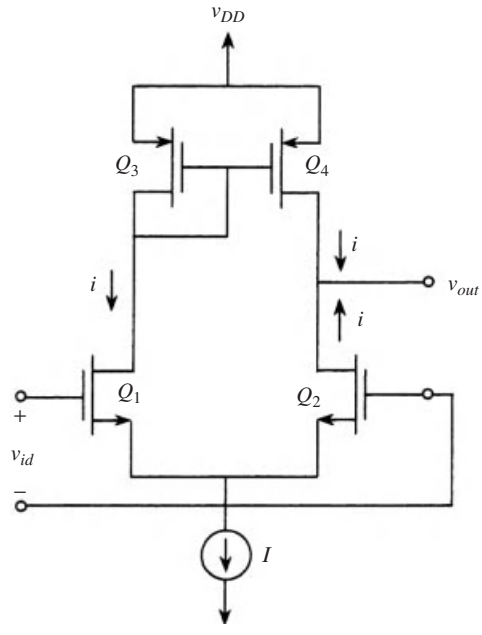


Figure 12.8 Active-loaded CMOS differential amplifier

the voltage gain is given by

$$A_v = \frac{v_{out}}{v_{id}} = g_m \frac{r_o}{2}$$
$$= \frac{V_A}{V_{GS} - V_t} \tag{12.35}$$

To calculate the common-mode gain and hence the common-mode rejection ratio, we include the bias source conductance G as shown in Figure 12.9(a) from which the equivalent circuit of Figure 12.9(b) is obtained, in which the subscript i is used to denote the parameters of the input differential stage devices while the subscript L is used to indicate the parameters of the load devices.

Under the assumptions $g_m, g_{mL} \gg G, g_{dsi}$ we obtain

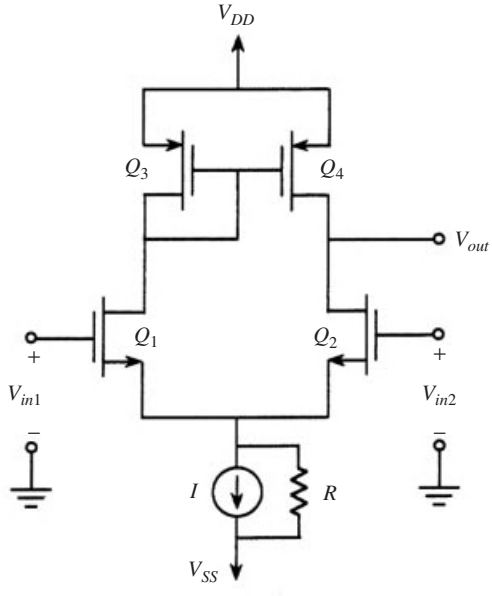
$$A_d \cong \frac{g_{mi}}{g_{dsL} + g_{dsi}} \tag{12.36}$$

and

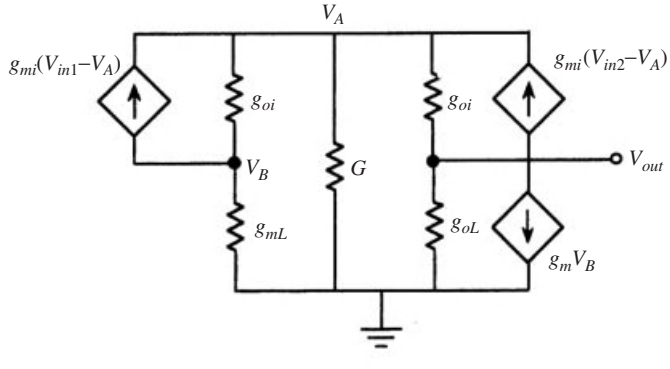
$$A_{CM} \cong \frac{-Gg_{dsi}}{2g_{mL}(g_{dL} + g_{dsi})} \tag{12.37}$$

so that

$$CMRR \cong 2\frac{g_{mi}g_{mL}}{Gg_{dsi}} \tag{12.38}$$



(a)



(b)

Figure 12.9 (a) CMOS differential amplifier. (b) Equivalent circuit

If $g_{dsL} = g_{dsi} = g_o$, then

$$A_d \cong \frac{g_{mi}}{2g_0}$$
$$A_{CM} \cong \frac{G}{4g_{mL}} \tag{12.39}$$

so that

$$CMRR \cong \frac{2g_{mi}g_{mL}}{Gg_0} \tag{12.40}$$

Also

$$r_{out} \cong \frac{1}{g_{dL} + g_{dsi}} \cong \frac{1}{2g_0} \tag{12.41}$$

12.5 The Two-stage CMOS Op Amp

In order to increase the gain, a two stage topology as shown in Figure 12.10 may be used, leading to the two-stage CMOS operational amplifier. It has the following properties:

1. Q_8 and Q_5 form a current mirror supplying the differential pair Q_1, Q_2 with the bias current.
2. The aspect ratio of Q_5 is chosen to give the required input stage bias.
3. Q_3 and Q_4 form a current mirror as an active load for the input differential pair.
4. The second stage is a common-source amplifier consisting of Q_6 actively loaded by the current source transistor Q_7.
5. The capacitor C_c is used for frequency compensation.
6. The resistor R_z is usually realized using transistors and is employed to control the location of the zero of the amplifier transfer function.

The equivalent circuit of the Op Amp is shown in Figure 12.11 in which C_1 is the total capacitance at the interface between the first and second stages, while C_2 is the total capacitance at the output node, including the load capacitance C_L. Thus C_1 and C_2 include the various parasitic capacitances discussed in an earlier section.

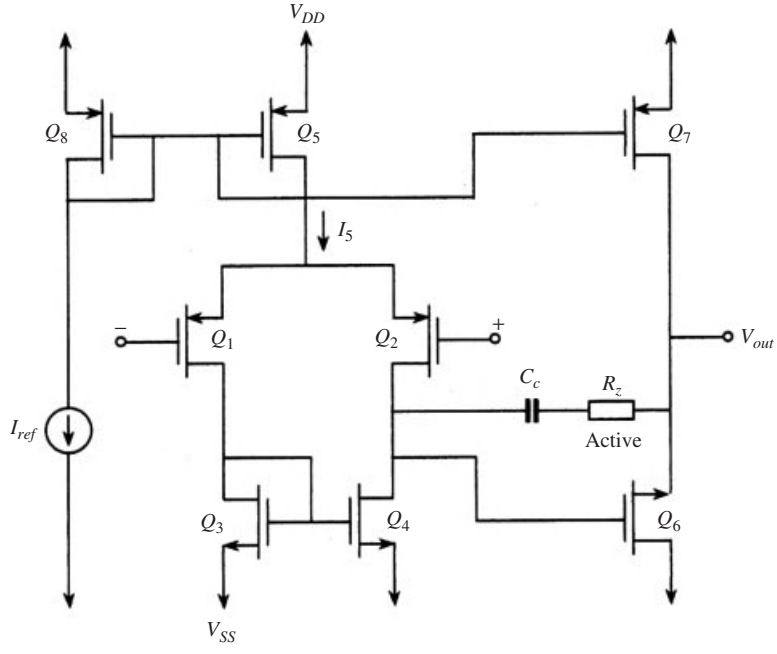


Figure 12.10 Basic two-stage CMOS Op Amp with compensation

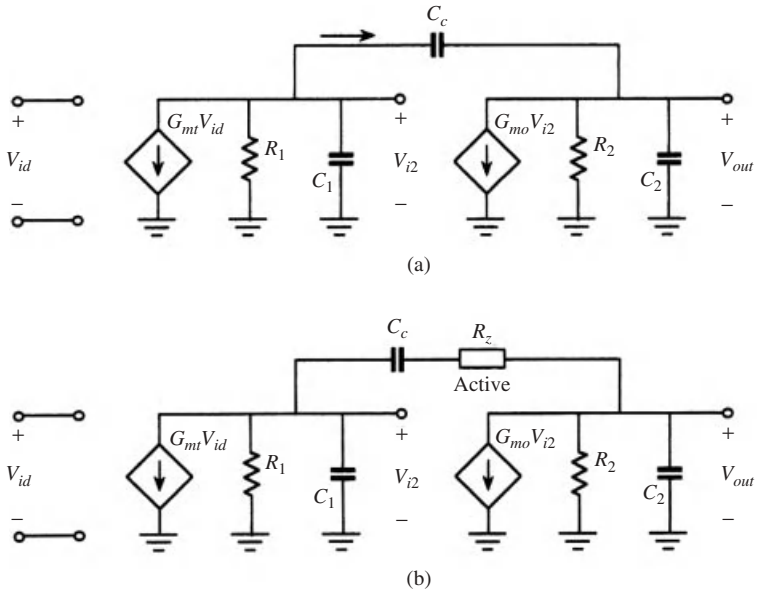


Figure 12.11 Equivalent circuit of the two-stage Op Amp in Figure 12.10: (a) without the nulling resistor, (b) with the nulling resistor

Also in Figure 12.11

$$\begin{aligned} G_{mi} &= g_{m1} = g_2 \\ R_1 &= r_{02} \| r_{o4} = \frac{1}{g_{ds2} + g_{ds4}} = \frac{2}{I_5(\lambda_2 + \lambda_4)} \\ G_{m0} &= g_6 \\ R_2 &= r_{o6} \| r_{07} = \frac{1}{g_{ds6} + g_{ds7}} = \frac{1}{I_6(\lambda_6 + \lambda_7)} \end{aligned} \tag{12.42}$$

where λ_i is the channel length modulation parameter for each transistor.

12.5.1 The dc Voltage Gain

Direct analysis of the circuit gives for the dc voltage gain

$$\begin{aligned} A_v &= \frac{g_{m2} g_{m6}}{(g_{ds2} + g_{ds4})(g_{ds6} g_{ds7})} \\ A_v &= \frac{2 g_{m2} g_{m6}}{I_5 I_6 (\lambda_2 + \lambda_4)(\lambda_6 + \lambda_7)} \end{aligned} \tag{12.43}$$

12.5.2 The Frequency Response

The frequency response of the amplifier can be evaluated using the equivalent circuit in Figure 12.11. Analysis of the circuit is undertaken with the assumptions of $C_1 \ll C_c$,

$C_1 \ll C_L$, $C_1 \ll C_2$ and $C_2 \approx C_L$. For the circuit in Figure 12.11(a) this gives poles at

$$s = -\omega_{p1} \cong \frac{1}{G_{m0} R_2 C_c R_1} \cong -\frac{G_{mi}}{A_v C_c} \tag{12.44}$$

$$s = -\omega_{p2} \cong -\frac{G_{m0} C_c}{C_1 C_2 + C_c (C_1 + C_2)} \cong -\frac{G_{m0}}{C_L} \tag{12.45}$$

and a zero at

$$s = z_1 = \frac{G_{m0}}{C_c} \tag{12.46}$$

To make ω_{p1} the dominant pole, we let it approximate the 3 dB point so that

$$A_v \omega_{p1} \cong A_v \omega_{3dB} = \omega_t \tag{12.47a}$$

or

$$GB = \omega_t = \frac{G_{m_i}}{C_c} \tag{12.47b}$$

which is the *unity gain frequency* or the *gain-bandwidth product*.

Now, since G_{mi} is of the same order as G_{m0}, the zero will be close to ω_t and introduces a phase shift that will decrease the phase margin, thus impairing the amplifier stability. If the zero is placed at least ten times higher than ω_t, then in order to achieve a desirable 60^0 phase margin, the second pole ω_{p2} must be placed higher than ω_t by a factor of at least 2.2. From these requirements the following relations are obtained

$$\frac{g_{m6}}{C_c} > 10 \left(\frac{g_{m2}}{C_c} \right) \tag{12.48}$$

that is

$$g_{m6} > 10 g_{m2} \tag{12.49}$$

Also

$$\frac{g_{m6}}{C_L} > 2.2 \frac{g_{m2}}{C_c} \tag{12.50}$$

which leads to the following condition

$$C_c > 0.22 C_L \tag{12.51}$$

12.5.3 The Nulling Resistor

Another method for eliminating the effect of the RHP zero is to add the nulling resistor R_z (made up of MOSFETs) as indicated in Figure 12.11(b) in series with the compensating capacitor. The new zero location becomes

$$s = z_1 = \frac{1}{C_c \left(\frac{1}{G_{m0}} - R_z \right)} \tag{12.52}$$

which $\rightarrow \infty$ as $R_z \rightarrow 1/G_{m0}$. A good choice would be $R_z > 1/G_{m0}$, placing the zero on the negative real axis so that it adds to the phase margin. A consequence of the addition of the nulling resistor is that a third pole results at the location

$$\omega_{p3} = -\frac{1}{R_z C_1} \tag{12.53}$$

Alternatively, the resistor R_z can be chosen such that the RHP zero coincides with the pole ω_{p2}, that is a pole-zero cancellation is achieved. This gives

$$R_z = \frac{1}{g_{m6}} \left(\frac{C_L + C_c}{C_c} \right) \tag{12.54}$$

The resistor is realized as a transistor Q_8 operating in the linear (triode) region in which $0 < V_{GS} - V_t < V_{DS}$. Figure 12.12 shows a complete Op Amp circuit, including the bias arrangement for Q_8 as well as for the entire amplifier in which we have changed the input devices to NMOS type.

The value of the nulling resistor is

$$R_z = \frac{1}{g_{ds8}} \tag{12.55}$$

where

$$g_{ds8} = K_8' \left(\frac{W}{L} \right)_8 (|V_{GS8}| - |V_t|) \tag{12.56}$$

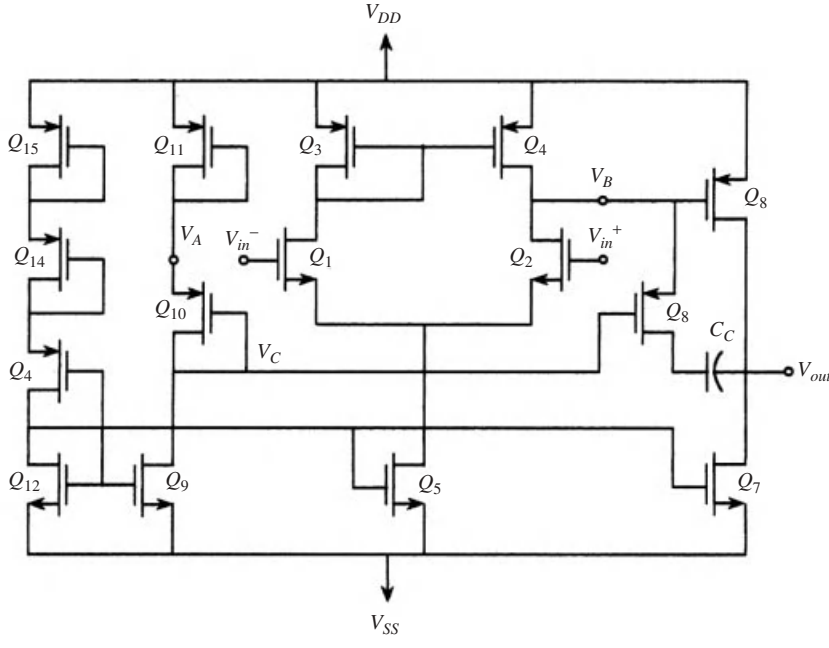


Figure 12.12 CMOS Op Amp with nulling resistor and compensation, including the bias circuits

The bias circuit for Q_8 consists of the devices Q_9, Q_{10} and Q_{11} as shown in Fig. 12.12. The circuit is designed such that $V_A = V_B$ resulting in

$$\left(\frac{W}{L}\right)_8 = \left[\left(\frac{W}{L}\right)_6 \left(\frac{W}{L}\right)_{10} \left(\frac{I_6}{I_{10}}\right)\right]\left[\frac{C_c}{C_L + C_c}\right] \tag{12.57}$$

In order to satisfy (12.55) V_{GS11} and V_{GS6} must be equal. This gives the required aspect ratio of Q_{11} as

$$\left(\frac{W}{L}\right)_{11} = \left(\frac{I_{10}}{I_6}\right)\left(\frac{W}{L}\right)_6 \tag{12.58}$$

With the bias circuit for the simulated resistor, the overall power consumption is given as

$$P_{diss} = (I_5 + I_6 + I_9 + I_{12})(V_{DD} + |V_{SS}|) \tag{12.59}$$

12.5.4 The Slew Rate and Settling Time

The slew rate (S_R) of the CMOS OP Amp is limited by the differential stage and is determined by the maximum current that can be sunk or sourced into C_c which is $I_{3(\max)} = I_5$. Thus

$$S_R = \frac{I_5}{C_c} \tag{12.60}$$

Therefore, in order to obtain the required slew rate, I_5 has to be fixed accordingly, since the value of C_c is determined from the requirement on the phase margin and stability of the amplifier. With

$$\omega_t = \frac{g_{m1}}{C_c} = \frac{I_5}{(|V_{GS}| - |V_t|)C_c} \tag{12.61}$$

where $|V_{GS}|$ is the gate-to-source voltage of Q_1 and Q_2, we have

$$S_R = (|V_{GS}| - |V_t|)\omega_t \tag{12.62}$$

If the slew rate is not specified, but instead the required settling time T_s is given, then an equivalent slew rate value is estimated based on the approximation that the amplifier slews half the supply rail voltage at a rate several times faster than the settling time. This leads to the estimate of the slew rate S_R' given by

$$S_R' = \alpha\left(\frac{V_{DD} + |V_{SS}|}{2T_s}\right) \tag{12.63}$$

where α is in the range 2–10. If both T_s and S_R are specified, then S_R' is calculated and the worse case value is used to evaluate I_5.

12.5.5 The Input Common-mode Range and CMRR

Now, the input common-mode range and CMRR are also determined by the differential stage. For the p-channel input differential stage, the lowest possible input-voltage (negative

CMR) at the gate of Q_1 or Q_2 is given by

$$V_{in(\min)} = V_{SS} + V_{GSS} + V_{SD1} - V_{sG1} \tag{12.64}$$

In saturation, the minimum value of V_{SD1} is

$$V_{SD1} = V_{SG1} - |V_{t1}| \tag{12.65}$$

So that

$$V_{in(\min)} = V_{SS} + V_{GS3} - |V_{t1}| \tag{12.66}$$

Using $V_{GS} = (2I_{DS}/2K)^{1/2} + V_t$ and $2I_1 = I_5$ leads to

$$V_{in(\min)} = V_{ss} + \left(\frac{I_5}{2K_3}\right)^{1/2} + V_{to3} - |V_{t01}| \tag{12.67}$$

where V_{t0} is the threshold voltage for $V_{DS} = 0$. In the above expression, the first two terms are determined by the designer. The last ones are fixed by the process and the way the substrate is connected for Q_1. Assuming an n-well process with the sources of Q_1 and Q_2 connected to this well, expression (12.67) becomes

$$V_{in(\min)} = V_{ss} + \left(\frac{I_5}{2K_3}\right)^{1/2} + V_{t03(\max)} - |V_{t01(\min)}| \tag{12.68}$$

in which the worst case V_t spread as specified by the process is used by the designer to adjust I_5 and K_3. In this case, the spread is a high n-channel threshold and a low p-channel threshold.

Similar analysis is used to obtain the highest possible input voltage (positive CMR) as

$$V_{in(\max)} = V_{DD} + V_{SD5} - |V_{SG1}| \tag{12.69}$$

or

$$V_{in(\min)} = V_{DD} - V_{SD5} - \left(\frac{I_5}{2K_1}\right)^{1/2} - |V_{t01(\max)}| \tag{12.70}$$

The above expressions allow the designer to maximize the common-mode range by making the aspect ratios of Q_1 (and Q_2) as well as Q_3 (and $Q4$) as large as possible and minimizing V_{SD5}. Also, a small I_5 leads to a large CMR.

The corresponding expressions for the n-channel differential input stage are obtained from the above by interchanging $V_{in(\min)}$ and $V_{in(\max)}$. Thus for the n-channel input differential stage we have

$$V_{in(\max)} = V_{DD} - \left(\frac{I_5}{2K_3}\right)^{1/2} - |V_{t03(\max)}| + |V_{t01(\min)}| \tag{12.71}$$

$$V_{in(\min)} = V_{ss} + V_{SD5}\left(\frac{I_5}{2K1}\right)^{1/2} + |V_{t01(\max)}| \tag{12.72}$$

The CMRR is determined by the differential stage and the expression is given by (12.40).

The two-stage CMOS Op Amp discussed so far is very popular in the design of analog filters and other signal processing systems. It gives good performance when the loads are capacitive of low value. If the two-stage Op Amp is required to drive large loads, a buffer stage must be added which provides a low output impedance for the Op Amp. Without this stage, a large capacitive load causes the non-dominant pole to decrease, thus decreasing the phase margin. Also without the buffer stage, a large resistive load will decrease the open-loop gain.

We now give a summary of the design equations of the two-stage Op Amp.

12.5.6 Summary of the Two-stage CMOS Op Amp Design Calculations

The process parameters, the supply voltage and the temperature range are fixed conditions for the Op Amp. In addition, the specifications are usually given in terms of specifications of the following parameters:

Dc gain $= A_v$
Unity-gain bandwidth $= f_t$

Input common-mode range:

Slew rate: S_R
Load capacitance: C_L
Settling time: T_s
Output voltage swing: $V_{0\max}$, $V_{0\min}$
Power dissipation $= P_{ss}$

The design steps with reference to Figure 12.12 are as follows:

1. Select the smallest channel length that will keep the channel length modulation parameter constant and give good matching for current mirrors.
2. Calculate the minimum compensation capacitor according to

$$C_c > 0.22C_L \tag{12.73}$$

   the lower limit gives 60° phase margin.
3. Find I_5 from the slew rate and/or settling time requirements as

$$I_5 = \max\left[(S_R C_c), \alpha\left(\frac{V_{DD} + V_{SS}}{2T_s} C_c\right)\right] \tag{12.74}$$

4. For an n-channel differential input stage, find $(W/L)_3$ from the specifications on the maximum input voltage according to the relation

$$(W/L)_3 = \frac{I_5}{K_1'[V_{DD} - V_{i\max} - |V_{t03}|_{\max} + V_{t1\min}]^2} \geq 1 \tag{12.75}$$

5. Find $(W/L)_2$ to meet the required value of the unity gain bandwidth f_t

$$g_{m2} = \omega_t . C_c \tag{12.76}$$

so that

$$(W/L)_2 = \frac{g_{m2}^2}{K_2' I_5} \tag{12.77}$$

6. For an n-channel input stage, find $(W/L)_5$ from the specification on the minimum voltage $V_{in(\min)}$, in the following two steps:
   (a) The saturation voltage of Q_5 is

$$V_{DS5(sat)} = V_{in(\min)} - V_{SS} - \left[\frac{I_5}{2K_1}\right]^{0.5} - V_{t1(\max)} \tag{12.78}$$

   (b) The aspect ratio of Q_5 is

$$(W/L) = \frac{2I_5}{K_5'[V_{DS5(sat)}]^2} \tag{12.79}$$

7. Find $(W/L)_6$ by choosing the second pole ω_{p2} to be $= 2.2\omega_t$, that is

$$g_{m6} = 2.2 g_{m_2}(C_L/C_C) \tag{12.80}$$

   and assuming

$$V_{DS6} = V_{DS6(\min)} = V_{DS6(sat)} \tag{12.81}$$

   so that

$$(W/L)_6 = \frac{g_{m6}}{K_6' V_{DS6(sat)}} \tag{12.82}$$

8. Find I_6 according to the following relation

$$I_6 = \max\left[\frac{g_{m6}^2}{2K_6'(W/L)_6}, \frac{(W/L)_6}{(W/L)_3} I_1\right] \tag{12.83}$$

9. Find $(W/L)_7$ to achieve the required current ratios

$$\frac{(W/L)_7}{(W/L)_5} = \frac{I_6}{I_5} \tag{12.84}$$

10. Check the gain and power consumption from

$$A_v = \frac{2g_{m2}g_{m6}}{I_5(\lambda_2 + \lambda_3)I_6(\lambda_6 + \lambda_7)} \tag{12.85}$$

$$P_{diss} = (I_5 + I_6)(V_{DD} + |V_{SS}|) \tag{12.86}$$

11. If the gain specifications is not met, then increase $(W/L)_2$ and/or $(W/L)_6$. Alternatively, I_5 and I_6 may be decreased.
12. If the power dissipation is too high, the only solution is to reduce I_5 and I_6. However, this may require a corresponding increase of some of the (W/L) ratios for satisfying the input and output swings.

The Bias Circuit

The design of the bias circuit is now considered. First, $V_{GSS} = V_{GS\,12}$ is calculated. Then, the following condition has to be satisfied

$$V_{DS\,15} + V_{DS\,14} + V_{DS\,13} + V_{DS\,12} = V_{DD} + |V_{SS}| \tag{12.87}$$

The aspect ratios of the transistors $Q_{12} - Q_{15}$ and the biasing current I_{12} are chosen such that a suitable configuration is obtained. A reasonable arrangement could be to set the aspect ratios of the p-channel transistors $Q_{13} - Q_{15}$ to unity, then calculate the current and aspect ratio of Q_{12}.

If compensation for the effect of the RHP zero is desired, using the nulling resistor, then the procedure is as follows:

1. In order to establish the biasing current (set $V_A = V_B$), Q_3 and Q_{11} and their drain currents are matched

$$\left(\frac{W}{L}\right)_{11} = \left(\frac{W}{L}\right)_3 \tag{12.88}$$

$$I_{11} = I_{10} = I_3 \tag{12.89}$$

2. The aspect ratio of Q_{10} is not dependent on the other components, so it can be chosen with minimum value.
3. The aspect ratio of Q_9 is determined by the ratios of the two currents I_9 and as

$$\left(\frac{W}{L}\right)_9 = \left(\frac{I_{10}}{I_5}\right)\left(\frac{W}{L}\right)_5 \tag{12.90}$$

4. From (12.57) the aspect ratio of Q_8 is determined.
5. Once the compensation circuit is designed, it may be useful to check the location of the RHP zero. First, $V_{GS\,8}$ is calculated as

$$|V_{GS\,8}| = |V_{GS\,10}| = \left[\frac{2I_{10}}{K'_{10}(W/L)_{10}}\right]^{1/2} + |V_T| \tag{12.91}$$

Then R_z is calculated using (12.55) and (12.56). This is used to calculate the zero location by (12.52) and if the zero cancellation has been successful, this value should be equal to (12.54).

12.6 A Complete Design Example

Consider the design of a two-stage CMOS Op Amp with the following specifications:

- $A_v >$ several thousands
- GB = 1 MHz
- $S_R > 3\,\text{V}/\mu\text{s}$
- CMR = ± 3 V
- $\text{C}_\text{L} = 22.5$ pF
- Supply voltage = ± 5 V
- Output voltage swing = ± 4 V
- $\text{P}_{diss} = 10$ mW

1. The process parameters are

$$K_p' = 2.4 \times 10^{-5}\,\text{A/V}^2,\ K_n' = 5.138 \times 10^{-5}\,\text{A/V}^2$$
$$\lambda_p = 0.01\,\text{V}^{-1}, \lambda_n = 0.02\,\text{V}^{-1}$$
$$V_{tp} = 0.9\,\text{V}, V_{tn} = 0.865\,\text{V}$$

2. A uniform channel length is chosen at 10 µm, say. *This taken for illustrative purposes only. Many present designs go all the way down to submicron, deep submicron and ultra deep submicron range with channel lengths as small as 65 nm. The future will bring even smaller values*.
3. The minimum value for C_C is calculated as

$$C_C = 0.22 \times C_L = 0.22 \times 22.5\,\text{pF} = 4.95\,\text{pF}$$

and adjusted to 6 pF.
4. The minimum value for the total current I_5 is

$$I_5 = S_R \times C_C = 3\frac{\text{V}}{\mu\text{s}} \times 6\,\text{pF} = 18\,\mu\text{A}$$

5. The aspect ratio of Q_3 is calculated as

$$\left(\frac{W}{L}\right)_3 = \left(\frac{18 \times 10^{-6}}{(2.4 \times 10^{-5})(5 - 3 - 0.9 + 0865)^2}\right) = 0.1943$$

which is increased to unity.
6. We obtain the transconductance g_{2m} which is necessary to calculate the aspect ratio of Q_2

$$g_m = GB \times C_C = (2 \times 10^6\pi)(6 \times 10^{-12}) = 37.69\,\mu\text{S}$$
$$\left(\frac{W}{L}\right)_2 = \frac{(37.69 \times 10^{-6})^2}{(5.138 \times 10^{-5})(18 \times 10^{-6})} = 1.5367$$

7. The saturation voltage of Q_5 is given by

$$V_{DS\,5(sat)} = \left[-3 + 5 - \sqrt{\frac{18 \times 10^{-6}}{(5.138 \times 10^{-5})(1.54)}} - 0865\right] = 6.57.53\,\text{mV}$$

The aspect ratio is

$$\left(\frac{W}{L}\right)_5 = \frac{36 \times 10^{-6}}{(5.138 \times 10^{-5})(657.53 \times 10^{-5})^2} = 1.6232$$

8. The transconductance g_{m6} and the aspect ratio of Q_6 are given by

$$g_{m6} = (2.2 \times 37.69\,\mu\text{s})\left(\frac{2.2}{6}\right) = 310\,\mu\text{s}$$
$$\left(\frac{W}{L}\right)_6 = \frac{311 \times 10^{-6}}{2.4 \times 10^{-5}} = 12.95$$

9. $I_6 = \dfrac{(311 \times 10^{-6})^2}{(2)(2.4 \times 10^{-5})(12.95)} = 155.54\,\mu\text{A}$ or

$$I_6 = \left(\frac{12.95}{1}\right)(9 \times 10^{-6}) = 116.5\,\mu\text{A}$$

and the larger value is chosen.

10. The aspect ratio of Q_7 is

$$\left(\frac{W}{L}\right)_7 = \frac{155.5}{18} \times 1.62 = 13.97$$

11. The gain obtained is calculated as

$$A = \frac{2(37.69 \times 10^{-6})(276.32 \times 10^{-6})}{(18 \times 10^{-6})(0.01 + 0.02)(115.54 \times 10^{-6})(0.01 + 0.02)} = 9334$$

The aspect ratios of the initial design are as follows

$$\left(\frac{W}{L}\right)_1 = \left(\frac{W}{L}\right)_2 = \frac{15}{10}$$

$$\left(\frac{W}{L}\right)_3 = \left(\frac{W}{L}\right)_4 = \frac{10}{10}$$

$$\left(\frac{W}{L}\right)_5 = \frac{16}{10}$$

$$\left(\frac{W}{L}\right)_6 = \frac{130}{10}$$

$$\left(\frac{W}{L}\right)_7 = \frac{150}{10}$$

The calculations for the RHP zero compensation using the nulling resistor are as follows:

1. $\left(\dfrac{W}{L}\right)_{11} = 1.5$
2. The currents are matched so that

$$I_{11} = I_{10} = I_3 = 9\,\mu\text{A}$$

3. The aspect ratio of Q_{10} is essentially free, so

$$\left(\frac{W}{L}\right)_{10} = 1$$

$$\left(\frac{W}{L}\right)_9 = \frac{9}{18}1.6 = 0.8$$

4. The aspect ratio of Q_8 is

$$\left(\frac{W}{L}\right)_8 = \sqrt{12 \times 1 \times \frac{155}{9}} \times \frac{6}{6 + 22.5} = 3$$

12.7 Practical Considerations and Other Non-ideal Effects in Operational Amplifier Design

In addition to the non-idea effects which have been discussed so far, there are others of which the designer should be aware and attempt to minimize. Some of these effects can be reduced by careful design, but for a considerable improvement, special circuit techniques have to be employed as will be discussed in the next chapter. Here, we point out the most important non-ideal effects other than those discussed earlier.

12.7.1 Power Supply Rejection

The power supply rejection ratio (PSRR) is of great importance in the use of MOS circuits to implement analog signal processing systems, in particular when switched-capacitor techniques are used. First, the clock signals couple to the power supply rails. Secondly, if digital circuits coexist on the same chip, digital noise also couples to the supply lines. If these types of noise couple to the signal paths, they are aliased into the useful frequency bands resulting in degradation of the signal to noise ratio.

In the case of an Op Amp, the PSRR is the ratio o the voltage gain from input to output to the gain from the supply to the output. The basic two-stage Op Amp is particularly prone to high frequency noise from the negative power supply. This is because as the frequency increases, the impedance of the compensation capacitor decreases, effectively connecting the drain of Q_6 to its gate. In this case, the gain from the negative supply to the output approaches unity. The same mechanism causes the gain from the positive supply to fall with frequency as the open-loop gain does, therefore, the positive PSRR remains relatively constant with frequency. The negative supply PSRR falls to about unity at the unity gain frequency of the amplifier.

12.7.2 dc Offset Voltage

This consists of two components, random and systematic. Random offset voltages result from mismatches of theoretically identical devices. The systematic type is a result of the circuit design and will always exist if all devices were perfectly matched. Techniques for reducing dc offset voltages will be discussed in a later chapter.

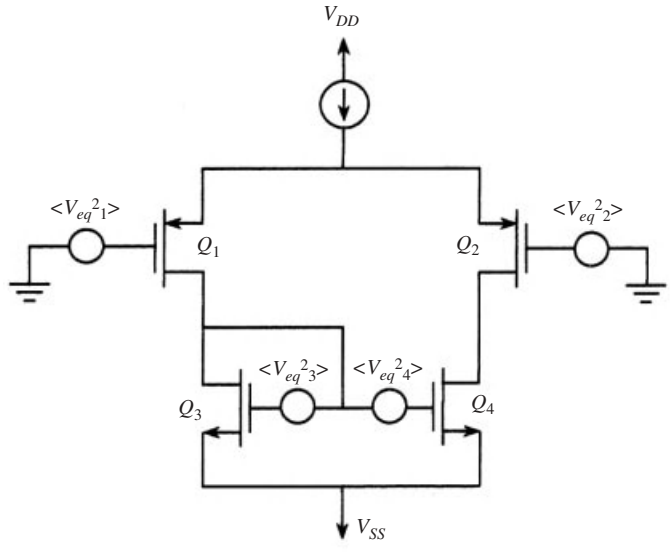
12.7.3 Noise Performance

Due to the relatively high $1/f$ noise of MOS transistors, the noise performance of CMOS amplifiers is an important design consideration. In Figure 12.13(a) the four transistors in the input differential stage contribute to the equivalent input noise as depicted in Figure 12.13(b).

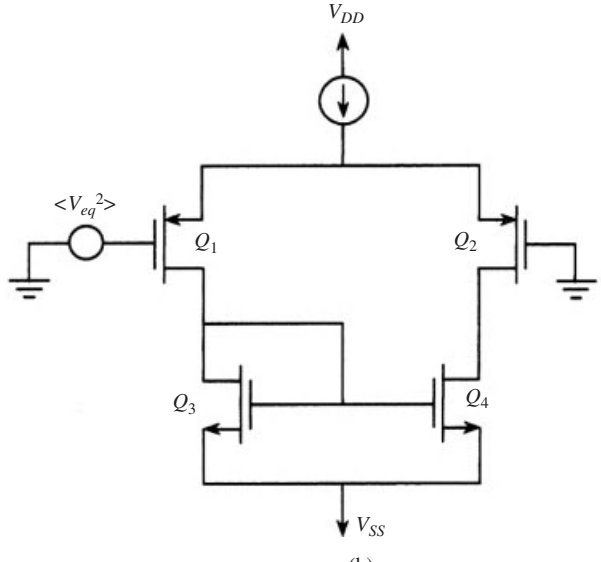
By direct calculation of the output noise for each circuit and equating the results, we obtain

$$\langle v_{eq}^2 \rangle = \langle v_{eq1}^2 \rangle + \langle v_{eq2}^2 \rangle + (g_{m3}/g_{m1})(\langle v_{eq3}^2 \rangle + \langle v_{eq4}^2 \rangle) \tag{12.92}$$

in which it is assumed that $g_{m1} = g_{m2}$, $g_{m3} = g_{m4}$. It follows that the input stage devices contribute directly to the noise, while the contribution of the load devices is reduced by the ratio of their transconductance to those of the input devices.



(a)



(b)

Figure 12.13 CMOS Op Amp input stage noise calculations: (a) device contributions, (b) equivalent input noise

12.7.3.1 Input Referred 1/*f* Noise

The equivalent flicker noise density of a typical MOS transistor is given by (10.40). Then, (12.92) can be used to find for the Op Amp.

$$\langle v_{1/f}^2 \rangle = \frac{2K_{fp}}{W_1 L_1 C_{ox}} \left(1 + \frac{K_{fn}\mu_n L_1^2}{K_{fp}\mu_p L_3^2}\right)\left(\frac{\Delta f}{f}\right) \tag{12.93}$$

where K_{fp} and K_{fn} are the flicker noise coefficients of the p-channel and n-channel, respectively, whose relative values are determined by the process details. The first term is the noise due to the input devices, while the second term is the increase due to the load devices. It is clear that the contribution to the noise by the load devices can be reduced by taking the channel lengths of the load devices longer than those of the input devices. Then, the input devices can be made wide enough to achieve the desired performance. Note that, however, increasing the width of the channel of the load devices does not reduce the $1/f$ noise.

12.7.3.2 Thermal Noise

The *input-referred thermal noise* of a MOS transistor is given by

$$\langle v_n^2 \rangle = 4kT \left(\frac{2}{3g_m} \right) \Delta f \tag{12.94}$$

Using the same procedure as in the case of flicker noise, we have for the Op Amp

$$v_{eq}^2 = 4kT \left(\frac{4/3}{\sqrt{2\mu_p C_{ox} (W/L)_1 I_D}} \right) \left(1 + \sqrt{\frac{\mu_n (W/L)_3}{\mu_p (W/L)_1}} \right) \tag{12.95}$$

The first term represents the thermal noise from the input devices, and the second term in the parentheses represents the increase in noise due to the loads. The latter will be small if the aspect ratios are chosen such that the transconductances of the input devices are much larger than those of the loads. Under this condition, the input noise is determined by the transcondunctances of the input devices.

12.8 Conclusion

The design of two-stage CMOS operational amplifiers was the focus of the discussion in this chapter, together with the inherent non-ideal effects. In analog integrated circuits for signal processing, Op Amp design techniques occupy much time and effort. In particular, the quest for high performance Op Amp designs is a major occupation of integrated circuit design engineers. The relatively simple design of this chapter is satisfactory in many applications, and it also forms the basis for the more elaborate techniques which will be discussed in the next chapter.

Problems

12.1 A basic differential pair employs NMOS transistors with r_{ds} 100 Ω, $K' = 100\,\mu A/V^2$ and $V_t = 1\,V$. Calculate the differential gain for bias current values of 50 μA, 100 μA and 200 μA.

12.2 Design the input differential stage of an operational amplifier to operate at $V_{GS} = 1.3\,V$ and provide a transconductance of 0.1 mA/V. For the device, $V_t = 1\,V$ and $K' = 10\,\mu A/V^2$. For the design, it is required to find the aspect ratios of the devices as well as the bias current.

12.3 Derive the exact expression for the transfer function of the two-stage Op Amp of Figures 12.10 and 12.11. Show that the pole and zero locations are approximately as given in Section 12.5 under the assumptions made.

12.4 Design a two-stage CMOS Op Amp with the following specifications:

- Low-frequency gain >2000
- Settling time $= 2\,\mu s$
- Unity gain frequency $= 1\,MHz$
- Load capacitance $= 10\,pF$
- Supply voltages $= \pm 5\,V$
- CMR $= \pm 4\,V$
- Output swing $= \pm 4\,V$
- Power dissipation $<20\,mW$
- Device parameters:

$$K_p' = 20\,\mu A/V^2,\ K_n' = 50\,\mu A/V^2$$

$$\lambda_p = 0.01\,V^{-1},\ \lambda_n = 0.01\,V^{-1}$$

13

High Performance CMOS Operational Amplifiers and Operational Transconductance Amplifiers

13.1 Introduction

Although the basic two-stage Op Amp discussed in the previous Chapter has a performance which is satisfactory in many applications, it suffers from a number of disadvantages which were pointed out. Improvements in the performance with regards to one or more of theses non-ideal effects, such as obtaining higher gain, better PSRR, reduced offset voltage, lower noise, better settling time and slew rate, require special techniques which may entail modification of the Op Amp structure. These techniques are discussed in this chapter [22–24]. We also give integrated circuit realizations of operational transconductance amplifiers (OTAs) which are used with great advantage in high-frequency applications, including submicron and deep submicron integrated circuit design [25, 26].

13.2 Cascode CMOS Op Amps

Figure 13.1 shows the first stage of a cascode Op Amp constructed with the main objective of increasing the gain. The two common-gate transistors Q_{1c} and Q_{2c} form the cascode for the differential pair Q_1, Q_2. The output resistance at Q_{2c} is

$$R_{o2c} \cong g_{m2c} r_{o2c} r_{o2} \tag{13.1}$$

which is much higher that the value without the cascode transistors. Naturally, in order to make full use of the high output resistance, we must also increase the active load resistance; thus a Wilson current mirror is used in Figure 13.1.

The output resistance of the Wilson current mirror was given in Chapter 11 as

$$R_{04c} \cong g_{m4c} r_{o4c} r_{o3} \tag{13.2}$$

Signal Processing and Integrated Circuits, First Edition. Hussein Baher.
© 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.

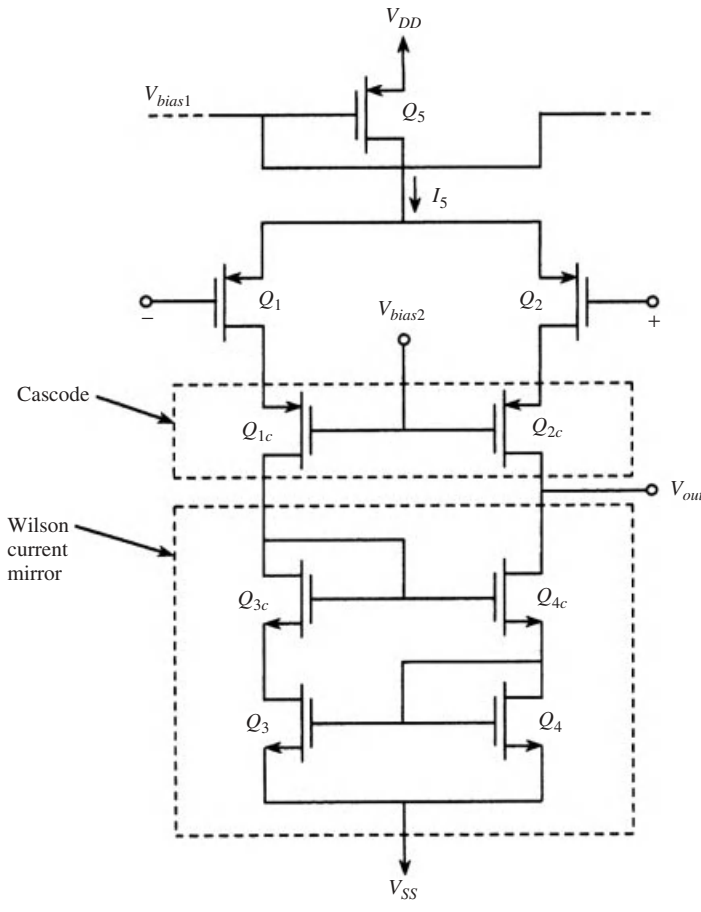


Figure 13.1 Input stage of a CMOS Op Amp employing the cascode configuration

Therefore, the output resistance of the stage is

$$R_o = R_{o2c} \| R_{o4c} \tag{13.3}$$

But the voltage gain of the stage is

$$A_1 = -g_m R_o \tag{13.4}$$

Therefore, the increase in the output resistance in reflected as an increase in the gain by the same factor. In fact, since the gain from this stage is quite high, a single stage Op Amp is possible using a modification of the cascode idea. This is discussed below, and it also leads to a high power supply rejection ratio and a better input common-mode range.

13.3 The Folded Cascode Op Amp

Starting from Figure 13.1, each of the six transistors below Q_1, Q_2 is replaced by its complement and the entire group is disconnected from V_{ss} and folded over to be connected

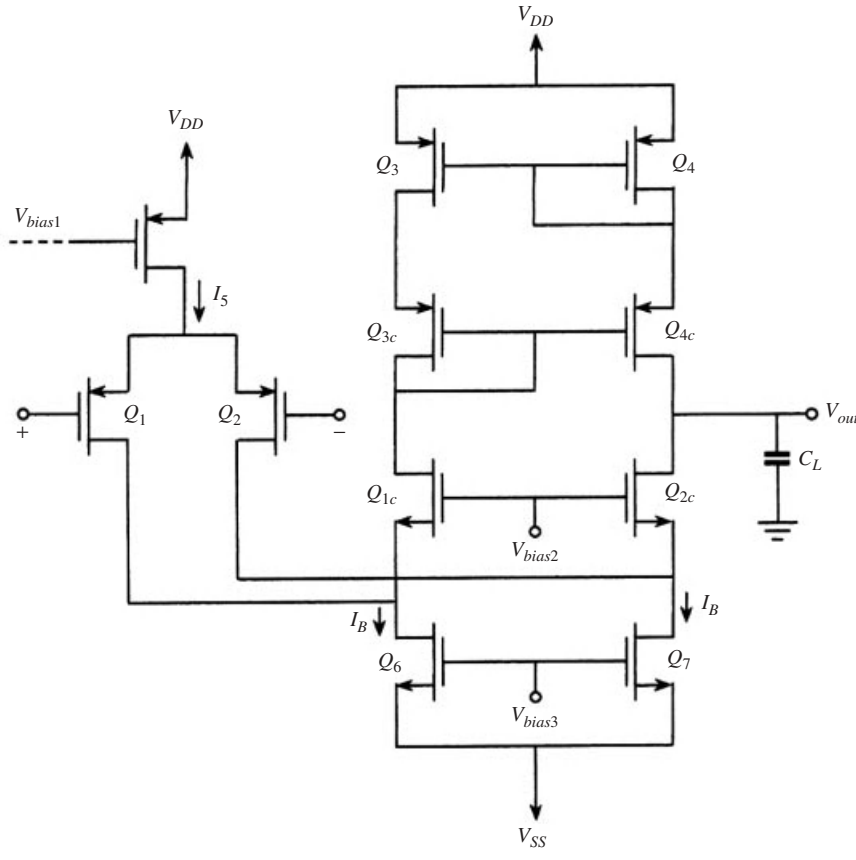


Figure 13.2 A folded cascode Op Amp

to V_{DD}, resulting in the folded cascode of Figure 13.2. The operation of the circuit is similar to the simple cascode except that the input common-mode range is larger since only three transistors are stacked in the input between the two power supplies by comparison with five in the original cascode. Q_6 and Q_7 are added current sources.

The voltage gain of the folded cascode is given by

$$A = g_{m1}R_o \tag{13.5}$$

where R_o is its output resistance given by

$$R_o = R_{o2c} \| R_{o4c} = [g_{m2c}r_{o2c}(r_{o7} \| r_{o2})] \| [g_{m4c}r_{o4c}r_{o3}] \tag{13.6}$$

and due to the high gain value, the circuit can be used as a single stage Op Amp. In fact if the devices are matched in pairs, with Q_6 and Q_7 used to establish $I_B = I_5$, the gain of the cascode can be put in the explicit form

$$A = \frac{|V_A|^2}{I_5} \frac{\sqrt{\mu_n \mu_p} C_{ox} \sqrt{W_1/L_1}}{3/\sqrt{(W_{2c}/L_{2c}) + \sqrt{\mu_n/\mu_p}/\sqrt{(W_{4c}/L_{4c})}}} \tag{13.7}$$

The dominant pole is determined by the total capacitance at the output node C_L which includes the load capacitance. This is given by

$$\omega_d = 1/R_o C_L \tag{13.8}$$

and the unity gain frequency is

$$\omega_t = A\omega_d = \frac{g_{m1}}{C_L} \tag{13.9}$$

The folded cascade has a better power supply rejection ratio than the two-stage Op Amp since the compensation capacitor and load capacitor are the same element in this case. Assuming that the load capacitance or part of it is not connected to the power supply, the circuit does not suffer from the degradation of the high frequency power rejection problem inherent in the compensated two-stage Op Amp discussed in the previous chapter. However, due to the fact that the cascode transistors are used at the output, the output swing of this circuit is lower than that of the two-stage amplifier; this disadvantage will be remedied in a later section. This Op Amp structure is commonly used in high-frequency switched-capacitor filters and other high-frequency applications.

13.4 Low-noise Operational Amplifiers

The signal to noise ratio (S/N) is of primary importance in communication circuits. This is closely related to the dynamic range which is the ratio of the maximum to the minimum signals which the circuit can process without distortion. The maximum is usually determined by the power supplies and the large signal swing limits. The minimum is determined by the noise or ripple injected by the power supply. In the design of Op Amps for good noise performance, two approaches are possible. The first consists in optimizing the device geometries and characteristics to yield the lowest noise possible; the relevant concepts were presented in Chapter 11. The second approach uses independent techniques such as correlated double sampling and chopper stabilization which also reduce the input offset voltage. We first give an example of the first approach, then introduce the new techniques.

13.4.1 Low-noise Design by Control of Device Geometries

A low-noise amplifier is shown in Figure 13.3(a) which is the two stage design with cascode devices Q_8, Q_9 to improve the PSRR. The input differential stage is composed of PMOS devices due to their better noise performance. The noise model of the Op Amp is shown in Figure 13.3(b) in which the noise due to the dc current sources is ignored since their gates are usually connected to a low impedance.

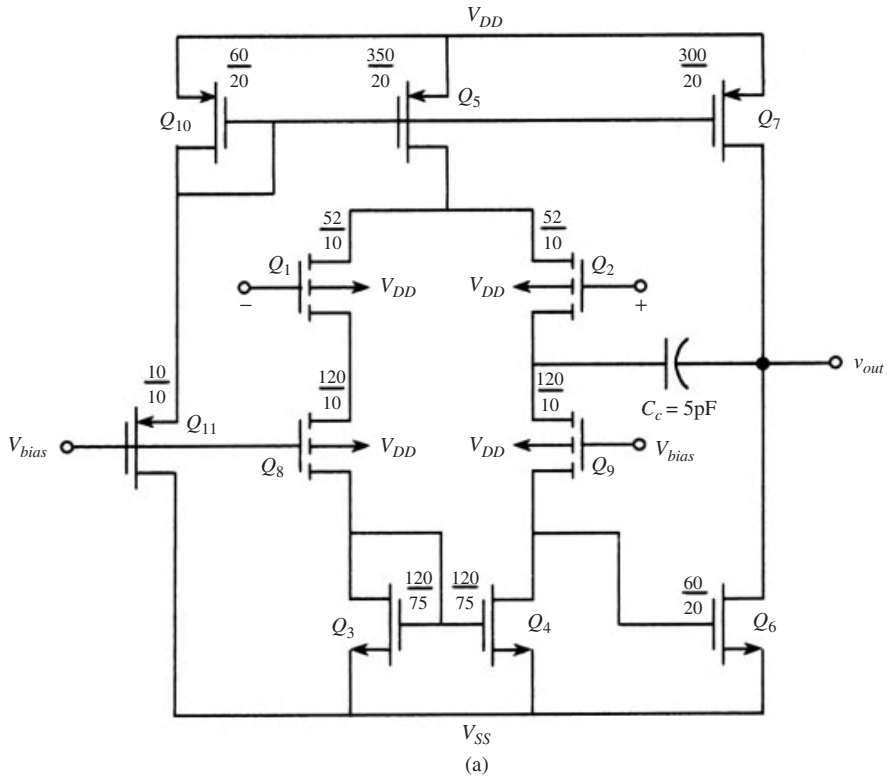
Also the contribution to the noise source at the gates of Q_8 and Q_9 is neglected due to the large impedances seen by the sources. Thus, the total output noise spectral density is

$$\langle v_n^2 \rangle = g_{m6}^2 R_2^2 \left\lfloor \langle v_{n6}^2 \rangle + R_1^2 \left(g_{m1}^2 \langle v_{n1}^2 \rangle + g_{m3}^2 \langle v_{n3}^2 \rangle + g_{m4}^2 \langle v_{n4}^2 \rangle \right) \right\rfloor \tag{13.10}$$

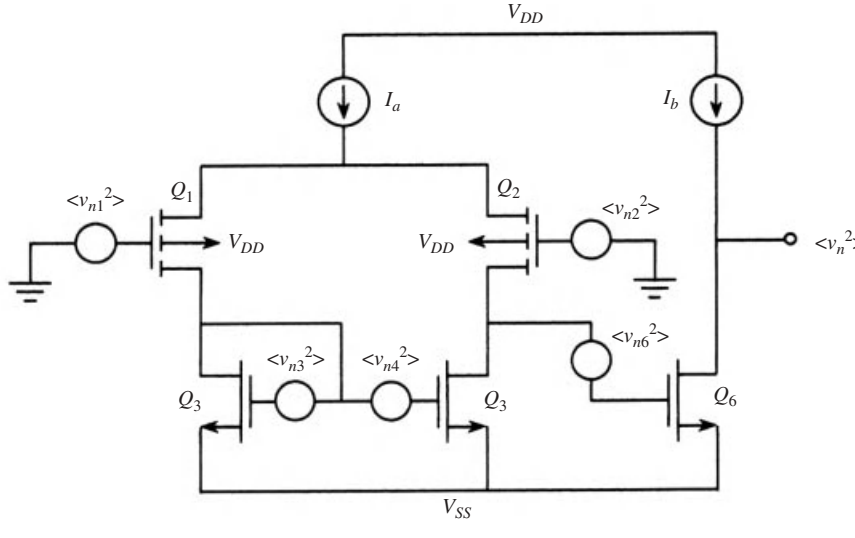
where R_1 and R_2 are the output resistances of the first and second stages, respectively. The equivalent input-referred noise spectral density is obtained from the above expression by dividing by the squared differential gain: $g_{m1}R_1 g_{m6}R_2$ to obtain

$$\langle v_{eq}^2 \rangle = \langle v_{n6}^2 \rangle / g_{m1}^2 R_1^2 + 2 \langle v_{n1}^2 \rangle \left\lfloor 1 + \left\{ (g_{m3}/g_{m1})^2 \right\} \left\{ \langle v_{n3}^2 \rangle / \langle v_{n1}^2 \rangle \right\} \right\rfloor \tag{13.11}$$

from which the noise contribution by the second stage is divided by the gain of the first and may be neglected.



(a)



(b)

Figure 13.3 (a) Example of a low noise Op Amp. (b) Noise model

Now, in order to minimize the noise, we take $g_{m1} > g_{m3}$ so that the input is dominated by the input devices which, inherently, have low noise. The thermal noise contribution can be reduced by increasing the transconductance of the output devices. This can be achieved by increasing the drain current and/or the aspect ratios. The *1/f* noise can be reduced by reducing both *W* and *L*. For the circuit of Figure 13.3(a), the dominant pole is at 100 Hz. With a flat noise of 130 nV/(Hz)$^{1/2}$ over 100 Hz, we have a noise voltage of 13 μV r.m.s. With a peak voltage of 4.3 μV, we obtain a dynamic range of 107 dB.

13.4.2 Noise Reduction by Correlated Double Sampling

This is a technique for reducing the 1/f noise density at low frequencies, as illustrated in Figure 13.4. The amplitude spectrum of the noise is multiplied by a function with amplitude $2\sin(\omega T/4)$. This suppresses the noise at zero frequency and at even multiples of the sampling frequency. The viability of this technique depends on the practicability of integrated circuit implementation of the sample-and-hold circuit and the summer without imposing undue demands on the time response of the Op Amp.

13.4.3 Chopper-stabilized Operational Amplifiers

This technique can be used with any Op Amp design to reduce the input offset voltage And *1/f* noise. The basic concept is illustrated in Figure 13.5 as applied to a two-stage

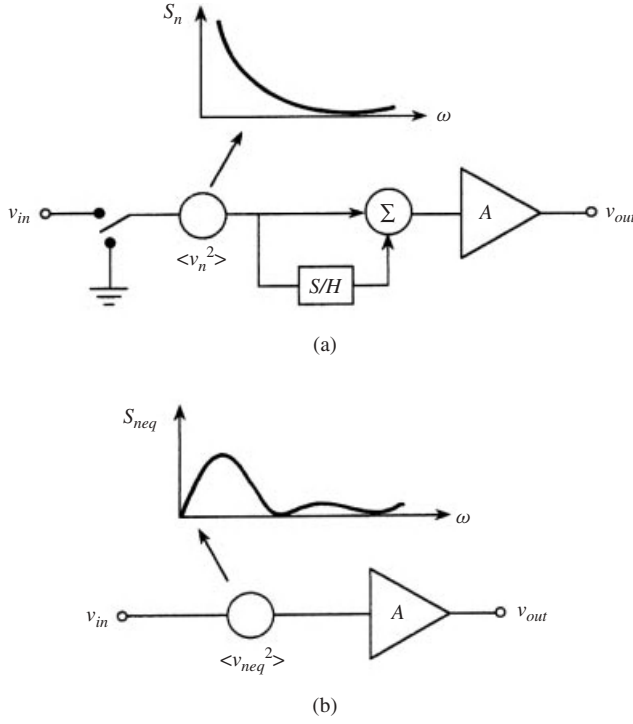


Figure 13.4 Noise reduction by correlated double sampling: (a) conceptual scheme, (b) equivalent input noise

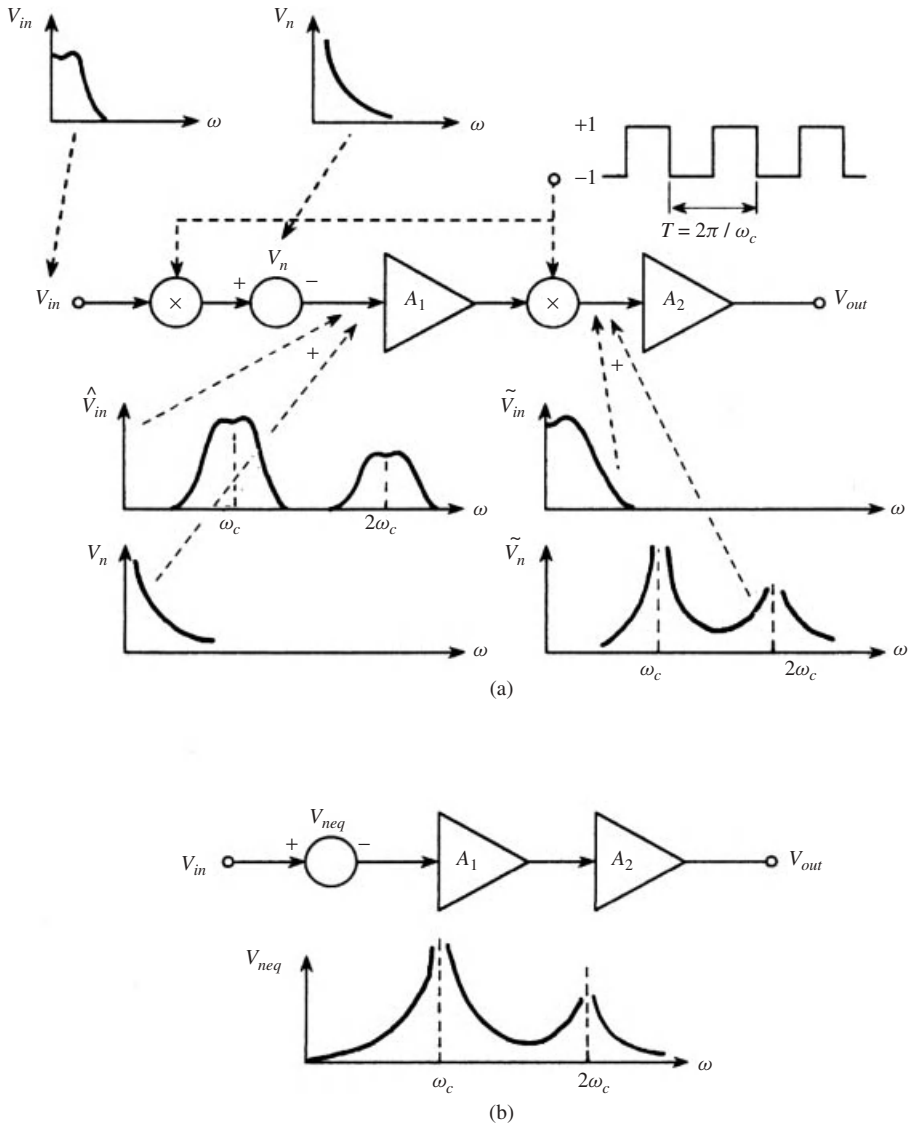


Figure 13.5 Noise reduction by chopper stabilization

amplifier. V_{in} is the input signal spectrum and V_n is the undesirable noise spectrum. The two multipliers are driven by a chopping square wave of amplitude $\pm 1\,V$. Figure 13.5 illustrates clearly the results of the chopping operation and if the chopping frequency is taken to be sufficiently higher than the baseband signal V_{in}, then the spectrum of the undesirable signal will be shifted to a location well outside the baseband. This unwanted signal is composed of the dc offset and the $1/f$ noise. Thus, the effect of these signals on the performance is reduced.

Figure 13.6 shows the implementation of chopper stabilization as applied to a CMOS Op Amp. The multipliers are realized by the switches which are controlled by

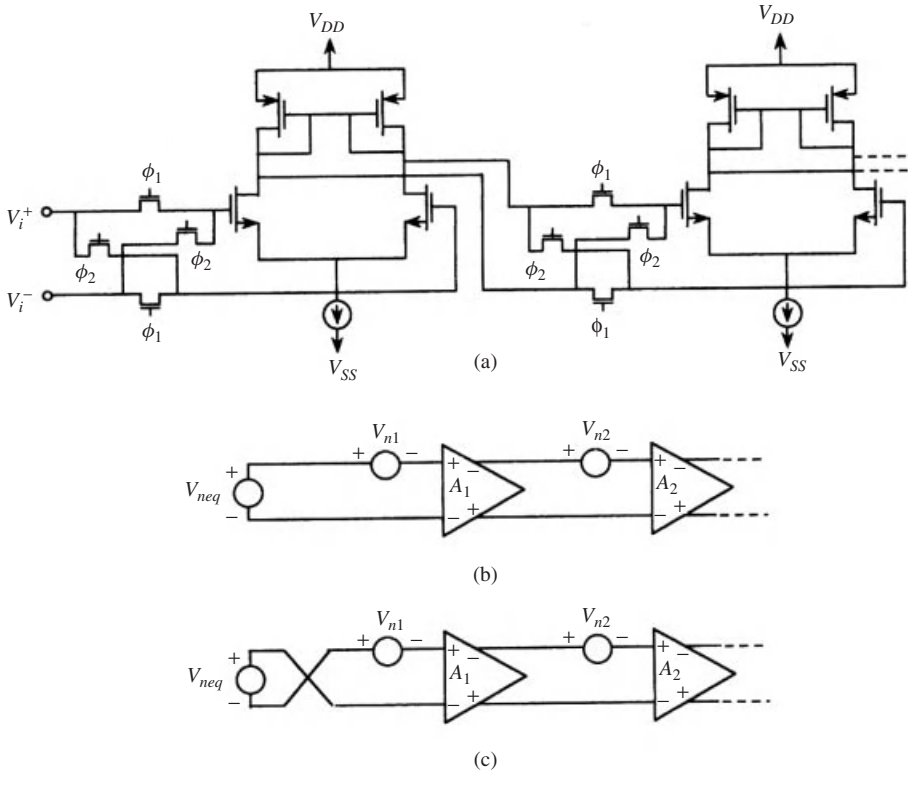


Figure 13.6 Application of chopper stabilization to a CMOS OP Amp: (a) circuit, (b) conditions with ϕ_1 ON and ϕ_2 OFF, (c) conditions with ϕ_1 OFF and ϕ_2 ON

a two-phase clock. With ϕ_1 ON and ϕ_2 OFF, we have from Figure 13.6(b)

$$V_{neq}(\phi_1) = V_{n1} + V_{n2}/A_1 \tag{13.12}$$

and when ϕ_1 is OFF and ϕ_2 is ON, we have from Figure 13.6(c)

$$V_{neq}(\phi_2) = -V_{n1} + V_{n2}/A_1 \tag{13.13}$$

Thus the average value of the input-referred noise over one period is

$$V_{neq}(av) = 1/2 \left\lfloor V_{neq}(\phi_1) + V_{neq}(\phi_2) \right\rfloor = V_{n2}/A_1 \tag{13.14}$$

Hence, the equivalent undesirable signal, in particular the *1/f* noise is cancelled. Also if A_1 is sufficiently high, the second-stage contribution to the noise is reduced.

13.5 High-frequency Operational Amplifiers

At high frequencies, the operational amplifiers have two important requirements: high gain and fast settling time. The former requirement has been discussed, and we now turn to the problem of designing amplifiers with fast settling time.

13.5.1 Settling Time Considerations

Consider a two-stage Op Amp of the type discussed in Chapter 12 with identical input devices having the same drain resistance and transconductance (Figure 13.7). We have seen that it has a dominant pole at $-1/r_o g_m C_C r_o$ and a non-dominant pole at $-g_m C_L$.

For a single-stage cascode Op Amp, the dominant pole is at

$$s_d = -1/r_o C_L g_m r_o \tag{13.15}$$

while the non-dominant pole is at

$$s_n = -g_m/C_p \tag{13.16}$$

where C_p denotes the total capacitance at the cascode node.

Now, if these amplifiers are used in a closed loop configuration, as they are invariably used in analog filters, then it can be shown that the settling time T_s is determined by the

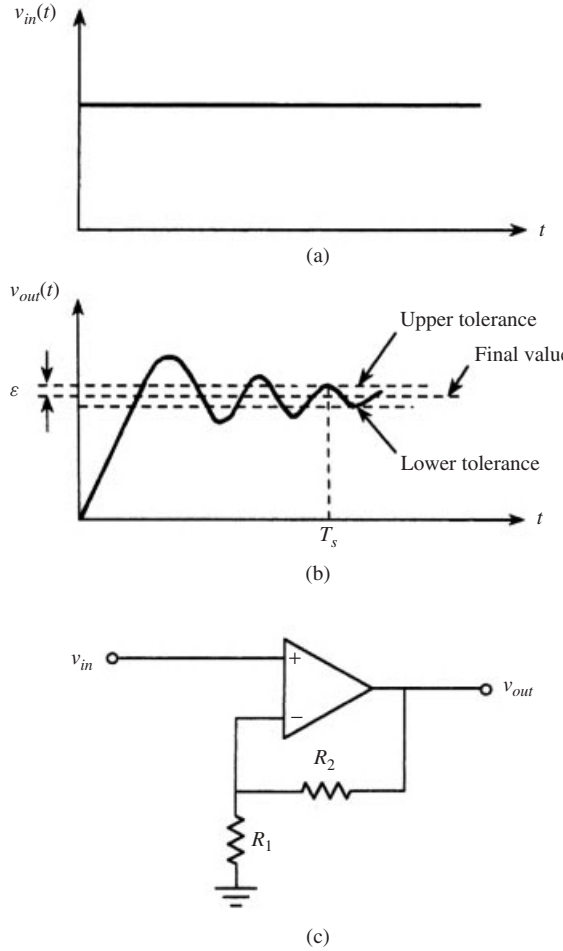


Figure 13.7 Settling time: (a) input step, (b) output, (c) circuit for measuring T_s

non-dominant pole s_n. Specifically, as the loop gain increases, s_d and s_n converge to form a complex pair at

$$s = -0.5 s_n \tag{13.17}$$

so that

$$T_s \approx 2/\mathrm{Re}(s_n) \tag{13.18}$$

Thus, comparing the pertinent expressions for the poles of the two-stage and single-stage cascode amplifiers we have

$$\frac{T_s(two - stage)}{T_s(\sin gle - stage, cascode)} \approx \frac{C_L}{C_p} \tag{13.19}$$

But C_p is of the order of 0.1–0.2 C_L so that

$$\frac{T_s(two - stage)}{T_s(\sin gle - stage, cascode)} \approx 5 - 10 \tag{13.20}$$

It follows that the single stage cascode has faster settling time by a factor of 5 to 10 than the two-stage design.

From the above considerations, it follows that in order to obtain high gain and high speed, the folded cascode design is a good choice.

13.6 Fully Differential Balanced Topology

Now, the single-ended operational amplifier architectures discussed so far have disadvantages when the noise from power supply lines and adjacent digital and switching circuits is to be minimized. This is particularly the case when digital and analog circuits exist on the same chip, but is true in general if power supply rejection is of prime importance. In these situations, the use of a fully differential balanced topology is advantageous. Figure 13.8 shows the three main types of Op Amp: (a) single-ended output, (b) differential output and (c) balanced fully differential output. The latter is the most useful when the noise considerations discussed above are of importance. Note that the balanced Op Amp is such that the two outputs, relative to ground, are accurately balanced. Thus, such an amplifier needs a fifth terminal to act as a reference for balancing the output.

Figures 13.9 and 13.10 show two possible implementations of the balanced Op Amps. Figure 13.9 is basically a single-ended Op Amp together with an inverter. Since both are available as standard cells in many IC design libraries, this is a relatively straightforward realization. However, it has the disadvantage that at high frequencies, the phase shift introduced by the inverter may destroy the balance of the two outputs.

The balanced design of Figure 13.10 requires a differential output Op Amp as well as additional circuits to sense the common-mode (average) output then compares it to ground and applies feedback to correct for its deviation from the desired zero value. Thus, this topology is symmetric and at high frequencies additional phase shift is present with equal amounts at both outputs and balancing is maintained.

Before giving a complete design for a balanced fully differential Op Amp, we summarize the advantages of such a design:

1. Noise from the power supply lines appears as a common-mode signal and, with proper design, can be reduced.

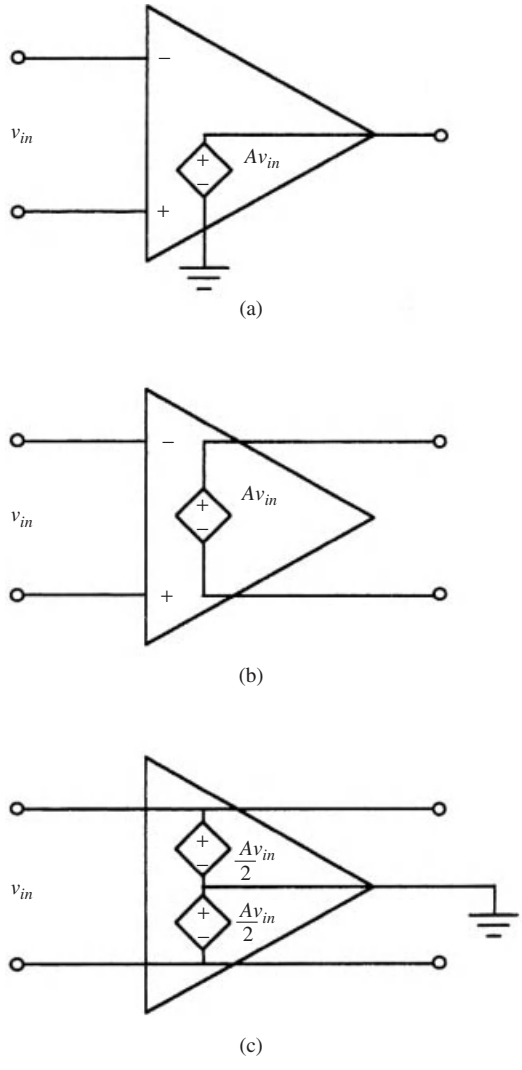


Figure 13.8 Three types of Op Amp: (a) single-ended, (b) differential output, (c) fully differential balanced output

2. The effective output swing is doubled, while retaining the same output circuit. This results in an increase of the dynamic range by about 6 dB relative to the single-ended case.
3. When the fully balanced differential-output OP Amp is used in conjunction with switches and capacitors to form switched-capacitor circuits, this results in a reduction of the clock feed-through noise since it appears as a common-mode signal. This is a distinct advantage at high frequencies because clock feed-through becomes more pronounced since we have to increase the device sizes to reduce charging time constants. This, in turn increases the charge injection into the signal paths. The particulars of this advantage will be discussed at the appropriate point in the book.
4. Symmetric offset voltages are reduced.

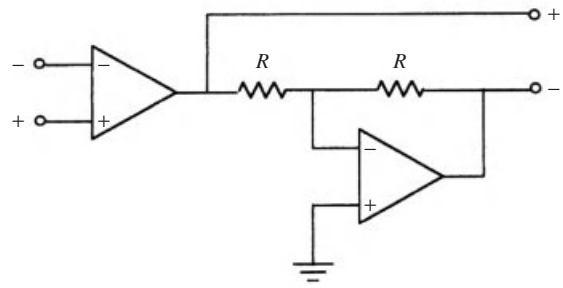


Figure 13.9 A differential output balanced Op Amp using an Op Amp and an inverter

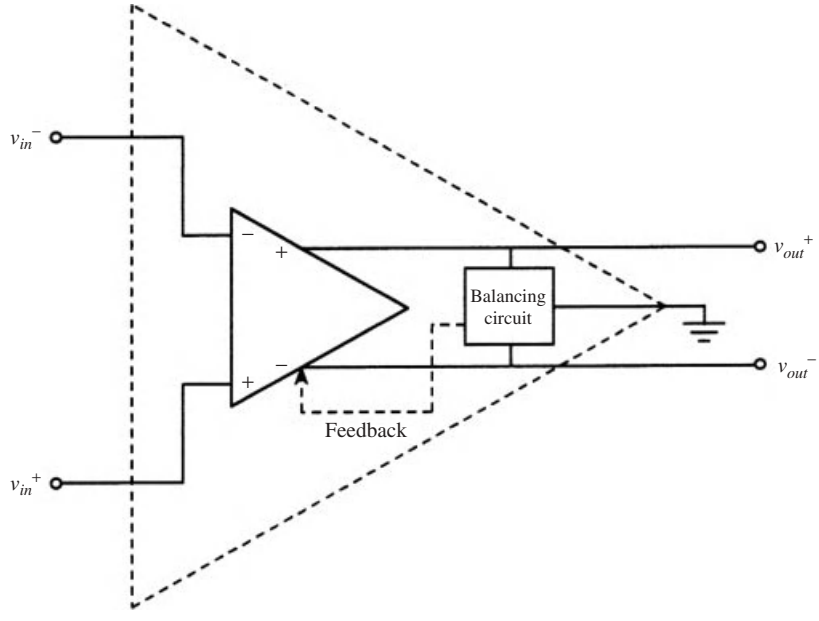


Figure 13.10 A balanced fully differential OP Amp using common-mode feedback

5. The chopper-stabilization technique can be used in conjunction with the fully differential topology to reduce the *1/f* noise, thus resulting in a superior performance suitable for use in high frequency precision VLSI communication circuits.

Having discussed the ideas and advantages of the fully differential balanced design, we consider the implementation of the idea as shown schematically in Figure 13.11.

First note that C_L functions as a compensation capacitor and consider the small-signal differential half-circuit shown in Figure 13.12 which gives for the gain

$$A_v = -(g_m r_{on} r_{op})(g_m r_{op})[1/(1 + r_{on}/R_s) + (r_{on}/R_{on})][1/(1 + r_{on}/R_L)] \qquad (13.21)$$

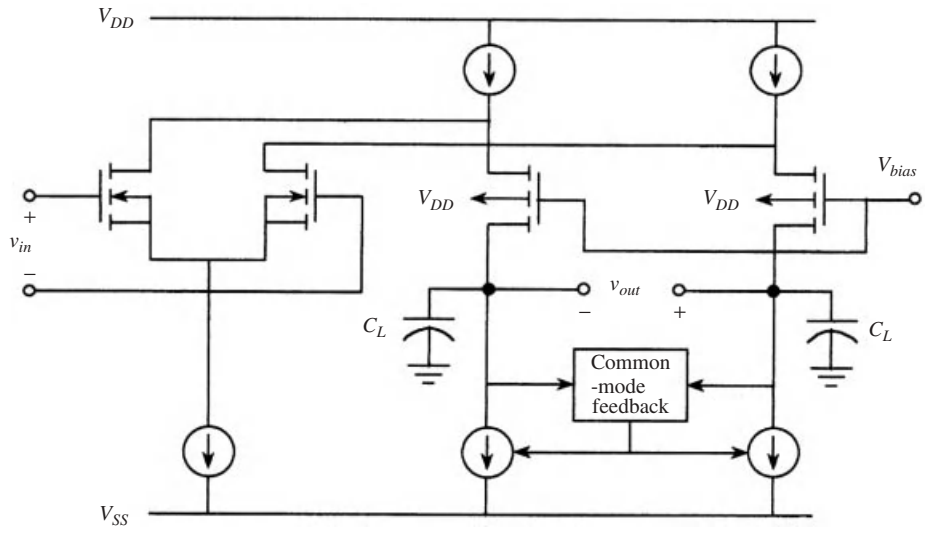


Figure 13.11 Details of the concept illustrated Figure 13.10

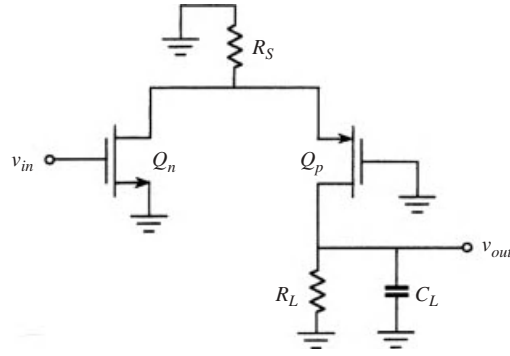


Figure 13.12 Small signal differential half-circuit of the Op Amp in Figure 13.11

where

$$R_{on} = (1/g_{mp})[1 + (R_L/r_{op})] \tag{13.22}$$

is the effective resistance looking into the source terminal of the p-channel cascode device.

Now, Figure 13.13 shows a fully differential folded cascode OP Amp with common-mode feedback balancing network and the following points are helpful in explaining its operation:

(a) Transistors QP_1 and QP_{1A} supply the bias current to the amplifier.
(b) QP_2 and QP_{2A} are the cascode elements.
(c) The channel length of QP_1 and QP_{1A} is taken longer than that of QP_2 and Q_{2A} in order to keep the output resistance relatively high.
(d) Transistors QN_1, QN_{1A}, QN_{2A} and QN_{2A} realize the high impedance current source loads.

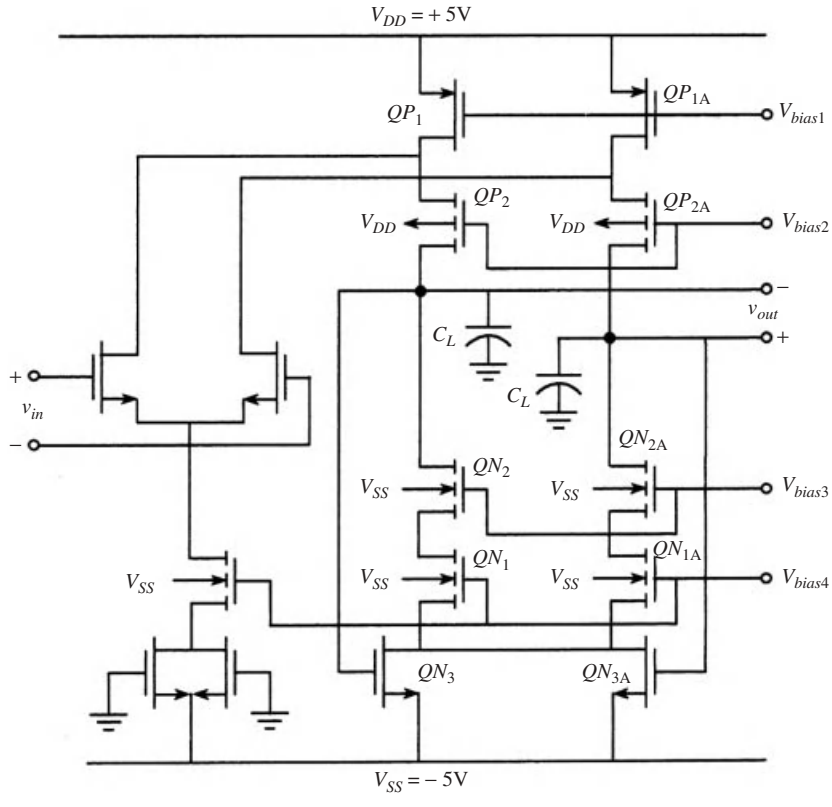


Figure 13.13 Fully differential balanced folded cascode CMOS Op Amp

(e) The common-mode feedback network is composed of the transistors QN_3 and QN_{3A}. These sample the common-mode output signal and feed back a corrective common-mode signal into the source terminals of QN_2 and QN_{2A}. This compensating signal is amplified by the cascade elements to restore the common-mode output voltage to its required original level (ground). Thus, the CMFB is essential for the precise definition of the common-mode output.

There is one drawback of this cascode amplifier, namely the reduced output swing. To appreciate this point, consider Figure 13.14(a). If the cascode devices Q_1 and Q_2 are biased from Q_3 and Q_4, the voltage across the cascode can swing only within $V_t + 2V_{Dsat}$ from the negative supply line, before Q_1 goes into the triode region. The swing can be improved by inserting a level-shifting dc source of strength V_t between the gates of Q_1 and Q_3 as shown in Figure 13.14(b). This forces Q_2 to be biased at the edge of saturation with $V_{DS} = V_{DSsat}$. In this case, the voltage across the cascode can swing to $2V_{Dsat}$ from the negative supply rail before Q_1 is pulled out of saturation.

Figure 13.15 shows a practical realization of the high-swing cascode circuit. All devices conduct equal biasing currents I_0. They all have the same aspect ratio except Q_5 for which $(W/L) = 0.25$ of the others. This results in Q_2 being biased at the edge of saturation with $V_{DS} = V_{DSsat}$. Under these conditions, the voltage across Q_1 and Q_2 can swing to within $2\Delta V$ of the negative supply voltage.

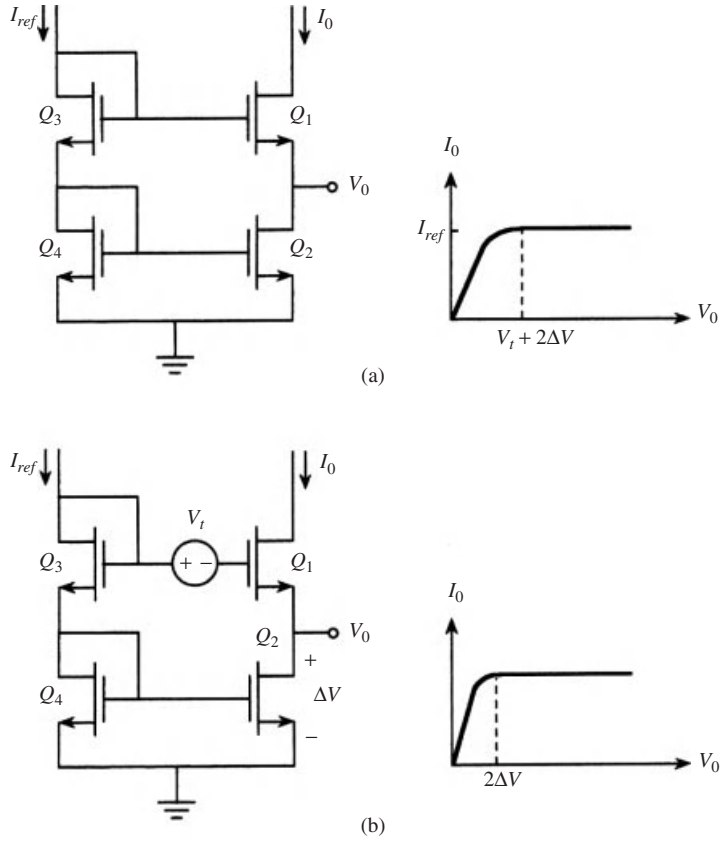


Figure 13.14 Illustrating the idea of high output swing cascode bias

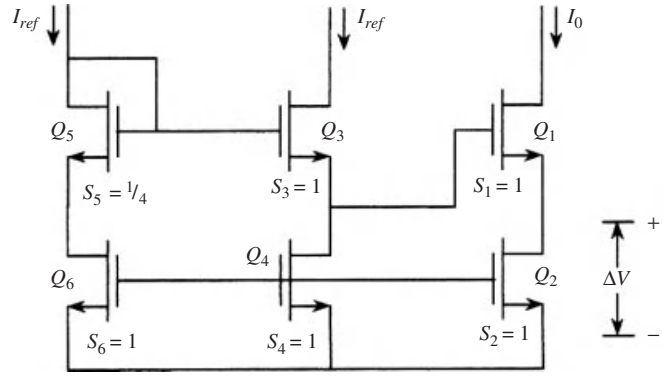


Figure 13.15 Implementation of the high output swing bias

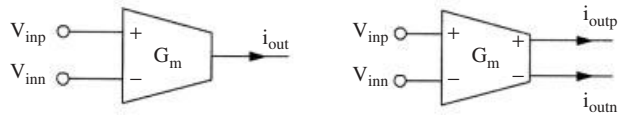


Figure 13.16 Symbols of the operational transconductance amplifier (OTA)

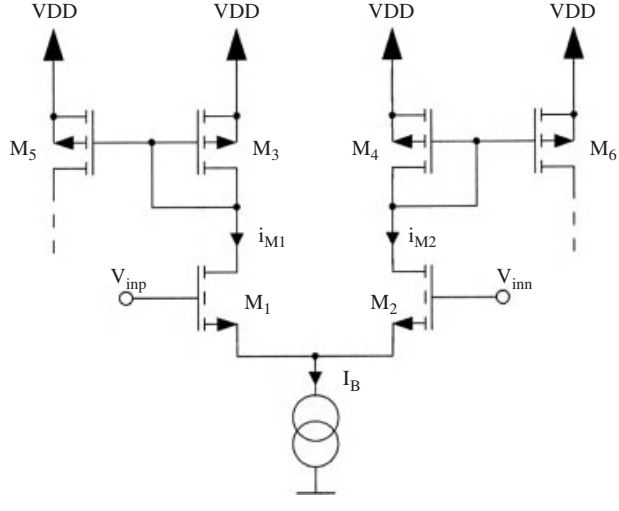


Figure 13.17 Basic realization of an OTA

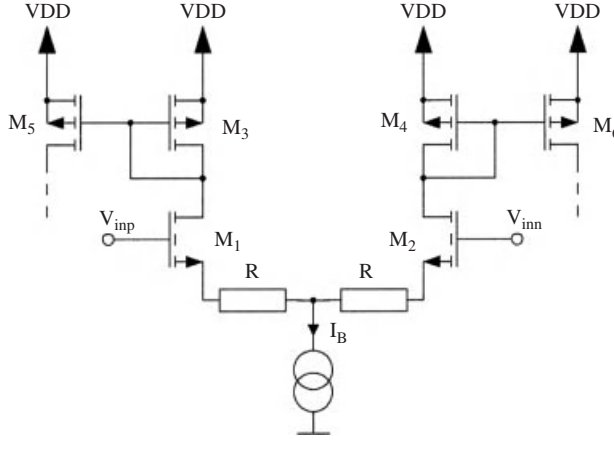


Figure 13.18 Transconductor with improved linearity

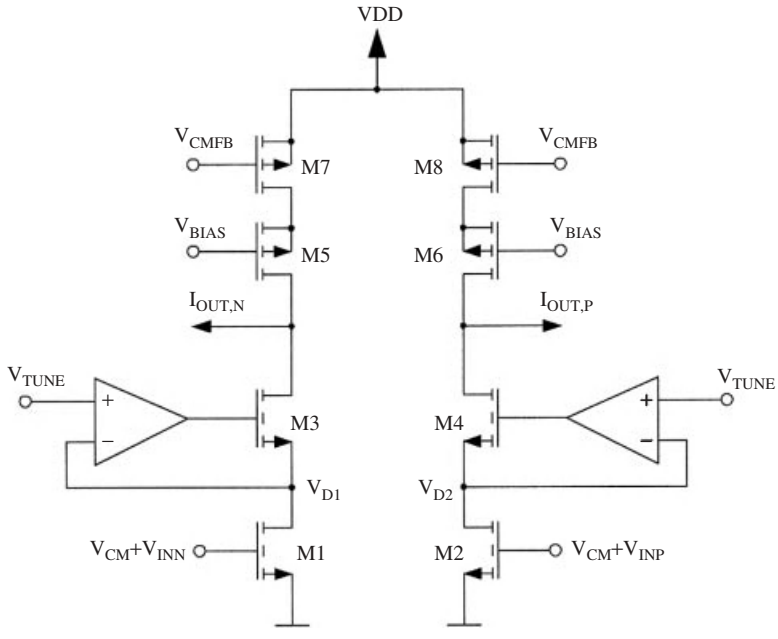


Figure 13.19 Transconductor in which the devices M and M operate in the triode region

13.7 Operational Transconductance Amplifiers

In Chapter 3, the need for operational transconductance amplifiers (OTAs), shown symbolically in Figure 13.16, and their advantages were discussed. Here, a few integrated realizations of OTAs are given [25, 26].

Figure 13.17 shows a simple transconductor, whose linearity can be improved by the addition of source resistors as shown in Figure 13.18.

Another technique to improve the linearity of the transconductor is shown in Figure 13.19 as a pseudo-differential topology in which the transistors M1 and M2 operate in the triode region. M3 and M4 together with the Op Amps form regulated gain control circuit fixing the drain currents of M1 and M2 at the required values. It is possible to obtain a third harmonic distortion value of $-60\,\text{dB}$ with this topology as compared with about $-38\,\text{dB}$ for the simple circuit of Figure 13.17.

A different method is shown in Figure 13.20 where cancellation of the distortion due to cross-coupled differential pairs is used to improve the linearity of the OTA. With this technique a third harmonic distortion value of $-48\,\text{dB}$ is possible.

13.8 Conclusion

This chapter introduced special design techniques of operational amplifiers with the objective of improved performance with regard to one or more of the design criteria such as high gain, low noise and fast settling time. In particular, we note that the use of the folded cascode balanced fully differential design and the use of chopper stabilization

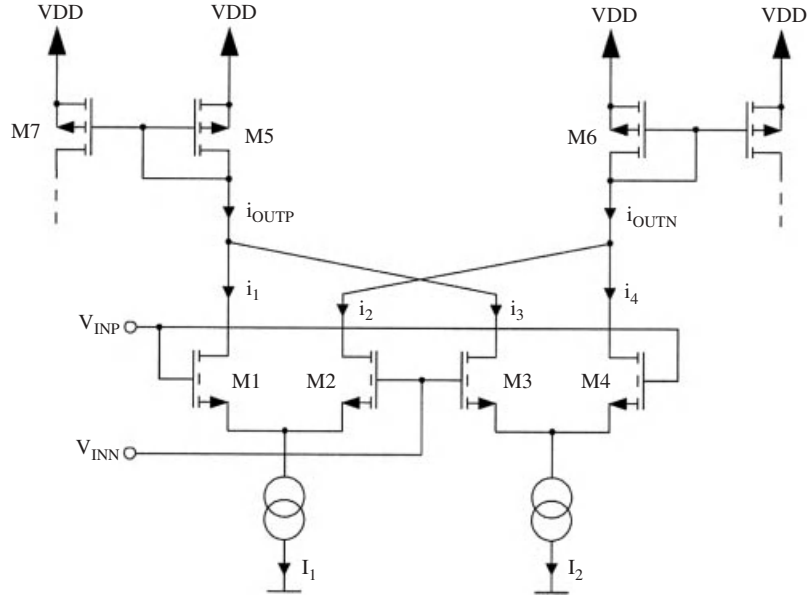


Figure 13.20 Distortion cancellation using cross-coupled differential pairs

result in an excellent Op Amp suitable for high precision applications in communication VLSI circuits. The chapter concluded with an account of operational transconductance amplifiers for use in $G_m - C$ circuits. These have been successfully incorporated in the deign of analog filters in the submicron, deep submicron and ultra deep submicron range [25].

Problems

13.1 A CMOS operational amplifier is to be designed in the folded cascode configuration according to the following specifications:

- Supply voltage $= +5\,V$
- Output voltage swing $= \pm 1\,V$
- Input common mode range $= \pm 1.5\,V$
- dc gain $\geq 70\,dB$.

For the design, an n-well process is used with minimum channel length of $10\,\mu m$, and the process parameters are

$$K_p' = 20\,\mu A/V^2, K_n' = 50\,\mu A/V^2$$
$$\lambda = 0.01\,V^{-1}, \chi = 0.1\,\mu m/V$$
$$V_{tn} = 1\,V, V_{tp} = -1\,V$$

13.2 Calculate the equivalent input noise voltage for the amplifier designed in Problem 11.5 at 100 Hz, 1 kHz and 50 kHz.

13.3 Apply chopper stabilization to the design of Problem 13.1 using two identical stages, with a chopping frequency of 20 MHz. Calculate the equivalent input noise voltage at the same frequencies as Problems 13.2 for the design and compare with that obtained without applying chopper stabilization.

13.4 Convert the Op Amp design in Problem 13.1 into a fully differential design using an additional inverter.

14

Capacitors, Switches and the Occasional Passive Resistor

14.1 Introduction

Many analog integrated circuits are constructed using operational Op Amps, OTAs, capacitors, switches and resistors. We have dealt with Op Amps and OTAs in the previous chapters, and it now remains to examine the design of the other building blocks. This chapter presents the various integrated circuit versions of capacitors and switches [24] and discusses the non-ideal effects in these components in relation to their use in analog and mixed-mode signal processing systems. Furthermore, resistors are sometimes also required as on-chip components. These can be realized as active devices as explained in Chapter 11; alternatively some situations call for the high degree of linearity associated with passive components. The possible passive MOS resistor structures are also reviewed.

14.2 MOS Capacitors

14.2.1 Capacitor Structures

An important building block of analog integrated circuits is the MOS capacitor. The most commonly used dielectric is S_iO_2 which is a very stable insulator with $\varepsilon_{ox} \cong 3.9$ and a high breakdown electric field of about 8×10^6 V/cm (although $S_{i3}N_4$ is also used). The choice of electrodes for the capacitor varies according to the available technology that is used for the fabrication of the entire integrated circuit. This leads to the following types:

1. Metal (or polysilicon) over diffusion structure. This is shown in Figure 14.1(a) and is formed by growing a thin S_iO_2 layer over a heavily-doped region in the substrate. In a metal gate process, the top plate of the capacitor is formed by covering the S_iO_2 with metal in the same processing step of providing the metalization for the gate and leads of the entire circuit. In a silicon gate process, heavily doped polycrystalline silicon (polysilicon) is used as the gate electrode and also to form the top plate of the capacitor. Ideally, if the electrodes are perfect conductors, the capacitance per unit area is

$$C_o = \varepsilon_{ox}/t \tag{14.1}$$

Signal Processing and Integrated Circuits, First Edition. Hussein Baher.
© 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.

However, the actual capacitance is voltage dependent and is of the form

$$C = C_o[1 + b(V_A - V_b)]^{1/2} \tag{14.2}$$

where b is a constant, inversely proportional to the doping density. For heavily doped n^+ layers, the voltage dependence of the capacitor is slight. Achieved capacitance values using this structure are in the range $0.35 - 0.5\,\text{fF}/\mu\text{m}^2$. The tolerance on the value of the capacitance is usually $\approx \pm (6 - 15)\%$. However, two identical capacitances can be matched to within 0.1–1.0%.

2. Polysilicon over polysilicon capacitors. In a siliconQ1 gate 'poly-poly' process, a second layer of high conductivity polysilicon is available for use in forming interconnect elements. These two layers can also be used as the top and bottom plates of a capacitor, as shown in Figure 14.1(b). A disadvantage of this structure is that, due to the irregularity of the polysilicon surface, the oxide thickness has a large random variation. Typical values achieved are $0.3–0.4\,\text{fF}/\mu\text{m}^2$.
3. Metal over polysilicon capacitor. This is shown in Figure 14.1(c) and its properties are similar to those of Figure 14.1(b).

14.2.2 Parasitic Capacitances

Parasitic capacitances are inherent in all the MOS capacitor structures discussed above. Figure 14.2 shows a model of the MOS capacitor including the main parasitics.

There is an inevitable large parasitic capacitance from the bottom plate to the substrate and hence to the substrate bias supply. For a metal (or polysilicon) over diffusion capacitor the bottom plate is embedded into the substrate and the stray bottom plate parasitic can be ≈15–30% of the nominal required value. For the structures of Figure 14.1(b, c), this capacitance is of the order of 5–20% of the required capacitance.

Another source of stray capacitance is due to the leads used to connect the capacitance plates in the rest of the circuit. This is the source of the top plate parasitic capacitance. Furthermore, in some analog integrated circuits, one or both plates of a capacitor are connected to the source or drain diffusion of a MOS switch. The pn junction of the diffusion contributes a depletion layer capacitance between the capacitor plate and substrate. All these parasitics depend on the capacitor size as well as the fabrication technology and layout.

14.2.3 Capacitor-ratio Errors

The performance of switched-capacitor circuits, which will be discussed in the next chapter, is determined by capacitor ratios, therefore the errors involved in these ratios constitute an important design consideration. The sources of these errors are now discussed briefly.

14.2.3.1 Random Edge Variation

Figure 14.3 shows a top view of a MOS capacitor, in which the edges of the electrodes exhibit random variation. The nominal capacitance value is

$$C = \frac{\varepsilon A}{t_{ox}} \tag{14.3}$$

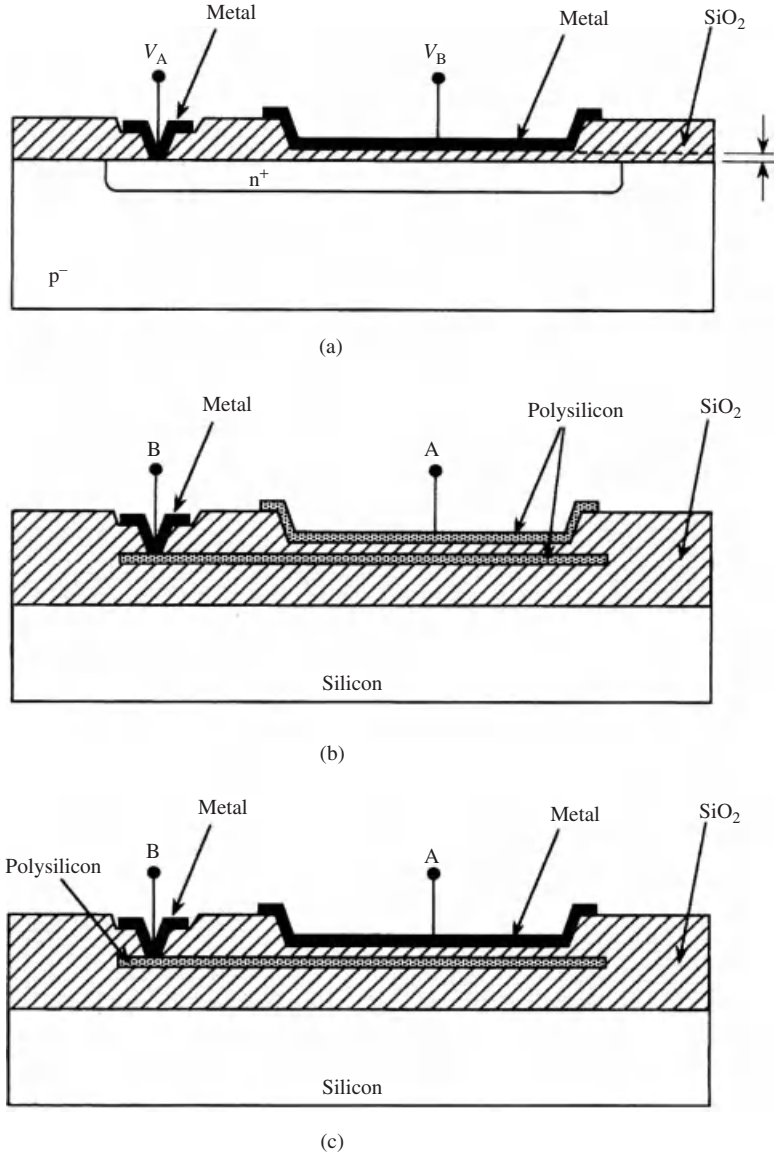


Figure 14.1 MOS capacitors: (a) metal (or polysilicon) over diffusion, (b) polysilicon over polysilicon, (c) metal over polysilicon

and due to the random variation in the area we have

$$\Delta C = \frac{\varepsilon}{t_{ox}}[(W + \Delta W)(L + \Delta L) - WL] \tag{14.4}$$

so that

$$\frac{\Delta C}{C} = \frac{\Delta W}{W} + \frac{\Delta L}{L} \tag{14.5}$$

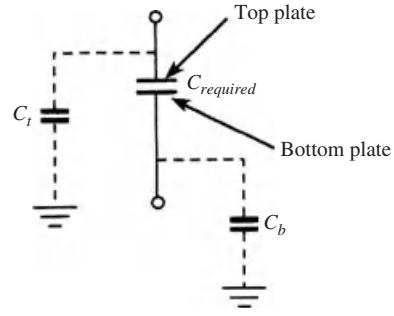


Figure 14.2 MOS capacitor model showing the parasitic capacitances

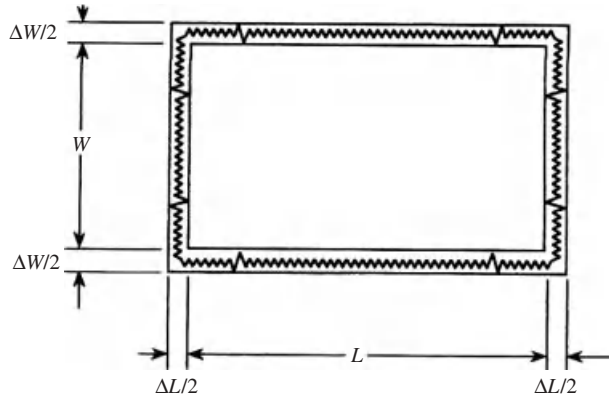


Figure 14.3 Schematic of random edge variations in a MOS capacitor

It can be assumed that ΔW and ΔL are independent random variables with equal standard deviation $\sigma_L = \sigma_W$. This leads to the standard deviation of ΔC being

$$\sigma_c = C\sigma_L\sqrt{W^{-2} + L^{-2}} \tag{14.6}$$

But C (and hence *WL*) is fixed, therefore the relative error σ_C/C is a minimum for $W = L$. In this case, the relative capacitance error is

$$\frac{\sigma_C}{C}|_{\min} = \sqrt{2}\sigma_L/L \text{ for } W = L \tag{14.7}$$

which means that the capacitor shape should be square.

The above considerations hold for capacitor ratios also. Suppose the nominal capacitance ratio is

$$\alpha = \frac{C_1}{C_2} \geq 1 \tag{14.8}$$

and the dimensions of C_1 are W_1, L_1 while those of C_2 are W_2, L_2. Assuming all dimensions to have the same standard deviation σ, then

$$\frac{\sigma_\alpha}{\sigma} = \sigma\sqrt{{L_1}^{-2} + {W_1}^{-2} + {L_2}^{-2} + {W_2}^{-2}} \tag{14.9}$$

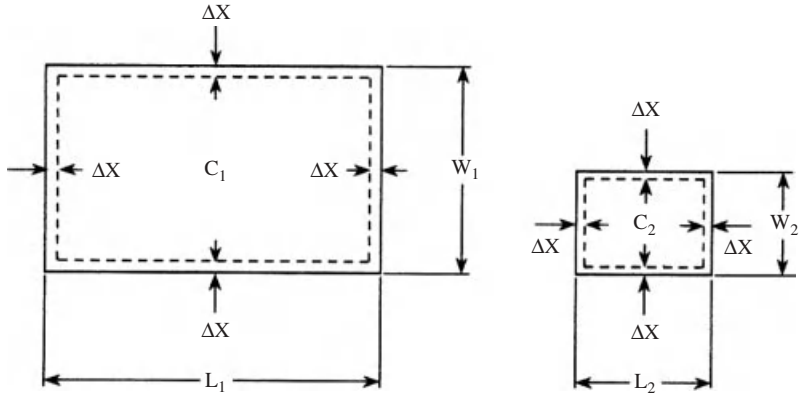


Figure 14.4 Undercut error

which is a minimum if

$$L_1 = W_1 = \sqrt{\sigma} L_2 = \sqrt{\sigma} W_2 \tag{14.10}$$

with a minimum value of

$$\frac{\sigma_\alpha}{\alpha}|_{\min} = \left(\frac{\sqrt{2}\sigma}{L_1}\right)\sqrt{1+\sigma} \tag{14.11}$$

which implies that for best accuracy, the capacitor ration should be 1.

14.2.3.2 Undercut Error

This results from the uncontrollable lateral etching of the plates of the capacitor along its perimeter in the fabrication process as shown in Figure 14.4. This gives rise to a decrease in the value of the capacitance, proportional to the perimeter. Again with the required capacitor ratio

$$\alpha_o = \frac{C_1}{C_2} = \frac{W_1 L_1}{W_2 L_2} \tag{14.12}$$

The undercut results in the actual ratio being

$$\alpha \cong \frac{W_1 L_1 - 2(W_1 + L_1)\Delta x}{W_2 L_2 - 2(W_2 + L_2)\Delta x} \tag{14.13}$$

where Δx is the depth of the undercut, which is assumed to be uniform along the perimeter.

As before, we can show that with $W_2 = L_2$ and

$$W_1 = L_2(\alpha - \sqrt{\alpha^2 - \alpha}) \tag{14.14}$$
$$L_1 = L_2(\alpha - \sqrt{\alpha^2 - \alpha})$$

the undercut error is zero. For this choice, the standard deviation is

$$\frac{\sigma_c}{\alpha} = \left(\frac{\sigma}{L_2}\right)\sqrt{6 - 2/\alpha} \tag{14.15}$$

However, a very common technique to avoid the undercut error is to connect identical unit capacitors in parallel to construct larger ones. This leads to the ratio of area to perimeter being the same for any two capacitors and the actual ratio is almost the same as the nominal one.

It is concluded that for best accuracy of capacitor ratios, the capacitors should be identical. So, of the ratio is different from one, the two capacitors should be constructed as integral multiples of a unit capacitor. If the errors in each unit are the same, the ratio will be free of error.

14.3 The MOS Switch

14.3.1 A Simple Switch

This is a key component in the design of switched-capacitor circuits, which are of the analog sampled-date type. Figure 14.5(a) shows the simplest MOSFET switch, while Figure 14.5(b) shows its equivalent circuit including the associated parasitic capacitances. Figure 14.5(c) shows the clock signal which drives the switch. The switch is in the ON state when the gate has a sufficiently high voltage (positive for NMOS and negative for PMOS). In this case, the voltage v_{DS} will cause a current i_D to flow between the switch terminals A and B. Since the gate voltage v_ϕ is usually much larger than the voltage across the terminals A and B, the MOSFET can be assumed to be in the triode region, so that for the current in the switch

$$i_D = K[2(v_{GS} - V_t)v_{DS} - {v_{DS}}^2] \tag{14.16}$$

Normally

$$|v_{GS} - V_t| \gg |v_{DS}| \tag{14.17}$$

and the switch behaves as a linear resistance

$$R_{on} \cong \frac{1}{2K(v_{GS} - V_t)} \tag{14.18}$$

14.3.2 Clock Feed-through

Due to the parasitic capacitances shown in Figure 14.5(b), an undesirable phenomenon may occur if the simple switch realization is used. To illustrate this effect, consider Figure 14.6 and suppose that the capacitance loading terminal A is C_A while that loading terminal B is C_B. Then the clock signal will be transmitted to nodes A and B as

$$v_A = \frac{C_{gs}}{C_{gs} + C_A}v_\phi \tag{14.19}$$

$$v_B = \frac{C_{gd}}{C_{gd} + C_B}v_\phi$$

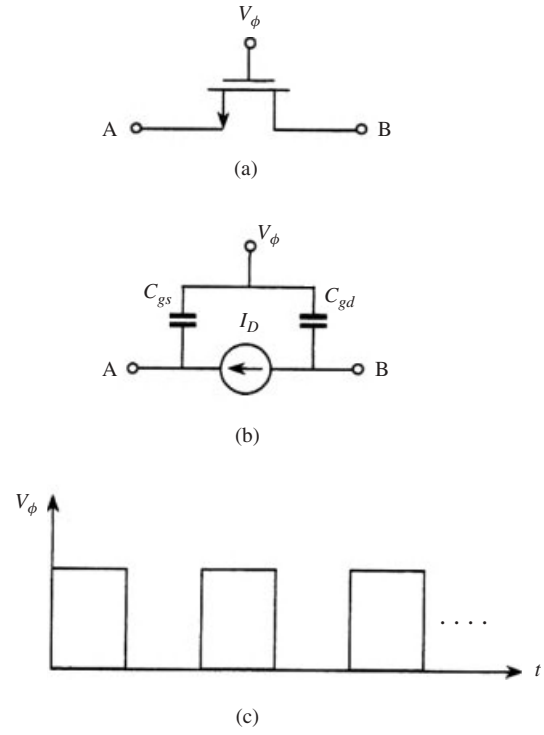


Figure 14.5 (a) MOS switch, (b) model with parasitics and (c) driving clock

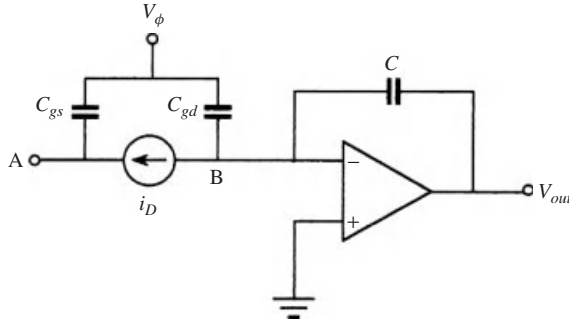


Figure 14.6 The clock feed-through effect

Typically, $C_{gs} \approx C_{gd}$ and $C_A \approx C_B = 100C_{gs} = 100C_{gd}$ so that $v_A \approx v_B \approx 0.01v_\phi$. This means that a signal of this value and a frequency equal to the clock frequency is transmitted to nodes A and B.

This is called *clock feed-through* and should be minimized. An obvious method would be to connect another transistor with an equal and opposite contribution to the clock feed-through. One possibility is to add a dummy transistor as shown in Figure 14.7 with its drain and source connected to the signal line and control voltage opposite in phase to the gate voltage of the main switch. This transistor performs the compensation operation only and has no switching function. Furthermore, the effect of feed-through

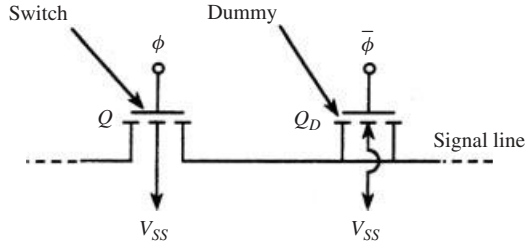


Figure 14.7 Reduction of clock feed-through using a dummy MOSFET

can be reduced by designing the switches as small as possible while the capacitors are designed as large as possible.

14.3.3 The CMOS Switch: Transmission Gate

A more elaborate compensation scheme employs both NMOS and PMOS transistors to from the CMOS switch or *transmission gate* shown in Figure 14.8. For this circuit the feed-through signals at each node cancel each other, in principle. Another advantage of this switch is that its ON resistance tends to be more linear as a result of the parallel connection of the NMOS and PMOS devices. Moreover, the dynamic signal range in the ON state is increased. This is because a very large input signal, and consequently an equally large output signal, causes one transistor to be OFF, since the gate to source voltage does not become sufficiently large. The same signal causes the complementary switch to be fully ON. At a given instant, at least one switch is ON.

For the design of a CMOS switch, complementary clock signals are needed but with the added flexibility of introducing slight delays relative to each other in order to minimize the clock feed-through. To achieve this, the system clock has to be inverted and delayed such that the gate voltages at the CMOS switch are the complements of each other. Figure 14.9 shows such a scheme.

Despite these precautions, in practice C_{gs} contains a non-linear component and is difficult to achieve compensation. Furthermore, the MOS switch introduces the two non-linear capacitances C_{sb} and C_{db}, which cause harmonic distortion and couple noise

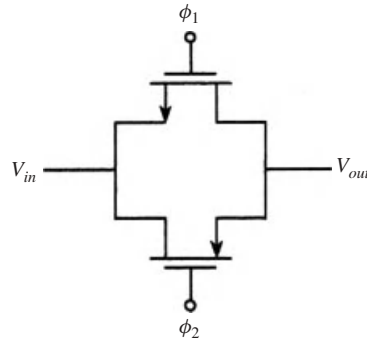


Figure 14.8 CMOS switch: transmission gate

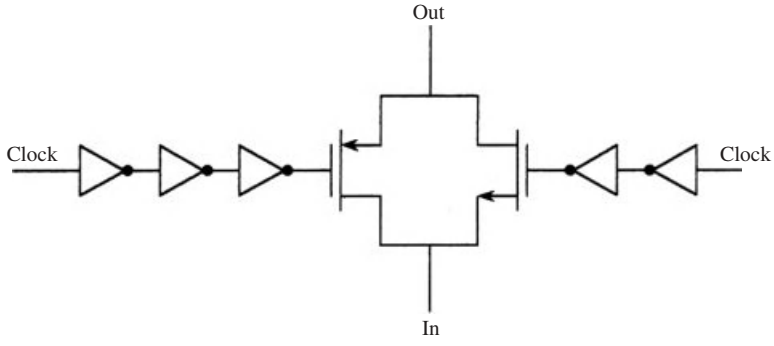


Figure 14.9 CMOS switch with inverter-delays

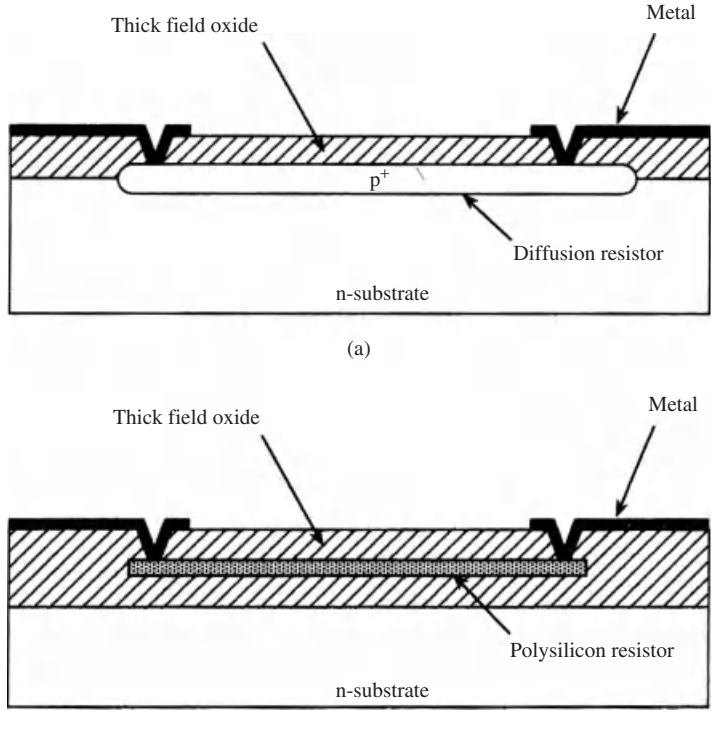


Figure 14.10 Two resistor types in the COMS process: (a) diffusion, (b) polysilicon

from the substrate into the signal path. Also, after the switch is turned OFF, the drain and source currents become the leakage currents associated with reverse-biased pn junctions. Although this current is very small, it may result in a steadily growing drain or source charge unless there is a dc path to ground connected at least occasionally from the drain and source terminals to ground. Finally, as in all MOS circuits, MOS switches are susceptible to thermal noise.

14.4 MOS Passive Resistors

An occasional passive resistor may be needed as an on-chip component. Therefore, for the sake of completeness, Figure 14.10 shows two types of resistor which are available in a CMOS process. The diffused resistor may be constructed using source-drain diffusion. It is voltage dependent, and the associated parasitic capacitance is also voltage-dependent. The polysilicon resistor shown in Figure 14.10(b) is surrounded by thick oxide. The associated parasitics are quite small and voltage-independent.

14.5 Conclusion

This chapter concludes the discussion of integrated circuit components for analog signal processing systems by studying the design of capacitors, switches and passive resistors. Particular attention was given to the non-ideal effects such as clock-feed-through in the switches and capacitor ratio errors. Methods for reducing these effects were also presented.

Part IV

Switched-capacitor and Mixed-mode Signal Processing

'But there are two considerations that are important as regards continuity. First, it is largely hypothetical. We do not observe any one thing continuously, and it is merely a hypothesis to assume that, while we are not observing it, it passes through conditions intermediate between those in which it is perceived. . . . Second, continuity is not a sufficient condition of material identity'.

Bertrand Russell
'The Relation of Sense-data to Physics'

15

Design of Microelectronic Switched-capacitor Filters

15.1 Introduction

The earliest filters to be firmly established in the electronic design repertoire were passive ones [13, 14] employing inductors, capacitors and transformers. These have become reference designs against which any other category of filter is measured and compared. This is due to a number of reasons. First, these passive structures have been shown to be capable of satisfying the most stringent specifications that may be encountered in all disciplines of electrical engineering design. Secondly, they have been found to possess low-sensitivity properties with respect to variations in element values; a highly desirable attribute from the practical viewpoint. Finally, they do not consume power.

However, as the operating frequencies fall into and below the audio range, the required inductors develop severe limitations and disadvantages: they become large-valued, bulky, with low quality factors and expensive to construct. For this reason, active filters were introduced with the primary objective of overcoming the practical disadvantages of inductors at the lower end of the frequency spectrum. These employ, resistors, capacitors and active devices such as transistors or Op Amps, as detailed in Chapter 3 where active structures which imitate the low sensitivity properties of passive models were introduced.

With the technological advances in integrated circuit design, the rather obvious step and natural tendency to implement analog active filters as monolithic integrated circuits were frustrated by practical factors. As shown in Chapter 3, in principle, analog filters can be realized by circuits whose responses depend on RC products; for example the integrator of Figure 2.20. The basic difficulty in implementing such a building block in integrated circuit form is that one would have to realize the RC product with high precision which, in turn, requires the realization of the absolute value of each component with an even greater precision. For example, if the errors in realizing an integrated resistor and a capacitor could be as large as 20% for each element, the error in the RC product may be as large as 40% which is, of course, unacceptable in any application. Furthermore, the required absolute values of the elements can be too large requiring a large area on the integrated circuit. In the event of implementing resistors using transistors, one has to

Signal Processing and Integrated Circuits, First Edition. Hussein Baher.
© 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.

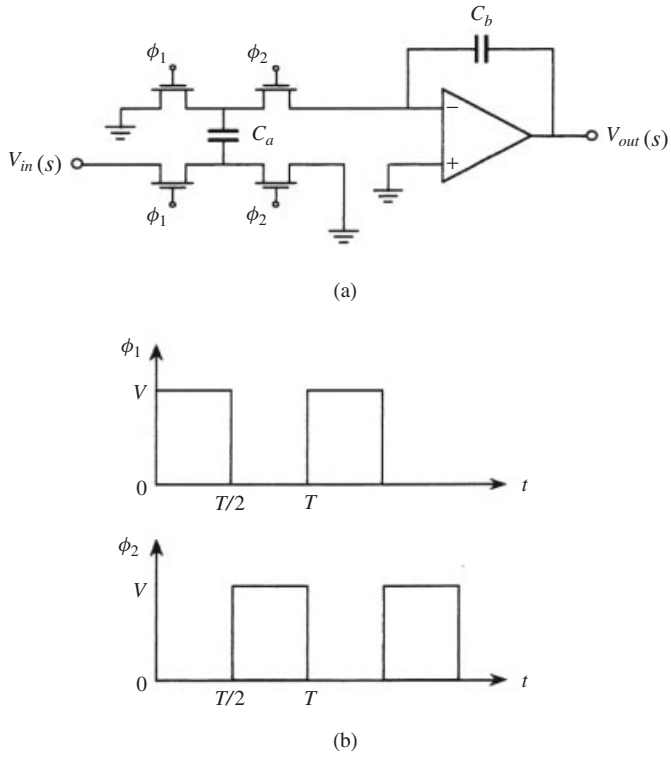


Figure 15.1 (a) A switched-capacitor circuit. (b) The bi-phase clock driving the switches

contend with the inherent non-linearities. There are, however, schemes for cancellation of these effects, but in most cases they require additional circuits which may be more complex than the filter itself. Another approach, which was discussed in Chapters 2 and 3, is to use $G_m - C$ circuits employing transconductors and capacitors only. The integrated circuit realizations of these transconductors were discusses in Chapter 13.

In this chapter, we consider another very successful approach, namely: that of switched-capacitor techniques [24]. These made possible the implementation of analog filters in monolithic integrated form using MOS technology; the same technology that had matured in the design of digital circuits. The key idea is deceptively simple. Consider the building block of Figure 15.1(a) which typifies the entire category of switched-capacitor filters. It is made up of an operational amplifier, two capacitors and analog switches. The switches are actuated periodically by a clock as shown in Figure 15.1(b) so that the input voltage is sampled then switched to the input of the operational amplifier half a period later. Under the given clocking scheme, we shall see that the transfer function V_{out}/V_{in} of this circuit is given by

$$T(s) = \frac{V_{out}(s)}{V_{in}(s)} = \frac{e^{-Ts/2}}{2(C_b/C_a)\sinh Ts/2} \tag{15.1}$$

The basic difference between this circuit and the continuous-time integrator of Figure 2.20 is that the transfer function in (15.1) is determined by the capacitor *ratio* C_b/C_a and *not* by *absolute* values as the function in (2.93). Hence a circuit composed of building

blocks with this key property, will have a response which is also determined by capacitor ratios, not by absolute element values. In integrated circuit implementations, it is possible to realize capacitor ratios with an accuracy that is orders of magnitude better than the accuracy of realizing absolute values. Furthermore, since the absolute capacitor values are of no consequence, these can be reduced at will to values that are limited only by the practical lower limits of the fabrication process. This allows the reduction of the overall area occupied by the capacitors in the integrated circuit.

In addition to the above key advantages, circuits of the generic type in Figure 15.1(a) are largely insensitive to the parasitic capacitances inherent in the integrated circuit manufacturing process.

In summary, the main features of switched-capacitor filters are the following:

- The building blocks are Op Amps, capacitors and analog switches.
- The circuits are of the analog sampled-data type: they operate directly on analog signals and produce outputs that are essentially analog; the filtering process is accomplished on samples of the signals. Thus, neither coding nor quantization is needed as in the case of digital filters which require the process of A/D conversion and subsequent D/A conversion if the filter is to produce analog outputs.
- The performance of the circuit depends on capacitor ratios, not on absolute values.
- The filters can be made programmable.
- The circuit structures can be made insensitive to parasitic capacitances.
- The filters can be designed to imitate the low-sensitivity properties of reference passive designs.
- High quality precision filters are possible for telecommunications applications.
- The filters can be fabricated using standard MOS technology, the same one used in digital circuit design. Hence these can be placed on the same chip with digital circuits.
- A switched-capacitor filter designed using CMOS circuits has a much simpler structure and consumes less power than a digital counterpart performing the same signal processing function.

In addition to the most important application in linear filtering of signals, switched capacitor circuits can be used as oscillators, modulators, A/D converters, speech processors in speech synthesis, rectifiers, detectors and comparators. Their applications in the design of sigma-delta data conversion is undertaken in a later chapter and gives rise to a mixed-mode processor. This chapter, however, gives a detailed account of the design of switched-capacitor filters relying very heavily on the results of Chapters 2–3 and the integrated circuit implementations in CMOS technology given in Chapters 10–14.

15.2 Sampled and Held Signals

In switched-capacitor filters the signals are sampled and held as shown in Figure 15.2.

Consider the signal $f(t)$, bandlimited to ω_m and let it be sampled as shown in Figure 15.2(a) such that : at the instant (nT), the signal is sampled and held at its value $f(nT)$ until the next sample is taken at $(n+1)T$. The sample values are defined at the beginning of each interval. Then the resulting signal is given by

$$f_s(t) = \sum_{n=-\infty}^{\infty} f(nT)\{u(t-nT) - [u(t-(n+1)T]\} \tag{15.2}$$

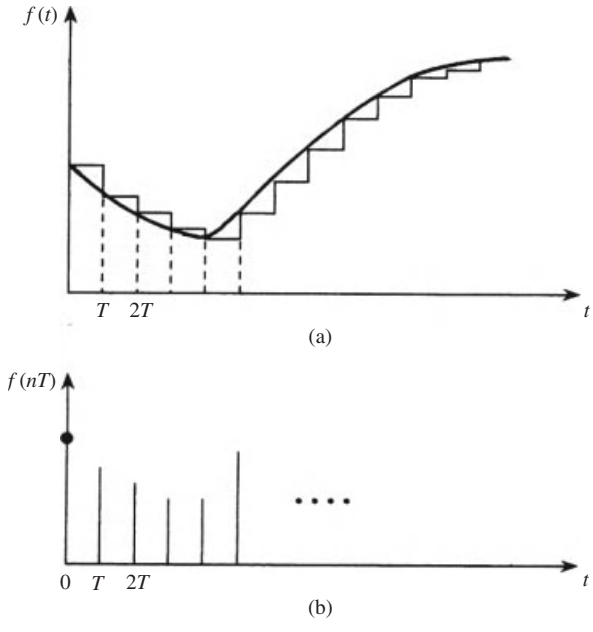


Figure 15.2 (a) Sampled and held signal. (b) Sample values defined at the beginning of each interval

where $u(t)$ is the unit step function. The function between the curled brackets is a unit pulse with width T starting at (nT) which has a Fourier transform

$$\Im\{u(t-nT)-u[(t-(n+1)T]\}=T\left(\frac{\sin\omega T/2}{\omega T/2}\right)e^{j\omega T/2}e^{-jn\omega T} \tag{15.3}$$

Therefore, the Fourier transform of $f_s(t)$ is given by

$$\hat{F}_s(\omega)=\left\{T\left(\frac{\sin\omega T/2}{\omega T/2}\right)e^{-j\omega T/2}\right\}\sum_{n=-\infty}^{\infty}f(nT)e^{-jn\omega T} \tag{15.4}$$

But

$$\sum_{n=-\infty}^{\infty}f(nT)e^{-j\omega T}=\Im\left\{\sum_{n=-\infty}^{\infty}f(nT)\delta(t-nT)\right\}=\frac{1}{T}\sum_{n=-\infty}^{\infty}F\left(\omega-\frac{2\pi n}{T}\right)$$

$$=\frac{1}{T}\sum_{n=-\infty}^{\infty}F(\omega-n\omega_o) \tag{15.5}$$

That is

$$\hat{F}_s(\omega)=\left\{T\left(\frac{\sin\omega T/2}{\omega T/2}\right)e^{-j\omega T/2}\right\}\sum_{n=-\infty}^{\infty}F(\omega-n\omega_o) \tag{15.6}$$

where the function

$$F_s(\omega)=\sum_{n=-\infty}^{\infty}F(\omega-n\omega_o) \tag{15.7}$$

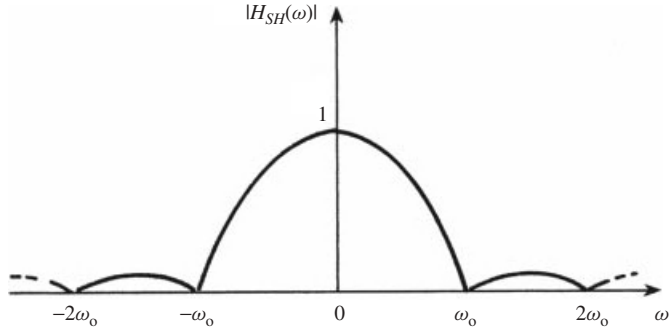


Figure 15.3 The amplitude of the sample and hold $\text{sinc}(x) = \sin x/x$ function

is the spectrum of the signal when impulse-sampled. In contrast, the spectrum

$$H_{SH}(\omega) = \left\{ T \left(\frac{\sin \omega T/2}{\omega T/2} \right) e^{-j\omega T/2} \right\} \tag{15.8}$$

is produced by the holding effect of the pulses. Therefore, this $(\sin x/x)$ function shown in Figure 15.3, multiplies the spectrum $F_s(\omega)$. Thus, this spectrum will be distorted in amplitude by the factor $(\sin \omega T/2)/\omega T/2$ while the factor $e^{-j\omega T/2}$ will introduce a linear phase shift, or a constant delay. This can be thought of as a system performing the sample and hold function whose transfer function is given by (15.8).

From the above discussion, it is noted that the sampling theorem is still valid in this case. However, the recovery of the sampled and held signal is not possible exactly due to the distortion introduced by the factor in (15.8). Therefore, this effect must be compensated for if perfect reconstruction of the signal is to be achieved.

We end this section by reference to Figure 15.4, which shows a general switched-capacitor filter including the interfaces with continuous-time environment. The functions of the constituent building blocks are as follows:

1. The anti-aliasing filter ensures that the input signal spectrum is bandlimited to half the sampling frequency.
2. The sample and hold stage ensures that the input to the sampled-data (switched-capacitor) filter is now of the sampled-data (discrete) type.
3. The smoothing filter provides a continuous time signal from the sampled and held one.
4. The amplitude equalizer is needed so that the reduction in the amplitude spectrum of the output due to the $\sin x/x$ function in (15.8) is undone. Naturally this should be a network which approximates the inverse of this effect, that is it approximates $(x/\sin x)$.

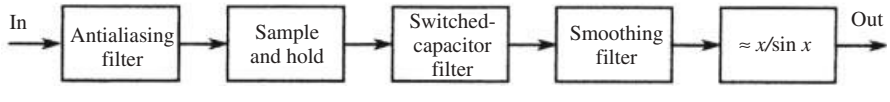


Figure 15.4 The switched-capacitor filter in continuous-time environment

15.3 Amplitude-oriented Filters of the Lossless Discrete Integrator Type

In this section, a very useful class of switched-capacitor filter is studied. This is characterized by its low-sensitivity properties relative to element value variations, a highly desirable attribute from the practical viewpoint. We have introduced this modelling method in Chapter 3 in relation to continuous-time filters, but it is repeated here for completeness. The building blocks and topologies discussed are also applicable to other types of filter discussed throughout this chapter.

15.3.1 The State-variable Ladder Filter

Consider the general passive ladder shown in Figure 15.5 where the branches are arbitrary impedances.

Write the state equations of the ladder, relating the series currents to the shunt voltages. For the sake of specificity, we assume n to be odd.

$$\begin{aligned} I_1 &= Z_1^{-1}(V_2 - V_g) \\ V_2 &= Z_2(I_1 - I_3) \\ I_3 &= Z_3^{-1}(V_2 - V_4) \\ &\cdots\cdots\cdots\cdots \\ &\cdots\cdots\cdots\cdots \\ I_n &= Z_n^{-1}V_{n-1} \end{aligned} \tag{15.9}$$

The transfer function of interest is given by

$$H_{21} = \frac{I_n}{V_g} \tag{15.10}$$

Regardless of the specific form of the branch impedances, any other circuit which implements the same set of equations (15.9) will produce the same transfer function. Note that we have deliberately refrained from using any specific frequency variable in the equations.

Now, in seeking an active implementation of (15.9) all we need are some building blocks capable of providing the frequency dependence of the branches as well as performing the mathematical operations of (15.9). A block diagram of the required implementation is shown in Figure 15.6 and is known as the *state variable leap-frog ladder*

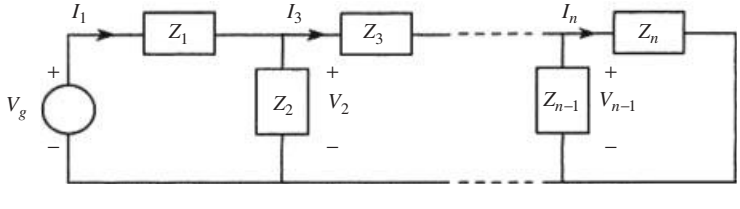


Figure 15.5 A general passive ladder

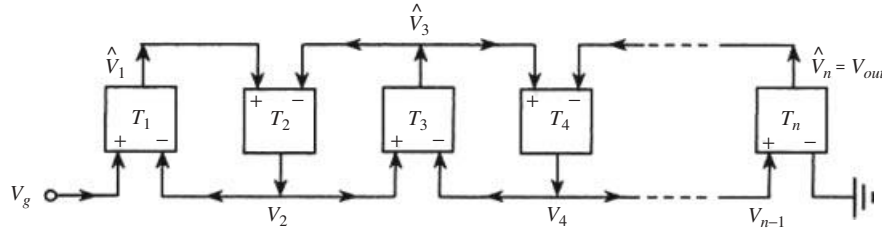


Figure 15.6 The state-variable (leap-frog) equivalent of the ladder in Figure 15.5

realization. This consists of differential-input boxes which possess voltage transfer functions $T_1, T_2, \ldots\ldots T_n$ connected such that:

$$\begin{aligned}
\hat{V}_1 &= T_1(V_g - V_2) \\
V_2 &= T_2(\hat{V}_1 - \hat{V}_3) \\
\hat{V}_3 &= T_3(V_2 - V_4) \\
&\cdots\cdots\cdots\cdots\cdots \\
&\cdots\cdots\cdots\cdots\cdots \\
\hat{V}_n &= T_n V_{n-1}
\end{aligned} \tag{15.11}$$

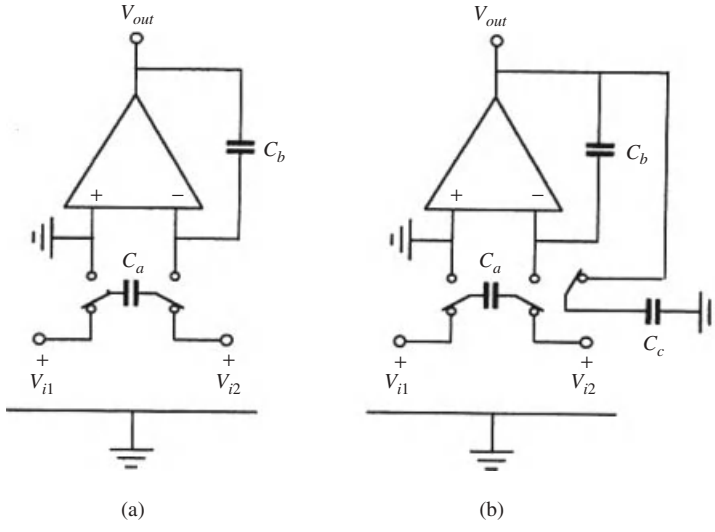
Let the currents $I_1, I_3, I_5, \ldots . I_n$ in (15.9) be simulated by the voltages $\hat{V}_1, \hat{V}_3, \hat{V}_5, \ldots . \hat{V}_n$. The internal structure of each box in Figure 15.6 is chosen such that

$$\begin{aligned}
T_1 &= \alpha Z_1^{-1} \\
T_2 &= \alpha^{-1} Z_2 \\
T_3 &= \alpha Z_3^{-1} \\
&\cdot \\
&\cdot \\
&\cdot \\
T_n &= \alpha Z_n^{-1}
\end{aligned} \tag{15.12}$$

where α is a constant. Then, the transfer function of the leap-frog ladder of Figure 15.6.

$$\hat{H}_{21} = \frac{V_{out}}{V_g} \tag{15.13}$$

only differs from H_{21} of the passive ladder by a constant. Consequently, given a ladder of the general form in Figure 15.5, with the specific type and values of elements, it is possible to determine the state variable simulation provided we can find the necessary building blocks with transfer functions satisfying (15.12). Conversely, if we decide on certain types of building block for the boxes in Figure 15.6 we can find the equivalent ladder. Starting from Figure 15.6, two basic types of building block are used as the boxes. These are shown in Figure 15.7 and are known as the *lossless discrete integrator* (LDI)



(a) (b)

Figure 15.7 The basic building blocks for a class of switched-capacitor filter. (a) Type A: lossless discrete integrator (LDI). (b) Type B: damped discrete integrator (DDI)

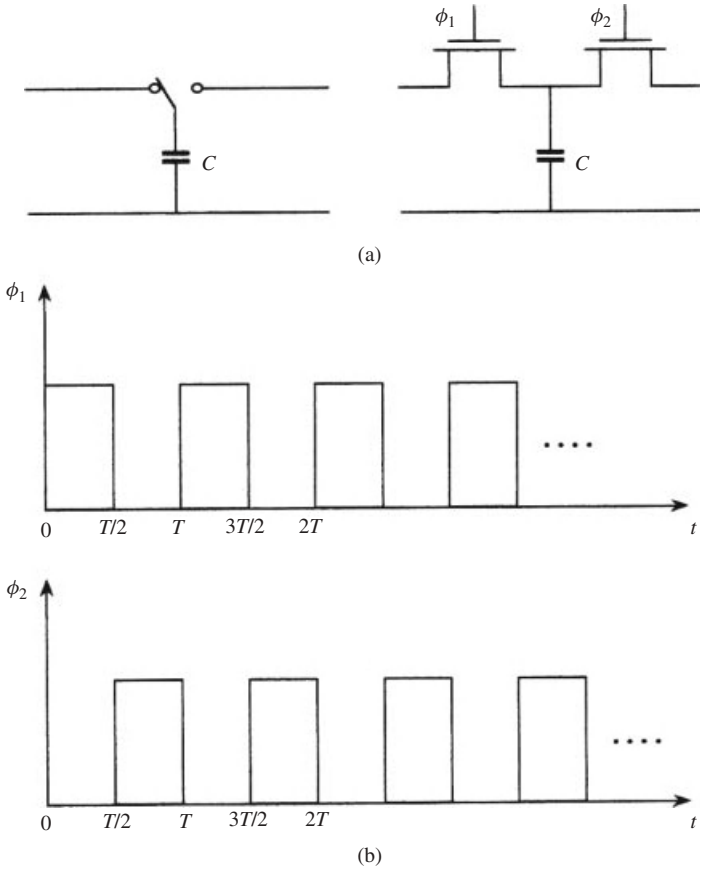


Figure 15.8 (a) A switched-capacitor. (b) The biphase clock driving the switches

and *damped discrete integrator* (DDI). These may be considered generic building blocks, in the sense that any other set which perform the same functions would be acceptable.

Type A. This is the building block of Figure 15.7(a) and is the LDI. The Op Amp is assumed ideal and the switches are driven by a biphase clock with non-overlapping pulses of period T as shown in Figure 15.8.

The capacitor C_a is switched to the inputs V_{i1} and V_{i2} at $t = (n-1)T$ and to the Op Amp inputs at $t = (n-1/2)T$. The output of the OP Amp is sampled at $t = (n-1)T$. Thus

$$V_{out}(nT) = V_{out}\{(n-1)T\} + \frac{C_a}{C_b}\left(V_{i1}\left\{\left(n-\frac{1}{2}\right)T\right\} - V_{i2}\left\{\left(n-\frac{1}{2}\right)T\right\}\right) \tag{15.14}$$

Taking the z-transform

$$V_{out}(z) = z^{-1}V_{out}(z) + z^{-1/2}\frac{C_a}{C_b}\{V_{i1}(z) - V_{i2}(z)\} \tag{15.15}$$

Therefore, this building block has the transfer function

$$\begin{aligned} T_A &= \frac{V_{out}(z)}{V_{i1}(z) - V_{i2}(z)} \\ &= \frac{z^{-1/2}}{\left(\frac{C_b}{C_a}\right)(1-z^{-1})} \\ &= \frac{1}{2\left(\frac{C_b}{C_a}\right)\gamma} \end{aligned} \tag{15.16a}$$

where

$$\gamma = \sinh(T/2)s \tag{15.16b}$$

Type B. This is the circuit of Figure 15.7(b) which is the same as that of the LDI but with a feedback capacitor C_C. This is known as the *damped discrete integrator* (DDI). With the same sequence of switching as Type A, we have

$$\begin{aligned} V_{out}(nT) = V_{out}\{(n-1)T\} &+ \frac{C_a}{C_b}\left(V_{i1}\left\{\left(n-\frac{1}{2}\right)T\right\} - V_{i2}\left\{\left(n-\frac{1}{2}\right)T\right\}\right) \\ &- \frac{C_c}{C_b}V_{out}\{(n-1)T\} \end{aligned} \tag{15.17}$$

and taking the z-transform

$$V_{out}(z) = z^{-1}V_{out}(z) + \frac{C_a}{C_b}z^{-1/2}\{V_{i1}(z) - V_{i2}(z)\} - \frac{C_c}{C_b}z^{-1}V_{out}(z) \tag{15.18}$$

$$
\begin{aligned}
T_A &= \frac{V_{out}(z)}{V_{i1}(z) - V_{i2}(z)} \\
&= \frac{z^{-1/2}}{\left(\frac{C_b}{C_a}\right)\left\{1 - z^{-1}\left(1 - \frac{C_c}{C_b}\right)\right\}} \\
&= \frac{1}{\left(\frac{2C_b - C_c}{C_a}\right)\gamma + \frac{C_c}{C_a}\mu}
\end{aligned}
\tag{15.19}
$$

with

$$\mu = \cosh(T/2)s$$

Now let these building blocks be used as the boxes in Figure 15.6 such that the *first* and *last* boxes are *Type B* while the *internal* ones are *Type A*. The resulting network takes the form shown in Figure 15.9 and is clearly of the analog sampled-data type. It is called a *switched-capacitor state-variable ladder*.

Examination of (15.9) and (15.12) shows that the switched-capacitor network has the equivalent network of Figure 15.10 where

$$
\begin{aligned}
L_k(or\ C_k) &= 2\left(\frac{C_b}{C_a}\right)_k,\ k = 2, 3, \ldots (n-1) \\
L_{1,n} &= \left(\frac{2C_b - C_c}{C_a}\right)_{1,n} \\
R_{g,L} &= \left(\frac{C_c}{C_a}\right)_{1,n}
\end{aligned}
\tag{15.20}
$$

Viewed as a doubly terminated ladder two-port, the elements of the equivalent network have the indicated frequency dependence. Direct analysis of the network [24] shows that its transfer function has the general form

$$H_{21}(\lambda) = \frac{(1 - \lambda^2)^{n/2}}{P_n(\lambda)} \tag{15.21}$$

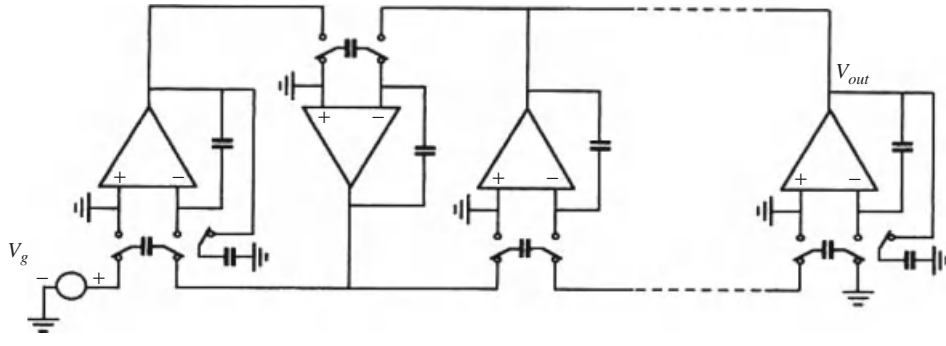


Figure 15.9 The switched-capacitor state variable (leap-frog) ladder filter using generic building blocks

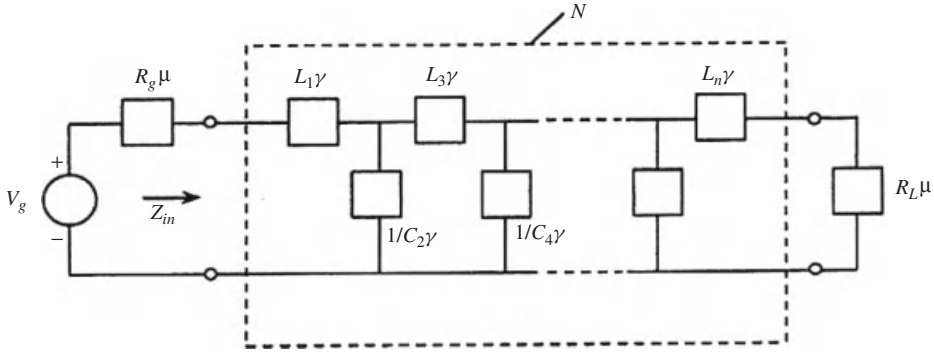


Figure 15.10 Equivalent network of the ladder in Figure 15.5

where

$$\lambda = \tanh(Ts/2) \tag{15.22}$$

For a maximally-flat amplitude response around the origin and one zero derivative at half the sampling frequency, we must have

$$|H_{21}|^2 = \frac{K}{1 + \left(\dfrac{\sin(\omega T/2)}{\sin(\omega_0 T/2)}\right)^{2n}} \tag{15.23}$$

where ω_0 is the 3 dB point. Clearly

$$\omega T/2 = \pi\omega/\omega_N \tag{15.24}$$

where ω_N is the radian sampling frequency. For an optimum equiripple response up to ω_0 we have

$$|H_{21}|^2 = \frac{K}{1 + \varepsilon^2 T_n^2\left(\dfrac{\sin(\omega T/2)}{\sin(\omega_0 T/2)}\right)} \tag{15.25}$$

where T_n is the Chebyshev polynomial of the first kind.

The element values are determined in the following steps:

1. Form the function

$$|S_{21}|^2 = \frac{4KR_L\cos^2(\omega T/2)}{1 + \left(\dfrac{\sin(\omega T/2)}{\sin(\omega_0 T/2)}\right)^{2n}} \tag{15.26}$$

or

$$|S_{21}|^2 = \frac{4KR_L\cos^2(\omega T/2)}{1 + \varepsilon^2 T_n^2\left(\dfrac{\sin(\omega T/2)}{\sin(\omega_0 T/2)}\right)} \tag{15.27}$$

2. Form the function

$$|S_{11}|^2 = 1 - |S_{21}|^2 \tag{15.28}$$

and it is factorized to obtain

$$S_{11}(\lambda) = \frac{N(\lambda)}{D(\lambda)} \tag{15.29}$$

where $D(\lambda)$ is a strictly Hurwitz polynomial for stability.

3. Use

$$\mu^2 = (1 + \gamma^2), \lambda = \gamma/\mu \tag{15.30}$$

to construct the function

$$Z_{in} = \mu \frac{1 + S_{11}}{1 - S_{11}} = \frac{\mu a_{n-1}(\gamma) + b_n(\gamma)}{\mu c_{n-2}(\gamma) + d_{n-1}(\gamma)} \tag{15.31}$$

where a, d are even polynomials, whereas oppositely b, c are odd polynomials.

4. Perform the continued fraction expansion of Z_{in} around γ with μ treated as a constant. The coefficients of γ in the expansion give $L_1, C_2, L_3, \ldots\ldots$ the element values of the network of Figure 15.10. Subsequently the capacitor ratios can be obtained from (15.20).
5. The dc gain can be adjusted by a suitable choice of K. Alternatively, this can be accomplished after the design has been completed by impedance-scaling the entire network by a suitable factor.
6. For even-degree filters, the same equivalent circuit results except that the load has frequency dependence R_g/μ. This means that in obtaining (15.31) we should put it in the same form *but with* μ *replaced by* $1/\mu$ and the continued fraction expansion is obtained in the same way. The final step will yield an element with impedance R_L/μ. In this case, the terminations may be unequal.

This design procedure is now illustrated by an example.

Design Example Consider the calculation of the element values for a seventh-order low-pass Chebyshev filter with 0.05 dB passband ripple and passband edge at 3.4 kHz for a clock frequency of 28 kHz. In this case, we have

$$f_0/f_N = 0.12$$
$$\sin(\pi f_0/f_N) = 0.37$$

Therefore, from (15.23)–(15.30) with $\theta = \pi f_0/f_N$, $K = 1/4, R_L = 1$,

$$|H_{21}|^2 = \frac{1/4}{1 + 0.01T_7^2(\sin\theta/0.37)}$$

$$|S_{21}|^2 = \frac{\cos^2\theta}{1 + 0.01T_7^2(\sin\theta/0.37)}$$

$$|S_{11}|^2 = \frac{\sin^2\theta + 0.01T_7^2(\sin\theta/0.37)}{1 + 0.01T_7^2(\sin\theta/0.37)}$$

The above expression is factored in the λ-domain choosing the left half plane poles (and also the left half-plane zeros for a minimum-phase response), then using (15.30) we obtain

$$Z_{in}(\gamma,\mu) = \frac{15074\gamma^7 + 7629.00\gamma^6\mu + 5924.\gamma^5 + 1889\gamma^4\mu + 622\gamma^3 + 115\gamma^2\mu + 16.\gamma + \mu}{3549\gamma^6 + 1680\gamma^5\mu + 1120\gamma^4 + 326\gamma^3 + 12\gamma\mu + 1}$$

Performing the continued fraction expansion around γ, we obtain for the element values of Figure 15.10.

$$L_1 = 4.53,\ C_2 = 4.12,\ L_3 = 5.18,\ C_4 = 4.4$$
$$L_5 = 4.77,\ C_6 = 3.74,\ L_7 = 2.05,\ R_L = 1$$

The actual capacitor ratios are then obtained from the above values and (15.20). The resulting network has a gain of 0.25 (−6 dB). To adjust the gain value to 1 (0 dB), the network is impedance-scaled by 0.5.

15.3.2 Strays-insensitive LDI Ladders

When designing switched-capacitor filters using silicon integrated circuits, as discussed in Chapter 10, the effect of stray parasitic capacitances must be taken into account. Attention has, therefore, been given to the use of building blocks which perform the same functions as those of Figure 15.7 but are *insensitive* to parasitic capacitances. This objective can be easily achieved by modifying the building blocks. Before showing this modification, let us examine the problem of parasitics in some detail. Consider Figure 15.11 which shows a typical circuit which is sensitive to these parasitics. The capacitance C_1 represents all the parasitics associated with node 1. This includes the source to drain diffusion capacitances of the transistors making up the switches. It also includes the capacitances of the leads connecting the top plate of C_A to the two transistors. The total stray capacitance is uncontrollable resulting in an error that can be as high as 50% of the required nominal value. The rather obvious solution to take C_A very large is impractical since this would result in inordinately large areas on the chip. Next, consider the circuit of Figure 15.11(b). The parasitic capacitance C_1 is charged from V_{in} then discharged to ground. Each of the parasitic capacitances C_2 and C_3 is grounded at both ends. Also C_4 is driven by the low impedance Op Amp output. It follows that none of theses parasitic capacitances affects the performance of the circuit of Figure 15.11(b). However, the circuit is sensitive to the parasitic capacitance between the clock signals to node 4. Minimization of this effect will be discussed later, and it will be shown that this can be achieved by modification of the clocking scheme.

Now, using building blocks that are strays-insensitive in the manner discussed above, we can obtain parasitic-insensitive leap-frog ladders according to the following procedure:

(a) Modify the state variable leap-frog block diagram of Figure 15.6 as shown in Figure 15.12. Clearly, this new structure still implements the same state equations. Each building block together with the summer at its input implements a typical equation in the set of equations (15.11) which simulate the ladder in Figure 15.5.
(b) Instead of the building blocks of Figure 15.7 we use the modified ones of Figures 15.13 and 15.14.

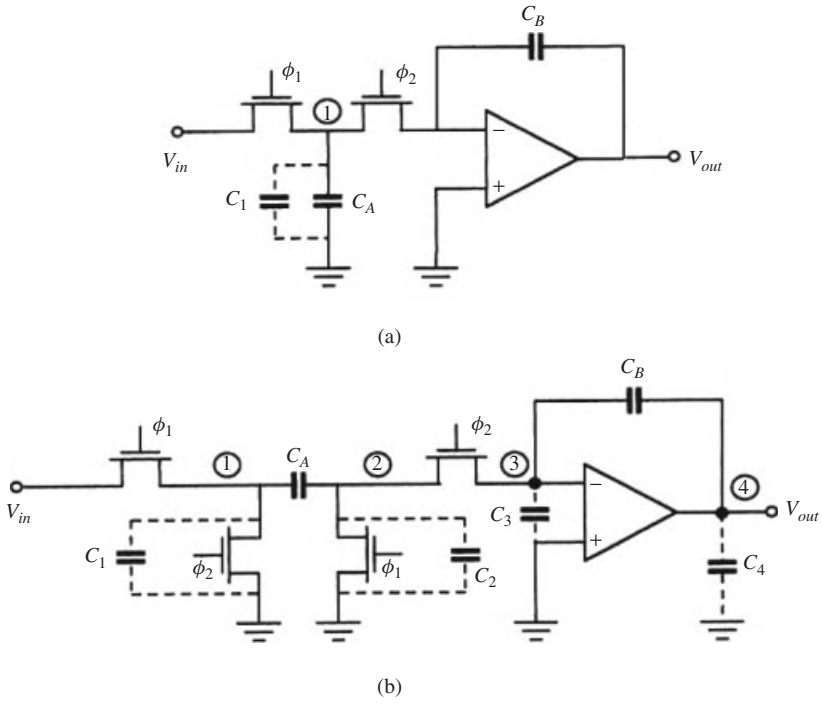


Figure 15.11 (a) Strays-sensitive building block (b) Strays-insensitive building block

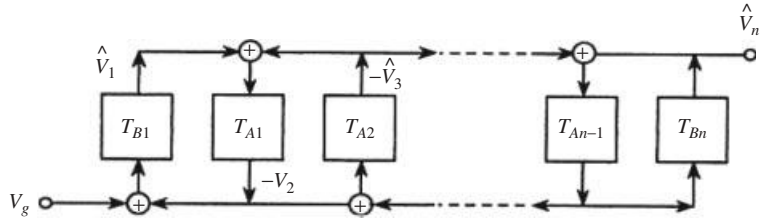


Figure 15.12 Alternative active simulation of the ladder in Figure 15.5

These are as follows:

Modified Type A. These are the circuits of Figure 15.13 with transfer functions

$$T_A = \frac{z^{-1}}{\frac{C_b}{C_a}(1 - z^{-1})} = \frac{z^{-1/2}}{2\left(\frac{C_b}{C_a}\right)\gamma} \tag{15.32}$$

for the circuit of Figure 15.13(a) and

$$\hat{T}_A = \frac{-1}{\frac{C_b}{C_a}(1 - z^{-1})} = \frac{-z^{-1/2}}{2\left(\frac{C_b}{C_a}\right)\gamma} \tag{15.33}$$

for the network of Figure 15.13(b).

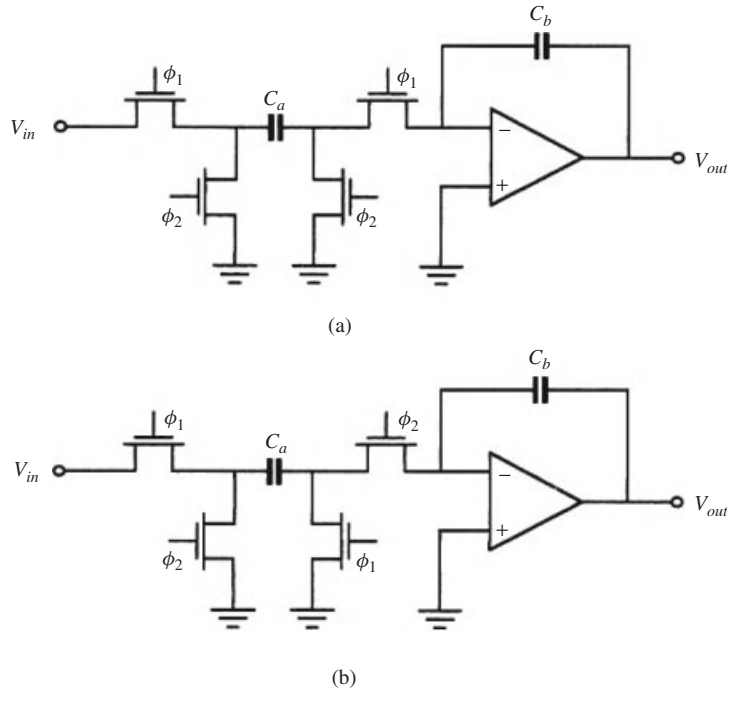


Figure 15.13 Strays-insensitive building blocks of type A: (a) positive, (b) negative

Modified Type B. These are the circuits of Figure 15.14 with transfer functions

$$T_B = \frac{z^{-1}}{\frac{C_b}{C_a}\left(1 - z^{-1} + \frac{C_c}{C_b}\right)} = \frac{z^{-1/2}}{\frac{2C_b + C_c}{C_a}\gamma + \frac{C_c}{C_a}\mu} \tag{15.34}$$

for the circuit of Figure 15.14(a) and

$$\hat{T}_B = \frac{-1}{\frac{C_b}{C_a}\left(1 - z^{-1} + \frac{C_c}{C_b}\right)} = \frac{-z^{1/2}}{\frac{2C_b + C_c}{C_a}\gamma + \frac{C_c}{C_a}\mu} \tag{15.35}$$

for the circuit of Figure 15.14(b).

(c) In the realization of Figure 15.12 using the modified building blocks, two points must be observed. First, positive and negative building blocks alternate. Second, the first and last building blocks are of *Type B* while the rest are of *Type A*. Clearly the summing operation is easily realized by adding another capacitor of the same value C_a to every building block, that is between the voltage to be summed and the node where the original C_a is connected as shown in Figure 15.15 for Type A blocks. The clock phases for negative Type A are shown between brackets. The same input arrangement is used with a Type B building block with a summer. A fifth-order example is shown in Figure 15.16.

Having modified the procedure as outlined above, it is a simple matter to show that the resulting filter has the equivalent circuit shown in Figure 15.17, which is the same as

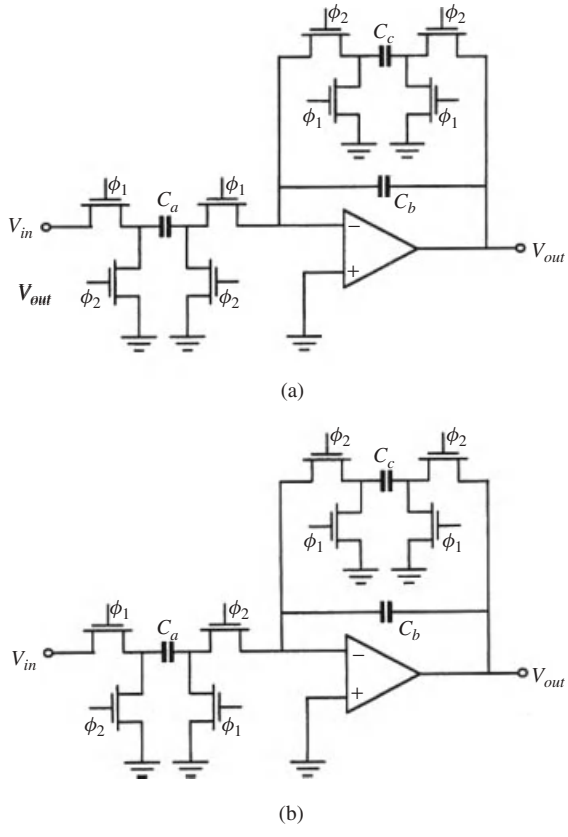


Figure 15.14 Strays-insensitive building blocks of Type B: (a) positive, (b) negative

that of Figure 15.10 except that all impedances are scaled by $z^{-1/2}$. This multiplies the transfer function by the same factor which has no effect on the amplitude response of the filter. The resulting structure, however, has the advantage of being completely insensitive to stray parasitic capacitances inherent in the IC fabrication process.

Henceforth, we shall use inductor and capacitor symbols to denote elements of impedance and admittance frequency dependence kx, respectively whether $x = \gamma$ *or* s *or* λ. *We shall therefore speak of a* γ*-domain inductor and capacitor and so on. This should not result in any confusion since the context is very clear and explicit*.

Figure 15.18 shows a typical response of a Chebyshev filter designed using the technique discussed so far.

15.3.3 An Approximate Design Technique

The earliest design techniques of switched-capacitor filters relied on the assumption of a very high sampling rate as compared with the bandwidth of the signal. This leads to a very simple design technique that gives a filter directly related to a passive lumped prototype. To see this, let the sampling frequency be such that

$$\omega_N \gg \omega \tag{15.36}$$

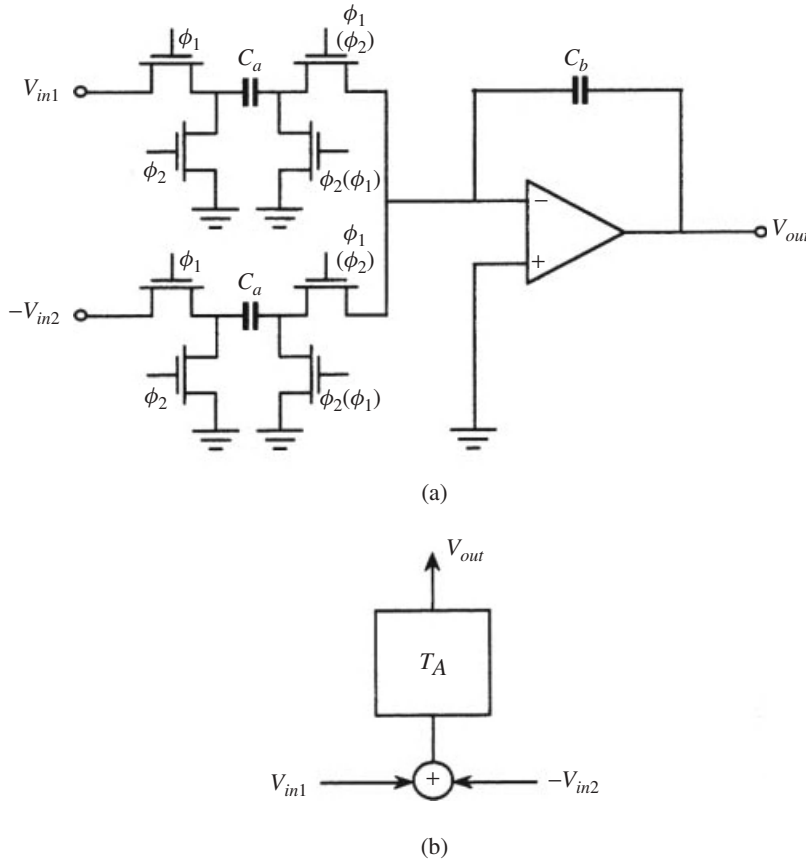


Figure 15.15 (a) LDI summer implementation of a typical section as in (b) for T_A positive. The clock phases for T_A negative are between brackets. The same input arrangement is used with *Type B* circuits

for all ω in the frequency band of interest. Then on the $j\omega$-axis

$$\sinh(j\pi\omega/\omega_N) \approx j\pi\omega/\omega_N$$
$$\cosh(j\pi\omega/\omega_N) \approx 1 \tag{15.37}$$

and we arrive at an approximate equivalent circuit of the filter of the form given in Figure 15.19 which is a true resistively- terminated LC ladder with element values $g_r = L_r$, C_r. In this case we have the element values of the switched-capacitor equivalent network in Figure 15.19 as

$$g_r' = g_r\pi/\omega_N = g_r T/2 \tag{15.38}$$

However, this method is inexact with unacceptable disadvantages. First, it forces the use of an excessively high sampling rate, so that the useful bandwidth of applications of the filters becomes severely limited. Second, the element values depend on the sampling frequency and they lead to large capacitor ratios. The flexibility of programming the filter using a variable clock frequency is lost.

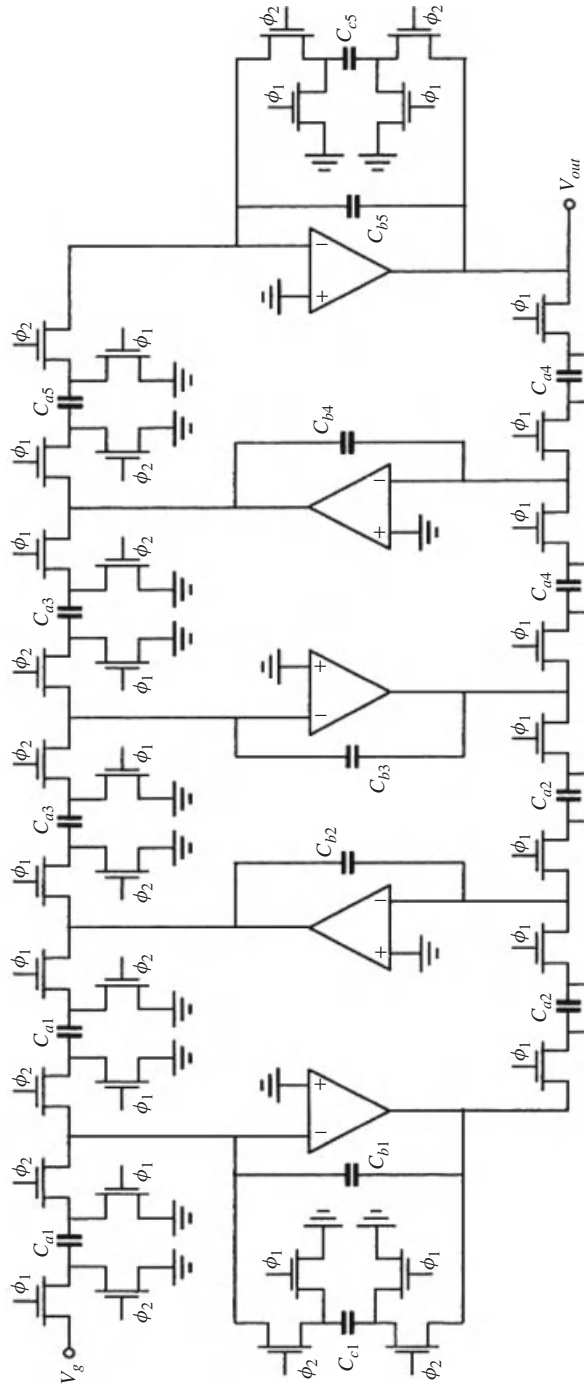


Figure 15.16 The circuit of a fifth-order low-pass filter with strays-insensitive building blocks

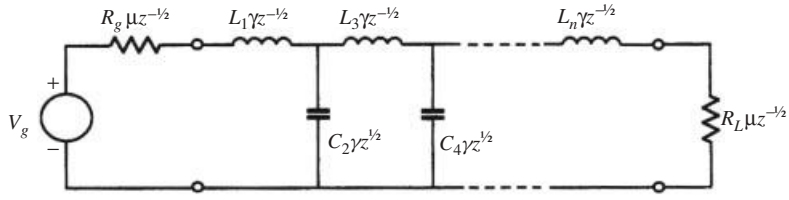


Figure 15.17 Equivalent circuit of stray-insensitive LDI ladder

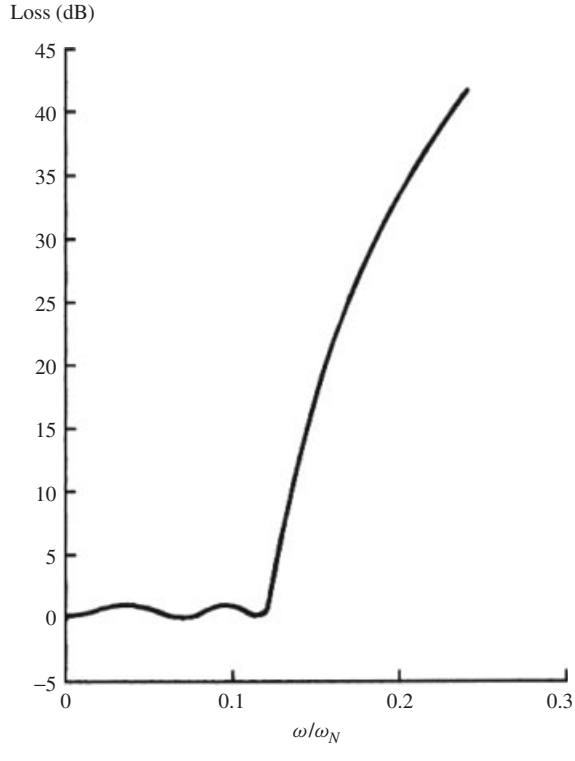


Figure 15.18 Amplitude response of a Chebyshev fifth-order LDI filter

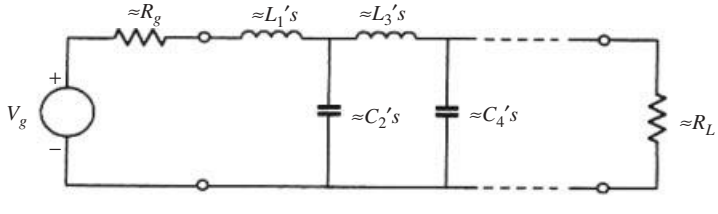


Figure 15.19 Approximate equivalent circuit of LDI filter under the assumption of very high sampling frequency

15.4 Filters Derived from Passive Lumped Prototypes

In this approach, LC ladder prototypes are first designed completely in the s-domain, such that when the bilinear transformation is applied, the filter meets the required specifications. Thus in the low pass LC prototype, the following transformation is used

$$s \rightarrow \lambda/\Omega_0 \tag{15.39}$$

which amounts to replacing every inductor in the prototype by an element with impedance L/Ω_0 while every capacitor is replaced by an element with admittance C/Ω_0. Here

$$\Omega_0 = \tan(\pi\omega_0/\omega_N) \tag{15.40}$$

Thus, we obtain the λ-domain ladder shown in Figure 15.20. Its transfer function is that of the corresponding lumped prototype with the bilinear transformation applied.

The problem now is to find the switched-capacitor building blocks for the realization of the λ – plane ladder. Actually, the required building blocks are basically the same ones used in the construction of the LDI ladder filters discussed in the previous section. There are some differences, however, in the addition of some extra capacitors for the introduction of finite zeros of transmission on the imaginary axis and in the input section. Thus, the circuits of Figures 15.13 and 15.14 are used, together with the more elaborate section shown in Figure 15.21 in the input stage. This will now be explained in detail.

First, we note that direct analysis of the circuit in Figure 15.21 shows that its transfer function is given by

$$\frac{V_{out}}{V_{in}} = \frac{-C_{a1}z^{1/2} + C_{a2}z^{-1/2} - 2C_{a3}\gamma - 2C_s\mu}{(2C_b + C_f)\gamma + C_f\mu} \tag{15.41}$$

The additional Op Amp is required for the buffer of the sample and hold circuit.

Suppose that, based on the specifications of the filter, we already have an nth order elliptic λ-domain ladder as shown in Figure 15.20 in which the order of the filter is assumed odd. We begin by some manipulations of the circuit to convert it to a form suitable for simulation using the switched-capacitor building blocks. First, we extract from each of the mid-shunt capacitors a negative capacitance of a value that resonates with the respective parallel inductor at $\lambda = \pm 1$. Now, the parallel combination of $L_i\lambda$ and $-L_i/\lambda$ gives an impedance

$$\frac{L_i(-L_i/\lambda)}{L_i\lambda - (L_i/\lambda)} = L_i\frac{\lambda}{1-\lambda^2} = L_i\gamma\mu \tag{15.42}$$

Then the network of Figure 15.22 is obtained.

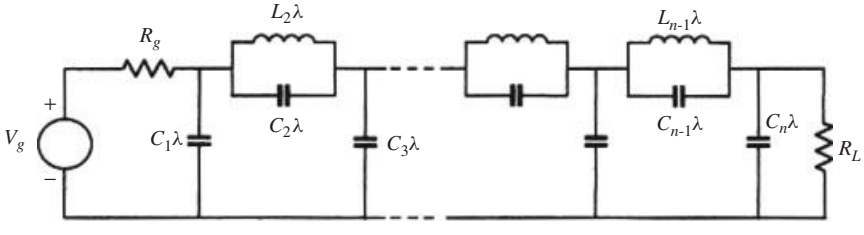


Figure 15.20 λ-domain ladder obtained from a lumped passive prototype

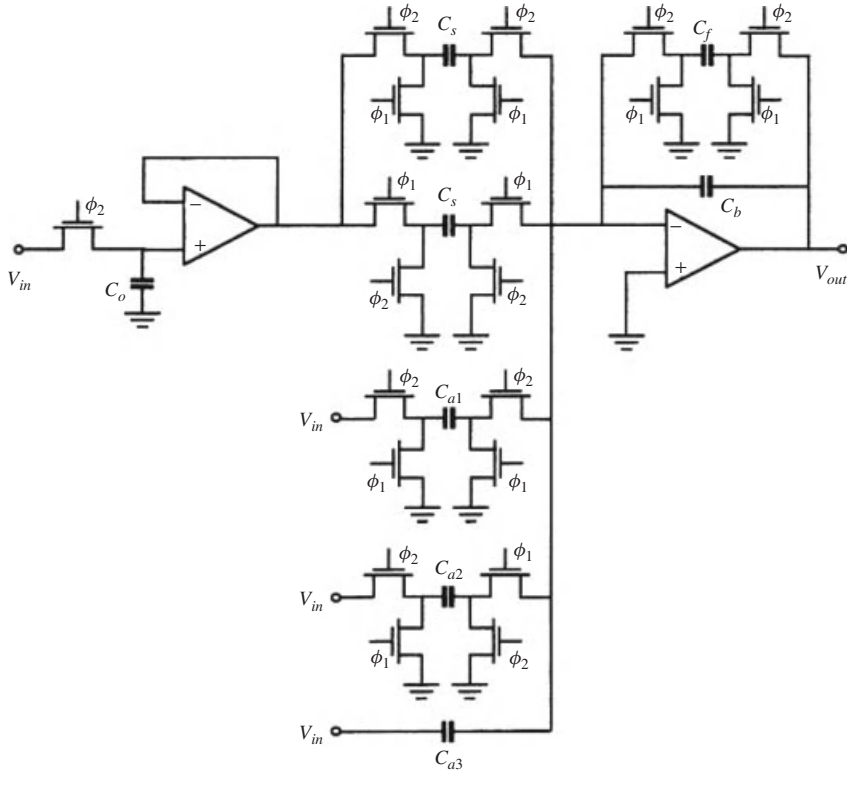


Figure 15.21 A composite LDI building block

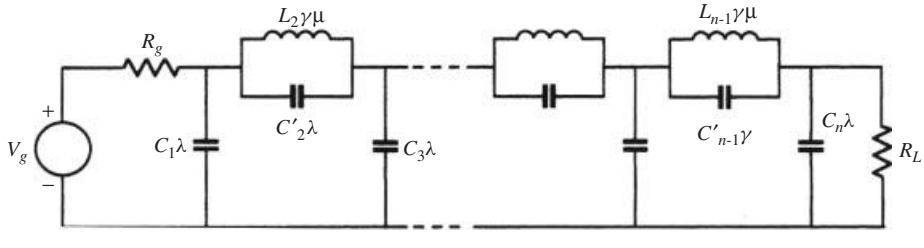


Figure 15.22 Modified network of Figure 15.20

Next we employ Norton's theorem to replace each of the bridging capacitors $C'_2, C'_4, \ldots$. by two voltage controlled current sources (VCCSs). Finally, we divide all impedances by μ (including the transimpedances of the VCCSs). This scaling does not affect the voltage transfer function of the network which is, in this case, a voltage transfer ratio. The resulting network is shown in Figure 15.23.

The operation of this network can be described by the following state equations:

$$V_j = \frac{(\mu I_{j-1}) - (\mu I_{j+1}) + \gamma C'_{j-1} V_{j-2} + \gamma C_{j+1} V_{j+2}}{\gamma C'_j} \quad \text{for } j = 3, 5, 7, \ldots . \qquad (15.43)$$

$$\mu I_j = \frac{V_{j-1} - V_{j+1}}{\gamma L_j} \quad \text{for } j = 2, 4, 6, \ldots . \tag{15.44}$$

$$V_1 = \frac{(\mu/R_g)V_g - (\mu I_2) + C_2'\gamma V_3}{C_1'\gamma + (\mu/R_g)} \tag{15.45}$$

$$V_n = \frac{(\mu I_{n-1}) + C_{n-1}'\gamma V_{n-2}}{C_n'\gamma + (\mu/R_L)} \tag{15.46}$$

Comparing the transfer functions of LDI building blocks in Figures 15.13 and 15.14 with Equations (15.43) and (15.44), it is clear that the equations can be realized using these building blocks. In contrast, Equations (15.45) and (15.46) of the terminating sections (at the input and output) can be realized using the building block of Figure 15.21 with a transfer function of the form given by (15.41). Let us examine the procedure in detail in relation to a fifth-order filter. In this case, the five equations are given by

$$V_1 = \frac{(\mu/R_g)V_g - (\mu I_2) + C_2'\gamma V_3}{C_1'\gamma + (\mu/R_g)} \tag{15.47}$$

$$\mu I_2 = \frac{V_1 - V_3}{\gamma L_j} \tag{15.48}$$

$$V_3 = \frac{(\mu I_2) - (\mu I_4) + \gamma C_{j-1}' V_1 + \gamma C_{4_5}}{\gamma C_3'} \tag{15.49}$$

$$\mu I_4 = \frac{V_3 - V_5}{\gamma L_4} \tag{15.50}$$

$$V_n = \frac{(\mu I_4) + C_4'\gamma V_3}{C_5'\gamma + (\mu/R_L)} \tag{15.51}$$

Figure 15.24 shows the switched-capacitor circuit of a fifth-order elliptic filter.

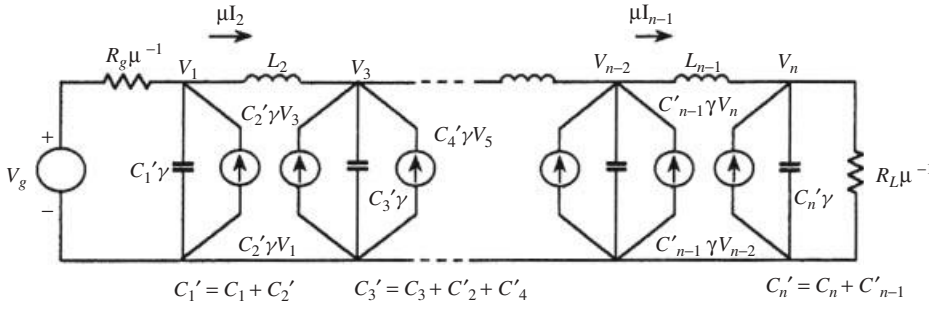


Figure 15.23 Network of Figure 15.22 after impedance scaling by $1/\mu$ and modification using VCCS-equivalents

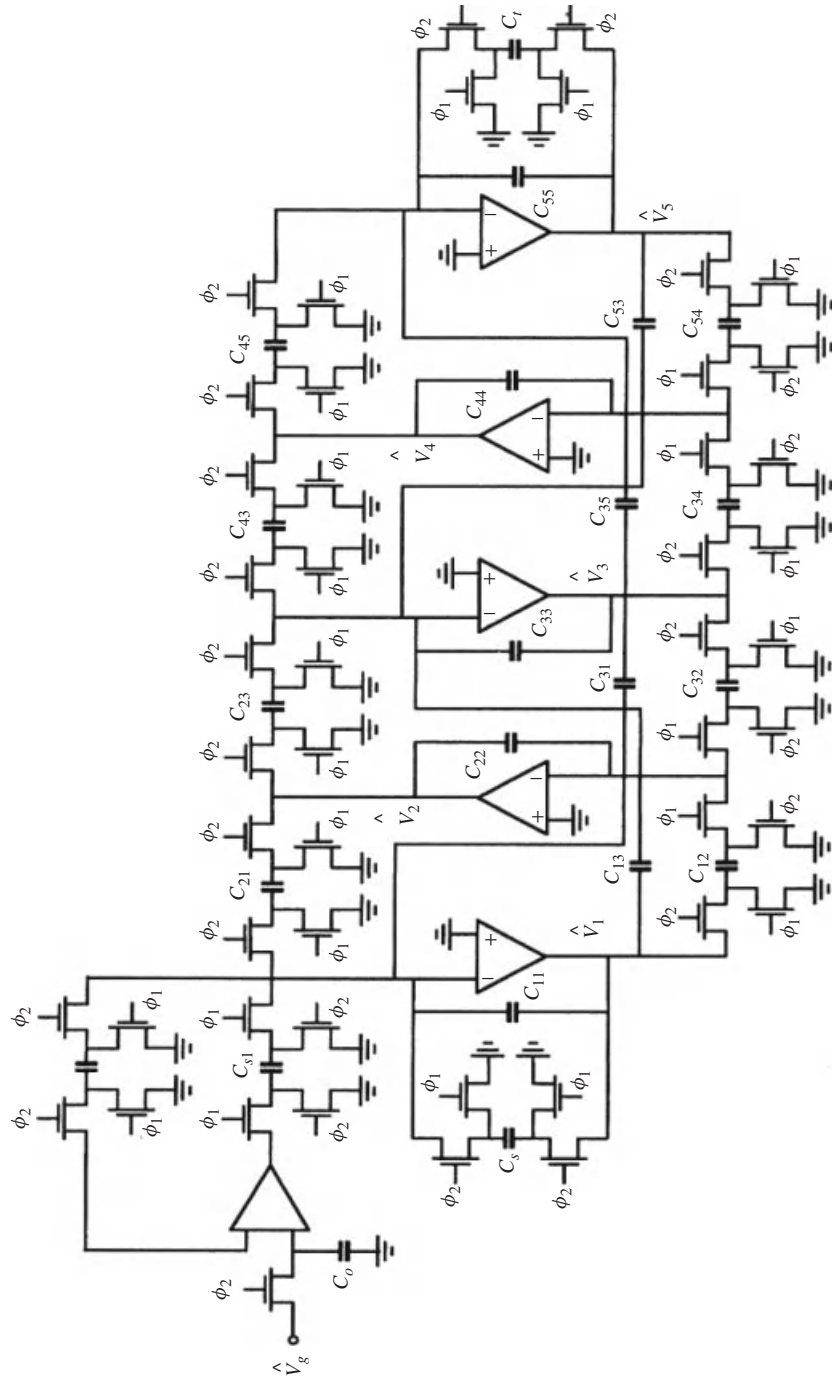


Figure 15.24 Fifth-order ladder filter with finite $j\omega$-axis zeros

To determine the capacitor ratios, the state equations of the SC circuit are written as follows:

$$\hat{V}_1 = \frac{-2C_{s1}\mu\hat{V}_g - C_{21}(z^{12}\hat{V}_2) - 2C_{31}\gamma\hat{V}_3}{(2C_{11} + C_s)\gamma + C_s\mu} \tag{15.52}$$

$$\hat{V}_2 = \frac{C_{21}\hat{V}_1 + C_{32}\hat{V}_3}{2C_{22}\gamma} \tag{15.53}$$

$$\hat{V}_3 = \frac{-2C_{13}\mu V_1 - C_{23}(z^{12}\hat{V}_4) - 2C_{31}\gamma\hat{V}_3}{2C_{33}\gamma} \tag{15.54}$$

$$\hat{V}_2 = \frac{C_{34}\hat{V}_3 + C_{32}\hat{V}_5}{2C_{44}\gamma} \tag{15.55}$$

$$\hat{V}_5 = \frac{-2C_{35}\mu\hat{V}_3 - C_{45}(z^{12}\hat{V}_4)}{(2C_{551} + C_\ell)\gamma + C_\ell\mu} \tag{15.56}$$

Next, all the variables appearing in (15.47)–(15.51) are simulated by voltages in expressions (15.52)–(15.56) and we establish the following correspondences

$$\hat{V}_g \Leftrightarrow V_g \tag{15.57}$$

$$\hat{V}_1 \Leftrightarrow V_1 \tag{15.58}$$

$$z^{1/2}\hat{V}_2 \Leftrightarrow \mu I_2 \tag{15.59}$$

$$\hat{V}_3 \Leftrightarrow V_3 \tag{15.60}$$

$$z^{1/2}\hat{V}_4 \Leftrightarrow \mu I_4 \tag{15.61}$$

$$\hat{V}_5 \Leftrightarrow V_5 \tag{15.62}$$

Then, the capacitor ratios are obtained by equating the coefficients of the transfer function of the building block to the coefficients in the state equations of the ladder prototype. The technique gives the capacitor ratios in the following order:

1. For building block 1:

$$\begin{aligned} ratio1 &= \frac{C_s}{2C_{11} + C_s} \\ ratio2 &= \frac{2C_{s1}}{C_s} \\ ratio3 &= \frac{C_{21}}{2C_{11} + C_s} \\ ratio4 &= \frac{2C_{31}}{2C_{11} + C_s} \end{aligned} \tag{15.63}$$

2. For building block n:

$$ratio1 = \frac{C_{n-1,n}}{C_{n,n} + C_\ell}$$
$$ratio2 = \frac{C_\ell}{2C_{n,n} + C_\ell} \tag{15.64}$$
$$ratio = \frac{2C_{n-2,2n}}{2C_{n,n} + C_\ell}$$

3. For building block i (i odd):

$$ratio1 = \frac{C_{i-1,i}}{2C_{i,i}}$$
$$ratio2 = \frac{C_{i+1,i}}{2C_{i,i}} \tag{15.65}$$
$$ratio3 = \frac{C_{i-2,i}}{2C_{i,i}}$$
$$ratio4 = \frac{C_{i+2i}}{2C_{i,i}}$$

4. For building block i (i even):

$$ratio1 = \frac{C_{i-1,i}}{2C_{i,i}}$$
$$ratio2 = \frac{C_{i+1,i}}{2C_{i,i}} \tag{15.66}$$

It is then possible to scale the capacitance values for minimum total capacitance and dynamic range. These ideas will be discussed in a later chapter. But this option can be bypassed in the design. It is a simple matter to write a MATLAB®-based computer program to implement the design procedure. It is left as an exercise to the reader to write such a program producing the capacitor values in the following order:

$$\begin{aligned}
&C_{s1}, C_{11}, C_{21}, C_{31}, C_s \\
&C_{i-1,i}, C_{i,i}, C_{i+1,i} \text{ for } i \text{ even} \\
&C_{i-1,i}, C_{i,i}, C_{i+1,i}, C_{i-2,i}, C_{i+2,i} \text{ for } i \text{ odd} \\
&C_{n-1,n}, C_{n,n}, C_{n-2,n} C_\ell (\text{with } C_\ell = 1)
\end{aligned} \tag{15.67}$$

This is illustrated by an example.

Example Consider the design of an elliptic filter with passband edge at 0.144 of the sampling frequency with 0.044 dB ripple, a stopband edge at 0.2 of the sampling frequency, with minimum attenuation of 39 dB.

We can use MATLAB® to obtain the required degree as 5 and the elliptic filter tables [16] to obtain the normalized prototype element values as

$$C_1 = 0.85535, C_3 = 0.15367, L_2 = 1.20763, C_3 = 1.48438, C_4 = 0.46265,$$
$$L_4 = 0.89794, C_5 = 0.63702, R_g = R_L = 1$$

These values are then divided by $\tan(\pi\omega_o/\omega_N)$ producing the λ-domain values. Then equations (15.63)–(15.67) are used to obtain the capacitor ratios in the final design. These calculations give for the building blocks:

Building block 1	0.40365, 2.0000, 0.40365,0.29011
Building block 2	0.40274, 0.40274
Building block 3	0.18998, 0.1998, 0.13654, 0.283263
Building block 4	0.54164, 0.54164

If the minimum allowed capacitance value is 1 pF (or one unit, say) with minimum capacitance scaling (which will be explained in the next chapter) we obtain the capacitance values in the order given in (15.67) as follows:

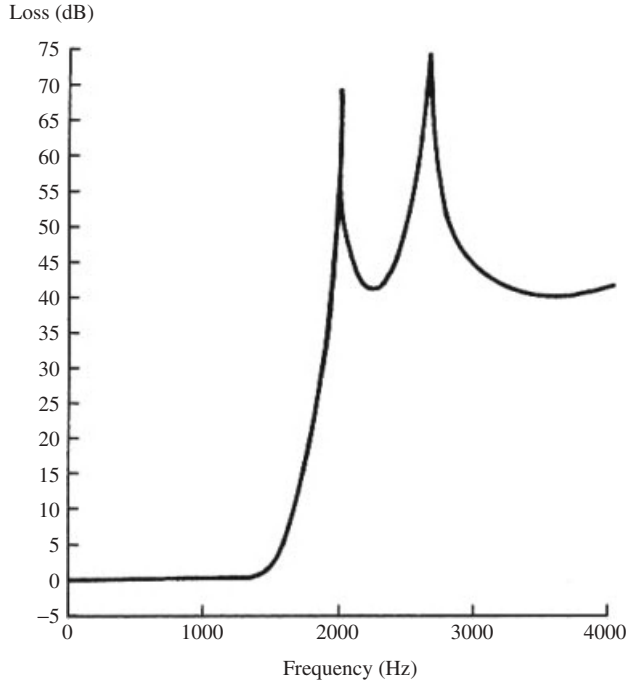


Figure 15.25 Response of the fifth-order filter of the Example

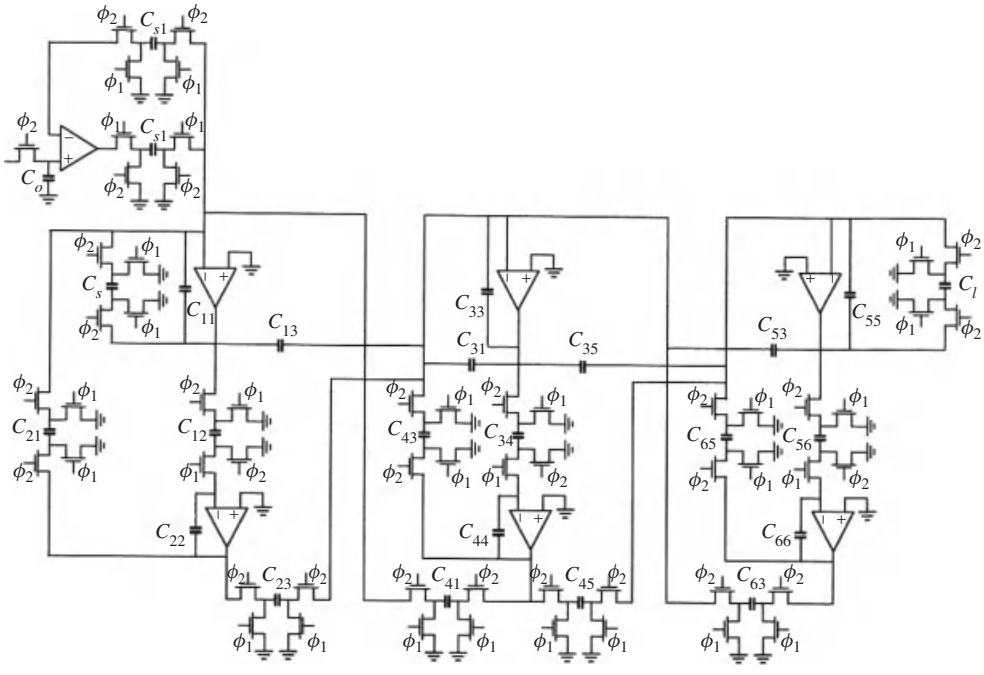


Figure 15.26 Sixth-order band-pass structure

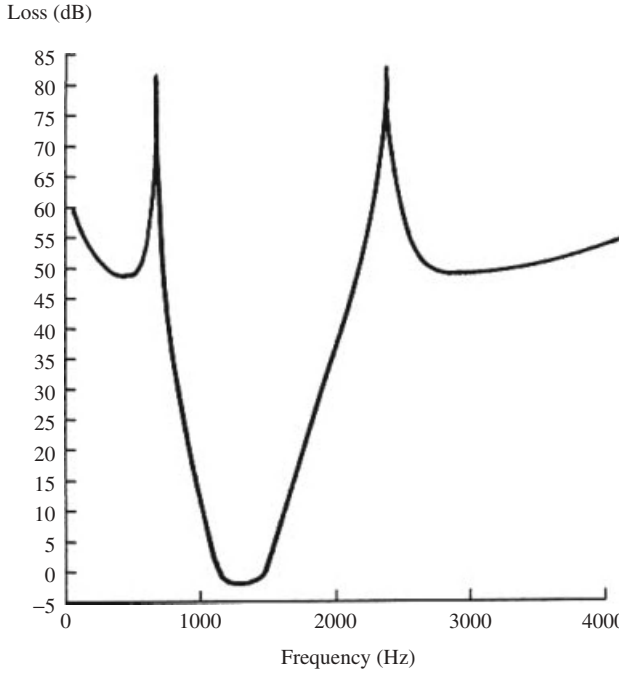


Figure 15.27 Typical sixth-order band-pass response

Block 1	2.78281, 2.05561, 2.78281, 1.00000, 2.78281
Block 2	1.00000, 1.24149,1.000000
Block 3	1.08329, 1.00000, 1.08329
Block 4	1.08329, 1.00000, 1.08329
Block 5	1.00000, 0.90132, 0.74644, 1.00000

and the total capacitance is 37.42644 pF (or units) where the final block is scaled such that $C_\ell = 1$; this can be further scaled at will. Figure 15.25 shows the response of the filter for a 10 kHz clock frequency.

The design of bandpass filters follows similar lines [24] and a typical sixth-order filter structure is shown in Figure 15.26 with a response shown in Figure 15.27.

15.5 Cascade Design

If the sensitivity properties of the filter are not of primary concern, as in the case of relatively low-order filters, the simplest method of realization is to decompose the transfer function into second order factors (and a possible first order one for odd degree cases), realize each factor by a simple circuit, then connect the resulting circuits in cascade. This is analogous to the same technique we have encountered in the design of digital filters and indeed analog continuous filters also. Thus, the transfer function is obtained in the z-domain according to the specifications as explained earlier and MATLAB® can be used in exactly the same procedure as in Chapter 5. Then it is written in the form

$$H(z) = \prod_{k=1}^{m} H_k(z) \tag{15.68}$$

where a typical quadratic factor is of the form

$$H_k(z) = \frac{a_{0k} + a_{1k}z^{-1} + a_{2k}z^{-2}}{b_{ok} + b_{1k}z^{-1} + b_{2k}z^{-2}} \tag{15.69}$$

and its switched capacitor realization is shown in Figure 15.28.

Naturally the same MATLAB® functions used for the factorization of the digital transfer functions can also be used here to determine the factors of the transfer function. The capacitor ratios can be obtained by comparing the above expression with the transfer function of the shown section given by

$$H_k(z) = \frac{I + (G - I - J)z^{-1} + (J - H)z^{-2}}{(1 + F) + (C + E - F - 2)z^{-1} + (1 - E)z^{-2}} \tag{15.70}$$

where it is assumed that $A = B = D = 1$

A first-order factor is of the form

$$H_k(z) = \frac{a_{0k} + a_{1k}z^{-1}}{b_{ok} + b_{1k}z^{-1}} \tag{15.71}$$

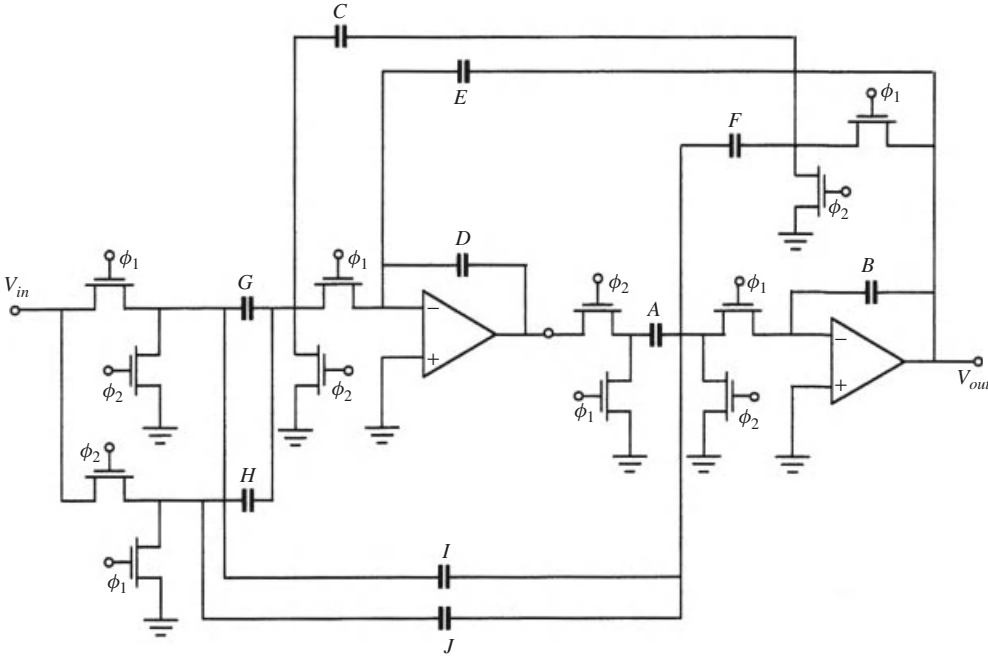


Figure 15.28 Second-order section for the realization of (15.70)

which can be realized as one of the four circuits in Figure 15.29 depending on the location of the pole-zero pair. Thus, for the circuit of Figure 15.29(a)

$$H_k(z) = \frac{(C_1 + C_3) - C_1 z^{-1}}{(1 + C_2) - z^{-1}} \tag{15.72}$$

while for the circuit of Figure 15.29(b)

$$H_k(z) = \frac{(C_1 + C_3) - C_1 z^{-1}}{(C_2 - 1)z^{-1} - 1} \tag{15.73}$$

and the circuit of Figure 15.29(c) has

$$H_k(z) = \frac{C_3 + C_1 z^{-1}}{z^{-1} - (1 + C_2)} \tag{15.74}$$

while for the circuit of Figure 15.29(d) we have

$$H_k(z) = \frac{C_3 + C_1 z^{-1}}{1 + (C_2 - 1)z^{-1}} \tag{15.75}$$

All the capacitor values above are relative to C_A.

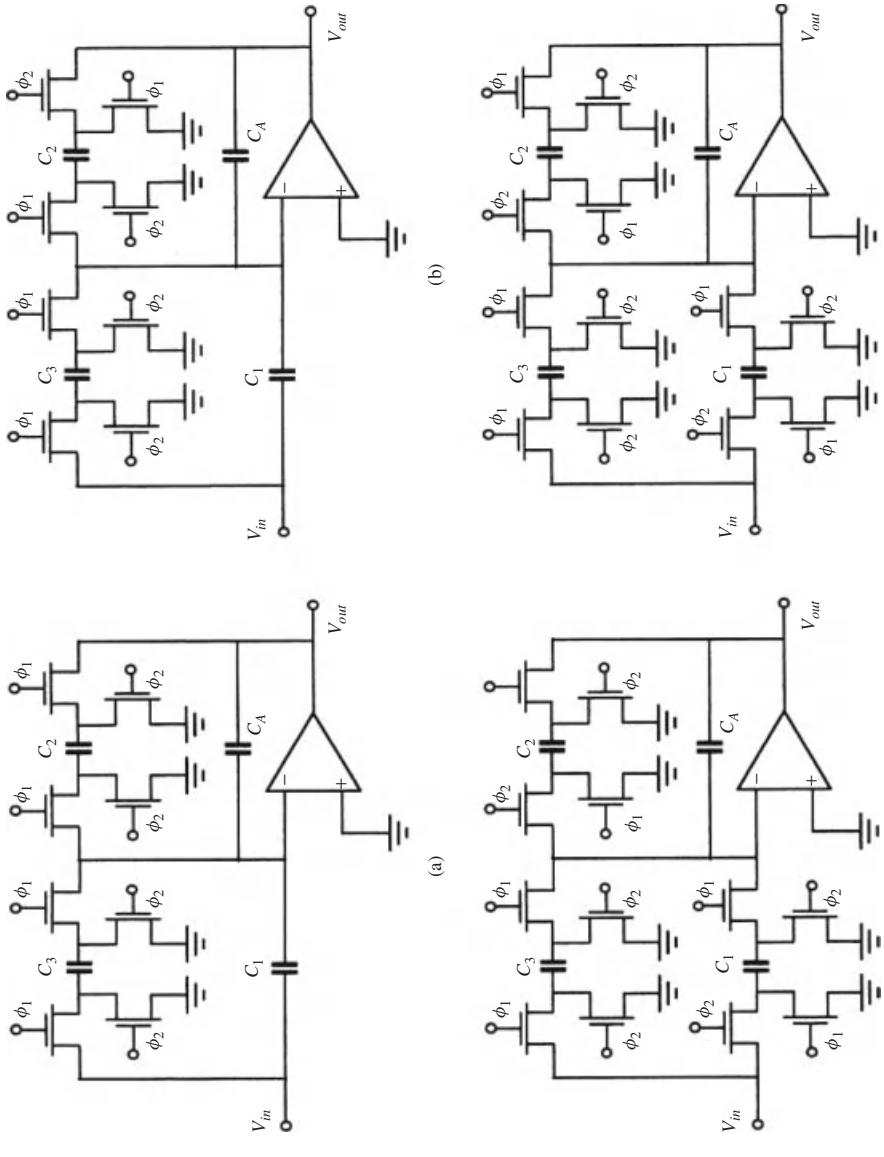


Figure 15.29 Circuits for the realization of (15.72)–(15.75) in the same order

In fact all the digital transfer functions used in Chapter 5 on the design of digital filters, can be implemented using an equivalent switched-capacitor filter using the present cascade design method.

15.6 Applications in Telecommunications: Speech Codecs and Data Modems

15.6.1 CODECs

Pulse code modulation (PCM) is used to transmit speech over telephone channels according to the scheme shown in Figure 15.30. Prior to being sampled at a rate of 8.0 kHz, the speech signal must be bandlimited to 3.4 kHz. It is also desirable to remove the 50 Hz or 60 Hz power line frequency. These functions are performed by the transmit filter which is a switched-capacitor band-pass type.

At the receiver end, a low-pass switched-capacitor filter is used. The idealized reponses of the transmit and receive filters are shown in Figure 15.31 which are designed with the following specifications.

There are two approaches to the design of of the integrated coder decoder (CODEC). The first is to implement the coder and decoder on one chip, with the transmit and receive filters on a second chip. Alternatively, the coder and transmit filter could be on one chip while the decoder and receiver filter are on a second chip. The separation of the rtwo filters in the second approach leads to reduced cross-talk and noise when the CODEC operates asynchronously. The implementation of the CODEC chip is one of the major success stories of switched-capacitor filters.

15.6.2 Data Modems

Switched-capacitor techniques have also been a very powerful tool for the implementation of the two-band low-pass filters needed for data modulators-denodulators (MODEMs).

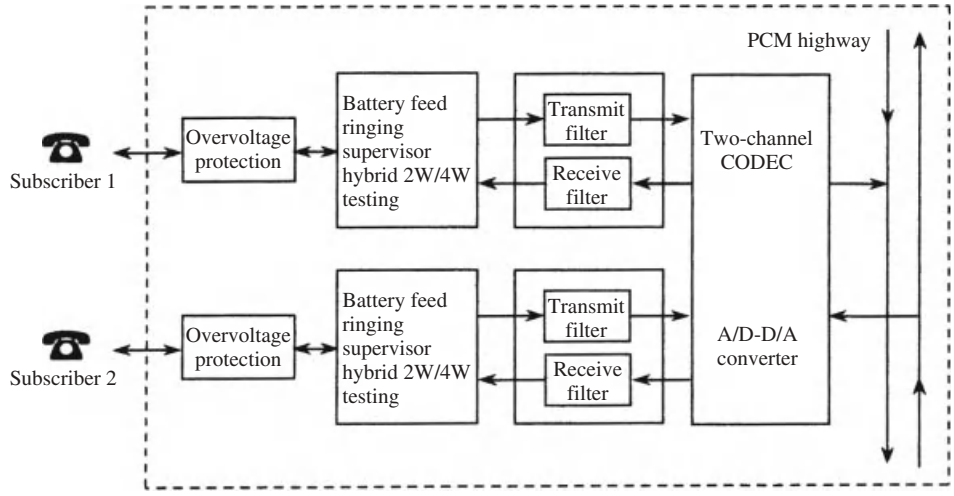


Figure 15.30 PCM speech CODEC

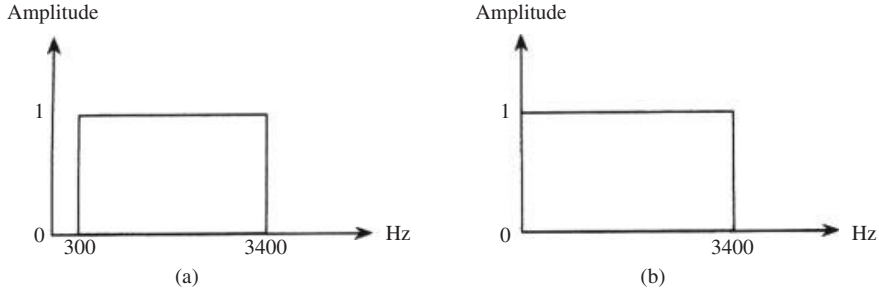


Figure 15.31 Idealized transmit and receive CODEC filter characteristics

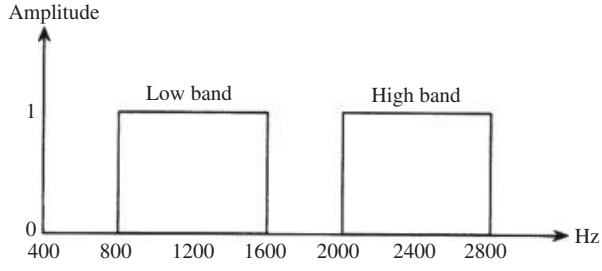


Figure 15.32 Idealized full-duplex modem filter responses

They are used to transmit and receive data simultaneously on telephone channels. These filters have the idealized amplitude responses shown in Figure 15.32. They must also approximate a linear phase characteristic in their passpands and have specific time responses. The set of two filters are made available on a single chip.

15.7 Conclusion

We have presented some basic design techniques for microelectronic switched-capacitor filters. These are sampled-data designs and have been used instead of continuous-time and digital filters in many applications. The chapter ended with typical applications in telecommunications. Since these filters are analog in nature, they are susceptible to a wide range of non-ideal effects which are discussed in the next chapter. The switched-capacitor technique can also be extended to design many other signal processors, a comprehensive example of which will be discussed in Chapter 17.

Problems

15.1 Realize the following transfer function as a switched-capacitor LDI ladder network

$$H(\lambda) = \frac{(1-\lambda^2)^{5/2}}{1 + 26.15\lambda + 294.10\lambda^2 + 2441.87\lambda^3 + 10018.20\lambda^4 + 44783.60\lambda^5}$$

15.2 The following set of specifications are to be met by the receive low-pass filter employed in a CODEC for PCM telephony:

Passband: 0–3.4 kHz, with 0.25 ripple.

Stopband edge at 4.6 kHz, with attenuation $\geq$30 dB.

Design an LDI Chebyshev filter to meet the above set of specifications.

15.3 The filter shown in Figure 15.33 is a third-order normalized elliptic low-pass prototype with cut-off at $\omega = 1$ with 0.25 dB ripple, and the ratio of stopband edge to passband edge = 1.5.
Transform the prototype into a switched-capacitor ladder with cut-off at 3 kHz and using a clock frequency of 20 kHz.

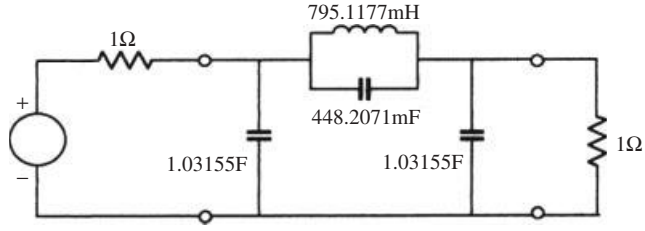


Figure 15.33

15.4 Design a maximally-flat filter in cascade form with the following specifications.

Passband: 0–1 kHz, with attenuation $\leq$0.1 dB.

Stopband edge at 3 kHz, with attenuation $\geq$30 dB.

Sampling frequency: 10 kHz.

15.5 Design an elliptic filter in cascade form to meet the specifications on the receive filter of a speech CODEC given in Problem 15.2.

15.6 Design an elliptic band-pass ladder filter to meet the specifications on the transmit filter of a PCM speech CODEC as given by:

Passband: 300–3.4 kHz with 0.25 dB maximum attenuation.

Lower stopband edge at 50 Hz with 25 dB minimum attenuation.

Upper stopband edge at 4.6 kHz with 32 dB minimum attenuation.

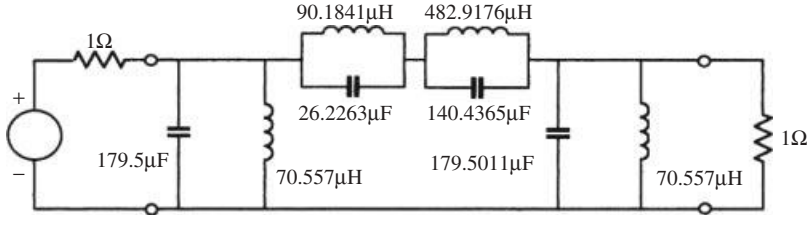


Figure 15.34

15.7 Figure 15.34 shows an elliptic band-pass filter designed to meet the following specifications:

Passband: 1.0–2.0 kHz, with 0.2 dB ripple.

Stopband edges at 0.5 and 3.0 kHz with minimum attenuation of 30 dB.

(a) Find the low-pass prototype from which the filter was obtained.

(b) For the given band-pass filter, use the bilinear transformation to obtain a switched-capacitor filter having the same passband edge frequencies with a clock frequency of 8 kHz. Find the resulting stopband frequencies of the switched-capacitor filter.

15.8 Evaluate the transfer function of the filter in Problem 15.7 and realize it in cascade form.

16

Non-ideal Effects and Practical Considerations in Microelectronic Switched-capacitor Filters

16.1 Introduction

In the previous chapter, the design techniques of switched-capacitor filters were delineated assuming idealized components. Further, Chapters 10–14 dealt with the integrated circuit building blocks together with their non-ideal behaviour. In this chapter, we combine the results to examine the effect of the non-ideal behaviour of the integrated circuit building blocks on the overall response of switched-capacitor filters [24, 27]. We also examine some practical issues, which may be of interest to the designer. Many of these effects are also applicable to analog continuous filters.

16.2 Effect of Finite Op Amp Gain

The building blocks of switched-capacitor filters are first or second order sections containing Op Amps, switches and capacitors. In the design methods presented in the previous chapters, it has been assumed that the Op Amps are ideal, having infinite gain values. However, as we know from Chapters 12–13, real CMOS Op Amps have large but finite gain values. This factor has to be taken into account in the final simulation of the switched-capacitor filter to determine its exact response, particularly for high-frequency designs where the operating frequencies approach the bandwidth of the Op Amp. The finite gains of the Op Amps result in distortion of the transfer functions of the building blocks and hence the overall filter transfer function. Furthermore, the strays-insensitive property of the building blocks used throughout, depends on the establishment of a virtual ground, which in turn is only closely approximated for very high-gain Op Amps. It is necessary, therefore, to model the finite gain effect and incorporate the model in the filter, then perform the analysis on the modified circuit to find out the response. This is illustrated here by the typical first-order transfer function of the building block shown in Figure 16.1.

Signal Processing and Integrated Circuits, First Edition. Hussein Baher.
© 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.

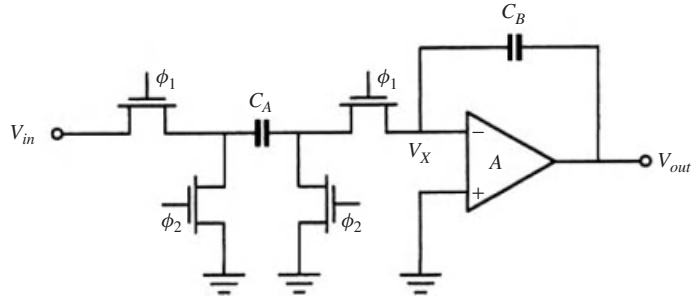


Figure 16.1 A first-order switched-capacitor circuit

We assume that the Op Amp has a finite gain of A and write

$$\begin{aligned} &C_A[-V_x(n) - V_{in}] \\ &\quad = -C_B\{[V_0(n) - V_x(n)] - [V_0(n-1) - V_x(n-1)]\} \end{aligned} \tag{16.1}$$

Substituting for $V_x = -V_{out}/A$, letting $C_B/C_A = \alpha$ and $1/A = \eta$ then taking the z-transform of both sides of the resulting equation, we can form the transfer function of the building block as

$$H(z) = \frac{z^{-1}}{\alpha(1+\eta)(1-z^{-1}) + \eta} \tag{16.2}$$

which approaches (15.32) as the gain approaches infinity, that is as $\eta \to 0$. It is easy to see that H is related to the ideal transfer function (15.32) by

$$H = \frac{H_{ideal}}{\alpha(1+\eta) + (\eta/2) + (\eta/2)\coth(Ts/2)} \tag{16.3}$$

Thus, with E denoting the ratio H/H_{ideal} and neglecting terms containing η^2 we have on the $j\omega$-axis

$$\begin{aligned} |E|^2 &\cong \frac{\alpha}{\alpha + \eta} \\ Arg(E) &\cong \frac{\eta}{2\alpha\tan(\omega T/2)} \end{aligned} \tag{16.4}$$

From the above expressions, the error in the magnitude is negligible for reasonably large values of $A(>1000)$, but is easily taken into account by noting that

$$|H|^2 = |H_{ideal}|^2(1 + \eta/\alpha) \tag{16.5}$$

which is equivalent to an error in the capacitor ratio by the factor η. However, the phase error is frequency dependent and can be quite large.

Obviously, the only 'cure' for the finite gain effects is to design Op Amps with high gain as discussed in Chapter 13. So the purpose of the above discussion is purely informative.

16.3 Effect of Finite Bandwidth and Slew Rate of Op Amps

Each Op Amp has a unity gain frequency which limits the frequency range over which the switched-capacitor filter is used. In particular, the unity gain frequency affects the settling time of the Op Amp as may be clearly seen by representing the Op Amp by its frequency-dependent transfer function, inserting this model in the switched-capacitor section, then finding the corresponding time response. For a first-order section of the type used in the previous analysis, the output will settle to its final value either exponentially or passing through a phase of damped oscillations. Thus the time T_{on} corresponding to the clock phase during which the switch is ON, must be large enough to allow the Op Amp to settle to its final value within a certain error. This implies that the settling time of each Op Amp must be less than half the sampling period, or

$$T_{settling} < 0.5T < 1/2f_N \tag{16.6}$$

Thus, the real solution to this problem is to design Op Amps with fast settling time. Calculations on both the positive and negative first-order sections have revealed that the unity gain frequency ω_t of the OP Amp must satisfy

$$\omega_t >> \omega_N/\pi \tag{16.7}$$

and a value of $\omega_t = 5\omega_N$ is found satisfactory in practice to result in negligible errors. In addition to the settling time, there is a delay caused by the finite slew rate of the Op Amp as shown in Figure 16.2.

Thus, instead of (16.6) we have

$$(T_{settling} + T_{slew}) < 0.5T \tag{16.8}$$

16.4 Effect of Finite Op Amp Output Resistance

The Op Amp must charge a load capacitance C_L through its output resistance R_{out}. The charging time constant can be found to be of the order of $2R_{out}C_L$, which must be less than $0.5T$ to give acceptable charge transfer. Naturally, buffered Op Amps can be used to reduce this effect.

16.5 Scaling for Maximum Dynamic Range

Consider a typical Op Amp in a switched-capacitor filter as shown in Figure 16.3. Let the transfer functions $\Delta Q/V$ of the branches connected to the output of the kth Op Amp be multiplied by a factor β_k. In Figure 16.3 these are branches B_4, B_5, B_6.

This leads to the scaling of the output voltage of the Op Amp by $1/\beta_k$. The input branches are left unchanged and their charges remain the same, which results in the charge in B_4 retaining its original value. The same is true for branches B_5 and B_6 since their capacitances are scaled by β_k while their voltages are scaled by the inverse of this factor. We conclude that multiplying all capacitor values which are connected, or switched, to the output of an Op Amp by a factor, scales its output voltage by the inverse of this factor, while leaving all charges flowing between this Op Amp and the rest of

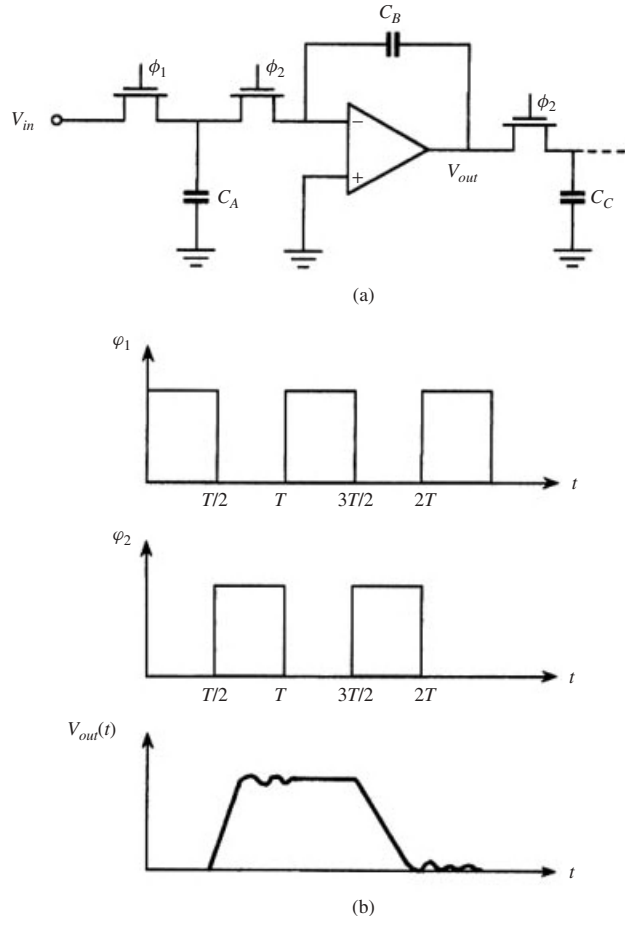


Figure 16.2 Pertinent to the discussion of settling time and slew rate effects

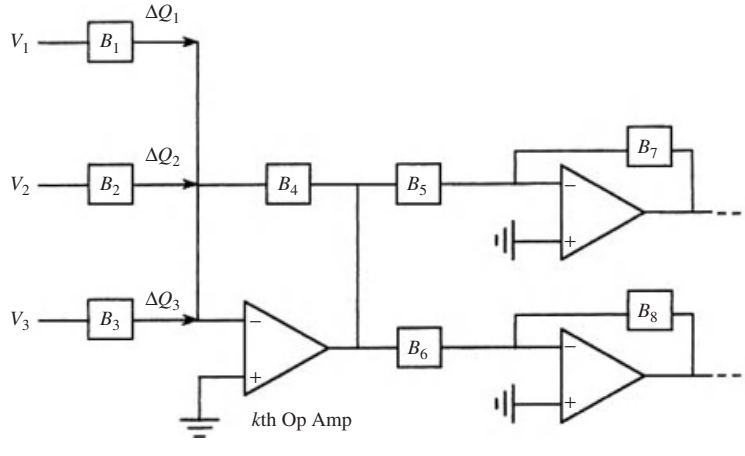


Figure 16.3 Pertinent to the discussion of maximum dynamic range and minimum capacitance scaling

the circuit unchanged. This process allows the improvement of the dynamic range of the filter according to the following procedure:

(a) Set $V_{in}(\omega)$ to the largest value for which the output Op Amp does not saturate.
(b) Calculate the maximum value V_{pk} for all internal Op Amp output voltages. These values usually occur near the passband edge of the filter.
(c) Multiply all capacitors connected or switched to the output terminal of Op Amp k by $\beta_k = V_{pk}/V_{k,\max}$ where $V_{k,\max}$ is the saturation voltage of Amp k.
(d) Repeat for all internal Op Amps.

16.6 Scaling for Minimum Capacitance

The scaling operation can be used to reduce the capacitance spread and hence the total capacitance used by the filter. This relies on the very simple principle that if all capacitors connected to the input terminal of an Op Amp are multiplied by the same number, then the output voltage remains the same. The procedure is as follows:

(a) Divide all capacitors in the filter into non-overlapping sets. Those in the ith set S_i are connected or switched to the input terminal of Op Amp i.
(b) Multiply all capacitors in S_i by $m_i = C_{\min}/C_{i,\min}$ where $C_{\min}$ is the smallest capacitor which the fabrication technology allows and $C_{i,\min}$ is the smallest capacitor in S_i.
(c) Repeat for all sets, including that associated with the output Op Amps.

Finally, we note that scaling for maximum dynamic range should be performed before scaling for minimum capacitance, since the former changes the voltages of the circuit nodes, while the latter does not.

16.7 Fully Differential Balanced Designs

It was shown in Chapter 13 that fully differential balanced MOS Amp designs offer several advantages over single-ended ones. We now justify and extend this concept to the entire switched-capacitor filter structure [28]. The dynamic range of the filter is determined by the ratio of the maximum signal swing giving acceptable distortion, to the noise level. It follows that an improvement in the signal swing results in improvement in the dynamic range. Since fully differential designs double the signal swing, they also improve the dynamic range. Moreover, since the signal paths are balanced, noise due to the power supply variation and clock charge injection are also reduced. This is in addition to the noise reduction inherent in the Op Amp design itself. The principle of conversion from a structure using single-ended output to a fully differential balanced structure is as follows:

(a) Sketch the single-ended circuit and identify the ground node(s).
(b) Duplicate the entire circuit by mirroring it at ground.
(c) Divide the gain of each active device by two.
(d) Reverse the signs of all duplicated active elements and merge each resulting pair into one balanced differential-input differential-output device.
(e) Simplify the circuit, if possible, by replacing any device which merely achieves sign inversion by crossed wires.

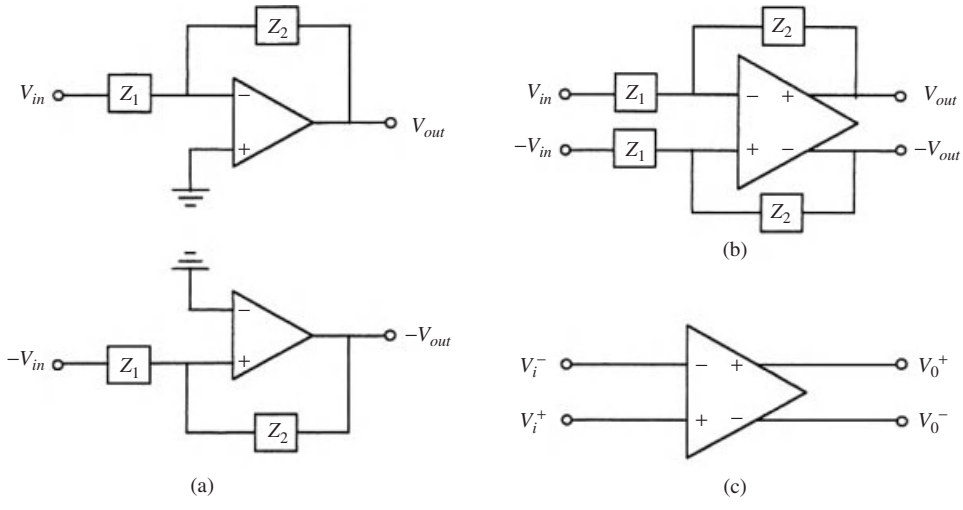


Figure 16.4 (a) A general Op Amp circuit and its mirror image. (b) Conversion into fully differential equivalent. (c) Symbol of the fully differential Op Amp

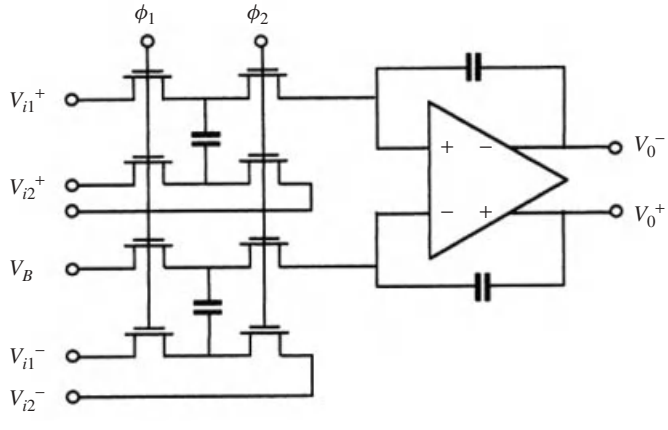


Figure 16.5 A fully differential switched-capacitor circuit

Figure 16.4 shows the application of the procedure to a general Op Amp circuit.

Figure 16.5 shows a fully differential first-order switched-capacitor section. In this circuit, v_{i1}^+ and v_{i1}^- constitute one of the differential input signals while v_{i2}^+ and v_{i2}^- form the other. With the assumption of zero common-mode signal, the common mode voltage of the Op Amp is V_B which can be set to a suitable value by an on-chip circuit.

An example of a fifth-order filter structure [28] using both chopper stabilization and fully-differential techniques is shown in Figure 16.6.

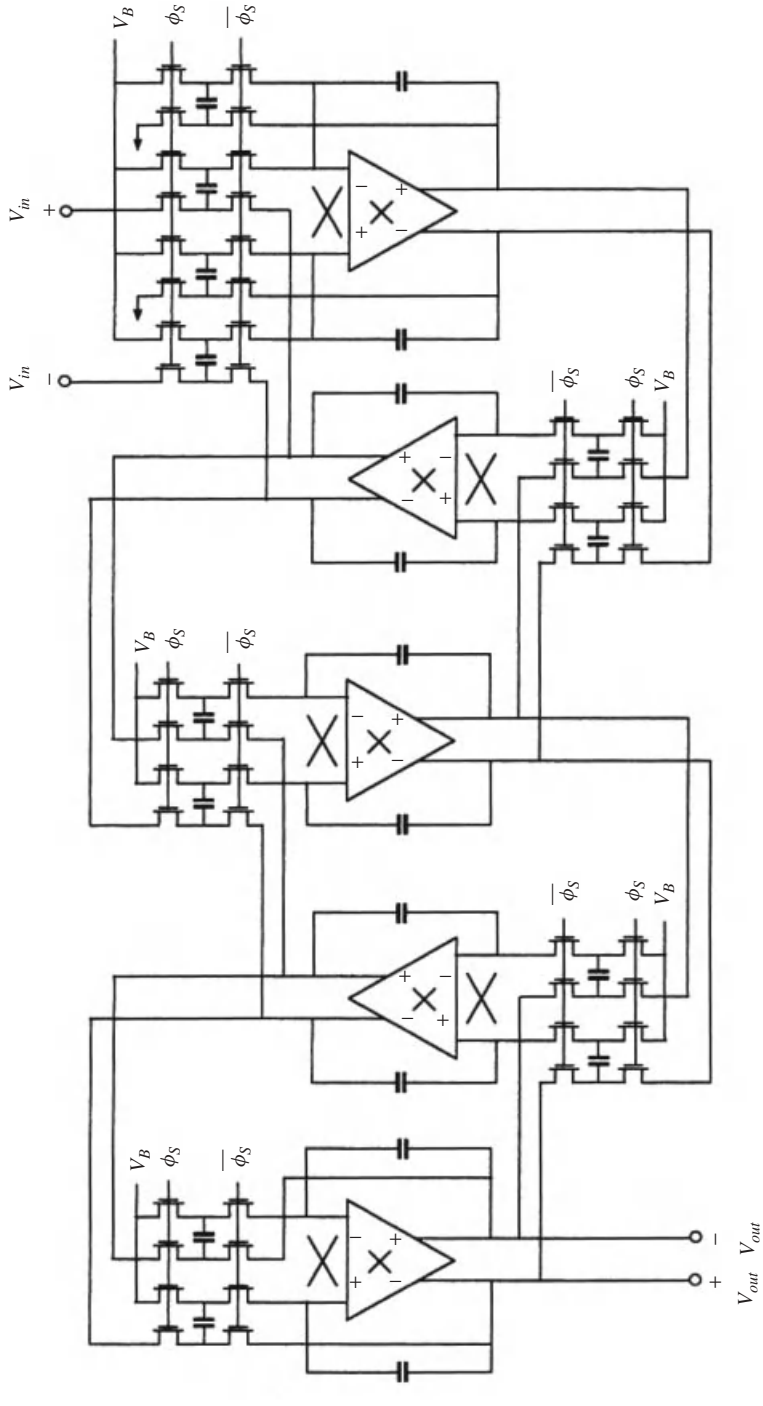


Figure 16.6 A fifth-order filter using chopper stabilization and fully differential topology

16.8 More on Parasitic Capacitances and Switch Noise

It has been emphasized that the use of parasitic-insensitive structures is essential for good performance of switched-capacitor filters. The adoption of switched-capacitor filter circuits which are parasitic insensitive has been a major factor in the development of high quality integrated filters. These circuits have the desirable property that their transfer functions are unaffected by additional capacitances from node to ground, assuming ideal Op Amps. However, this immunity does not extend to the effect of ungrounded parasitic capacitances C_{xc} and C_{yc} associated with the control terminals of the switches as shown in Figure 16.7. These parasitic capacitances can be shown to cause dc offset voltages, modification of the transfer function and distortion.

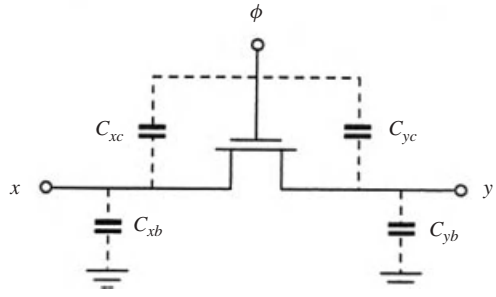


Figure 16.7 A switch with the associated parasitic capacitances

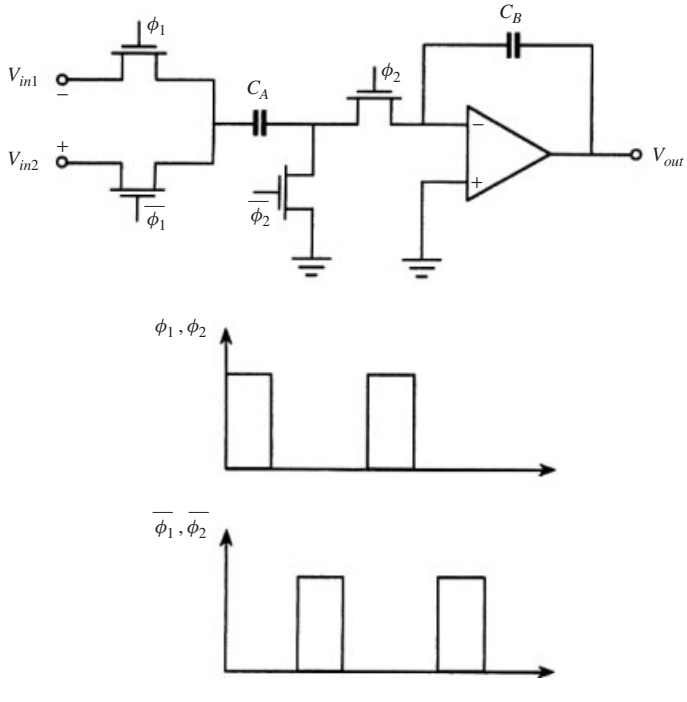


Figure 16.8 Typical first-order section with conventional two-phase clocking scheme

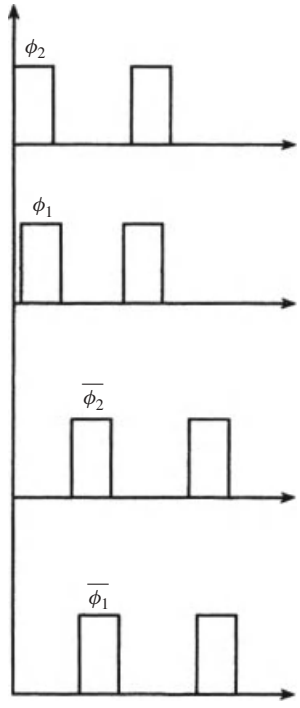


Figure 16.9 A more elaborate clocking scheme for the circuit in Figure 16.8

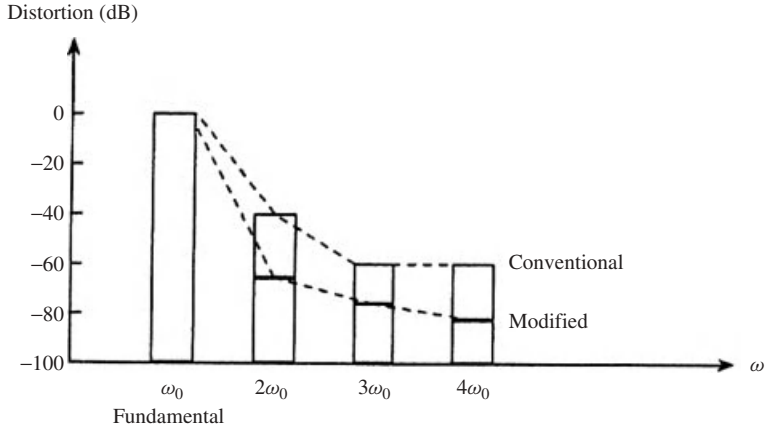


Figure 16.10 Comparison of the distortion values of the two clocking schemes

Figure 16.8 shows a typical first-order section together with the conventional two-phase clocking scheme. A four-phase scheme [29] is possible as shown in Figure 16.9 with the objective of reducing the remaining parasitic effects. This is achieved by slightly delaying the switching waveforms of two of the switches. Figure 16.10 shows the improvement in performance due to the different waveform arrangement.

16.9 Pre-filtering and Post-filtering Requirements

As pointed out earlier, a switched-capacitor filter working in continuous-time environment, should be preceded by an analog continuous-time filter which ensures that aliasing does not occur. This is called an *antialiasing filter* (AAF) and is realized on the same chip using an active RC circuit. A possible realization is the Sallen–Key second-order section shown in Figure 16.11. It is designed to have a typical response as shown in Figure 16.12. in relation to the response of the switched-capacitor filter.

Also due to the (sin x/x) effect present at the output of the filter, an amplitude equalizer circuit with response approximating the inverse function may be needed if the passband edge of the filter is not small compared with the sampling frequency. If the sampling frequency is greater than ten times the passband edge, then a simpler continuous-time filter similar to the pre-filter can be used. If the sampling frequency is not very large compared with the pass-band edge of the SC filter, then the required degree of the AAF could be large resulting in an increased area on the chip. To ease the requirement on the selectivity of the AAF and hence reduce the degree, a decimator [30] can be used as shown in Figure 16.13(a). This introduces zeros at multiples of the sampling frequency as shown in Figure 16.13(b).

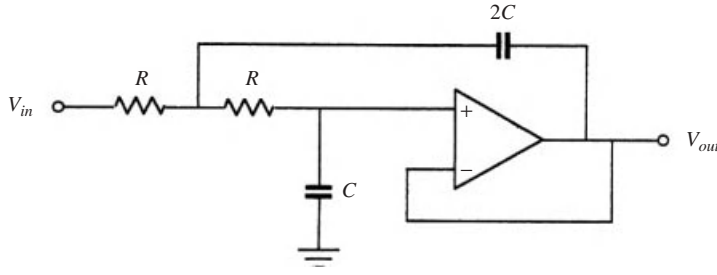
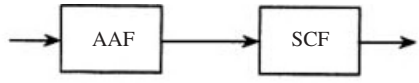
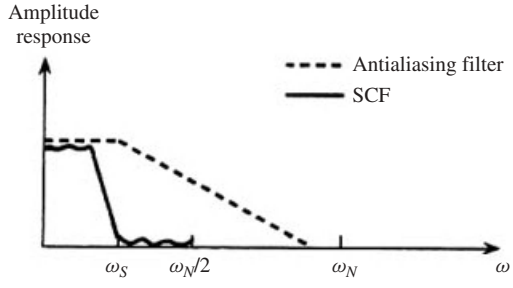


Figure 16.11 Sallen–Key second-order section



(a)



(b)

Figure 16.12 Response of the antialiasing filter relative to that of the switched-capacitor filter

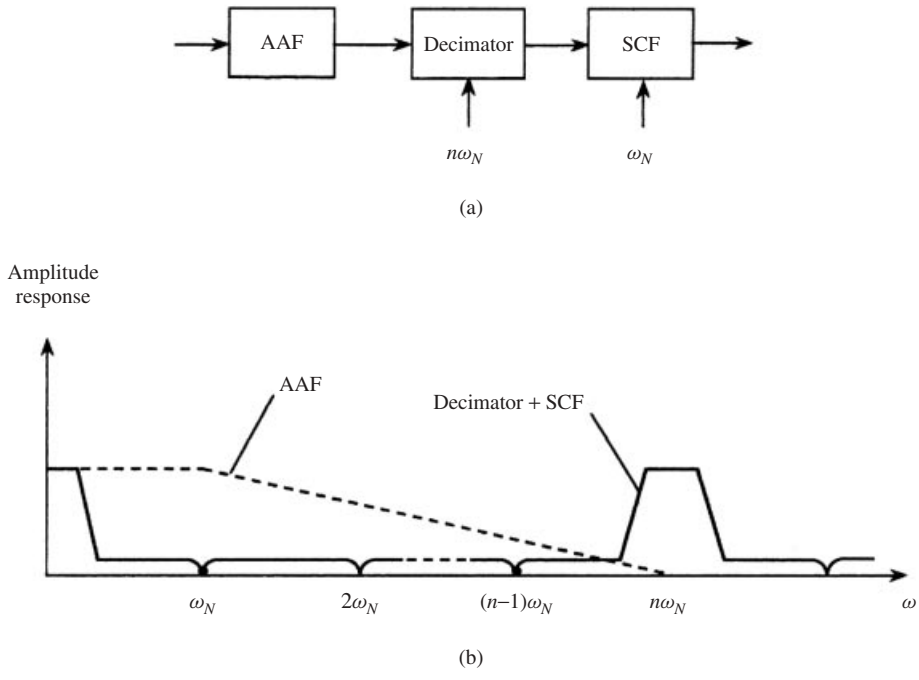


Figure 16.13 The use of a decimator and AAF for pre-filtering: (a) scheme, (b) the amplitude response of the AAF relative to that of the decimator

The required degree of the AAF in this case is obviously lower than that which would be required for the direct implementation of Figure 16.12. Thus in Figure 16.13, aliasing is eliminated in two stages. A circuit which implements the scheme of Figure 16.13 is shown in Figure 16.14 together with the clock forms and amplitude response.

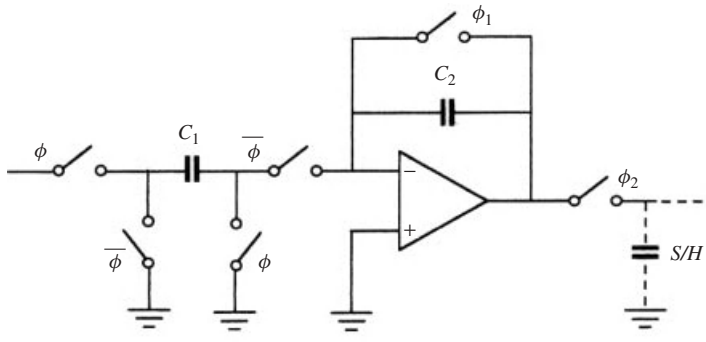
The latter is given by

$$|H(\omega)| = \frac{C_1}{C_2}\left|\frac{\sin(\pi\omega/\omega_N)}{\sin(\pi\omega/n\omega_N}\right| \tag{16.9}$$

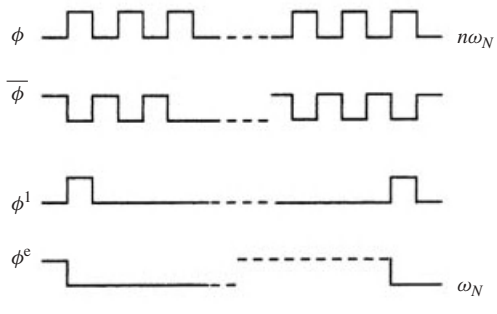
These techniques are only successful at relatively low frequencies and, for a complete non-contrived solution to the antialiasing and smoothing filters problem, good continuous-time filtering techniques seem to be the only real answer. These were discussed in Chapter 3.

16.10 Programmable Filters

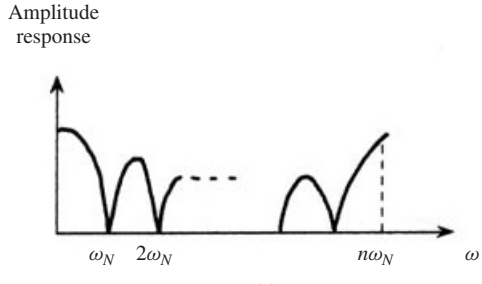
A switched-capacitor filter can be made programmable by simply varying the clock frequency, as in the case of digital filters. As we have seen, the response of the filter contains the key points such as the passband edge and stopband edge. These are basically determined as ratios of the actual frequencies to the clock frequency. Thus, if the clock frequency is multiplied by a factor, the filter key frequencies and hence the entire frequency axis will be multiplied by the same factor. The clock can be programmed digitally.



(a)



(b)



(c)

Figure 16.14 (a) A possible decimator, (b) clock forms, (c) amplitude response

Another method is to replace the capacitors by *capacitor arrays* which are switched into place in various combinations; thus controlling the response of the filter. Such arrays can also be programmed digitally.

A third approach is *mask programmability*. In this case, the components, such as Op Amps and switches are provided on the chip without interconnection, and a separate portion is dedicated to the capacitors. Other ancillary components such as resistors and clock generators are also provided. At the final mask stages, the appropriate connections are made together with the selection of the capacitors.

16.11 Layout Considerations

Analog integrated circuits are particularly sensitive to the geometrical and physical arrangements of their components. This layout of the integrated circuit affects the performance of the switched-capacitor filter by influencing a number of design parameters. Among these are the noise injection from power lines, clock lines, ground lines and the substrate, clock feed-through noise, accuracy of matching components and the high-frequency response. This is particularly the case when both digital and analog circuits exist on the same chip, which is a very common situation in communication systems.

We have seen how to obtain designs which minimize noise and clock feed-through, so that care in the layout should be exercized in order not to lose these attributes. So the following rather common-sense precautions should be observed in addition to the more fundamental design techniques discussed before, such as the use of fully-differential balanced topology, chopper stabilization and delayed clocking schemes for reduction of clock feed-through.

First, power lines, clock lines and ground lines should be kept free from noise and noise coupling between lines must be minimized. Secondly, separate power lines for digital and analog signals should be used if possible. Also separate bonding pads for these circuits can be used, together with separate pins which can be connected externally together. Finally, external decoupling capacitors can be used between the pins, thus reducing the impedance coupling to the external supply. With these precautions, the bias voltage lines for the substrates and wells can be connected to the supply pads without introducing digital noise into the substrate or wells. If a certain signal or clock line is particularly troublesome, it can be shielded as shown in Figure 16.15. The shielding is formed from two ground metal lines and a grounded polysilicon layer. Obviously, this arrangement can also be used to isolate the analog and digital signals and to prevent noise coupling to and from the substrate.

Noise coupling into the substrate can be reduced by using a clean power supply for biasing, by shielding the substrate from all capacitors, by placing grounded wells below them and by establishing a good bond using gold if possible between the back surface of the substrate and the package header.

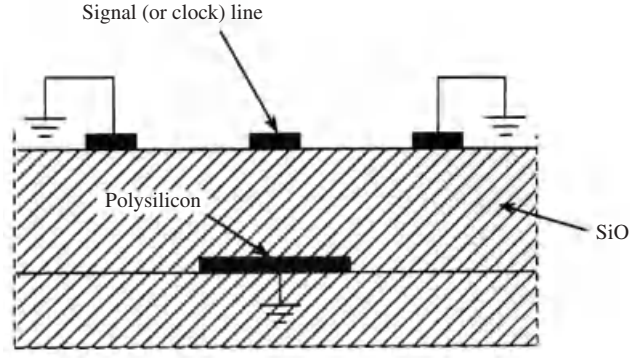


Figure 16.15 A possible shielding scheme for signal or clock lines

For the reduction of noise from the substrate, the fully differential balanced topology discussed before is used together with shielding.

The most significant capacitances coupling the substrate to the circuit are those between the substrate and the bottom plates. Therefore, the bottom plate should not be connected or switched to the inverting input of an Op Amp since this terminal has a high noise gain to the output. In addition, the lines connecting the input nodes to any capacitors should be as short as possible and made of polysilicon or metal; diffusion type lines should be avoided since they are, capacitively, too strongly coupled to the substrate.

The crossing of input-node lines with other signal lines should be avoided. The lines should be shielded and guard rings used to shield the input devices of an Op Amp. Also the number of components connected to an input node should be kept to a minimum. Only one switch should be used at the input terminal and implemented with minimum area transistors.

16.12 Conclusion

In this chapter, we considered many practical issues, characteristic of integrated circuit realization of switched-capacitor filters. These include high frequency designs and fully differential circuits as well as the effects of the non-ideal behaviour of Op Amps and switches on the overall performance of the designed filter. Scaling for maximum dynamic range and minimum capacitance, prefiltering, postfiltering and layout considerations were also discussed.

17

Integrated Sigma-Delta Data Converters: Extension and Comprehensive Application of Analog and Digital Signal Processing

17.1 Motivation and General Considerations

It was shown in Chapter 4 that the process of conventional analog to digital conversion requires a number of high-precision operations involving band-limiting filters, samplers, quantizers and encoders. We now ask the question: is it possible to employ our knowledge of both analog and digital processors to develop an A/D converter that does not require high precision components and can be easily integrated? In particular, we know from Chapter 15 that switched-capacitor circuits have many advantages when it comes to integrated circuit implementation in MOS technology and together with digital circuits on the same chip. Can we also use switched-capacitor techniques to give an answer to this question?

It turns out that the development of switched-capacitor circuits have allowed the introduction of an ingenious method of A/D conversion that has a number of advantages over the conventional method [12, 31–35].

The general structure of the converter is shown in Figure 17.1. The high resolution is obtained by *oversampling* the analog input signal, that is, it is sampled at a much higher rate than the critical Nyquist rate together with a coarse quantizer that is usually a simple two-level device or a comparator, which employs feedback loops to generate a one-bit data stream.

The system shown in Figure 17.1 has the following main parts:

(i) The noise shaper or modulator: This is the only analog part of the circuit and has a transfer function, which *pushes the quantization noise to the higher-frequency band*, well outside the signal baseband, thus allowing the removal of the noise by low-pass

Signal Processing and Integrated Circuits, First Edition. Hussein Baher.
© 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.

filtering. This analog part is the critical stage in the design because it determines the maximum signal to noise ratio that can be achieved by the converter.

(ii) The decimator and low-pass filter: These remove the out-of-band noise using digital low-pass filtering and resample the signal at the Nyquist rate.

The illustrated scheme has the following advantages over the conventional multi-bit approach:

(a) The converter circuits are particularly amenable to implementation using CMOS VLSI techniques because it neither requires any high-precision analog components nor requires any trimming.
(b) The analog circuit part of the converter is small and can be implemented on the same chip with the digital part and interface.
(c) No band-limiting (anti-aliasing) filter is needed because the oversampling nature of the converter makes it easy to remove the out-of-band frequencies in the digital processing stage following the analog part.
(d) The sigma-delta (Σ-Δ) converter makes full use of the rapid advances in VLSI technology regarding both high speed and reduced size.

Clearly, Σ-Δ converters represent a striking example of modern trends in which both analog and digital signal processors exist on the same integrated circuit chip.

We now explain the principles of Σ-Δ conversion using the simplest first-order converter and then extend the concepts to higher-order converters for the purpose of improving the performance.

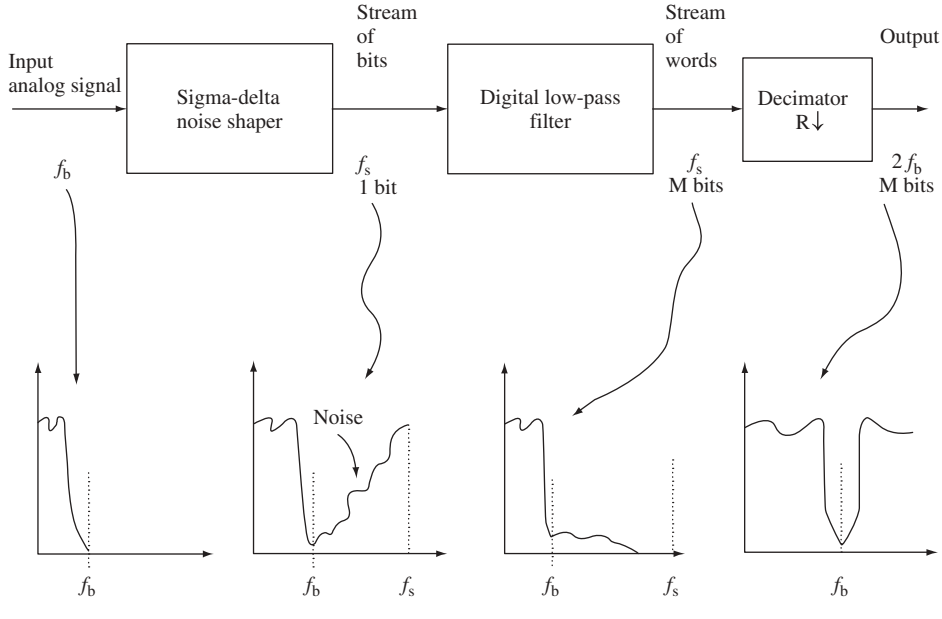


Figure 17.1 General structure of a Σ-Δ converter

17.2 The First-order Converter

The concept of Σ-Δ conversion originated as a modification of delta modulation. The latter was introduced to overcome the problem of coding dc inputs and the accumulation of errors. A simple delta modulator is shown in Figure 17.2(a). The circuit modulates the differential change in the input signal and the receiver is just an integrator. To obtain a Σ-Δ modulator from this circuit, the integrator is placed before the quantizer instead of being in the feedback loop as shown in Figure 17.2(b), which is equivalent to delta modulating an integrated version of the input. Clearly, the integrator in the receiver now becomes redundant and only a low-pass filter is required.

The block diagram of Figure 17.3 illustrates the simplest form of a Σ-Δ converter and is called a *first-order noise shaper*. A mathematical or functional model of this noise shaper is shown in Figure 17.4 in which the z-domain representation is employed. This model is used because switched-capacitor circuits are usually used for the implementation and because the circuit is of the *analog sampled data or discrete* type; both the analysis and design are performed in the z-domain. An example of the switched-capacitor circuit used to implement the first-order converter is shown in Figure 17.5.

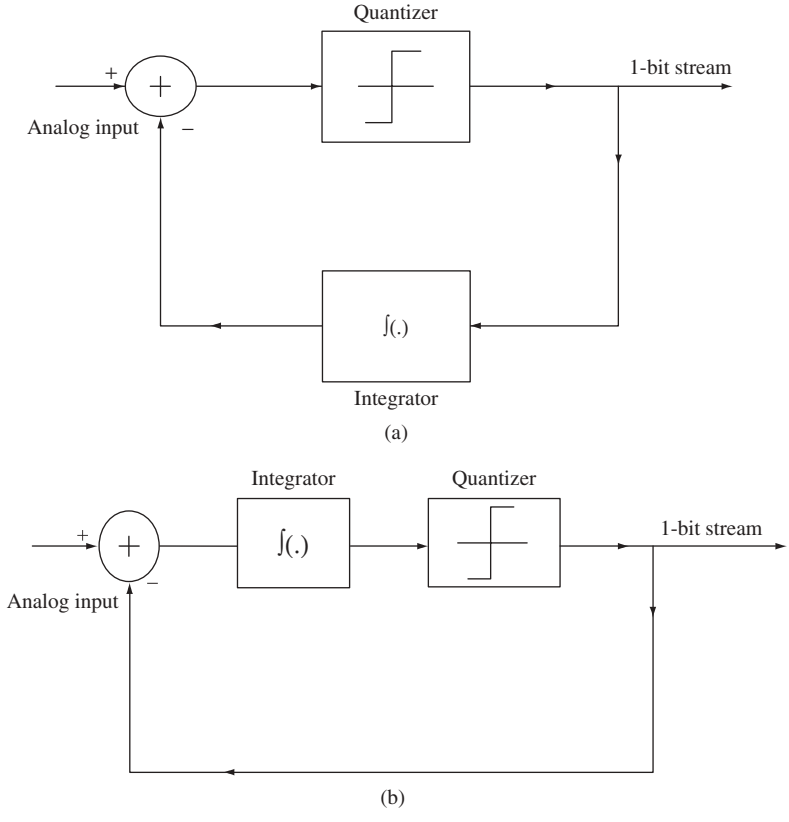


Figure 17.2 (a) Delta (Δ) modulator. (b) First-order Σ-Δ modulator obtained from the Δ modulator

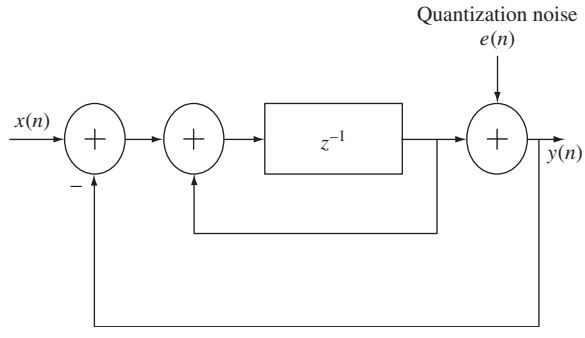


Figure 17.3 First-order Σ-Δ modulator

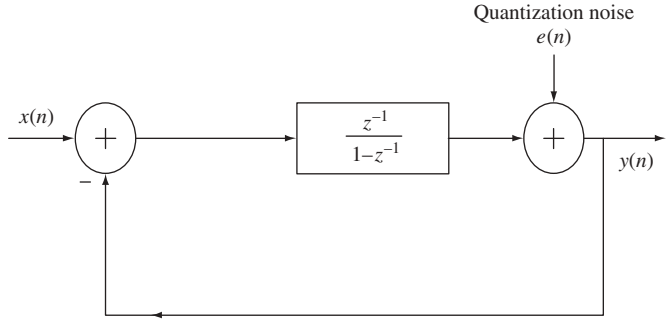


Figure 17.4 Functional model of a first-order Σ-Δ modulator

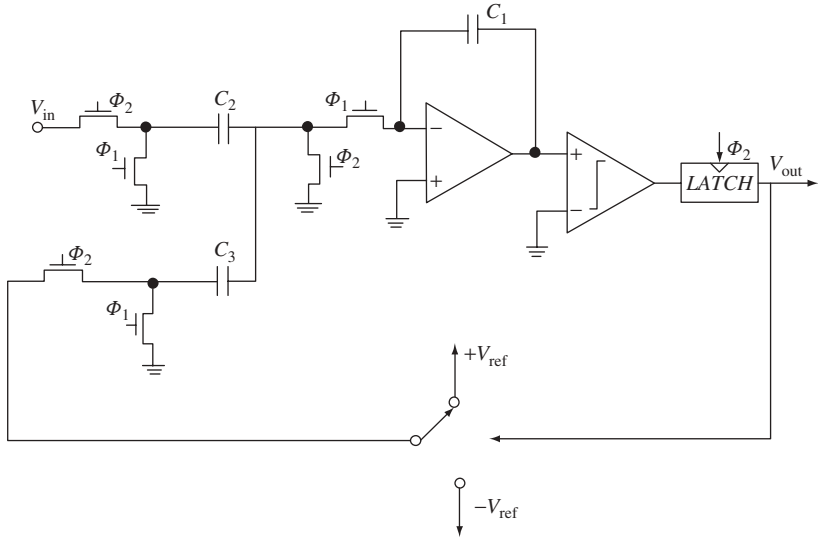


Figure 17.5 Switched capacitor circuit first-order Σ-Δ modulator

The switches are driven by a bi-phase clock with phases Φ_1 and Φ_2, which means that when the group of switches marked Φ_1 are off, the other group marked Φ_2 are on and vice versa. This so-called integrator has a transfer function

$$T(z) = \frac{z^{-1}}{1 - z^{-1}}, \tag{17.1}$$

which is an ideal accumulator with unity gain. The quantizer is modelled by an additive noise source having a peak value of $(\Delta/2)$, where Δ is the quantizer step that is twice the comparator output level. The noise source is assumed *white*, an assumption that, though not true for all input signals, simplifies the analysis and linearizes the converter. Moreover, the linear model of the quantizer becomes more accurate for higher-order converters and leads to good approximations. For this model we can write

$$Y(z) = [X(z) - Y(z)]\frac{z^{-1}}{1 - z^{-1}} + E(z) \tag{17.2}$$

so that the input–output relationship becomes

$$Y(z) = z^{-1}X(z) + [1 - z^{-1}]E(z) \tag{17.3}$$

from which it is clear that the output consists of two components: (i) the input signal $X(z)$ delayed by one clock cycle and (ii) the quantizer noise $E(z)$ *shaped* by the function $(1 - z^{-1})$. Substituting $\exp(-j\omega T)$ for z^{-1} we obtain

$$1 - \exp(-j\omega T) = \exp(-j\omega T/2)[\exp(j\omega T/2) - \exp(-j\omega T/2)] \tag{17.4}$$

so that

$$|1 - \exp(-j\omega T)| = 2\ \sin(\pi f/f_s), \quad T = 1/f_s, \tag{17.5}$$

where f_s is the sampling frequency. Equation (17.5) is the *noise-shaping function* for the converter and its frequency response is shown in Figure 17.6. Clearly, this noise-shaping function attenuates the noise in the baseband of the signal and amplifies it in the higher band. However, for the converter, the total power of the noise remains constant because the area under the noise-shaping function is unity.

Now, assuming that the quantization noise is uniform and limited to the range from $(-\Delta/2)$ to $(\Delta/2)$ as shown in Figure 17.7, the r.m.s value of the noise is given by

$$\sqrt{\int_{-\Delta/2}^{\Delta/2} P(E)E^2\,\mathrm{d}E} = \frac{\Delta}{\sqrt{12}}. \tag{17.6}$$

The spectral density of the noise can be written as the product of the quantizer error spectrum (assumed uniform) and the square of the noise-shaping transfer function, thus

$$|E(z)(1 - z^{-1})|^2 = \frac{4\Delta^2}{12f_s}\sin^2(\pi f/f_s) \tag{17.7}$$

The r.m.s noise is obtained by integrating the noise spectral density *in the baseband* assuming that the digital low-pass filter will remove all the higher-frequency components.

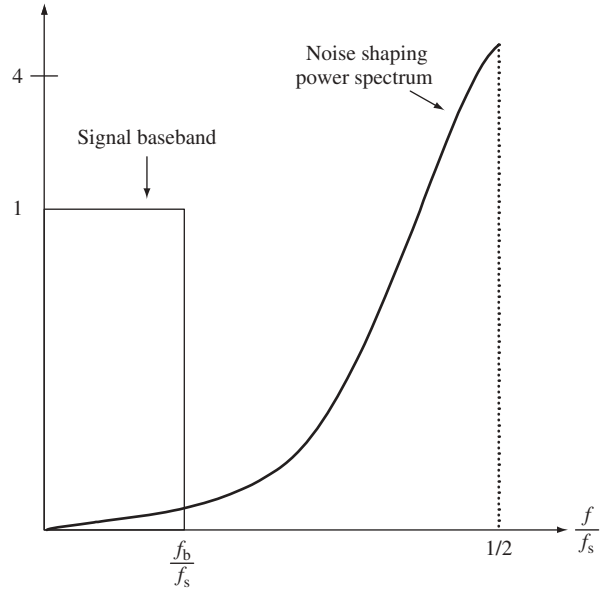


Figure 17.6 Noise-shaping function of the first-order converter

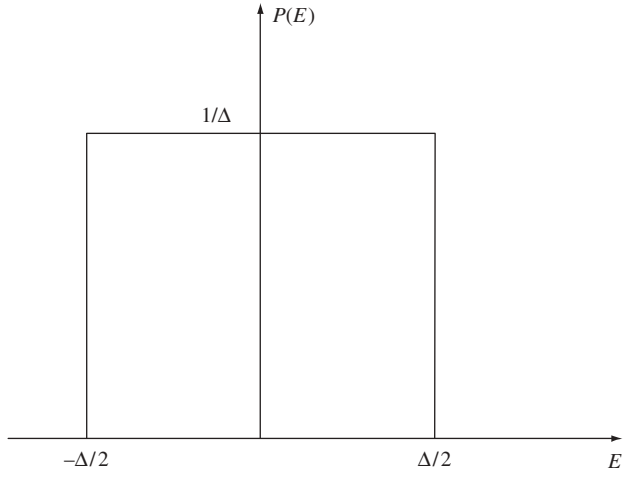


Figure 17.7 Probability density of the quantization error

Usually, the baseband f_b is much lower than the sampling frequency, and the function $\sin(\pi f_b/f_s)$ can be approximated by its argument $(\pi f_b/f_s)$. Thus

$$\text{r.m.s noise} = \left[2\int_0^{f_b} \frac{4\Delta^2}{12f_s}(\pi f/f_s)\right]^{1/2} \tag{17.8}$$

which gives

$$\text{r.m.s. noise} = \frac{\Delta\pi}{6}[2f_b/f_s]^{3/2} = \frac{\Delta\pi}{6}\left[\frac{1}{R}\right]^{3/2} \tag{17.9}$$

where $R = f_s/2f_b$ is the *oversampling ratio*. The above expression reveals that the quantization noise decreases by 9 dB with every doubling of the sampling frequency (or oversampling ratio). However, it has been found that the quantization noise is highly correlated with the input signal and the spectrum of the noise is not white, but coloured. For this simple design, a fairly high sampling ratio is required to achieve good noise reduction. More elaborate noise shapers have been developed to further reduce the noise with the same oversampling ratio. Higher-order designs have been found to achieve higher resolution for higher-signal bandwidth. The second-order converter is introduced next as an illustration of the benefits to be gained by increasing the complexity of the circuit.

17.3 The Second-order Converter

Having introduced the principles of operation of Σ-Δ converters in relation to the simple first-order modulator, we now consider the obvious improvement that should result from increasing the order of the converter. The second-order design studied has excellent properties and is widely used in digital signal acquisition systems. The structure is shown in Figure 17.8 in which two feedback loops are used to reduce the noise in the baseband.

The converter is described by the relation

$$Y(z) = \left[\{X(z) - Y(z)\}\frac{1}{1-z^{-1}} - Y(z)\right]\frac{z^{-1}}{1-z^{-1}} + E(z) \tag{17.10}$$

so that the input–output relationship becomes

$$Y(z) = z^{-1}X(z) + (1-z^{-1})^2E(z) \tag{17.11}$$

where $(1-z^{-1})^2$ is the noise-shaping function. Hence, on the real-frequency axis

$$|(1-z^{-1})|^2 = 4\ \sin^2\left[\frac{\pi f}{f_s}\right] \tag{17.12}$$

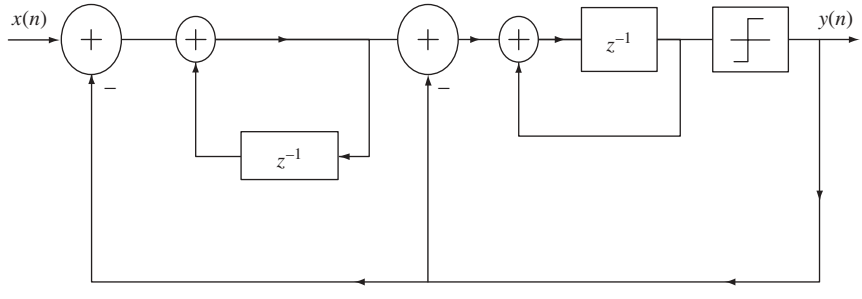


Figure 17.8 A second-order Σ-Δ converter

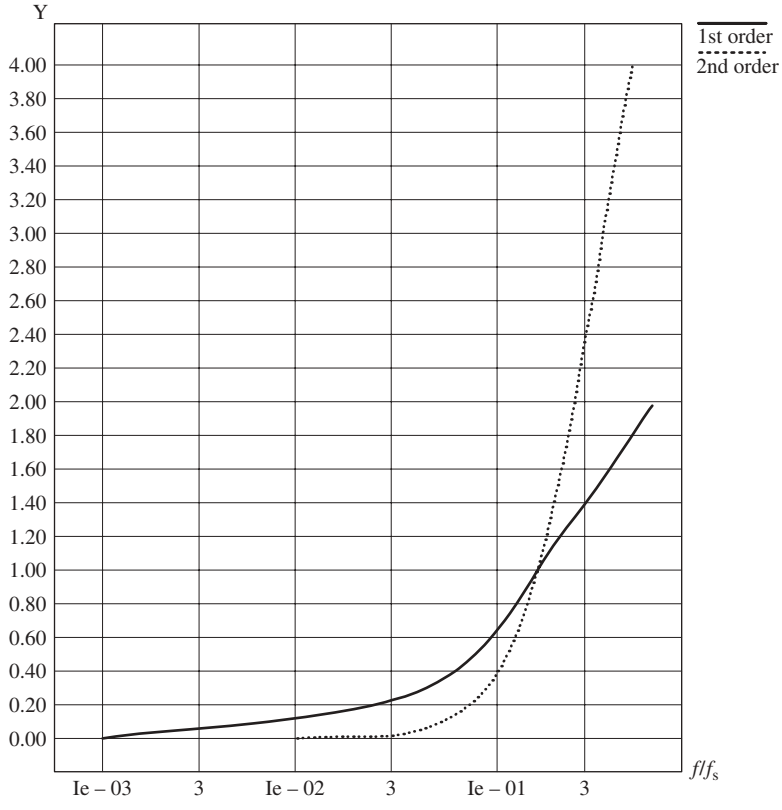


Figure 17.9 Noise-shaping functions of the first- and second-order converters

and the noise spectral density is given by

$$|E(z)(1-z^{-1})^2|^2 = \frac{8\Delta^2}{12f_s}\sin^4\left[\frac{\pi f}{f_s}\right]. \tag{17.13}$$

The r.m.s. noise can be obtained by integrating the noise in the baseband. Thus

$$\text{r.m.s. noise} = \left[2\int_0^{f_b} |E(z)(1-z^{-1})^2|^2\,\mathrm{d}f\right]^{1/2} \tag{17.14}$$

which gives upon using the same approximation as in the first-order converter

$$\text{r.m.s. noise} \cong \frac{\Delta\pi^2}{\sqrt{60}}\left\{\frac{2f_b}{f_s}\right\}^{5/2} \tag{17.15}$$

or

$$\text{r.m.s noise in dB} = 20\log\left\{\frac{\Delta\pi^2}{\sqrt{60}}\right\} + 50\log\left\{\frac{2f_b}{f_s}\right\}, \tag{17.16}$$

and in terms of the oversampling ratio R, the preceding expression becomes

$$\text{r.m.s noise in dB} = 20\log\left\{\frac{\Delta\pi^2}{\sqrt{60}}\right\} + 50\log(1/R). \tag{17.17}$$

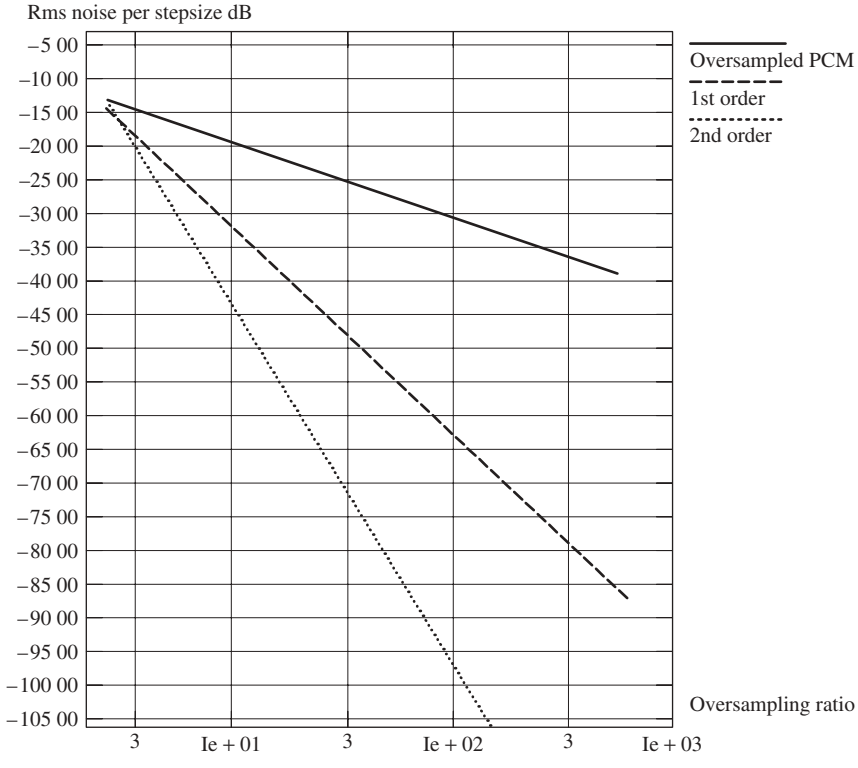


Figure 17.10 Achievable resolutions for the first- and second-order converters

The above expression reveals that the noise decreases by 15 dB for every doubling of the oversampling ratio as compared with 9 dB for the first-order converter studied in the previous section. Figure 17.9 shows a comparison between the two converters in terms of the noise-shaping functions. Figure 17.10 shows the resolutions obtained in both cases.

A better structure for the second-order converter is shown in Figure 17.11. It has the advantage that the delay in the first integrator makes it easier to implement the circuit using switched-capacitor techniques as shown in Figure 17.12. Furthermore, for the original structure, the output signal of each integrator is several times the full-scale output $(-\Delta/2, \Delta/2)$, which is a problem for CMOS operational amplifiers with restricted dynamic range. The modified circuit has smaller ranges, usually only the full-scale input.

Other higher-order structures are also in widespread use, notably the cascades of first-order converters commonly referred to as *MASH converters* by the compact disc industry.

Finally, we note that the performance of a Σ-Δ converter is limited by the non-ideal effects inherent in the electronic components. The analysis of any converter is usually performed using the FFT algorithms discussed earlier in this book. Thus, we see that this type of processor combines the analysis and design techniques of both analog and digital circuits. Hence it is a good example of the entire discipline of signal processing and, hopefully, demonstrates the relevance of the approach in this book combining both analog and digital techniques in one volume.

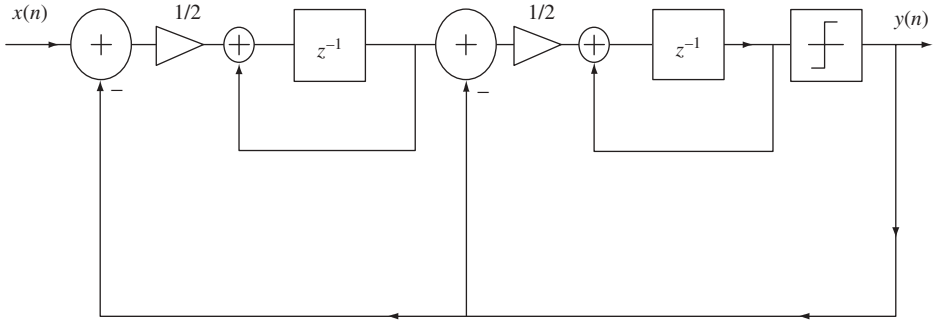


Figure 17.11 A modified second-order structure

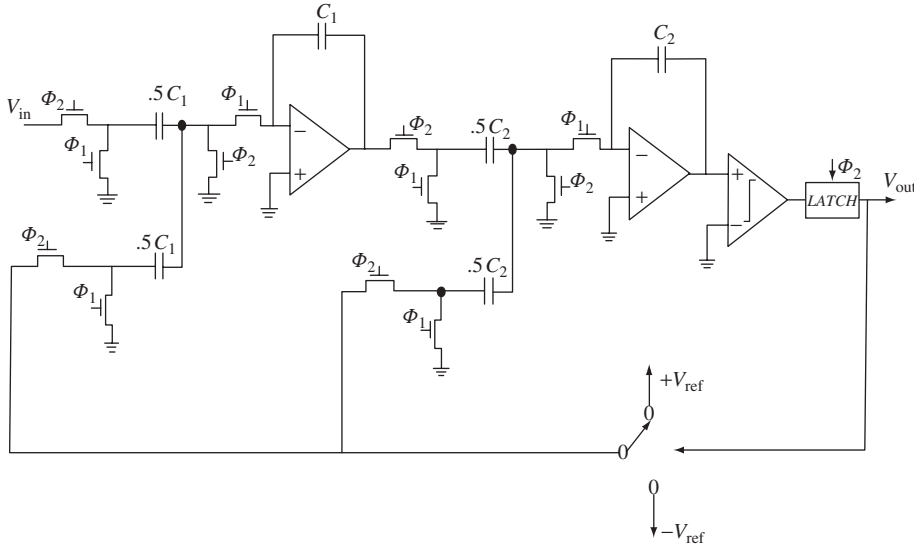


Figure 17.12 Switched capacitor implementation of the second-order converter

17.4 Decimation and Digital Filtering

17.4.1 Principles

The analog part of the converter has been discussed. We now turn to the digital part, which consists of a decimator and a digital filter. Remember that we have *oversampled* the analog signal in the first place. Decimation is the process of reducing the sampling rate of an input stream of samples. With reference to Figure 17.13, if we have an input stream of samples $x(n)$ with frequency $f_s = 1/T_s$, applying this signal to the decimator we should obtain an output stream $y(m)$ with frequency $f_s' = 1/T_s'$ where

$$M = \frac{T_s'}{T_s} = \frac{f_s}{f_s'} \tag{17.18}$$

is the decimation factor.

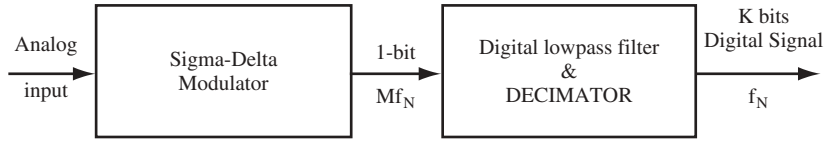


Figure 17.13 Sigma-delta converter including decimation and digital filtering

With reference to Figure 17.14

$$y(m) = \sum_{n=-\infty}^{\infty} h(n)x(n-m) \tag{17.19}$$

Assuming that the input has been band-limited to the range $[-f_s/2, f_s/2]$, in order to lower the sampling rate without aliasing, it is necessary to filter the input signal with a low-pass filter. For an idealized low-pass filter with

$$\begin{aligned} |H(j\omega)| &\approx 1 \quad |\theta| \le 2\pi f_s'/2T_s = \pi/M \\ H(j\omega) &\approx 0 \quad \text{elsewhere} \end{aligned} \tag{17.20}$$

the sampling rate can be reduced by extracting every Mth output sample to obtain $y(m)$ as shown in Figures 17.14 and 17.15. The signal just after low-pass filtering is

$$w(n) = \sum_{k=-\infty}^{\infty} h(k)x(n-k) \tag{17.21}$$

and

$$y(m) = w(mM) = \sum_{k=-\infty}^{\infty} h(k)x(mM-k) \tag{17.22}$$

It is clear that the system is not time-invariant, that is, $x(n-\delta)$ does not lead to $y(m-\delta/M)$ unless $\delta = rM$. We define the signal

$$\begin{aligned} w'(m) &= w(n) \quad \text{for } n = \pm rM \\ &= 0 \qquad \text{otherwise} \end{aligned} \tag{17.23}$$

$$w'(n) = w(n)\left[\frac{1}{M}\sum_{i=0}^{M-1} e^{j2\pi in/M}\right] \tag{17.24}$$

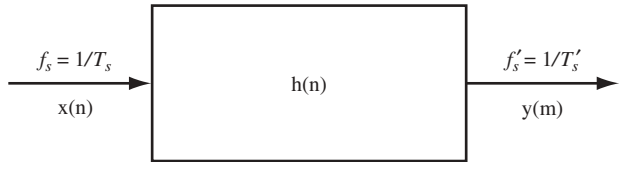


Figure 17.14 A basic decimator

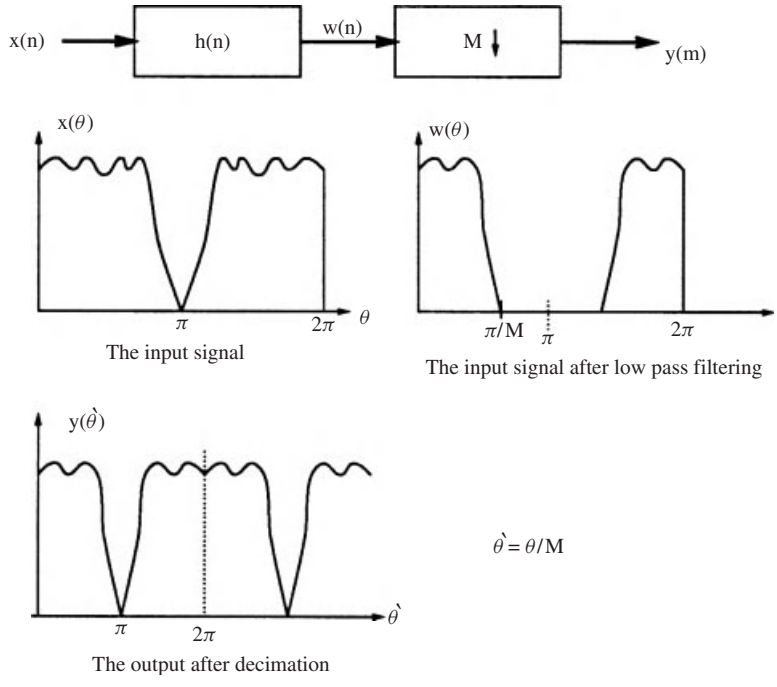


Figure 17.15 Signals before and after decimation and filtering

where the term between the square brackets is the DFT of a train of pulses with a period of M samples.

$$y(m) = w'(mM) = w(mM) \tag{17.25}$$

$$Y(z) = \sum_{m=-\infty}^{\infty} y(m)z^{-m} = \sum_{m=-\infty}^{\infty} w'(mM)z^{-m} \tag{17.26}$$

Putting $k = mM$

$$y(z) \sum_{k=-\infty}^{\infty} w'(k)z^{-k/M} \tag{17.27}$$

where k is an integral multiple of M. But since $w'(k) = 0$ for k not an integer multiple of M, then the equation is also valid for all k. Therefore

$$\begin{aligned} Y(z) &= \sum_{k=-\infty}^{\infty} w(k) \left[\frac{1}{M} \sum e^{j2\pi ik/M} \right] z^{-k/M} \\ &= \frac{1}{M} \sum_{i=0}^{M-1} \left[\sum_{-\infty}^{\infty} w(m) e^{j2\pi ik/M} z^{-k/M} \right] \\ &= \frac{1}{M} \sum_{i=0}^{M-1} W(e^{-j2\pi i/M} z^{1/M}) \end{aligned} \tag{17.28}$$

Since $W(z) = H(z)X(z)$, we have

$$Y(z) = \frac{1}{M}\sum_{i=0}^{M-1} H\left(e^{-j2\pi i/M} z^{1/M}\right) X\left(e^{-j2\pi i/M} z^{1/M}\right) \tag{17.29}$$

$$y(e^{j\theta'}) = \frac{1}{M}\sum_{i=0}^{M-1} H\left(e^{j(\theta'-2\pi i)/M}\right) X\left(e^{j(\theta'-2\pi i)/M}\right) \tag{17.30}$$

where

$$\theta' = 2\pi f T_s' \tag{17.31}$$

Writing $Y(z)$ explicitly as

$$Y(e^{j\theta'}) = \frac{1}{M}\left[H(e^{j\theta'/M})X(e^{j\theta'/M}) + H\left(e^{j\frac{(\theta'-2\pi)}{M}}\right) X\left(e^{j\frac{(\theta'-2\pi)}{M}}\right) + \ldots\ldots\ldots\ldots\right] \tag{17.32}$$

showing that this is the Fourier transform of the output $y(m)$ in terms of the aliased components of the filter output $x(n)$. If the filter has a cutoff frequency $f_c = f_s/2M = \theta = \pi/M$ then we only have the first term in the range $\theta' \leq \pi$, that is

$$Y(e^{j\theta'}) \approx \frac{1}{M} X(e^{j\theta'/M}) \text{ for } \theta' \leq \pi \tag{17.33}$$

which highlights the importance of proper design of the low-pass filter preceding decimation.

17.4.2 Decimator Structures

There are various structures which may be used for the decimator. The simplest is a direct FIR filter as shown in Figure 17.16 in which the operations of multiplication and addition are performed at the rate of f_s.

A more efficient realization can be achieved by commuting the compression branch with the gain branches as shown in Figure 17.17. Here the multiplications and additions are made at a rate of f_s/M.

Another form of this design exploits the symmetry of FIR linear phase filters as shown in Figure 17.18. In this case, the number of multiplications of the previous design is reduced by half.

An efficient method of designing a decimator is to use a multistage structure as shown in Figure 17.19. The decimation factors are $M_1, M_2, M_3, \ldots\ldots\ldots M_N$ with $M = \prod_{i=1}^{N} M_i$. This has the advantages of reducing the computations, reducing storage requirements, simplifying filter design and reducing the finite word-length effects and coefficient sensitivity. This design is highly efficient for the design of high decimation factors, which is always the case.

As we have seen for any digital filter the required degree increases with decreasing the transition band Δf as illustrated in Figure 17.20 (i.e. high selectivity filters). If the filtering operation is accomplished in stages, then the specifications on the transition band can be relaxed in the earlier stages allowing some aliasing in the transition band, which

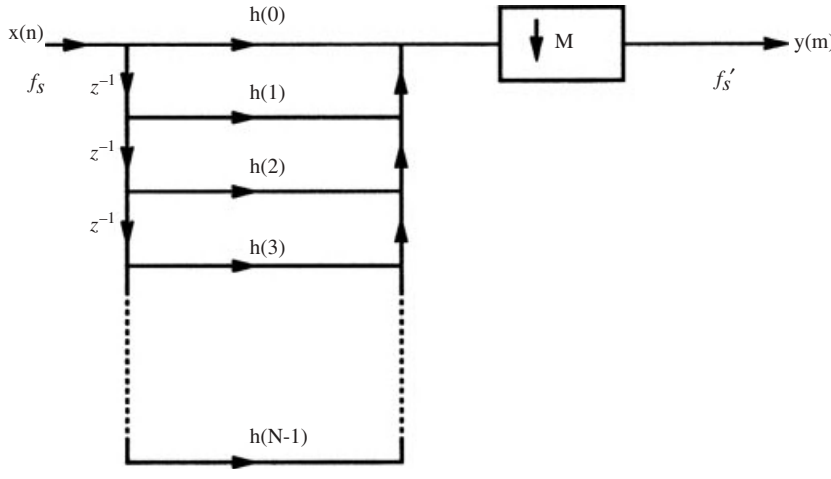


Figure 17.16 Direct FIR decimation structure

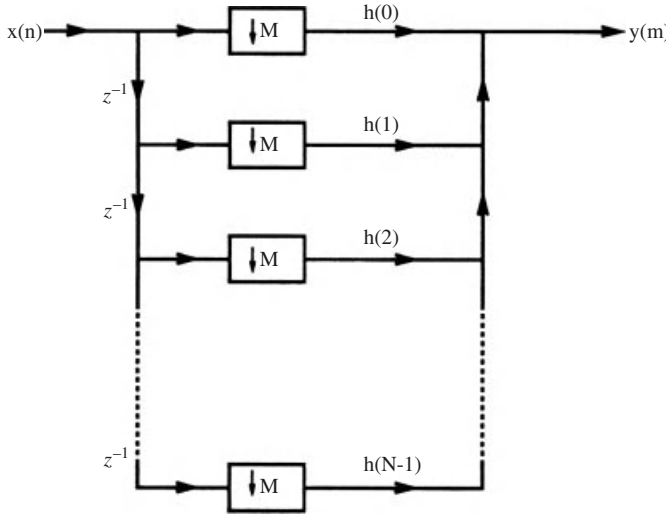


Figure 17.17 A more efficient decimator structure

will be removed at a later stage. Hence, at the later stages the transition band can be reduced without increasing the order of the filter. The overall reduction in computation is significant, and the decimator area on a chip is reduced.

It is always desirable to design the decimator with minimum hardware. Comb filters offer an attractive design method because they do not require multipliers or coefficient ROMs. The transfer function of a first order comb filter is given by

$$H(z) = \sum_{i=0}^{N-1} z^{-i} = \frac{1}{N}\left(\frac{1 - z^{-N}}{1 - z^{-1}}\right) \tag{17.34}$$

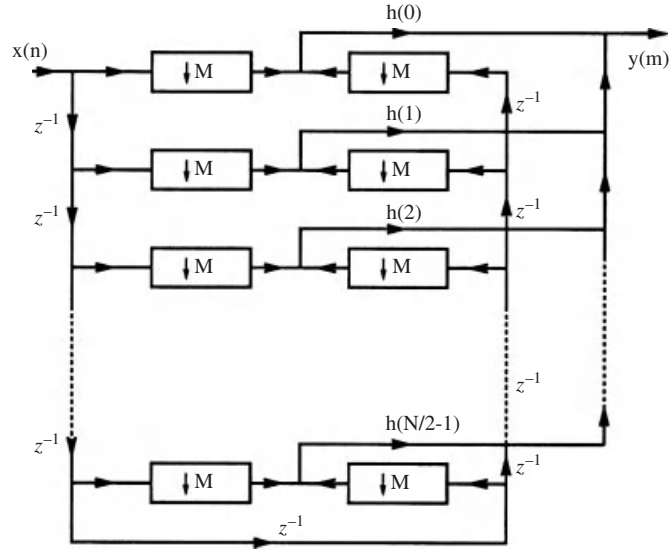


Figure 17.18 Design exploiting the symmetry of linear phase filters

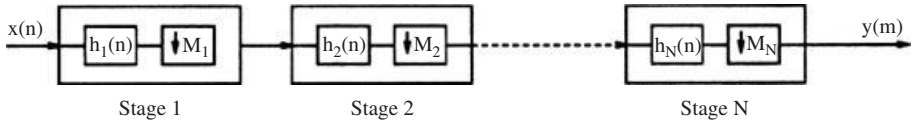


Figure 17.19 Multistage decimator

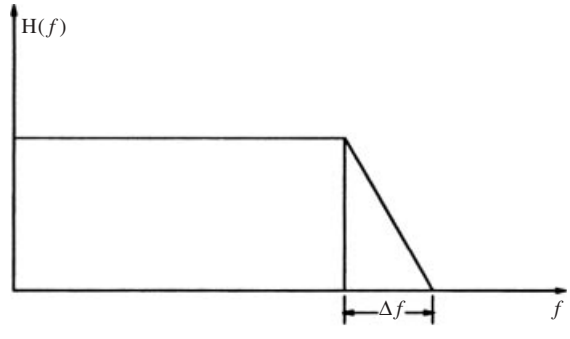


Figure 17.20 Illustrating the transition band of the filter decimator

and the amplitude frequency response is given by

$$|H(j\omega)| = \frac{1}{N}\frac{\sin(N\omega T/2)}{\sin(\omega T/2)} \tag{17.35}$$

but if the baseband is small compared with f_s, $\sin(\omega T/2) \approx \omega T/2$ leading to

$$|H(j\omega)| \approx \frac{\sin(N\omega T/2)}{N\omega T/2} \approx \sin c(N\omega T/2) \tag{17.36}$$

One disadvantage of this structure is that it attenuates the baseband frequencies; therefore, it is only used for a decimation ratio of 4 or lower. Then a digital filter can be used to reduce aliasing. Higher order comb filters can be obtained by raising the function to the required order. Possible implementations of third-order comb filters are shown in Figures 17.21 and 17.22, each is obviously a cascade of FIR and IIR sections. In the latter, the FIR part operates at a sampling rate of f_s/N. Figure 17.23 shows the frequency response of a third-order comb filter with decimation ratio of 64.

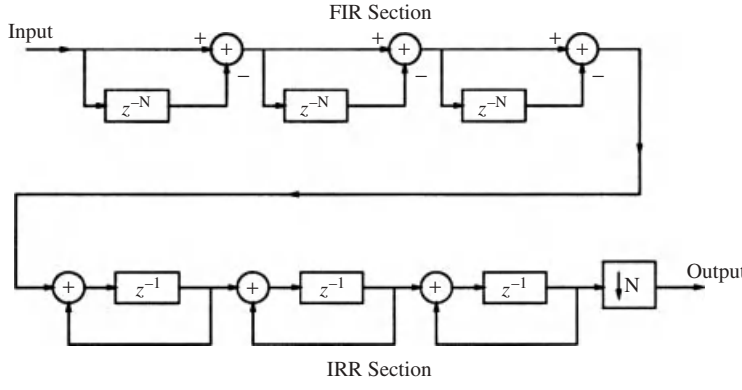


Figure 17.21 Implementation of comb filter decimator

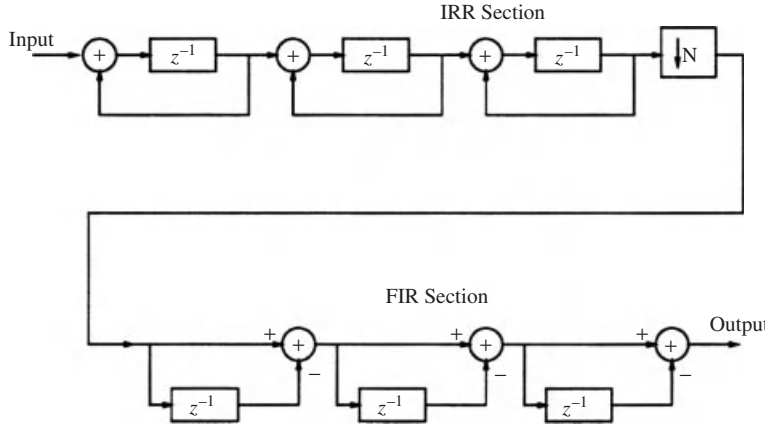


Figure 17.22 Implementation of comb filter decimator

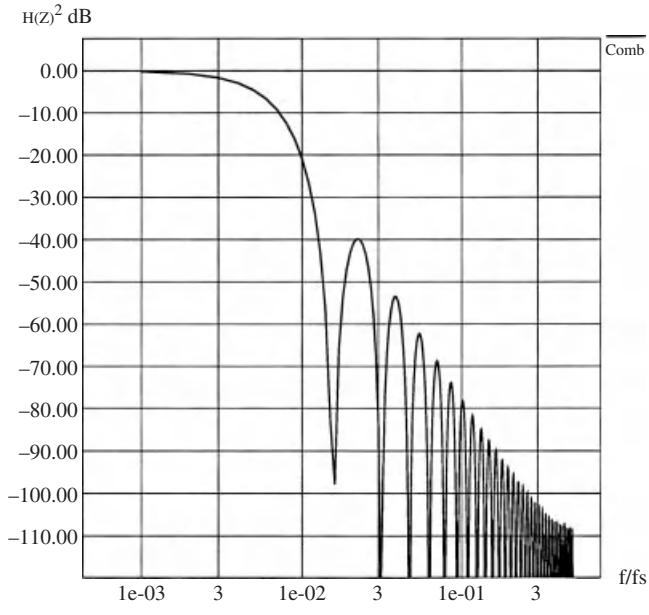


Figure 17.23 Response of a third-order comb filter decimator

17.5 Simulation and Performance Evaluation

Sigma-delta converters are realized as integrated circuits. They contain both analog and digital parts. The modulator is a *non-linear dynamic system*. In the analysis of such a system various approximations are made to be able to treat the system as *quasi-linear* within the range of operation. Therefore, after the design, simulation must be carried out to ascertain that the approximations made are justified within the operating range and conditions and that the system remains stable. There is also the added complication that the modulator uses switched-capacitor circuits with all the non-ideal effects of switches, operational amplifiers and capacitors discussed in Chapter 16. These non-ideal effects must be incorporated in the simulation of the converter.

The simulation of sigma-delta converters is accomplished as follows:

1. A mathematical model is devised which incorporates the non-ideal effects in the circuits used.
2. A test signal is applied to the converter and the output stream of bits is stored.
3. The stream of bits is used to estimate the performance parameters such as the signal to noise ratio, non-linearity and dynamic range. This usually uses FFT periodograms as explained in Chapters 6 and 7.

Testing a designed sigma-delta converter *chip* is done in the same way except that the bit stream is obtained from the converter using a *computer interface*, and the same analysis technique is applied.

Usually, the digital low-pass filter and decimator are also modelled and the bit stream from the converter is applied to them and the output is analysed to evaluate the overall performance of the converter. This also facilitates the choice of the optimum decimator prior to fabrication. The details of this simulation method are now outlined.

(a) The analog converter is simulated by applying samples of a sine wave and performing an FFT on the output stream to estimate the spectrum and to check the range of inputs over which the converter is stable. The required S/N ratio is usually higher than 80 dB, so it is necessary to use a window with very small side lobes. It is also necessary to use a long FFT with these windows since they have a wide main lobe.
(b) A low-pass filter is applied to measure the S/N ratio and the harmonic distortion. Low-pass filtering can be performed either on the bit stream output of the converter or in the frequency domain by multiplying the FFT of the signal by the response (transfer function) of the filter.
(c) Measuring the S/N ratio is accomplished by the sinusoidal minimum square method. This is discussed below.
(d) Sensitivity analysis is carried out to evaluate the dependence of the S/N and the non-linearity on the performance of the integrated circuit.

Now, in the sinusoidal minimum square error method for S/N ratio calculation we have

$$x(n) = A\cos(2\pi f_x nT) \tag{17.37}$$

which is applied to the converter and after low-pass filtering, the output signal $y(n)$ is obtained which consists of a sinusoid with frequency f_x, its harmonics and the noise. This can be written as

$$y(n) = \hat{y}(n) + e(n) \tag{17.38}$$

where

$$\hat{y}(n) = a_0 + a_1\cos(2\pi f_x nT + \phi_1) + \sum_{k=2}^{K} a_k\cos(2\pi k f_x nT + \phi_k) \tag{17.39}$$

The mean square error is

$$e^2 = E[e^2(n)] = E[(y(n) - \hat{y}(n))^2] \tag{17.40}$$

and the methods of power spectrum estimation using the FFT algorithms of Chapters 6 and 7 are used with an N-point DFT to reach the conclusion that the input power can be calculated as

$$P_{in} = E[y^2(n)] = \frac{1}{N}\sum_{n=0}^{N-1} y^2(n) = \frac{1}{N^2}\sum_{n=0}^{N-1} |DFT\{y(n)\}|^2 \tag{17.41}$$

The output power is that of a sinusoid with frequency f_x

$$P_{out} = a_1^2/2 \tag{17.42}$$

whereas harmonic power is

$$P_h = \frac{1}{2}\sum_{k=2}^{K} a_k^2 \tag{17.43}$$

and the noise power is

$$P_e = P_{in} - P_{out} - P_h - a_0^2 \tag{17.44}$$

The performance parameters can now be calculated as

$$S/N = \frac{P_{out}}{P_e} \tag{17.45}$$

$$\frac{S}{N+H} = \frac{P_{out}}{P_e + P_h} \tag{17.46}$$

Finally, several improvements in the design techniques of sigma-delta converters and methods for modelling, optimization of the power consumption can be found in the references [33–36].

17.6 A Case Study: Fourth-order Converter

In principle, increasing both the order of the converter and the sampling frequency will result in improved performance. However, this improvement is limited in practice by two factors. The first is the uncertainty as to the stability of the higher-order converter used. The second is the technological limit on the use of higher sampling rates due to the non-ideal effects inherent in the switched-capacitor circuits used. In this section we give a complete evaluation of a fourth-order converter [31, 32] with the objective of highlighting the problems involved and the analysis methods used for the evaluation of the performance.

The fourth-order converter considered here is the multistage structure shown in Figure 17.24 and is seen to consist of two second-order converters. The additional scaling factors in the signal path are used to control the signal level to avoid driving the Op Amps into saturation.

Two types of scaling are applied:

(a) For each second order section, the gains K_1 and P_1 are added in order to prevent the integrator output level from reaching the saturation level. These gains do not affect the digital output of the quantizer since they are equivalent to a gain of K_1P_1 before the quantizer. However, the input to the next stage is scaled by the factor K_1P_1 and in order to have total cancellation of the noise, this should be compensated in the digital gain of the second stage g_1.
(b) The second type of scaling is between the stages with a factor of J_1, its purpose is to adjust the level of input signal to the second stage such that it is not driven into noisy oscillation mode.

Taking into account these two scaling factors, the total gain at the input of the second stage is $K_1P_1J_1$. In calculating the transfer function of the circuit with these factors taken into account, we first note that they do not affect the digital output of the quantizer and the output can still be written as

$$Y_1(z) = X_1(z)z^{-2} + E_1(z)\left(1 - z^{-1}\right)^2 \tag{17.47}$$

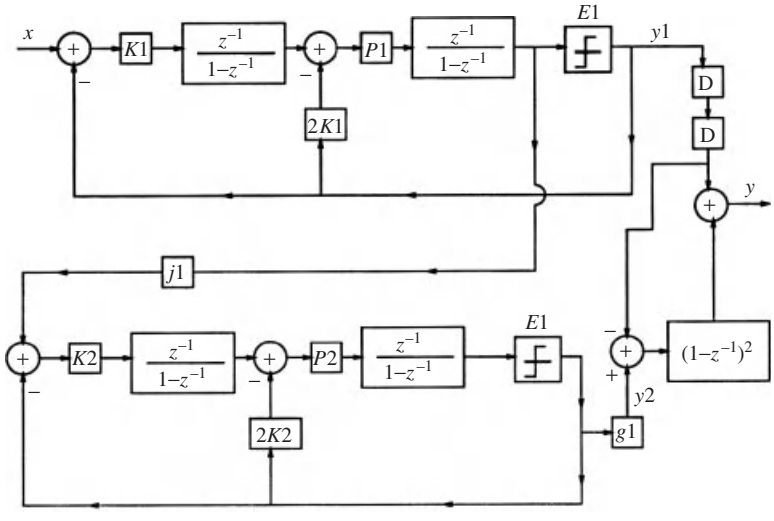


Figure 17.24 Fourth-order converter

and the input to the quantizer in the case of no scaling is

$$Y_1(z) - E_1(z) = X_1(z)z^{-2} + E_1(z)\left[\left(1 - z^{-1}\right)^2 - 1\right] \tag{17.48}$$

This signal can now be scaled by $K_1P_1J_1$ and applied to the second converter whose output becomes

$$\begin{aligned} Y_2 &= g_1\left[X_2(z)z^{-2} + E_2(z)\left(1 - z^{-1}\right)\right] \\ &= g_1\left[J_1K_1P_1X_1(z)z^{-4} + J_1K_1P_1E_1(z)\left\{\left(1 - z^{-1}\right)^2 - 1\right\}z^{-2} + E_2(z)\left(1 - z^{-1}\right)^2\right] \end{aligned} \tag{17.49}$$

and the final output is

$$\begin{aligned} Y(z) &= \left[Y_2(z) - Y_1(z)\right]\left(1 - z^{-1}\right)^2 + Y_1(z)z^{-2} \\ &= X_1(z)z^{-4}\left[1 - z^{-1}\right]^2(1 - g_1K_1P_1J_1) + E_1(z)\left[1 - z^{-1}\right]^4 z^{-2} \\ &\quad (g_1K_1P_1J_1 - 1) - (g_1K_1P_1J_1 - 1) - (g_1K_1P_1J_1 - 1)z^{-2}\left(1 - z^{-1}\right)^2 E_1(z) \\ &\quad + X_1(z)z^{-4} + g_1E_2(z)\left(1 - z^{-1}\right)^4 \end{aligned} \tag{17.50}$$

Ideally, for complete cancellation of noise we must have

$$g_1 = 1/K_1P_1J_1 \tag{17.51}$$

and the output becomes

$$Y(z) = X_1(z)z^{-4} + g_1E_2(z)\left(1 - z^{-1}\right)^4 \tag{17.52}$$

from which the noise in the baseband can be calculated. Assuming white noise with variance σ_q the power spectral density of the noise is

$$S_q(f) = \frac{\sigma_q^2}{f_s} \tag{17.53}$$

and the noise P_n in the baseband is

$$\begin{aligned} P_n &= 2\int_0^{f_b} \frac{\sigma_q^2}{f_s} {g_1}^2 (2\sin(\pi f/f_s))^8 \cong 2\int_0^{f_b} \frac{\sigma_q^2}{f_s} {g_1}^2 (2\pi f/f_s)^8 \\ &\cong \frac{g_1^2 \pi^8}{9} \sigma_q^2 \left(\frac{2f_b}{f_s}\right)^9 \end{aligned} \tag{17.54}$$

$$P_n(\text{dB}) = 20\log\left(\frac{\sigma_q \pi^2 g_1}{3}\right) + 90\log\left(\frac{2f_b}{f_s}\right) \tag{17.55}$$

For an input sinusoidal signal of amplitude A, the signal to noise ratio is

$$S/N = 10\log\left(\frac{A^2}{2}\right) - 20\log\left(\frac{\sigma_q \pi^4 g_1}{3}\right) - 90\log\left(\frac{2f_b}{f_s}\right) \text{dB} \tag{17.56}$$

The simulation methods discussed earlier are used to verify these results. Figure 17.25 shows the power spectrum of the output for an input sinusoid with frequency $f_s/1024$ and amplitude -6 dB relative to the quantizer step. Figure 17.26 shows the dynamic range

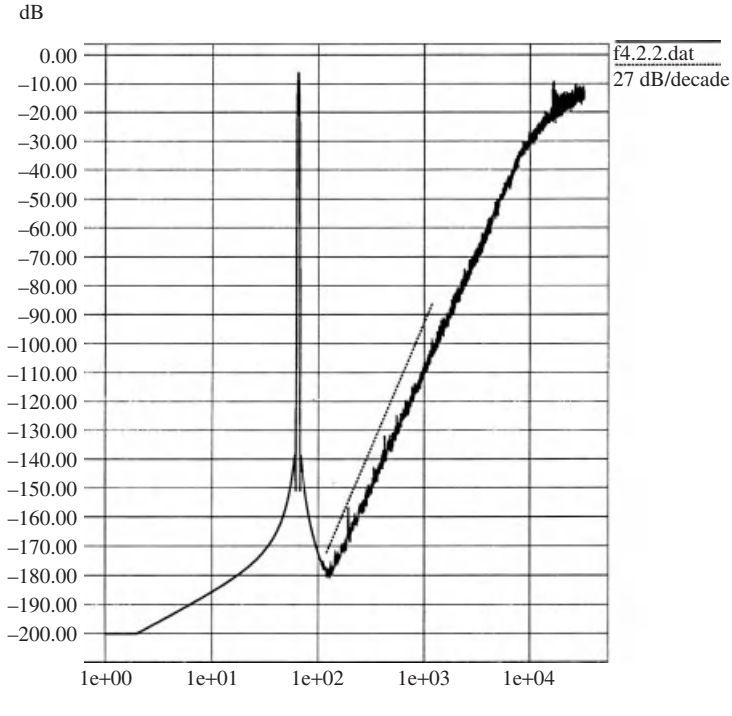


Figure 17.25 Power spectrum output of the fourth-order converter

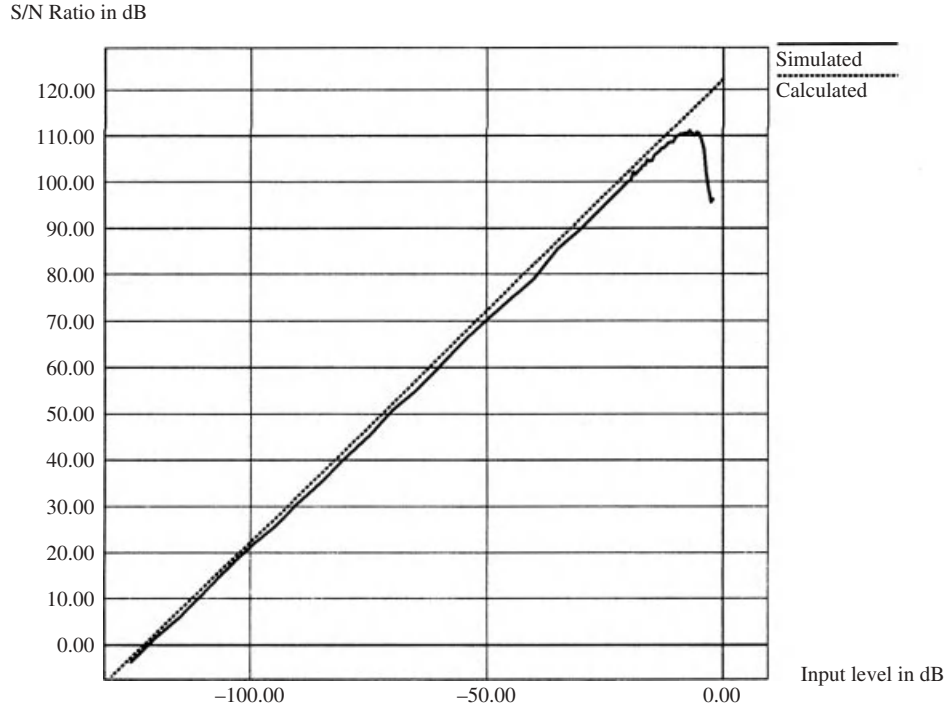


Figure 17.26 Dynamic range of the fourth-order converter

of the converter for an oversampling ratio $R = 64$ and $g_1 = 4$ together with an ideal decimator for both simulation and theoretical calculations.

The main cause of performance degradation from component mismatch occurs when the digital gain g_1 does not equal the inverse of the gain $J_1K_1P_1$. This will cause second-order noise to reach the final input. Other sources of deviation from the required performance, are those inherent in the non-ideal effects of the operational amplifiers and switches which were discussed in Chapter 16.

17.7 Conclusion

This chapter represents an *extensive application* of *virtually all the concepts and techniques of signal processing in the analog and digital domains*. We have employed oversampling, digital filtering, switched-capacitor techniques, spectrum analysis, FFT algorithms and integrated circuits to put together a successful design technique of analog to digital converters with many advantages. The converter is basically of the *mixed-mode* type since it uses both analog and digital circuits on the same integrated circuit chip. It is, therefore, a fitting culmination and conclusion for this book which treats both analog and digital techniques. It is also an illustration of the present trend in signal processing where analog and digital circuits complement each other. The chapter concluded with a case study which highlights the approach which may be taken for extending well-established results to design a new processor with improved performance.

Answers to Selected Problems

Chapter 2

2.1 (a) $v(t) = \dfrac{2}{\pi} + \dfrac{4}{\pi}\sum_{r=1}^{\infty} \dfrac{1}{(1-4r^2)} \cos\left(\dfrac{2r\pi t}{T}\right)$

(b) $v(t) = \dfrac{V_0}{\pi} + \dfrac{V_0}{2}\sin\left(\dfrac{2\pi t}{T}\right) + 2\dfrac{V_0}{\pi}\sum_{r=1}^{\infty} \dfrac{1}{(1-4r^2)} \cos\left(\dfrac{4r\pi t}{T}\right)$

(c) $v(t) = \dfrac{-8V_0}{\pi^2}\sum_{r=1}^{\infty} \dfrac{1}{(2r-1)^2} \cos\left(\dfrac{2(2r-1)\pi t}{T}\right)$

(d) $f(t) = \dfrac{1}{2} + \dfrac{4}{\pi^2}\sum_{r=1}^{\infty} \dfrac{1}{(2r-1)^2} \cos\left(\dfrac{2(2r-1)\pi t}{T}\right)$

(e) $f(t) = \dfrac{2}{\pi}(\sinh\,\pi)\left(\dfrac{1}{2} + \sum_{r=1}^{\infty} \dfrac{(-1)^r}{(1+r^2)}(\cos rt - r\sin rt)\right)$

(f) $f(t) = 2\sum_{n=1}^{\infty} \dfrac{(-1)^{n+1}}{\pi n} \sin\left(\dfrac{2n\pi t}{T}\right).$

2.2 $F(\omega) = T\left(\dfrac{\sin(\omega T/2)}{\omega T/2}\right)^2.$

2.3 $G(\omega) = \dfrac{1}{(1+j\omega)^3}$

2.4

$$G(s) = \frac{s^2+2s+3}{s(s+1)^2(s^2+s+1)(s+3)},$$

$$g(t) = L^{-1}[G(s)] = \{1 - 1.5e^{-t} - 0.0714e^{-3t} + 0.5714e^{-o.5t}\cos(\sqrt{3}/2)t$$

$$- 0.4948e^{-0.5t}\sin(\sqrt{3}/2)t\}u(t)$$

2.5 (a) Wide-sense stable, (b) unstable, (c) unstable, (d) unstable, (e) BIBO stable, (f) unstable.

Signal Processing and Integrated Circuits, First Edition. Hussein Baher.
© 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.

Chapter 3

3.1 Degree = 11.

3.2 Degree = 6.

3.3 Degree = 8.

3.4 Degree = 4.

$$H(s) = s^2/(1.419 \times 10^{11} + 5.328 \times 10^5 s + 41 + 7.506 \times 10^{-5} s^3 + 2.817 \times 10^{-9} s^4)$$

3.5 Degree = 22.

Chapter 4

4.1 (a) $1 + z^{-3} + z^{-4} + z^{-5} + z^{-6}$

(b) $1 + z^{-1} - z^{-2} - z^{-3}$

(c) $\sum_{n=0}^{\infty} nz^{-n}$

(d) $\sum_{n=0}^{\infty} n^2 z^{-n}$

(e) $\dfrac{z^{-1}\sin\alpha}{[1 - (2\cos\alpha)z^{-1} + z^{-2}]}$

(f) $\dfrac{z}{(z-1)^{k+1}}$

(g) $\dfrac{z\sin\alpha}{z^2 - 2\cos\alpha z + 1}$, $|z| > 1$

(h) $\dfrac{z(z - \cos)\alpha}{z^2 - 2\cos\alpha z + 1}$, $|z| > 1$

(i) $\dfrac{ze^{-\alpha}\sin\beta}{z^2 - 2e^{-\alpha}\cos\beta z + e^{-2\alpha}}$

(j) $\dfrac{z(z - e^{-\alpha}\cos\beta)}{z^2 - 2e^{-\alpha}\cos\beta z + e^{-2\alpha}}$

4.2 (a) $f(n) = -8u_0(n) + 4\{(-2)^{2n} + 2^n\}u_1(-n)$

4.4 $f(n) = u_1(n) - u_1(n-9) - u_1(n-4) + u_1(n-13)$

4.5 (a) $u_1(n)$ (b) $u_0(n) + u_1(n-1) - 0.1875(0.25)^{n-1}u_0(n-1)$ (c) $u_1(n-5)/2^{n-4}$

4.6 (a) $g(n) = 3f(n) + 7f(n-1) + 5g(n-1)$

(b) $g(n) = f(n) + 0.2f(n-1) + g(n-1)$

4.7 (a) $H(z) = \dfrac{3 + 7z^{-1}}{1 - 5z^{-1}}$ (b) $H(z) = \dfrac{1 + 0.2z^{-1}}{1 - z^{-1}}$

4.8 (a) $H(z) = \dfrac{z(1+3z)}{(z-1)^2}$; double pole on the unit circle: unstable.

(b) $H(z) = \dfrac{z+2}{z^2 - z - 4}$, poles at 2.5615, 1.561 outside the unit circle: unstable.

(c) $H(z) = \dfrac{0.1z^3 + 0.5z^2 - 0.6z}{z^3 - 0.3z^2 - 0.5z - 0.7}$ one of the poles at 1.201, outside the unit circle: unstable.

4.9 (a) The poles are inside the unit circle: stable, (b) the poles are inside the unit circle: stable.

Chapter 5

5.1 Degree = 10.
For parallel realization:

$$H(z) = \frac{-3.0188 - 6.775z}{0.177 - 0.177z + z^2} + \frac{3.673 - 9.49z}{0.063 - 0.154z + z^2} + \frac{-0.7858 + 3.634z}{0.38 - 0.20072z + z^2} + \frac{-0.0203 + 12.675z}{0.011533 - 0.147z + z^2} + \frac{-0.2784 - 0.0307z}{0.732 - 0.25173z + z^2}$$

5.2 Degree = 5.
For cascade realization:

$$H(z) = \left[\frac{0.73(1+z)^2}{1.286z^2 + 0.685z + 1}\right]\left[\frac{0.935(1+z)^2}{2.91z^2 - 0.17234z + 1}\right]\left[\frac{1.3719(1+z)}{2.74z + 1}\right]$$

5.3 Degree = 8.

$$H(z) = 1.486\left[\frac{(z-1)^2}{1.065z^2 - 2.6336z + 1}\right] \times \left[\frac{(z-1)^2}{1.47z^2 - 2.367z + 1}\right] \times \left[\frac{(z-1)^2}{4.984z^2 + 0.116z + 1}\right] \times \left[\frac{(z-1)^2}{1.377z^2 + 1.142z + 1}\right]$$

5.4 Degree = 6.

Chapter 6

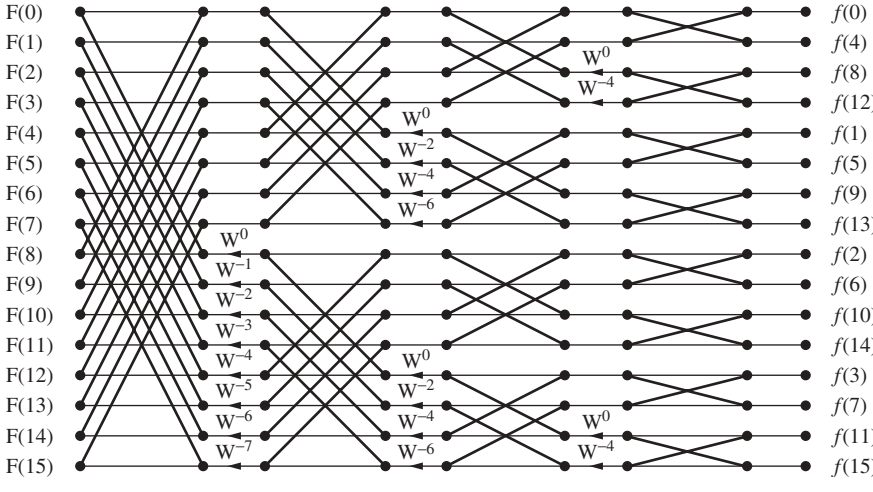
6.1

$$\text{(a) } [\mathrm{F}(k)] = 4.0 \begin{bmatrix} 1.5 \\ 1.161 + \mathrm{j}0.231 \\ 0.427 + \mathrm{j}0.177 \\ -0.143 - \mathrm{j}0.096 \\ -0.250 - \mathrm{j}0.251 \\ -0.064 - \mathrm{j}0.096 \\ 0.073 + \mathrm{j}0.177 \\ 0.046 + \mathrm{j}0.231 \\ 0.0 \\ 0.046 - \mathrm{j}0.231 \\ 0.073 - \mathrm{j}0.177 \\ -0.064 + \mathrm{j}0.096 \\ -0.250 + \mathrm{j}0.251 \\ -0.143 + \mathrm{j}0.096 \\ 0.427 - \mathrm{j}0.177 \\ 1.161 - \mathrm{j}0.231 \end{bmatrix}$$

$$\text{(b) } [\mathrm{F}(k)] = 4.0 \begin{bmatrix} 1.5 \\ 0.658 - \mathrm{j}0.984 \\ -0.177 - \mathrm{j}0.427 \\ 0.169 + \mathrm{j}0.034 \\ 0.250 - \mathrm{j}0.250 \\ -0.022 - \mathrm{j}0.113 \\ 0.177 + \mathrm{j}0.073 \\ 0.196 + \mathrm{j}0.131 \\ 0.0 \\ 0.196 + \mathrm{j}0.131 \\ 0.177 - \mathrm{j}0.073 \\ -0.022 + \mathrm{j}0.113 \\ 0.250 + \mathrm{j}0.250 \\ 0.169 - \mathrm{j}0.034 \\ -0.177 + \mathrm{j}0.427 \\ 0.658 + \mathrm{j}0.984 \end{bmatrix}$$

$$\text{(c) } [\mathrm{F}(k)] = 4.0 \begin{bmatrix} 1.5 \\ -0.658 - \mathrm{j}0.984 \\ -0.177 + \mathrm{j}0.427 \\ -0.169 + \mathrm{j}0.034 \\ 0.250 + \mathrm{j}0.250 \\ 0.022 + \mathrm{j}0.113 \\ 0.177 - \mathrm{j}0.073 \\ -0.196 - \mathrm{j}0.131 \\ 0.0 \\ -0.196 + \mathrm{j}0.131 \\ 0.177 + \mathrm{j}0.073 \\ 0.022 + \mathrm{j}0.113 \\ 0.250 - \mathrm{j}0.250 \\ -0.169 - \mathrm{j}0.034 \\ -0.177 - \mathrm{j}0.427 \\ -0.658 + \mathrm{j}0.984 \end{bmatrix}$$

6.2



Chapter 7

7.1 $P(f,t) = \dfrac{1}{(2\pi)^{1/2}} \exp[-(f+1)^2/2]$, Mean $= -1$, Autocorrelation $= 2$

7.5 $R_{ff}(t) = \begin{cases} \dfrac{t+2\alpha}{\alpha^2}, -2\alpha < t < 0 \\ \dfrac{2\alpha - t}{\alpha^2}, 0 < t < 2\alpha \end{cases}$

Chapter 8

8.1 (a) 1.0168×10^{-5}, (b) 2.542×10^{-6}

Chapter 15

15.1 With equal terminations of 0.5, and using the strays-insensitive circuits

Building block	Cb/Ca	Cc/Ca
1	0.3473	0.5
2	0.7111	–
3	0.3576	–
4	0.9446	–
5	0.1380	0.5

15.2 Degree = 11.

Building block	Cb/Ca	Cc/Ca
1	0.673	0.5
2	0.164	–
3	0.689	–
4	0.175	–
5	0.678	–
6	0.182	–
7	0.653	–
8	0.1865	–
9	0.611	–
10	0.192	–
11	0.511	0.5

15.3 Degree = 3.
In the order given in (15.67), and taking the first capacitor to be one unit, the capacitor values are:

Building block 1: 1.0000, 1.2725, 1.0000, 0.76024, 1.0000
Building block 2: 1.2866, 1.0000, 1.28164
Building block 3: 1.0000, 1.2725, 0.76024, 1.0000

15.4 Degree = 4.
With reference to Figures 15.28–15.29 and Equations (15.69)–(15.75)

$H_1(z)$
Section 1
a0=0.000000 a1=−0.000000 a2=0.000000
b0=1 b1=−0.000000 b2=−1.000000

$H_2(z)$
Section 2
a0=1.000000 a1=−8589934592.000000 a2=−1.000002
b0=1 b1=−2.000000 b2=1.000000

Section 1
F=0.0
E=2.000000
C=2.000000
I=0.000000
H=0.000000
J=0.000000
G=0.000000

Section 2
E=0.0
F=0.000000
C=0.000000
I=1.000002
H=8589934589.999998
J=8589934590.999998
G=0.000000

15.5 Degree = 3.
With reference to Figures 15.28–15.29 and Equations (15.69)–(15.75)
$H_1(z)$:
a0=1.099004 a1=2.145666 a2=1.099004
b0=1 b1=1.680843 b2=1.662832

$H_2(z)$
a0=3.347499 a1=3.347499 a2=0.000000
b0=1 b1=5.694998 b2=0.000000

Second-order section
E=0.0
F=0.662832
C=4.343674
I=1.099004
H=0.000000
J=1.099004
G=4.343674

Linear section type 4
C2=1.175593
C1=0.587796
C3=0.587796

References

The following references complement the material in this book by giving more detailed accounts of some subject matter than is possible to accomplish in one volume. The list is not intended to be comprehensive or to mention all contributions in the field; such an effort would be pretentious and its execution would be impossible. The references are chosen also for the fact that they themselves provide lists of references of historical, academic or practical importance. Alternatively, a recent reference is chosen beacuase it contains material which has its origins in much older references which are not easily available.

1. Keyes, R. (2008) Moore's law today. *IEEE Circguits and Systems Magazine*, **8** (2), 53–54.
2. Chang, L. *et al*. (2003) Moore's law lives on. *IEEE Circuits and Systems Magazine*, **19** (1), 35–42.
3. IEEE (2003) Special issue on nanoelectronics and nanoscale processing. *Proceedings of IEEE*, **91** (11).
4. Gielen, G. and Rutenbar, R. (2000) Computer-aided design of analog and mixed-signal integrated circuits. *Proceedings of IEEE*, **88** (12), 1826–1852.
5. Millozzi, P. *et al*. (2000) A design system for RFIC: challenges and solutions. *Proceedings of IEEE*, **88** (10), 1613.
6. IEEE (2000) Special issue on low power systems. *Proceedings of IEEE*, **88** (10).
7. Benini, L. *et al*. (2001) Designing low-power circuits: practical recipes. *IEEE Circuits and Systems Magazine*, **1** (1), 6–25.
8. IEEE (2001) Special issue on digital, technology directions and signal processing. *IEEE Journal of Solid State Circuits*, **40** (1).
9. Baschirotto, A. *et al*. (2006) Baseband analog front-end and digital back-end for reconfigurable multistandard terminals. *IEEE Circuits and Systems Magazine*, **6** (1), 8–28.
10. Lanczos, C. (1966) *Discourse on Fourier Series*, Hafner, New York.
11. Papoulis, P. (1984) *Signal Analysis*, McGraw–Hill, New York.
12. Baher, H. (2001) *Analog and Digital Signal Processing*, 2nd edn, John Wiley & Sons, Ltd, Chichester.
13. Baher, H. (1984) *Synthesis of Electrical Networks*, John Wiley & Sons, Ltd, Chichester.
14. Rhodes, J.D. (1976) *Theory of Electrical Filters*, John Wiley & Sons, Inc., New York.
15. Abramovitz, M. and Stegun, I.A. (eds) (1970) *Handbook of Mathematical Functions*, Dover, New York.
16. Saal, R. (1977) *Handbook of Filter Design*, AEG Telefunken, Heidelberg.
17. Baher, H. and Beneat, J. (1993) Design of analog and digital data transmission filters. *IEEE Transactions on Circuits and Systems*, **CAS-40** (7u), 449–460.
18. Huang, H. *et al*. (2011) The sampling theorem with constant-amplitude variable-width pulses. *IEEE Transactions on Circuits and Systems: I. Regular Papers*, **58** (6), 1178–1190.
19. Baher, H. (1993) *Selective Linear-phase Switched-capacitor and Digital Filters*, Kluwer Academic, Dordrecht.
20. Rabiner, L. and Cold, C. (eds) (1976) *Digital Signal Processing*, IEEE, London.
21. Haykin, S. (2001) *Adaptive Filter Theory*, Prentice Hall, London.
22. Allen, P. and Holberg, D. (2002) *CMOS Analog Circuit Design*, 2nd edn, Oxford University Press, Oxford.
23. Gray, P. *et al*. (2009) *Analysis and Design of Analog Integrated Circuits*, 5th edn, John Wiley & Sons, Ltd, Chichester.

Signal Processing and Integrated Circuits, First Edition. Hussein Baher.

© 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.

24. Baher, H. (1996) *Microelectronic Switched-capacitor Filters: with ISICAP, a Computer-aided Design Package*, John Wiley & Sons, Ltd, Chichester.
25. Kolm, R. (2008) Analog filters in deep submicrom and ultra deep submicrom technologies. Doctoral thesis. Vienna University of Technology.
26. Sanchez-Sinencio, E. and Silva-Martinez, J. (2000) CMOS transconductance amplifiers, architectures and active filters: a tutorial. *IEE Circuits, Devices and Systems*, **147** (1), 3–12.
27. Gregorian, R. and Temes, G. (1986) *Analog MOS Integrated Circuits for Signal Processing*, John Wiley & Sons, Ltd, Chichester.
28. Hsieh, K. *et al*. (1981) A low-noise chopper stabilised differential switched-capacitor filter technique. *IEEE Journal of Solid State Circuits*, **SC-16** (6), 708–715.
29. Haigh, D.G. and Singh, B. (1983) A switching scheme for switched-capacitor filters which reduces the effect of parasitic capacitances associated with switch control terminals. *Proceedings of IEEE International Symposium on Circuits and Systems*, **1983**, 586–589.
30. Grunigen, D. *et al*. (1982) Integrated switched-capacitor low-pass filter with combined antialiasing decimation filter for low frequencies. *IEEE Journal of Solid State Circuits*, **SC-17** (6), 1024–1029.
31. Afifi, E. (1992) A novel multistage sigma-delta analog-to-digital converter. Master thesis. Worcester Polytechnic Institute.
32. Baher, H. and Afifi, E. (1995) A fourth-order switched-capacitor cascade structure for sigma-delta converters. *International Journal of Circuit Theory and Applications*, **23**, 3–21.
33. Malobert, F. (2001) High-Speed Data Converters For Communication Systems. *IEEE Circuits and Systems Magazine*, **1** (1), 26–36.
34. Suarez, G. *et al*. (2007) Behavioural modelling methods for switched-capacitor sigma delta modulators. *IEEE Transactions on Circuits and Circuits*, **54** (6), 1236–1244.
35. Baschirotto, A. *et al*. (2003) Behavioral modelling of switched-capacitor sigma delta modulators. *IEEE Transactions on Circuits and Systems I*, **50** (3).
36. Karnstedt, C. (2010) Optimizing power of switched capacitor integrators in sigma-delta modulators. *IEEE Circuits and Systems Magazine*, **10** (4), 64–71.

Index

Signal Processing and Integrated Circuits, First Edition. Hussein Baher.

© 2012 John Wiley & Sons, Ltd. Published 2012 by John Wiley & Sons, Ltd.